GTR: A Grafting-Then-Reassembling Framework for Dynamic Scene Graph Generation

Jiafeng Liang[1] , **Yuxin Wang**[1] , **Zekun Wang**[1] , **Ming Liu**[1,2*] , **Ruiji Fu**[3] , **Zhongyuan Wang**[3] and **Bing Qin**[1,2]
[1]Harbin Institute of Technology, Harbin, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Kuaishou Technology, Beijing, China
{jfliang, yuxinwang, zkwang, mliu, qinb}@ir.hit.edu.cn, {furuiji, wangzhongyuan}@kuaishou.com

Abstract

Dynamic scene graph generation aims to identify visual relationships (subject-predicate-object) in frames based on spatio-temporal contextual information in the video. Previous work implicitly models the spatio-temporal interaction simultaneously, which leads to entanglement of spatio-temporal contextual information. To this end, we propose a **G**rafting-**T**hen-**R**eassembling framework (**GTR**), which explicitly extracts intra-frame spatial information and inter-frame temporal information in two separate stages to decouple spatio-temporal contextual information. Specifically, we first graft a static scene graph generation model to generate static visual relationships within frames. Then we propose the temporal dependency model to extract the temporal dependencies across frames, and explicitly reassemble static visual relationships into dynamic scene graphs. Experimental results show that GTR achieves the state-of-the-art performance on Action Genome dataset. Further analyses reveal that the reassembling stage is crucial to the success of our framework.

1 Introduction

Scene graph is a structured representation of an image that clearly represents entities (nodes) and relationships between them (edges) through a series of triples. Such structured representations play an important role in many downstream tasks, such as visual question-answering [Garcia and Nakashima, 2020; Luo *et al.*, 2022], visual reasoning [Shi *et al.*, 2019], image captioning [Zhang *et al.*, 2021] and vision-and-language navigation (VLN) [Hong *et al.*, 2020]. Scene graph can be applied to individual images (as static scene graph) or to each frame of a video (as dynamic scene graph). Most methods for generating static scene graphs begin by detecting objects in the image using an object detector, and then obtain the relationships between the objects. However, these methods cannot be directly applied to dynamic scene graph generation because they neglect the natural temporal dependence of relationships across frames (shown in Figure 1).

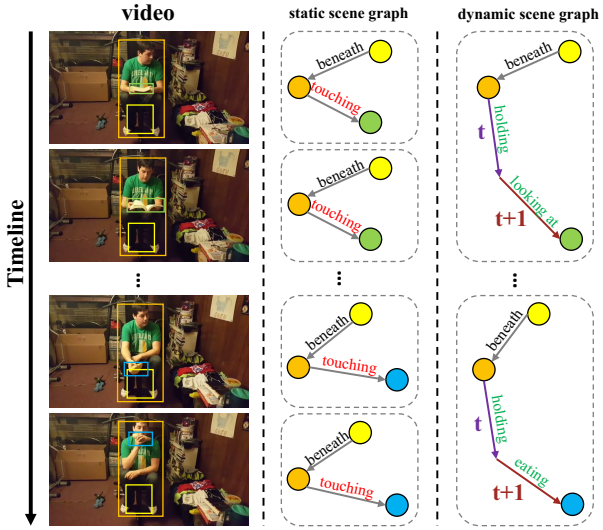

Figure 1: The difference between static scene graph and dynamic scene graph. Dynamic scene graph (**column 3**) has an additional temporal dimension compared to static scene graph (**column 2**). For these four frames in a video (**column 1**), dynamic scene graph generation can generate consecutive actions based on the temporal dependencies across frames.

Most previous methods [Cong *et al.*, 2021; Li *et al.*, 2022] for dynamic scene graph generation utilize the Transformer [Vaswani *et al.*, 2017] to encode the visual features of entity pairs and decode their relationships. Although these methods achieve remarkable performance, they implicitly model spatio-temporal interaction simultaneously and require a large amount of video data for training, which presents two challenges. First, the one-stage modeling process cannot extract spatial and temporal contextual information separately, leading to entanglement between them (shown in Figure 2 (a)). Second, these methods require large amounts of video data to learn spatio-temporal interactions, resulting in expensive annotation and training costs.

To address the aforementioned challenges, we propose a novel two-stage **G**rafting-**T**hen-**R**eassembling framework (**GTR**) for dynamic scene graph generation. **In the first stage**, we graft a pre-trained static scene graph generation model to

* Corresponding Author.

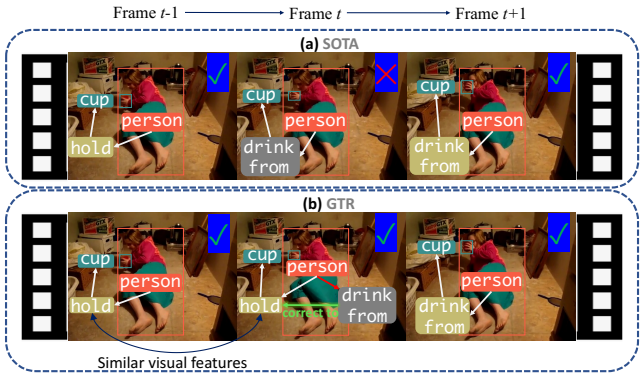

Figure 2: Dynamic scene graphs generated by sota model [Cong *et al.*, 2021] and GTR. **(a)** Relation predicates with similar visual features (**Frame *t* and Frame *t*+1**) cannot be distinguished due to the entanglement of spatio-temporal contextual information. **(b)** GTR rectifies errors by seeking the correct relational predicate with similar visual features in other frames.

generate static visual relationships. Since the model already has the basic static information generation ability, we only need a small amount of the video data and take each frame as an image to fine-tune the model for learning consecutive action relationship types. With the help of grafting, the intra-frame static visual relationships can be obtained without the need for expensive training and manual annotation of large amounts of video data.

