2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

# Cross-Modality Time-Variant Relation Learning for Generating Dynamic Scene Graphs

Jingyi Wang, Jinfa Huang, Can Zhang, and Zhidong Deng*

***Abstract*— Dynamic scene graphs generated from video clips could help enhance the semantic visual understanding in a wide range of challenging tasks such as environmental perception, autonomous navigation, and task planning of self-driving vehicles and mobile robots. In the process of temporal and spatial modeling during dynamic scene graph generation, it is particularly intractable to learn time-variant relations in dynamic scene graphs among frames. In this paper, we propose a Time-variant Relation-aware TRansformer (TR$^2$), which aims to model the temporal change of relations in dynamic scene graphs. Explicitly, we leverage the difference of text embeddings of prompted sentences about relation labels as the supervision signal for relations. In this way, cross-modality feature guidance is realized for the learning of time-variant relations. Implicitly, we design a relation feature fusion module with a transformer and an additional message token that describes the difference between adjacent frames. Extensive experiments on the Action Genome dataset prove that our TR$^2$ can effectively model the time-variant relations. TR$^2$ significantly outperforms previous state-of-the-art methods under two different settings by 2.1% and 2.6% respectively.**

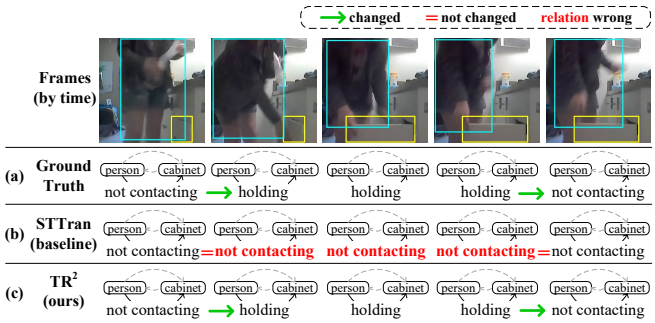


Fig. 1. An example of the existing method fails to judge the change of relations in dynamic scene graphs. The person is bounded by blue boxes and the cabinet is bounded by yellow boxes. The person is opening the cabinet from the second frame to the fourth one and withdraws her hand at the fifth frame eventually. (a) The ground truth that represents the temporal change of the relation between the person and the cabinet. (b) The generation results obtained by STTran [15], where the same relation and scene graphs are retained by mistake. (c) Our TR$^2$ succeeds in judging the change of relations in the second frame and the fifth frame.

## I. INTRODUCTION

Scene graphs represent various entity nodes and relations among nodes as edges in the data format of graphs [1]. Entities and their relations constitute `<subject-predicate-object>` triplets in scene graphs. Scene graphs could help perform tasks related to visual understanding [2]–[6]. Furthermore, dynamic scene graph generation delivers frame-level scene graphs for video clips. The temporal information in dynamic scene graphs could be used for dynamic visual understanding [7], [8] and particularly conducts the decision-making process and task planning of robots particularly [9]–[14]. Compared to methods of image scene graph generation, dynamic scene graph generation methods focus on the modeling of temporal information [15]–[17].

There are significant advances in the area of dynamic scene graph generation in recent years, where the learning of relation features is exploited to perform relation classification. However, the existing methods behave ambiguously when they should judge if the relation between a `subject-object` pair differs from that in the last frame. On the one hand, the subtle movement of entities may imply the change of relation between frames. However, the subtle change would be hard for the visual backbone to recognize. Therefore, the existing methods prefer maintaining the same relations with the former frames. For example, in Fig. 1 (b), the existing method STTran [15] fails to capture the change of contacting relations between the person and the cabinet and insists that the person is not contacting the cabinet. On the other hand, sometimes the form or the position of entities changes a lot compared with the last frame but the corresponding relations remain the same in fact. Encountering such a situation, the existing methods may be confused by the obvious movement of entities. After that, the existing methods predict wrong relations with fake changes.

Existing methods have poor performance on the judgment of the change of relations because of their negligence in the modeling of the difference of relation features in adjacent frames. Without full use of the labels and information about relation changes that are implied in the dataset, existing methods are judging relations of each frame independently despite the temporal information fusion [15], [16].

To address this problem about the ambiguity for time-variant relations, the improvement should focus on the modeling and guidance of the temporal difference of relation features, which corresponds to the difference of relation labels. Accordingly, we propose a Time-variant Relation-aware TRansformer (TR$^2$) for dynamic scene graph generation. Explicitly, TR$^2$ extracts the difference of relation

Jingyi Wang is with Department of Computer Science, Tsinghua University, Beijing 100084, China (email: wang-jy20@mails.tsinghua.edu.cn)

Jinfa Huang and Can Zhang are with the School of Electronic and Computer Engineering, Peking University, China (emails: jinfahuang@stu.pku.edu.cn, zhangcan@pku.edu.cn)

*Zhidong Deng is with Beijing National Research Center for Information Science and Technology (BNRist), THUAI, Department of Computer Science, State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China (email: michael@tsinghua.edu.cn)

features in adjacent frames and constrains it with the situation of change of corresponding relations. We perform cross-modality knowledge distillation for the learning of time-variant relations. Furthermore, TR$^2$ guides the difference of relation features with text embeddings of the prompted sentences like "a photo of a `subject predicating` an `object`" labelled with text relation. Implicitly, the relation feature fusion module in TR$^2$ performs intra-frame and inter-frame information fusion with a transformer. Besides, TR$^2$ emphasizes the relation change with a message token which takes the influence degree of the last frame on the current frame into account. With the above explicit and implicit modeling of the time-variant relations, TR$^2$ could judge the temporal modification of scene graphs correctly.

