Revisiting Weakly-Supervised Video Scene Graph Generation via Pair Affinity Learning

Minseok Kang[1], Minhyeok Lee[1], Minjung Kim[2], Jungho Lee[1], Donghyeong Kim[1], Sungmin Woo[1], Inseok Jeon[1], and Sangyoun Lee[1⋆]

[1] Yonsei University, Seoul, South Korea
[2] LG Electronics, Seoul, South Korea
louis0503@yonsei.ac.kr

https://github.com/minseokii/PAWS

Abstract. Weakly-supervised video scene graph generation (WS-VSGG) aims to parse video content into structured relational triplets without bounding box annotations and with only sparse temporal labeling, significantly reducing annotation costs. Without ground-truth bounding boxes, these methods rely on off-the-shelf detectors to generate object proposals, yet largely overlook a fundamental discrepancy from fully-supervised pipelines. Fully-supervised detectors implicitly filter out non-interactive objects, while off-the-shelf detectors indiscriminately detect all visible objects, overwhelming relation models with noisy pairs. We address this by introducing a learnable pair affinity that estimates the likelihood of interaction between subject–object pairs. Through Pair Affinity Learning and Scoring (PALS), pair affinity is incorporated into inference-time ranking and further integrated into contextual reasoning through Pair Affinity Modulation (PAM), enabling the model to suppress non-interactive pairs and focus on relationally meaningful ones. To provide cleaner supervision for pair affinity learning, we further propose Relation-Aware Matching (RAM), which leverages vision-language grounding to resolve class-level ambiguity in pseudo-label generation. Extensive experiments on Action Genome demonstrate that our approach consistently yields substantial improvements across different baselines and backbones, achieving state-of-the-art WS-VSGG performance.

Keywords: Video Scene Graph Generation · Weakly Supervised Learning

1 Introduction

Video scene graph generation (VSGG) aims to parse video content into structured triplets in the form of ⟨subject, predicate, object⟩, capturing spatial and temporal interactions between entities and supporting downstream tasks such as video understanding [7,20], visual reasoning [17,6,8], and video question answering [29,5,27]. Despite its significance, fully-supervised VSGG (FS-VSGG) methods face a critical scalability bottleneck: every frame requires exhaustive bounding box annotations for all objects together with their relationship triplets.

⋆ Corresponding author. syleee@yonsei.ac.kr

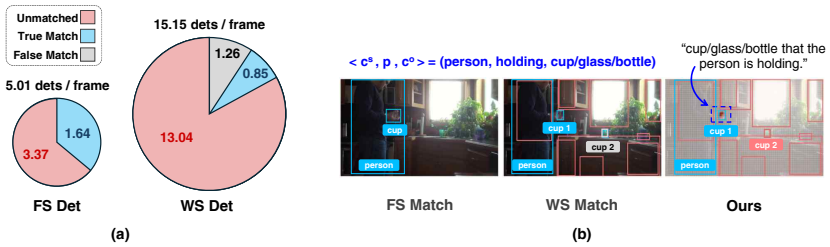

Fig. 1: **Detection distribution and matching in FS-VSGG and WS-VSGG. (a)** Per-frame detection statistics in the weakly annotated training set. Detections are categorized as True Match (spatial overlap with a ground-truth), False Match (class-matched without overlap), or Unmatched. **(b)** Detection and matching comparison for the triplet ⟨person, holding, cup/glass/bottle⟩. Fully-supervised detector produces relation-relevant proposals, whereas off-the-shelf detector generates many irrelevant objects, leading to false matches (e.g., cup 2) under class-level matching. Our method retains only the interaction-consistent instance (cup 1).

To alleviate this burden, recent works have explored weakly-supervised VSGG (WS-VSGG), where supervision is provided in a reduced or indirect form. Representative approaches include PLA [1], which leverages sparse temporal supervision with unlocalized triplets annotated only on the middle frame, where each triplet specifies the subject class, predicate, and object class without bounding boxes. Without ground-truth bounding boxes, WS-VSGG methods inevitably rely on an off-the-shelf object detector [32] to generate object proposals, leading most approaches to adopt two-stage architectures [3,4] that first detect objects and then predict relations. However, existing WS-VSGG methods [1,30] largely adopt this pipeline without accounting for a fundamental shift in detection characteristics between the fully-supervised and weakly-supervised regimes. As illustrated in Figure 1(a), a detector fine-tuned on scene graph datasets produces a curated set of proposals centered around relationally relevant objects. An off-the-shelf detector in WS-VSGG, by contrast, produces a far larger set of proposals dominated by objects irrelevant to any interaction. This shifts the burden of filtering non-interactive pairs from the detector to the relation prediction model.