In the second stage, to resolve the entanglement of spatio-temporal contextual information, we propose the temporal dependency model (TDM), which contains a temporal attention module and a context attention module. Specifically, after obtaining the visual and semantic features of each frame, we extract temporal dependencies across frames based on these frame features using the temporal attention module. In addition, the mask strategy is designed to capture fine-grained temporal dependencies in the temporal attention module, which can effectively distinguish the cross-frame temporal dependencies of different entity pairs. The context attention module aims to explicitly reassemble static visual relationships into the dynamic scene graph based on temporal dependencies. We consider that there is a positive inductive bias in videos, that is consecutive visual relationships often occur in sequence (*i.e.*, $\langle person - holding - cup \rangle$ and $\langle person - drinking\ from - cup \rangle$ have a high probability of occurring in sequence). For the reason, when static visual relationships within frames are incorrect due to a lack of temporal contextual information, the context attention module can rectify the error by replacing it with the correct relation predicate that has similar visual features in other frames based on the temporal dependence (shown in Figure 2 (b)). In addition, to increase the availability of relation predicates, we design a noise filter (NFT) based on visual feature similarity, which can effectively filter out redundant relation predicates.

To evaluate the performance of the proposed framework, we conduct extensive experiments on Action Genome [Ji *et al.*, 2020]. Our experiments show that GTR achieves state-of-the-art results using only 60% of the video data for training, (*i.e.*, 71.2% Recall@10 for predicate classification task), which is 1.8% higher than the previous best result. Besides, extensive analysis experiments demonstrate that GTR has excellent performance in capturing the spatio-temporal interaction.

Our contributions are summarized as follows:

- We propose GTR, a novel two-stage framework to explicitly capture spatio-temporal interactions for accurate dynamic scene graphs generation.
- Our framework does not require a large amount of video data for training, saving expensive manual video data annotation costs.
- Experimental results show that our framework has significant performance improvements compared to the one-stage approach. Further analysis indicates that our proposed reassembling stage is the key to success.

2 Related Work

2.1 Static Scene Graph Generation

Scene graph generation task was first proposed by Johnson *et al.* [2015], advancing the state of the art in downstream computer vision tasks, natural language processing tasks and multimodal tasks. Currently, the methods of static scene graph generation are mainly based on CNN [Zhang *et al.*, 2017b; Li *et al.*, 2017a; Woo *et al.*, 2018], RNN [Chen *et al.*, 2019; Tang *et al.*, 2019] and TransE [Zhang *et al.*, 2017a; Wan *et al.*, 2018; Gkanatsios *et al.*, 2019; Hung *et al.*, 2019]. CNN-based methods attempt to extract visual features of entity pairs in images by convolution and classify relationships based on these features. Samy Bengio *et al.* [2018] propose a relational embedding module to improve scene graph generation by explicitly modeling inter-dependency among the entire object instances. RNN-based methods attempt to infer relationships of entity pairs based on visual contextual information. Tang *et al.* [2019] propose to compose dynamic tree structures that place the objects in an image into a visual context, helping scene graph generation. TransE-based methods attempt to infer relationships between subject and object by computing the distance between them in the semantic vector space. Wan *et al.* [2018] provide a fully convolutional module to extract the visual embeddings of a visual triple and apply hierarchical projection to combine the structural and visual embeddings of a visual triple. However, as downstream video tasks are widely studied, the static scene graph is not sufficient for their needs. As a result, dynamic scene graph generation has started to be gradually studied by scholars.

2.2 Dynamic Scene Graph Generation

The difference between dynamic scene graph generation and static scene graph generation is that dynamic scene graph generation is for videos, which have an additional time dimension compared to images, making the task more challenging. Currently, there is not much research work [Cong *et al.*, 2021; Li *et al.*, 2022; Gao *et al.*, 2022; Qian *et al.*, 2019; Teng *et al.*, 2021] on this task. Cong *et al.* [2021] proposed a Spatial-Temporal Transformer, which encodes spatial context within single frames and decodes relationships based on temporal

dependencies. Li *et al.* [2022] propose a Transformer-based anticipatory pre-training paradigm that uses unlabeled frames for pre-training to improve dynamic scene graph generation. However, these methods implicitly model the spatio-temporal interaction simultaneously and require a large amount of video data for training. Thus, we propose a novel two-stage framework to improve the problem of spatio-temporal contextual information entanglement by explicitly reassembling static visual relationships into dynamic scene graph based on temporal dependencies. We also grafted a pre-trained static scene graph generation model into our framework, leading to outstanding performance without the need for extensive video data training.

2.3 Transformer for Time Series Modeling

Currently, most dynamic scene graph generation models are based on the Transformer [Vaswani *et al.*, 2017] and demonstrate a powerful understanding of the dependencies between long sequences of data [Zhou *et al.*, 2022; Tuli *et al.*, 2022; Zerveas *et al.*, 2021]. Recently, Transformer has also started to be widely applied to computer vision tasks. Girdhar *et al.* [2019] propose an Action Transformer model for recognizing and localizing human actions in video clips. Arnab *et al.* [2021] propose a pure Transformer-based model to classify videos by encode spatio-temporal tokens from the video. Due to the powerful time-series modeling capabilities of Transformer, our framework models temporal dependencies in the video based on the Transformer architecture.