We evaluate our TR$^2$ on the Action Genome (AG) benchmark [18]. Extensive experiments demonstrate the effectiveness of our explicit and implicit modeling of time-variant relations. Our main contributions are summarized as follows:

1) For the first time, cross-modality guidance is performed for time-variant relations in dynamic scene graph generation. We use the difference of text embeddings of prompted sentences about relation labels as the supervision signal for relations.
2) We design a relation feature fusion module with a message token to model the relation features and their temporal differences implicitly.
3) We propose a new framework called Time-variant Relation-aware TRansformer (TR$^2$), which achieves new state-of-the-art performances on the public AG dataset for dynamic scene graph generation.

## II. Related Work

### A. Scene Graph Generation

A scene graph expresses the entities and their relations in a single image in a graphical structure where nodes indicate the entities and edges indicate relations among the entities [1], [19]. [20] and [21] jointly consider local visual features and introduce a box attention mechanism. The interactions between local features are used to improve scene graph generation (SGG) performance [22], [23]. [24] found the strong regularization for relationship prediction provided by statistical co-occurrences of the Visual Genome dataset [25].

Based on image SGG, Ji et al. [18] released the dataset Action Genome which is the benchmark dataset for video SGG now. Video SGG methods provide frame-level or clip-level scene graphs [26]. [15] encodes the spatial context within single frames and decodes with a temporal module. [16] proposed a hierarchical relation tree and aggregates the context information efficiently with the tree. [27], [28] generate clip-level scene graph generation based on tracklet computations. [17] handles the biases in Video SGG with the help of meta learning.

### B. Scene Graphs for Robot Planning

Scene graph representations naturally express objects and predicates, which provides feasibility for task planning of robots [14]. Scene understanding helps various visual tasks [29]–[34]. Therefore, robot planning with scene graphs attracts great attention in recent years [10]–[12]. For example, [9] presents a goal-directed Programming by Demonstration system at the level of scene graphs that focuses on the poses of objects and considers the robot as an operator in the scene. [13] pays attention to robot planning under partial observability and uses local scene graphs of single images to build and augment global scene graphs toward context-aware robot planning under partial observability. Recently, [14] demonstrates the graph-based planning scheme in complex sequential manipulation tasks by attaching extra task-dependent attributes information on nodes as attributes to constrain the possible interactions with the node.

### C. Knowledge Distillation

The guidance on the relation features in our TR$^2$ model is inspired by knowledge distillation. Knowledge distillation is first proposed in [35] which means distilling the knowledge from a larger deep neural network into a small network. Feature-based knowledge distillation uses the output of intermediate layers, i.e. feature maps as the knowledge to supervise the training of the student model [36]–[39]. Most of the previous knowledge distillation methods work offline [35], [40]–[43] and so does our TR$^2$ model. In offline knowledge distillation, the teacher model is trained on a large set of training data before distillation. The intermediate features of the large teacher model are distilled into the student model to guide its training process. Cross-modal distillation transfers knowledge between different modalities. For example, [41] proposed the probabilistic knowledge distillation and transferred knowledge from the textual modality into the visual modality.

## III. Time-variant Relation-aware Transformer

Let $G = \{G_t\}_{t=1}^{T}$ denote dynamic scene graphs of a video clip, where $t$ indicates the frame index and $T$ stands for the total number of labeled frames in the video. Suppose $\{i_t\}_{t=1}^{T}$ expresses the frames in the video. $G_t = \{V_t, E_t\}$ stands for the scene graph of frame $i_t$, where $V_t$ is the set of entities as nodes and $E_t$ is the set of relations as edges among nodes in $V_t$ of frame $i_t$. Entities in $V_t$ and relationships in $E_t$ form multiple <subject-predicate-object> triplets. With the images of keyframes in a video as the input, our TR$^2$ model produces the corresponding dynamic scene graphs as the output. In the following, we introduce the overall model framework of TR$^2$ and then clarify its key modules.

### A. The Overall Framework

In this paper, we propose TR$^2$ for dynamic scene graph generation. Fig. 2 shows the overall framework of our TR$^2$ model.