3 Method

In this section, we first present the definition of dynamic scene graph generation task (Section 3.1). Then, we describe our **G**rafting-**T**hen-**R**eassembling framework (**GTR**) in detail. As shown in Figure 3, our framework consists of two stages: the grafting stage (Section 3.2) and the reassembling stage (Section 3.3). In the first stage, we graft a static scene graph generation model to generate static visual relationships within frames. In the second stage, we extract the temporal dependencies between frames by proposed **T**emporal **D**ependency **M**odel (**TDM**) and reassemble static information into dynamic scene graphs. Meanwhile, we introduce a **N**oise **F**il**t**er (**NFT**) to remove redundant candidate static relation predicates.

3.1 Task Definition

Given a video $V = \{F_1, F_2, ..., F_t\}$, the goal of the dynamic scene graph generation task is to parse the video content as a set of scene graphs $G^{vid} = \left\{G_1^{vid}, G_2^{vid}, .., G_t^{vid}\right\}$. The G_t^{vid} is the scene graph based on the *t*-th frame F_t, defined as $G_t^{vid} = \{B_t, E_t, R_t\}$, where $B_t = \{b_1, b_2, ..., b_i\}$ denotes the bounding box set, $E_t = \{e_1, e_2, ..., e_j\}$ denotes the entity set and $R_t = \{r_1, r_2, ..., r_k\}$ denotes the relation predicate set.

3.2 Grafting Stage

In this stage, we adopt a static scene graph generation model as initialization and fine-tune it with video data to generate static visual relationships within frames in the video. We convert each frame in the video to an image as data for fine-tuning, *i.e.*, $\{F_1, F_2, ..., F_t\} \rightarrow \{I_1, I_2, ..., I_t\}$, which allows the model to learn consecutive action relationship types that are unique to the video. In this way, we extend the image pre-training model from image task to video task to leverage its extensive pre-training knowledge. Specifically, for a given video *V*, we obtain a set of static scene graphs $G^{img} = \left\{G_1^{img}, G_2^{img}, ..., G_t^{img}\right\}$ by static scene graph generation model and extract the static visual relationships $VR^{img} = \{\{S_1, R_1, O_1\}, \{S_2, R_2, O_2\}, ..., \{S_t, R_t, O_t\}\}$ from them, where S_t, O_t, R_t denotes the categories of the entity pairs (subject and object) and static relation predicates between them in *t*-th frame. Moreover, to increase the amount of relation predicates available for the next stage, we generate the top-*k* possible predicates between entity pairs based on the likelihood score during the inference process.

3.3 Reassembling Stage

This stage aims to model the inter-frame dependencies and reassemble the static visual relationships into dynamic scene graphs. To this end, we propose the Temporal Dependency Model (TDM) and the Noise Filter (NFT). The TDM consists of two parts: a temporal attention module and a context attention module.

Feature Extractor For the given video, we use a pre-trained Faster R-CNN to extract frame-level feature following [Cong *et al.*, 2021]. To comprehensively describe the entity pairs, we consider both visual and semantic features. Specifically, as depicted in the bottom middle part of Figure 3, the visual feature of the entity pair *p* in *t*-th frame contains subject's feature v_i^t, object's feature v_j^t and their union bounding box feature $v_{i,j}^t$. The semantic feature of entity pair *p* in *t*-th frame contains semantic embeddings of the subject and object categories, *i.e.*, c_i^t, $c_j^t \in \mathbb{R}^{d_s}$. Then, the feature $f_p^t \in \mathbb{R}^{d_f}$ for entity pair *p* is:

$$f_p^t = [\mathrm{MLP}_v([v_i^t; v_j^t; v_{i,j}^t]); \mathrm{MLP}_s([c_i^t; c_j^t])] \tag{1}$$

where MLP_v and MLP_s are two trainable MLPs, [;] denotes the concatenation.

Temporal Dependency Model (TDM) It is a natural feature of video that the relationships between entity pairs in different frames are correlated. The temporal attention module aims to extract potential temporal dependencies across frames in the video. Specifically, the video feature $F_p \in \mathbb{R}^{n \times d_f}$ is presented as:

$$F_p = \left\{f_{p,1}^1, f_{p,2}^1, ..., f_{p,N(t)}^t\right\} \tag{2}$$

where *N*(*t*) denotes the number of entity pairs in the *t*-th frame. The correlation score between frames can be calculated as:

$$Q^{frm} = K^{frm} = F_p + E_p, V^{frm} = F_p \tag{3}$$

$$S_p^{frm} = \mathrm{softmax}(\frac{Q^{frm}(K^{frm})^{\mathrm{T}}}{\sqrt{d_k}}) \tag{4}$$

where E_p is constructed with learned embedding parameters that is used to inject time positions in entity pair features. Intuitively, temporal dependencies only occur between same entity pairs in different frames. For example, in *t*-th frame, there is a visual relationship

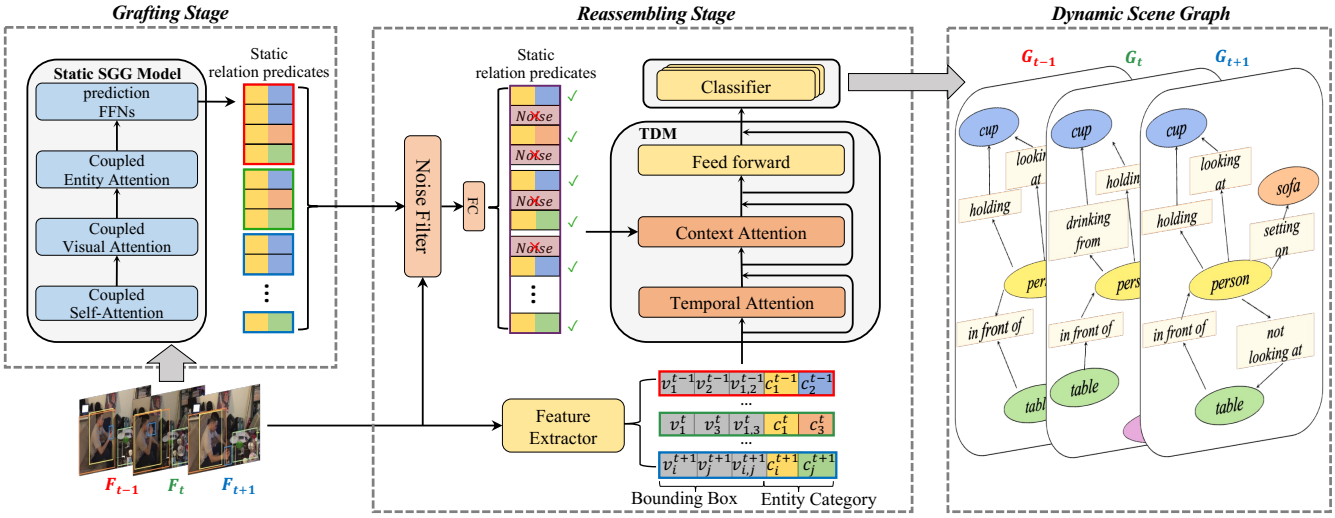

Figure 3: The architecture of GTR, which consists of a grafting stage and a reassembling stage. **In the grafting stage**, the static scene graph generation (SGG) model is grafted into GTR and fine-tuned with the video data to generate static visual relationships. **In the reassembling stage**, the extracted video frame features are fed to the temporal attention module to extract the temporal dependencies in the video, and the context attention module generate the dynamic scene graph by reassembling the static visual relationship based on the temporal dependencies.

$\langle person - drinking\ from - cup \rangle$, which is only temporal dependent from visual relationship $\langle person - holding - cup \rangle$ in (t-1)-th frame, and temporal independent from visual relationship $\langle person - watching - television \rangle$ in (t-1)-th frame. Thus, to obtain fine-grained temporal dependence, we mask the correlation scores between different entity pairs, *i.e.*, $S_p^{frm} \rightarrow \text{Mask}_p \left[S_p^{frm}\right]$. Then, we utilize the masked correlation scores to weight temporal dependencies across frames:

$$H_p^{frm} = \text{Mask}_p \left[S_p^{frm}\right] V^{frm} \tag{5}$$

After obtaining the temporal contextual information in the video, we reassemble the static visual relationships into dynamic scene graphs by context attention module. We consider that there is a positive inductive bias in videos, that is, consecutive visual relationships often occur in sequence (*e.g.*, high probability of $\langle person - holding - cup \rangle$ and $\langle person - drinking\ from - cup \rangle$ sequence occurring in one video). Therefore, we treat static relation predicates in all frames and entity pairs in current frame as candidates and targets respectively to explicitly model the correlation between them based on temporal dependencies (*e.g.*, target: $\langle person - ? - cup \rangle$, candidates: $\langle holding, drinking\ from, touching, ... \rangle$). For the entity pair p_n, its candidate static relation predicates R_n are presented as:

$$R_n = \left\{ r^1_{p_n,1}, r^1_{p_n,2}, \ldots, r^t_{p_n,c} \right\} \tag{6}$$

where c denotes the number of candidate static relation predicates in t-th frame. The matching scores between p_n and R_n is obtained by:

$$Q^{frm} = H_p^{frm}, K^{stc} = V^{stc} = W_R R_n \tag{7}$$

$$S_p^{stc} = \text{softmax}(\frac{Q^{frm}(K^{stc})^{\mathrm{T}}}{\sqrt{d_k}}) \tag{8}$$

where W_R is a trainable weight. We decompose the representation S_p^{stc} into three parts based on the three different types of relation predicates (attention, spatial, contacting) by mask operation. The weighed representation after matching is:

$$H_{att}^{frm} = \text{Mask}_{att} \left[S_p^{stc}\right] V^{stc} \tag{9}$$

$$H_{spa}^{frm} = \text{Mask}_{spa} \left[S_p^{stc}\right] V^{stc} \tag{10}$$

$$H_{con}^{frm} = \text{Mask}_{con} \left[S_p^{stc}\right] V^{stc} \tag{11}$$

Noise Filter (NFT) Static relation predicates are generated from a static scene graph generation model and fed to the temporal dependency model. To improve the availability of static relation predicates, we design NFT to remove redundant ones. Specifically, for an entity pair p in frame F_i, We calculate the cosine similarity score between the visual content in the bounding box of entity pair p in frame F_i and in frame F_j:

$$\text{Sim}(F_i, F_j) = \frac{p^i_{\text{bbox}} \cdot p^j_{\text{bbox}}}{\|p^i_{\text{bbox}}\| \cdot \|p^j_{\text{bbox}}\|} \tag{12}$$

The relation predicates in F_j are regarded positive if the corresponding similarity confidence is greater than the threshold and can be used as candidate relation predicates for the entity pair p.