First, the frames are fed into an object detector. The detector detects entities in frames and provides the bounding boxes, categories, and visual features of the detected entities. The visual features from the detector serve as a component of the representations of relations in frames, i.e., the representations of `predicate` in

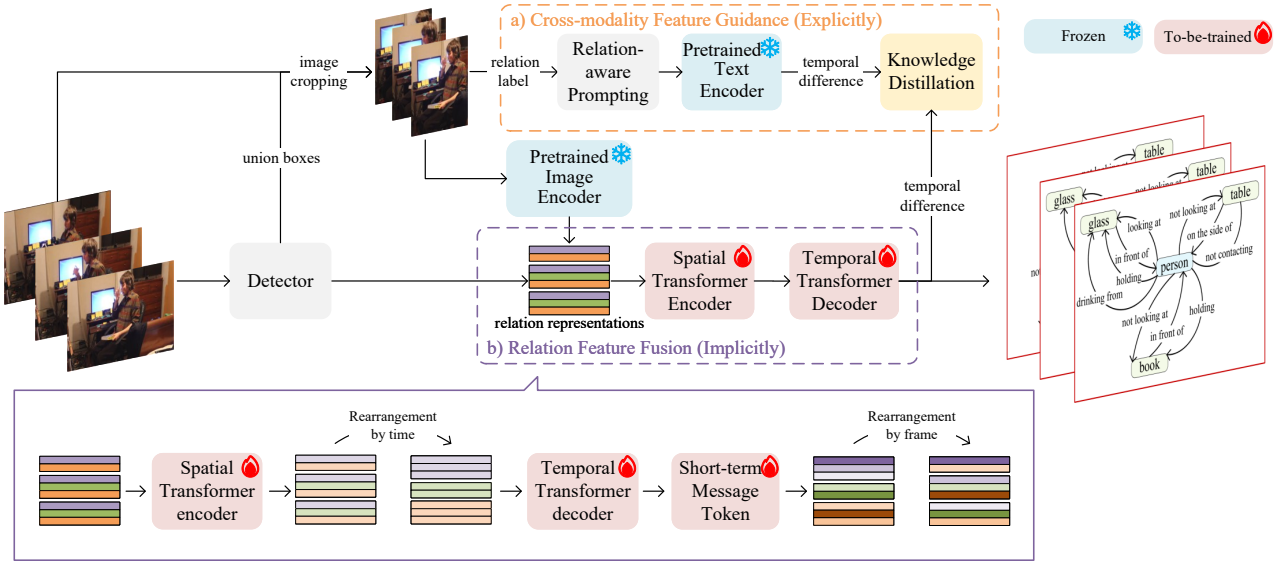


Fig. 2. Our TR$^2$ framework. The detector provides the bounding box, the prediction of categories, and visual features of entities in frames, respectively. The bounding boxes of the subject and the object in a pair are combined as the union bounding box, which is used to crop the original frames to get the region of interest. The cropped images are fed into the pretrained image encoder and become relation representation components in order to adapt the cross-modality distillation process. The guidance module in the orange dotted box a) is applied in the training phase only. The spatial and temporal transformer module in purple boxes b) perform intra-frame and inter-frame information fusion on the relation representation. The short-term message token emphasizes the influence of last frames on current frames. Parameters of modules with the fire symbol are updated during training and that of modules with the snow symbol are frozen. In the relation feature fusion module, colored lines in a block indicate an input unit of the transformer module. Lines in the same or similar color mean that they are representations of the same entity in different frames. Best viewed in color.

<subject-predicate-object> triplets. We combine the detected bounding boxes of the subject and the object in <subject-predicate-object> triplets as union boxes.

Second, we crop the input images with the union boxes to keep the regions of interest that are corresponding to the triplets in the union boxes. We use the image encoder of a large-scale vision-and-language model to encode the cropped images to adapt properly to the guidance given in III-C. The prompted relation labels of the cropped image would participate in the cross-modality feature guidance.

Third, the relation representations obtained in the last two steps would be fed into the relation feature fusion module. The fusion module detailed in III-B carries out intra-frame and inter-frame feature fusion. The relation features after information fusion are classified and form scene graphs with the entity nodes detected by the object detector. In the training phase, the output features of the relation feature fusion module are guided in the cross-modality feature guidance module that would be in details in III-C.

### B. Relation Feature Fusion

The purple box b) in Fig. 2 shows the relation feature fusion module based on a spatial-temporal transformer. First, TR$^2$ performs intra-frame spatial feature fusion on the relation representations. The spatial encoder operates among relations in a frame without position embedding. Second, we rearrange the relation representations by entity and time before the temporal module. In this step, the relation features of a pair of entities in all frames that the pair appears are gathered. Then, the temporal decoder performs long-term inter-frame fusion on the temporal sequence of relation features of each entity pair with temporal position embedding. After that, we model the time-variant relations implicitly with a short-term message token. Specifically, features are calculated with the message token like

$$e_{r_t} = \text{Concat}\left(e_{f_t}, e_{f_{t-1}} \cdot m_{t-1}\right) \tag{1}$$

where $e_f$ stands for the output of the temporal decoder. $t$ and $t-1$ in the subscripts indicate the current frame and the last frame. $m_{t-1}$ is our short-term message token that evaluates how much the last frame affects the current frame. The message token is calculated with $m_{t-1} = g(\text{Concat}(e_{f_t}, e_{f_{t-1}}))$ where $g$ is a feed-forward network. Then we concatenate the product of the message token $m_{t-1}$ and the features of the last frames to the features of the current frames. $e_r$ denotes the concatenate result. At last, the relation representations are rearranged by frame again to facilitate the scene graph output for each frame.