3.4 Training

We consider the training loss for both object distribution and relation predicate classification. For frame F_t, which contains n_o entities and n_p entity pairs, there are n_r predicates between an entity pair, (*e.g.*, $\langle person - touching - food \rangle$ and $\langle person - holding - food \rangle$ occur simultaneously). Specifically, we optimize the model parameters θ by minimising cost

as follows:

$$\hat{\theta} = \underset{\theta}{argmin} \sum_{p_n=1}^{n_p} \sum_{m=1}^{n_r} \mathcal{L}_{\text{match}} \left(r^*_{p_n,m}, r_{\theta(p_n,m)} \right) + \sum_{l=1}^{n_o} \mathcal{L}_{\text{match}} \left(o^*_l, o_{\theta(l)} \right) \tag{13}$$

where $r^*_{p_n,m}$ and o^*_l denote ground-truth relation predicates and entity categories respectively. $\mathcal{L}_{\text{match}} \left(r^*_{p_n,m}, r_{\theta(p_n,m)} \right)$ and $\mathcal{L}_{\text{match}} \left(o^*_l, o_{\theta(l)} \right)$ are entropy-based log-likelihood matching cost functions, which are defined as:

$$\mathcal{L}_{\text{match}} \left(r_{p_n,m}, r^*_{\theta(p_n,m)} \right) = - \mathbb{1}_{\left\{ c^*_{p_n,m} \neq \emptyset \right\}} \log P \left(c^r_{\theta(p_n,m)} = c^{r*}_{p_n,m} \right) \tag{14}$$

$$\mathcal{L}_{\text{match}} \left(o^*_l, o_{\theta(l)} \right) = -\mathbb{1}_{\left\{ c^*_l \neq \emptyset \right\}} \log P \left(c^o_{\theta(l)} = c^{o*}_l \right) \tag{15}$$

where $\mathbb{1}_{\{\cdot\}}$ is an indicator function, $c^r_{\theta(p_n,m)}$ and $c^o_{\theta(l)}$ denote logits of relation predicates and entities distribution.

4 Experiments

4.1 Dataset

We train and evaluate our model on the Action Genome (AG) dataset [Ji *et al.*, 2020], which describes relationships over time. It regularly selects some frames in the video and annotates the position information of entities and the relationships between entity pairs. AG contains 17M human-object relationship instances with 35 object categories and 25 relation predicates categories. These 25 relation predicates are subdivided into three different types, which are *attention* relationships, *spatial* relationships and *contact* relationships.

4.2 Evaluation Metrics

Following the scene graph generation task, we evaluate our framework in three different modes: Predicate classification (PREDcls): given the ground-truth bounding box and entity categories, predict relation predicates between entities. Scene Graph classification (SGcls): given the ground-truth bounding box, predict entity categories and relation predicates between entities. Scene graph detection (SGdet): given an image, predict bounding boxes and categories of entities, and relation predicates between entities. For SGcls and SGdet, we can not obtain the full ground-truth entities information directly, so we utlize a detector to detect objects. The detection strategy is that the predicted box is considered correctly when it has at least 0.5 IoU (Intersection over Union) overlap with the ground-truth box. We adopt the widely used Recall@K metrics (K = [10, 20, 50]) to evaluate our model, which calculates the recall in the most important (top-1) relation predicates.

4.3 Implementation Details

In this section, we introduce the details of the experimental setting and dataset.

Detector. Following previous work, we adopt Faster R-CNN with ResNet-101 as the backbone network to detect objects.

Parameter Settings. For the feature detector, we map the visual features to a vector of dimension 512 and the semantic features of the object categories to a vector of dimension 300. The MLPs in the paper are three-layer fully connected network and the hidden layer dimension is set to 512.

Training Details. In the grafting stage, we adopt the original RelTR model [Cong *et al.*, 2022], changing only the output number of the classifier. The Action Genome dataset [Ji *et al.*, 2020] is converted to COCO-format for fine-tuning. The RelTR model is fine-tuned for total 20 epochs with mini-batch size 8 in this stage. The initial learning rates of the classifier are unchanged and the learning rates of the other layers are multiplied by 0.9 of the initial learning rate. In the reassembling stage, we train the temporal dependency model (TDM) by SGD optimizer for total 15 epochs with mini-batch size 1. The initial learning rate is set to 1e-5 and adjusted to 5e-6 after the 5 epochs of training and to 1e-6 after 10 epochs of training. In the noise filter (NFT), we set the similarity threshold to 0.9.

4.4 Comparisons with State-of-the-Arts

To verify the superiority of GTR, we compared it with 8 state-of-the-art scene graph generation methods on the Action Genome dataset [Ji *et al.*, 2020].

VRD [Lu *et al.*, 2016] propose a relationship detection method, training two separate vision models, one to recognise objects and the other to recognise relationships.

Motif Freq [Zellers *et al.*, 2018] investigate the problem of producing structured graph representations of visual scenes and propose a new stacked motif networks for capturing higher order motifs.