### C. Cross-Modality Feature Guidance

As we mentioned before, we perform cross-modality feature guidance on relation features in adjacent frames explicitly. In this way, we alleviate the problem that existing methods behave ambiguously when dealing with time-variant relations. To be specific, we use the knowledge of a pretrained large-scale vision-and-language model as the supervision signal to guide TR$^2$ besides the original target of scene graph labels. Different from the original feature-based knowledge distillation, we perform temporal difference on the guiding and guided features before distillation. We clarify the specific guidance process as follows.

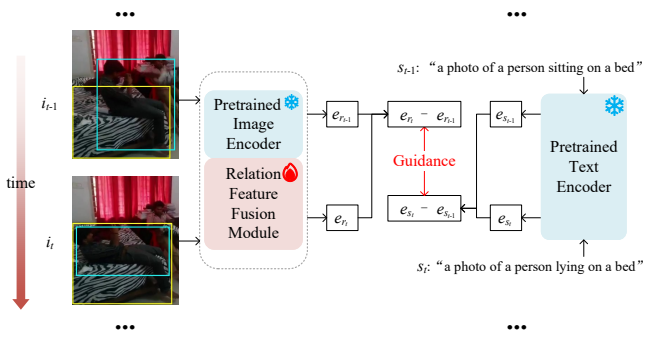


Fig. 3. The cross-modality feature guidance module. We use a pretrained text encoder to extract the text embeddings of prompted sentences that describe the relations in scene graphs. For every two adjacent labeled frames, we perform knowledge distillation from the temporal difference of text embeddings to the temporal difference of relation features. Parameters of the pretrained image and text encoders are frozen.

The orange box a) in Fig. 2 shows the input of the cross-modality guidance module and the specific process is shown in Fig. 3. $i_{t-1}$ and $i_t$ in Fig. 3 are two adjacent labeled frames. After the relation feature fusion module, we get the representations of the relation between the person and the bed, denoted by $e_{r_{t-1}}$ and $e_{r_t}$ for $i_{t-1}$ and $i_t$, respectively. At the same time, we mine the information in the text embeddings of prompted text labels of time-variant relations. To be specific, we extract the words of the subjects, objects, and their relations in scene graphs. Using these words, we construct the sentence of "a photo of a `subject predicating` an `object`" with "a photo of" as the prompt. For example, the prompted sentence for $i_{t-1}$ is "a photo of a person sitting on a bed" and the prompted sentence for $i_t$ is "a photo of a person lying on a bed". Feeding the prompted sentences into the text encoder of the pretrained large-scale vision-and-language model, we get the text features of the sentences denoted by $e_{s_{t-1}}$ and $e_{s_t}$. With relation features after the relation feature fusion module and text embeddings of prompted sentences, we perform temporal difference and obtain $e_{r_t} - e_{r_{t-1}}$ and $e_{s_t} - e_{s_{t-1}}$. TR$^2$ employs $e_{s_t} - e_{s_{t-1}}$ as the supervision signal to guide $e_{r_t} - e_{r_{t-1}}$. Similar operations are performed on every two adjacent frames. In this way, the teacher model helps TR$^2$ to be sensitive to time-variant relations. As a result, TR$^2$ alleviates the ambiguity towards the temporal change of relations.

### D. Training

The training objective function consists of three components, i.e., the entity classification loss, the relation classification loss, and the knowledge distillation loss.

First of all, after the pretrained object detector, TR$^2$ utilizes cross-entropy loss to measure the prediction of entity classification. We utilize $L_{obj}$ to denote this entity loss term. Second, as for the classification of relations among entities, we use binary cross-entropy loss to deal with the multi-label predicates through treating the judgment of each label as binary classification. After that, we gather them in a focal loss form. We adopt $L_{rel}$ to express this relation loss term. Third, in the training phase, cross-modality feature guidance is done. We exploit mean-squared loss for the distillation here, i.e.,

$$L_{guidance} = \frac{1}{T-1} \sum_{i=2}^{T} \left[ (e_{r_t} - e_{r_{t-1}}) - (e_{s_t} - e_{s_{t-1}}) \right]^2 . \quad (2)$$

Then the total loss can be formulated as

$$L = \lambda L_{obj} + L_{rel} + L_{guidance} \quad (3)$$

where $\lambda$ is the weight of the entity detection loss that distinguishes the detection stage from the relation learning stage.

## IV. EXPERIMENTS

We present the experimental results of TR$^2$ on dynamic scene graph generation. First, we introduce the dataset preparation and experiment settings. Second, the implementation details of TR$^2$ in experiments are provided. Finally, we show and analyze the experimental results that include but are not limited to overall assessments and ablation studies. We present some visualization cases in the accompanying video.

### A. Experimental Setup

*1) Dataset:* Our experiments are conducted on the AG dataset [18], which is the benchmark dataset of dynamic scene graph generation. AG is built upon the Charades dataset [44] and provides frame-level scene graph labels with a total of 234,253 frames in 9,848 video clips. In AG, there contain 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.