MSDN [Li *et al.*, 2017b] propose a new end-to-end neural network model to exploit the interconnections across different semantic levels.

VCTREE [Tang *et al.*, 2019] propose to compose dynamic tree structures that place the objects in an image into a visual context, helping scene graph generation.

RelDN [Zhang *et al.*, 2019] improve the relationship detection network and propose a corresponding contrastive loss construction method that accurately identifies the specific relationship between two entities.

GPS-Net [Lin *et al.*, 2020] propose a graph property sensing network that fully explores the edge direction information, the difference in priority between nodes and the long-tailed distribution of relationships.

STTran [Cong *et al.*, 2021] propose a spatial-temporal transformer model to identify the relationships between entities.

AP-Net [Li *et al.*, 2022] propose anticipatory pre-training paradigm based on transformer to model the temporal correlation of visual relationships, consider both global and local information.

As shown in Table 1, our framework outperforms the previous state-of-the-art method in all evaluation metric, improves it 1.8% on PREDcls-R@10, 1.5% on SGcls-R@10 and 1.6% on SGdet-R@10. From the experimental results, we can observe that our framework shows greater improvement with R@10 than with R@20 and R@50, which can indicate that our framework has high prediction efficiency in both PREDcls, SGcls and SGdet, *i.e.*, more correct relation predicates can be

Methods	Predicate Classification			Scene Graph Classification			Scene Graph Detection			Mean
	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50	
VRD	51.7	54.7	54.7	32.4	33.3	33.3	19.2	24.5	26.0	36.6
Motif Freq	62.4	65.1	65.1	40.8	41.9	41.9	23.7	31.4	33.3	45.1
MSDN	65.5	68.5	68.5	43.9	45.1	45.1	24.1	32.4	34.5	47.5
VCTREE	66.0	69.3	69.3	44.1	45.3	45.3	24.4	32.6	34.7	46.9
ReIDN	66.3	69.5	69.5	44.3	45.4	45.4	24.5	32.8	34.9	48.1
GPS-Net	66.8	69.9	69.9	45.3	46.5	46.5	24.7	33.1	35.1	48.6
STTran	68.6	71.8	71.8	46.4	47.5	47.5	25.2	34.1	37.0	50.0
AP-Net	69.4	73.8	73.8	47.2	48.9	48.9	26.3	36.1	38.3	51.4
GTR (w/o RS)	68.3	71.9	71.9	46.1	46.8	46.8	25.0	34.1	37.2	49.8
GTR	**71.2**	**74.5**	**74.5**	**48.7**	**49.7**	**49.7**	**27.9**	**37.0**	**39.9**	**52.6**

Table 1: Experimental results on Action Genome [Ji *et al.*, 2020]. "RS" denotes Reassembling Stage. The best result is in **bold**.

NFT	Mask Strategy	TDM	SGdet	
			R@20	R@50
✔	✔	✔	**37.0**	**39.9**
✘	✔	✔	36.1	39.0
✘	✘	✔	35.5	38.5
✘	✘	✘	34.1	37.2

Table 2: Results of ablation study.

Context Attention	Temporal Attention	SGdet	
		R@20	R@50
✔	✔	**37.0**	**39.9**
✘	✔	35.1	37.9
✔	✘	34.7	37.5

Table 3: Results of the ablation of context/temporal attention.

generated with a small number of recalls. It is worth noting that the complete GTR has a significant performance improvement compared to the GTR without the reassembling stage phase, which indicates that the modeling temporal contextual information is the key to our framework.

4.5 Ablation Study

In this section, we conduct experiments to verify the effectiveness of the components in the reassembling stage.

Impact of Noise Filter (NFT). NFT is proposed to filter the redundant static relation predicates generated by the grafting stage. We investigate the impact of NFT by removing it. The results in the second row of Table 2 demonstrate the effectiveness of using NFT to remove redundant static relation predicates, leading to a significant increase in their availability.

Impact of Mask Strategy. To capture the fine-grained temporal dependencies in the video, we propose the mask strategy in the temporal attention module. As shown in the third row of Table 2 result, the performance of the framework degrades to a certain extent when removing mask strategy, demonstrate that mask strategy can improve the extraction process of temporal dependencies, leading to enhancement of temporal attention module performance.

Method	Precision	
	$holding \longrightarrow drinking\ from$	$holding \longrightarrow eating$
STTran	21/30	19/30
GTR	**25/30**	**25/30**

Table 4: The precision of distinguish similar consecutive actions. We select two samples (*i.e.*, *holding* and *drinking from*, *holding* and *eating*), each containing 30 sets. Compared to STTran [Cong *et al.*, 2021], GTR can predict the consecutive actions more accurately based on the accurate spatio-temporal interaction.

Impact of Temporal Dependency Model (TDM). We investigate the impact of TDM by removing it. Dropped results in the fourth row of Table 2 demonstrate that TDM can comprehensively understand the temporal contextual information in the video. Moreover, we further investigate the impact of the two modules in the TDM (shown in Table 3), demonstrating that both temporal attention and context attention are critical.

4.6 Analysis

In this section, we conduct further analytical experiments to evaluate our framework.