*2) Evaluation Tasks and Metrics:* In the same way as [16], [18], we make the evaluation of TR$^2$ on the AG dataset under three tasks below: predicate classification (PredCls), scene graph classification (SgCls), and scene graph detection (SgDet). In PredCls, the bounding boxes and object categories are provided and the model needs to predict predicate categories. In SgCls, only the bounding boxes of entities are given. In SgDet, the model needs to detect entities and predict the predicates. In line with [15], [16], *Recall@K* (i.e. $R@K$, $K = [10, 20, 50]$) is adopted as the evaluation metric. As for the predicate choice of predictions for each pair, we borrow the settings from both the **With Constraints** and the **No Constraints** strategies from [15] and **top** $k$ **predictions** ($k = 6$) from [16] to make fair and sufficient comparison with baselines. In these predicate choice settings, **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. Eclectically, **top** $k$ **predictions** picks the first $k$ predictions sorted by scores for each pair.

TABLE I

RECALL (%) COMPARISON RESULTS OF OUR TR$^2$ AND BASELINES IN THE WITH CONSTRAINTS AND THE NO CONSTRAINTS SETTINGS

| Method | With Constraints | | | | | | | | | No Constraints | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PredCls | | | SgCls | | | SgDet | | | PredCls | | | SgCls | | | SgDet | | |
| | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 |
| M-FREQ [45]*CVPR'18* | 62.4 | 65.1 | 65.1 | 40.8 | 41.9 | 41.9 | 23.7 | 31.4 | 33.3 | 73.4 | 92.4 | 99.6 | 50.4 | 60.6 | 64.2 | 22.8 | 34.3 | 46.4 |
| RelDN [46]*CVPR'19* | 66.3 | 69.5 | 69.5 | 44.3 | 45.4 | 45.4 | 24.5 | 32.8 | 34.9 | 75.7 | 93.0 | 99.0 | 52.9 | 62.4 | 65.1 | 24.1 | 35.4 | 46.8 |
| TRACE [16]*ICCV'21* | 64.4 | 70.5 | 70.5 | 36.2 | 37.4 | 37.4 | 19.4 | 30.5 | 34.1 | 73.3 | 93.0 | 99.5 | 36.3 | 45.5 | 51.8 | 27.5 | 36.7 | 47.5 |
| STTran [15]*ICCV'21* | 68.6 | 71.8 | 71.8 | 46.4 | 47.5 | 47.5 | 25.3 | 34.1 | 37.0 | 77.9 | 94.2 | 99.1 | 54.0 | 63.7 | **66.4** | 24.6 | 36.2 | 48.8 |
| Ours | **70.9** | **73.8** | **73.8** | **47.7** | **48.7** | **48.7** | **26.8** | **35.5** | **38.3** | **83.1** | **96.6** | **99.9** | **57.2** | **64.4** | 66.2 | **27.8** | **39.2** | **50.0** |

TABLE II

RECALL (%) COMPARISON RESULTS WITH TOP 6 PREDICTIONS FOR EACH PAIR

| Method | PredCls | | SgCls | | SgDet | |
|---|---|---|---|---|---|---|
| | R@20 | R@50 | R@20 | R@50 | R@20 | R@50 |
| M-FREQ [45]*CVPR'18* | 85.9 | 89.4 | 44.9 | 47.2 | 34.5 | 43.7 |
| RelDN [46]*CVPR'19* | 89.6 | 93.6 | 46.8 | 49.1 | 35.2 | 44.9 |
| TRACE [16]*ICCV'21* | 90.8 | 94.0 | 48.1 | 50.3 | 37.3 | 47.4 |
| STTran [15]*ICCV'21* | 90.2 | 92.1 | 60.6 | 61.4 | 36.0 | 47.2 |
| MVSGG [17]*ECCV'22* | 90.5 | 94.1 | 47.7 | 50.0 | 36.8 | 46.7 |
| Ours | **93.5** | **95.0** | **62.4** | **63.1** | **39.1** | **48.7** |

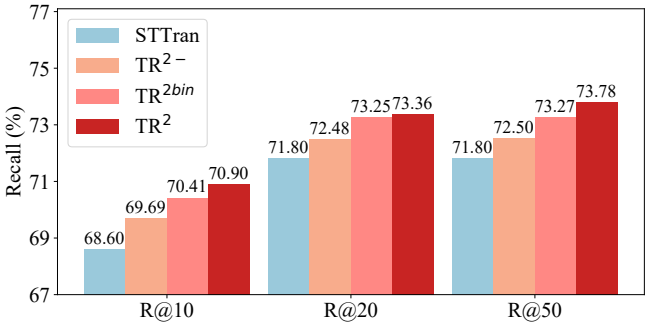


Fig. 4. The ablation study of the cross-modality feature guidance module. TR$^{2-}$ means TR$^2$ without the guidance module. TR$^{2bin}$ considers the guidance as a binary classification task. TR$^2$ denotes the vanilla model and STTran [15] is also illustrated for comparison.

### B. Implementation Details

The object detector is implemented with a Faster-RCNN [47] network based on ResNet-101 [48]. As for the relation feature fusion module, the spatial module is a 1-layer encoder and the temporal module is a 3-layer decoder with 8 heads. The feed-forward dimension is 2048 and the dropout is 0.1 in this transformer. In the cross-modality guidance module, we use CLIP which is a large-scale pretrained vision-and-language model to obtain the text embeddings of prompted sentences. Specifically, we use the VIT-B-32 model provided by CLIP[1]. We adopt the AdamW optimizer [49] to train our TR$^2$ with the initial learning rate of $1e^{-5}$.

### C. Overall Performance Assessments

The comparison results of TR$^2$ and baselines are summarized in Table I and Table II, respectively. In Table I, we report the performance of TR$^2$ in the settings of **With Constraints** and **No Constraints**. The values of $R$@20 and $R$@50 of PredCls task in **With Constraints** are almost the same because this setting only chooses one predicate for each pair so there are a few cases that could reach 50 predictions in total. In Table II, we report the results under the setting of **top** $k$ **predictions** with $k = 6$.

Analyzing the experimental results in Table I and Table II, it is obvious that TR$^2$ achieves the new state-of-the-art results on almost all the metrics and settings. Compared to M-FREQ [45] and RelDN [46], the effect of the basic temporal modeling in TR$^2$ is proved, which is also adopted in STTran [15] and TRACE [16]. Furthermore, with the guidance on the temporal difference between adjacent frames, TR$^2$ performs better than STTran [15] and TRACE [16]. In the SgCls task, the $R$@50 value of STTran looks slightly higher than that of TR$^2$ because of the inherent gap inside the detector. The PredCls task attaches great importance to the relation classification and the temporal difference of dynamic scene graphs. In the PredCls task, TR$^2$ yields 2.1% improvement over STTran [15], which is our main baseline, and 4.5% improvement over TRACE [16] in the **With Constraints** setting. TR$^2$ outperforms previous state-of-the-art methods by 2.6% in the PredCls task under the **No Constraints** setting.

[1]https://github.com/openai/CLIP

### D. Ablation Study

In order to confirm the effect of each module in our TR$^2$, we perform ablation experiments on different guidance settings and the relation feature fusion module. Experiments of our ablation studies are all done on the PredCls task for the AG dataset in the **With Constraints** setting.

**Guidance Settings.** As for the guidance module, which corresponds to our explicit modeling of the time-variant relations in dynamic scene graphs, we conduct the ablation study in the following two aspects:

**a) The effect of the guidance module.** We perform ablation studies on TR$^2$ without guidance and with simpler guidance. The results are presented in Fig. 4. TR$^{2-}$ in Fig. 4 removes the guidance module and is supervised by the classification of relations alone. TR$^{2bin}$ considers the guidance of temporal change as a binary classification task, that is, we use the temporal difference of relation features to classify if the relation change or not. As shown in Fig. 4, TR$^{2-}$ performs better than STTran [15] owing to the temporal decoder that

TABLE III

ABLATION STUDY OF TEMPORAL DIFFERENCES PRECEDING GUIDANCE MODULE

| Temporal Difference | R@10 | R@20 | R@50 |
|---|---|---|---|
| - | 70.1 | 73.0 | 73.0 |
| ✓ | 70.9(+0.8) | 73.8(+0.8) | 73.8(+0.8) |

TABLE IV

ABLATION STUDY ON THE RELATION FEATURE FUSION MODULE OF TR$^2$. TOKEN MEANS THE SHORT-TERM MESSAGE TOKEN.

| Spatial | Temporal | R@10 | R@20 | R@50 |
|---|---|---|---|---|
| - | - | 69.5 | 72.3 | 72.3 |
| - | ✓ | 70.2(+0.7) | 73.1(+0.8) | 73.1(+0.8) |
| ✓ | - | 69.7(+0.2) | 72.5(+0.2) | 72.5(+0.2) |
| ✓ | Decoder | 70.5(+1.0) | 73.3(+1.0) | 73.4(+1.1) |
| ✓ | Token | 70.0(+0.5) | 72.8(+0.5) | 72.9(+0.6) |
| ✓ | Decoder+Token | 70.9(+1.4) | 73.8(+1.5) | 73.8(+1.5) |

passes long-term information and the message token that emphasizes short-term influence. With the simple coarse-grained binary guidance, TR$^{2bin}$ outperforms TR$^{2-}$ and could model the change of relations already. Furthermore, with fine-grained prompted text embeddings that represent the relations, TR$^2$ yields better results than TR$^{2bin}$.

In addition, we tried several different kinds of prompts and the aggregated one of multiple prompts to find that different prompts make little difference. This result demonstrates that it is our leverage of the label text that improves the generation performance rather than the prompting words.

**b) The temporal difference before guidance.** The knowledge distillation module in TR$^2$ focus on the change of relations in dynamic scene graphs. We also had a trial of guidance without the temporal difference, that is, distillation on temporal items directly and separately. In this experiment, the $L_{guidance}$ is updated as $L^{'}_{guidance}$ with

$$L^{'}_{guidance} = \frac{1}{T-1}\sum_{i=2}^{T}(e_{r_t} - e_{s_t})^2 . \quad (4)$$

The comparison results in Table III show the importance of temporal difference in the guidance module. Guidance on the temporal difference of features is more beneficial than guiding separate temporal items directly.