Performance of Spatio-temporal Interaction. We evaluate the performance of spatio-temporal interaction by having the GTR distinguish consecutive actions with similar visual content. We selected two most common samples for evaluation in Action Genome [Ji *et al.*, 2020], where each sample contains two consecutive actions with highly similar visual feature, *i.e.*, $holding \longrightarrow drinking\ from$ and $holding \longrightarrow eating$. The results are shown in Table 4. Compared to STTran [Cong *et al.*, 2021], our GTR achieves superior performance in distinguishing consecutive actions with temporal dependencies demonstrate that our framework can better capture spatio-temporal interactions.

Number of Video Data. we investigate the effect of the training video data magnitude on the performance of GTR. we initially started the experiment using 40% of the data[1],

[1]Video data magnitude below 40% can result in entity and relation predicate categories not being fully covered.

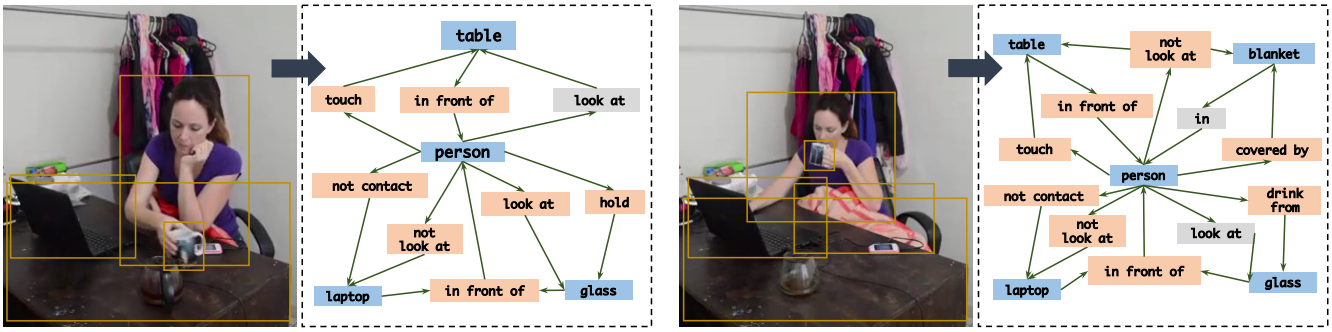

Figure 4: Qualitative results of our framework. We generate the scene graphs with top-1 confidence relation predicate. Incorrect relation predicates are colored with gray.

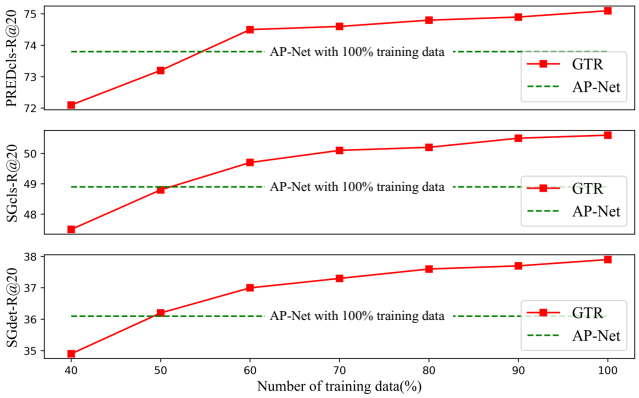

Figure 5: Results of our proposed framework training with different numbers of video data. The green dashed line indicates the results obtained by the previous sota method with full video data training.

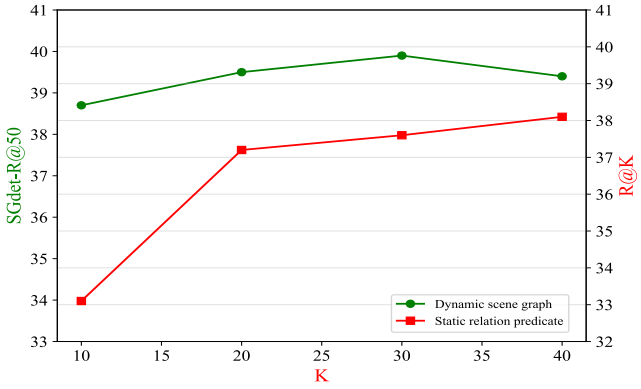

Figure 6: Results of our proposed framework with different numbers of static relation predicates. The red fold line represents the correct percentage of the K candidate static relation predicates recalled. The green fold line represents the result of the framework on SGdet-R@50.

increasing it by 10% each time. As shown in Figure 5, GTR outperform the previous state-of-the-art method in SGdet with 50% data training. When the training data reaches 60%, GTR can outperform the previous state-of-the-art method in all modes. These observations validate our motivation that GTR can achieve excellent performance without the need for extensive video data training.

Number of Candidate Static Relation Predicates. We investigate the effect of the number of candidate relation predicates on our framework by adjusting the number of predicate recalled in the grafting stage. The results are shown in Figure 6. As the number of candidates K increases (the relation predicates may be recurring), the performance of our framework (green fold line) gradually improves and the best performance is achieved at K=30. However, when the number of recalled candidates K=40, the accuracy of the static relation predicates (red fold line) still improves but the performance of our framework decreases, indicating that the framework captures redundant relation predicates, which will weaken the performance of our framework.

4.7 Qualitative Results

The qualitative results are shown in Figure 4. We select two consecutive frames from the video, where the blue boxes are the correct detection results and the pink boxes are the correct relation predicates and the gray boxes are the incorrect relation predicates. The scene graphs are generated with the top-1 confidence relation predictions. The case shows that our model can detect most of the relationships. Some relation predictions could not be predicted due to unclear visual features (*e.g., look at*). Compared to previous model [Cong *et al.*, 2021], our framework is able to accurately predict actions based on time dependence even the action is not obvious (*e.g., touch*).