**Relation Feature Fusion Module.** In order to illustrate the effect of each component of the relation feature fusion in TR$^2$, we conduct experiments with/without spatial or temporal modules. The results are shown in Table IV. According to Table IV, we come to the conclusion that every module related to the relation feature fusion of TR$^2$ is effective. Each module is necessary for the final best result with the $R$@10 of 70.9. As for the temporal modules, we ablate the decoder and the use of the message token separately. The temporal decoder could catch long-term information, while the message token emphasizes short-term information. The third row to the fifth one in Table IV confirm the benefit of taking long-term and short-term temporal information separately. Equipping with the long-term decoder and the short-term message token at the same time, the performance would be better as shown in the last row of Table IV.

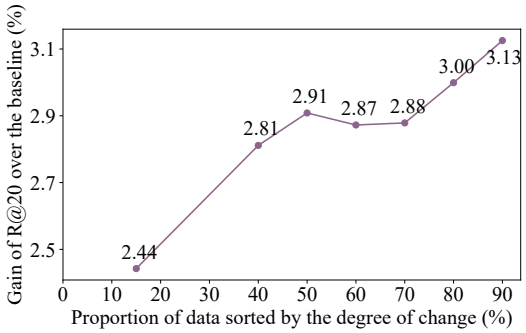


Fig. 5. The more relation changes there exist in data, the more gain our TR$^2$ model outperforms the baseline (STTran).

### E. Performance on different ratings of relation change data

As discussed above, our TR$^2$ model focuses on the change of relations between adjacent frames in dynamic scene graphs. Correspondingly, we evaluate the performance of TR$^2$ in terms of data with changes of different ratings. In particular, we sorted the data in AG [18] by the degree of change, i.e., the proportion of frames where the relations changed. The experimental results are given in Fig. 5, where the horizontal axis means the proportion of data sorted by change, while the vertical one indicates the gain value of $R$@20 that TR$^2$ outperforms STTran [15]. When using data of a higher proportion, there exist more changes among frames, leading to harder difficulty for models to be tackled. The gain values of the top 30 percent of the data are combined because they keep the same relationships all the time with the degree of change as 0. Obviously, the harder the data are, the higher gain that TR$^2$ exceeds STTran [15]. This experiment proves the superiority of TR$^2$ for modeling time-variant relations in dynamic scene graphs.

## V. CONCLUSION

In this paper, we propose a new cross-modality time-variant relation learning method for generating dynamic scene graphs. We perform cross-modality feature guidance on the time-variant relations explicitly. We use a relation feature fusion module with a message token to model the relation features implicitly. The experimental results show that the proposed method has the best performance on the AG dataset, which outperforms existing SOTA models by 2.1% and 2.6% under two different settings, respectively. We perform experiments to illustrate the superiority of TR$^2$ for modeling time-variant relations in dynamic scene graphs. The interpretability of temporal modeling in dynamic scene graphs would be an interesting topic worthy of further study.

## ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation of China (NSFC) under Grant No. 62176134, by a grant from the Institute Guo Qiang (2019GQG0002), Tsinghua University, and by research and application on AI technologies for smart mobility funded by SAIC Motor.

## References


[1] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and F.-F. Li, "Image retrieval using scene graphs." in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2015, pp. 3668–3678.

[2] L. Gao, B. Wang, and W. Wang, "Image captioning with scene-graph based semantic concepts." in *ICMLC*. ACM, 2018, pp. 225–229.

[3] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning." in *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, 2019, pp. 10 685–10 694.

[4] L. Chen, W. Ma, J. Xiao, H. Zhang, and S.-F. Chang, "Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1036–1044.

[5] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick, "Inferring and executing programs for visual reasoning." in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2017, pp. 3008–3017.

[6] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering." in *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 10 797–10 806.

[7] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "Tvqa: Localized, compositional video question answering." in *The 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Association for Computational Linguistics, 2018, pp. 1369–1379.

[8] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "Tvr: A large-scale dataset for video-subtitle moment retrieval." in *European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., vol. 12366. Springer, 2020, pp. 447–463.

[9] Z. Zhen, Z. Zheming, S. Zhiqiang, and O. C. Jenkins, "Semantic robot programming for goal-directed manipulation in cluttered scenes," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 7462–7469.

[10] Z. Yifeng, T. Jonathan, B. Stan, and Y. Zhu, "Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6541–6548.

[11] N. Son-Tung, S. O. Ozgur, N. H. Valentin, and M. Toussaint, "Self-supervised learning of scene-graph representations for robotic sequential manipulation planning," in *4th Annual Conference on Robot Learning (CoRL)*, 2020, pp. 2104–2119.

[12] C. Agia, K. M. Jatavallabhula, M. Khodeir, O. Miksik, V. Vineet, M. Mukadam, L. Paull, and F. Shkurti, "Taskography: Evaluating robot task planning over large 3d scene graphs," in *Conference on Robot Learning*. PMLR, 2022, pp. 46–58.

[13] A. Saeid, C. Kishan, and Z. Shiqi, "Reasoning with scene graphs for robot planning under partial observability," in *2022 IEEE International Conference on Robotics and Automation (ICRA)*, 2022, pp. 5560–5567.

[14] Z. Jiao, Y. Niu, Z. Zhang, S.-C. Zhu, Y. Zhu, and H. Liu, "Sequential manipulation planning on scene graph," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[15] Y. Cong, W. Liao, H. Ackermann, M. Y. Yang, and B. Rosenhahn, "Spatial-temporal transformer for dynamic scene graph generation," in *2021 IEEE International Conference on Computer Vision (ICCV)*, Jul. 2021, pp. 16 372–16 382.

[16] Y. Teng, L. Wang, Z. Li, and G. Wu, "Target adaptive context aggregation for video scene graph generation," in *2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 688–13 697.

[17] L. Xu, H. Qu, J. Kuen, J. Gu, and J. Liu, "Meta spatio-temporal debiasing for video scene graph generation," in *European Conference on Computer Vision (ECCV)*, 2022.

[18] J. Ji, R. Krishna, F.-F. Li, and J. C. Niebles, "Action genome: Actions as compositions of spatio-temporal scene graphs." in *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 10 233–10 244.

[19] A. Goel, B. Fernando, F. Keller, and H. Bilen, "Not all relations are equal: Mining informative labels for scene graph generation," in *2022 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 596–15 606.

[20] S. Woo, D. Kim, D. Cho, and I. S. Kweon, "Linknet: Relational embedding for scene graph." in *NeurIPS*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 558–568.

[21] A. Kolesnikov, A. Kuznetsova, C. Lampert, and V. Ferrari, "Detecting visual relationships using box attention." in *ICCV Workshops*. IEEE, 2019, pp. 1749–1753.

[22] Y. Li, W. Ouyang, X. Wang, and X. Tang, "Vip-cnn: Visual phrase guided convolutional neural network." in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2017, pp. 7244–7253.

[23] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, and C. C. Loy, "Zoom-net: Mining deep feature interactions for visual relationship recognition." in *European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, vol. 11207. Springer, 2018, pp. 330–347.

[24] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context." in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2018, pp. 5831–5840.

[25] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations." *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[26] S. Feng, S. Tripathi, H. Mostafa, M. Nassar, and S. Majumdar, "Exploiting long-term dependencies for generating dynamic scene graphs," *arXiv preprint arXiv:2112.09828*, 2021.

[27] G. Jung, J. Lee, and I. Kim, "Tracklet pair proposal and context reasoning for video scene graph generation," *Sensors*, vol. 21, no. 9, p. 3164, 2021.

[28] K. Gao, L. Chen, Y. Niu, J. Shao, and J. Xiao, "Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs," in *2022 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 19 497–19 506.

[29] W. H. Smith, M. Milford, M.-M. Klaus D., E. Shoaib, and R. B., "Openscenevlad: Appearance invariant, open set scene classification," in *2022 IEEE International Conference on Robotics and Automation (ICRA)*, 2022, pp. 4578–4584.

[30] G. Walter, V. Sagar, H. Ioannis, and P. Ingmar, "Semantically grounded object matching for robust robotic scene rearrangement," in *2022 IEEE International Conference on Robotics and Automation (ICRA)*, 2022, pp. 11 138–11 144.

[31] Y. Zhu, D. Ren, D. Qian, M. Fan, X. Li, and H. Xia, "Star topology based interaction for robust trajectory forecasting in dynamic scene," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 05 2021, pp. 3255–3261.

[32] Y. Liang, B. Chen, and S. Song, "Sscnav: Confidence-aware semantic scene completion for visual semantic navigation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 05 2021, pp. 13 194–13 200.

[33] M. Schwarz and S. Behnke, "Stillleben: Realistic scene synthesis for deep learning in robotics," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 502–10 508.

[34] Y. Di, H. Morimitsu, Z. Lou, and X. Ji, "A unified framework for piecewise semantic reconstruction in dynamic scenes via exploiting superpixel relations," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 737–10 743.

[35] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[36] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets." in *International Conference on Learning Representation (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2015.

[37] K. Xu, L. Rui, Y. Li, and L. Gu, "Feature normalized knowledge distillation for image classification." in *European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., vol. 12370. Springer, 2020, pp. 664–680.

[38] X. Wang, T. Fu, S. Liao, S. Wang, Z. Lei, and T. Mei, "Exclusivity-consistency regularized knowledge distillation for face recognition." in *European Conference on Computer Vision (ECCV)*, ser. Lecture

Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., vol. 12369. Springer, 2020, pp. 325–342.
[39] H. Chen, Y. Wang, C. Xu, C. Xu, and D. Tao, "Learning student networks via feature embedding," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 25–35, 2020.
[40] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," *arXiv preprint arXiv:1707.01219*, 2017.
[41] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer." in *European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11215. Springer, 2018, pp. 283–299.
[42] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer." in *International Conference on Learning Representation (ICLR)*. OpenReview.net, 2017.
[43] S.-I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant." in *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2020, pp. 5191–5198.
[44] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding." in *European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, vol. 9905. Springer, 2016, pp. 510–526.
[45] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context." in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2018, pp. 5831–5840.
[46] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro, "Graphical contrastive losses for scene graph parsing." in *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, 2019, pp. 11 535–11 543.
[47] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks." in *NeurIPS*, 2015, pp. 91–99.
[48] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.
[49] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization." in *International Conference on Learning Representation (ICLR)*, 2019.