5 Conclusion

In this paper, we propose a Grafting-Then-Reassembling (GTR) framework for dynamic scene graph generation to decouple spatio-temporal contextual information in video. We firstly graft a static scene graph generation model to generate static visual relationships within frames. Then, we introduce the temporal dependency model to extract temporal dependencies across frames. Finally, we explicitly reassemble the static visual relationships into dynamic scene graphs. Experimental results on the benchmark dataset demonstrate the effectiveness of our proposed framework.

Acknowledgements

We thank anonymous reviewers for their insightful feedback that helped improve the paper. The research in this article is supported by the National Key Research and Development Project (2021YFF0901600), the National Science Foundation of China (U22B2059, 61976073, 62276083), and Shenzhen Foundational Research Funding (JCYJ20200109113441941), Major Key Project of PCL (PCL2021A06).

References

[Arnab *et al.*, 2021] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6816–6826. IEEE, 2021.

[Chen *et al.*, 2019] Yunian Chen, Yanjie Wang, Yang Zhang, and Yanwen Guo. Panet: A context based predicate association network for scene graph generation. In *IEEE International Conference on Multimedia and Expo, ICME 2019, Shanghai, China, July 8-12, 2019*, pages 508–513. IEEE, 2019.

[Cong *et al.*, 2021] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 16352–16362. IEEE, 2021.

[Cong *et al.*, 2022] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *CoRR*, abs/2201.11460, 2022.

[Gao *et al.*, 2022] Kaifeng Gao, Long Chen, Yulei Niu, Jian Shao, and Jun Xiao. Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19475–19484. IEEE, 2022.

[Garcia and Nakashima, 2020] Noa Garcia and Yuta Nakashima. Knowledge-based video question answering with unsupervised scene descriptions. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII*, volume 12363 of *Lecture Notes in Computer Science*, pages 581–598. Springer, 2020.

[Girdhar *et al.*, 2019] Rohit Girdhar, João Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 244–253. Computer Vision Foundation / IEEE, 2019.

[Gkanatsios *et al.*, 2019] Nikolaos Gkanatsios, Vassilis Pitsikalis, Petros Koutras, Athanasia Zlatintsi, and Petros Maragos. Deeply supervised multimodal attentional translation embeddings for visual relationship detection. In *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*, pages 1840–1844. IEEE, 2019.

[Hong *et al.*, 2020] Yicong Hong, Cristian Rodriguez Opazo, Yuankai Qi, Qi Wu, and Stephen Gould. Language and visual entity relationship graph for agent navigation. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[Hung *et al.*, 2019] Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. Union visual translation embedding for visual relationship detection and scene graph generation. *CoRR*, abs/1905.11624, 2019.

[Ji *et al.*, 2020] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10233–10244. Computer Vision Foundation / IEEE, 2020.

[Johnson *et al.*, 2015] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3668–3678. IEEE Computer Society, 2015.

[Li *et al.*, 2017a] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiaoou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7244–7253. IEEE Computer Society, 2017.

[Li *et al.*, 2017b] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1270–1279. IEEE Computer Society, 2017.

[Li *et al.*, 2022] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13864–13873. IEEE, 2022.

[Lin *et al.*, 2020] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3743–3752. Computer Vision Foundation / IEEE, 2020.

[Lu *et al.*, 2016] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905

of *Lecture Notes in Computer Science*, pages 852–869. Springer, 2016.

[Luo *et al.*, 2022] Siwen Luo, Feiqi Cao, Felipe Nunez, Zean Wen, Josiah Poon, and Soyeon Caren Han. Scenegate: Scene-graph based co-attention networks for text visual question answering. *CoRR*, abs/2212.08283, 2022.

[Qian *et al.*, 2019] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi, editors, *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 84–93. ACM, 2019.

[Shi *et al.*, 2019] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8376–8384. Computer Vision Foundation / IEEE, 2019.

[Tang *et al.*, 2019] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6619–6628. Computer Vision Foundation / IEEE, 2019.

[Teng *et al.*, 2021] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13668–13677. IEEE, 2021.

[Tuli *et al.*, 2022] Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *Proc. VLDB Endow.*, 15(6):1201–1214, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[Wan *et al.*, 2018] Hai Wan, Yonghao Luo, Bo Peng, and Wei-Shi Zheng. Representation learning for scene graph completion via jointly structural and visual embedding. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 949–956. ijcai.org, 2018.

[Woo *et al.*, 2018] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 558–568, 2018.

[Zellers *et al.*, 2018] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5831–5840. Computer Vision Foundation / IEEE Computer Society, 2018.

[Zerveas *et al.*, 2021] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In Feida Zhu, Beng Chin Ooi, and Chunyan Miao, editors, *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 2114–2124. ACM, 2021.

[Zhang *et al.*, 2017a] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3107–3115. IEEE Computer Society, 2017.

[Zhang *et al.*, 2017b] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed M. Elgammal. Relationship proposal networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5226–5234. IEEE Computer Society, 2017.

[Zhang *et al.*, 2019] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11535–11543. Computer Vision Foundation / IEEE, 2019.

[Zhang *et al.*, 2021] Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. Consensus graph representation learning for better grounded image captioning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3394–3402. AAAI Press, 2021.

[Zhou *et al.*, 2022] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 27268–27286. PMLR, 2022.