In conventional two-stage WS-VSGG training pipelines, detector outputs are matched to triplet annotations to assign relation labels. Given unlocalized triplets $\mathcal{G}^u$ for the middle frame t and detected object proposals $\mathcal{D}_t$, detections are matched to triplets based on class identity to construct pseudo-localized training pairs. As illustrated in Fig. 2(a), only matched pairs $\mathcal{P}^+$ receive supervision while unmatched pairs $\mathcal{P}^-$ are entirely discarded. While this convention is viable in FS-VSGG, it creates a severe train-test distribution gap in WS-VSGG. The relation prediction model is trained exclusively on interactive pairs, yet at inference it must operate over a detection space dominated by non-interactive pairs that were never observed during training.

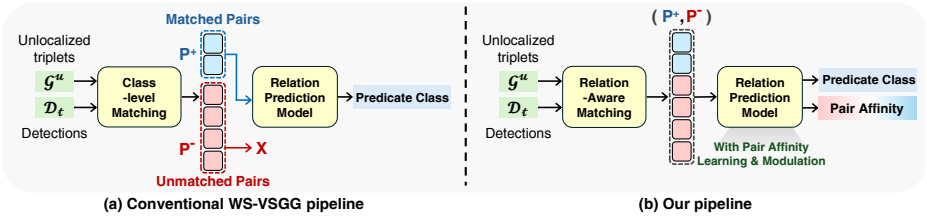

Fig. 2: **Comparison of WS-VSGG training pipelines.** $\mathcal{G}^u$: annotation of unlocalized triplets, $\mathcal{D}_t$: detected object proposals. (a) Only matched pairs $\mathcal{P}^+$ are used for training; unmatched pairs $\mathcal{P}^-$ are discarded. (b) RAM refines the matching, and both $\mathcal{P}^+$ and $\mathcal{P}^-$ are used to learn predicate classification and pair affinity jointly.

We address this by introducing a learnable **pair affinity** that captures the boundary between interactive and non-interactive pairs, whose scalar output measures whether a subject–object pair actively participates in an interaction regardless of the specific predicate. To jointly learn pair affinity alongside predicate classification, we introduce separate representations for the two objectives within the relation prediction model, each with distinct supervision signals. Unlike conventional training pipelines that discard unmatched pairs entirely (Figure 2(a)), we retain them as negative supervision for pair affinity learning, together with matched pairs as positive (Figure 2(b)). At inference, the learned pair affinity score is incorporated into triplet ranking to suppress non-interactive pairs.

While pair affinity score suppresses non-interactive pairs at the output level, the model's internal contextual reasoning remains susceptible to their dominance. Modern VSGG architectures [3,4,9,18] perform contextual reasoning through spatial and temporal attention layers, where each pair refines its representation by attending to other pairs, yet non-interactive pairs equally participate and dilute relational context. To prevent non-interactive pairs from dominating this process, we propose **Pair Affinity Modulation (PAM)**, which gates attention between pairs based on their affinity, allowing interactive pairs to preferentially attend to each other. PAM is applied to both spatial and temporal attention layers, improving contextual reasoning within and across frames as a plug-and-play module compatible with existing attention-based architectures.

Both pair affinity learning and modulation are supervised using the matched-/unmatched partition. Since only unlocalized triplets are available, matching relies solely on class identity, which can generate incorrect pseudo labels when multiple instances of the same class coexist in a frame (Fig. 1(b)). These false matches corrupt both predicate classification and pair affinity supervision by introducing spurious positive labels for non-interacting pairs.

We propose **Relation-Aware Matching (RAM)** to resolve this class-level matching ambiguity by leveraging vision-language (VL) models [13,15] to ground relational descriptions to specific object instances. RAM constructs relation-aware text queries from each triplet and uses cross-modal attention from the VL model to identify the most relation-consistent detection. As illustrated in Fig-

ure 1(b), class-level matching indiscriminately assigns all same-class detections as positive, including non-interactive instances (cup 2), while RAM retains only the relation-consistent one (cup 1). RAM is a one-time preprocessing step that refines pseudo-labels for training, providing cleaner supervision for pair affinity learning with no overhead at inference.

Together, our contributions form a unified framework that addresses the detection distribution shift in WS-VSGG. Extensive experiments on the Action Genome dataset demonstrate consistent improvements across multiple baselines and backbones, achieving state-of-the-art performance.

2 Related Work

2.1 Fully-Supervised Video Scene Graph Generation

VSGG extends image-level scene graph generation to the temporal domain, producing scene graphs that capture dynamic relationships between objects across video frames. Existing methods operate at different granularities: video-level methods [22,26,21] generate a global scene graph for an entire clip, while frame-level methods have been actively explored since the introduction of Action Genome (AG) [7], which provides dense frame-level scene graph annotations over Charades videos and has served as the predominant benchmark for this setting. Among frame-level methods [25,18,9,27], STTran [3] captures within-frame object interactions via a spatial transformer and propagates relational context across adjacent frames via a temporal transformer. TRACE [25] adaptively aggregates context based on target relevance, and DSG-DETR [4] captures long-term temporal context through class-wise inter-frame graphs. More recently, OED [27] reformulates the task as a set prediction problem, enabling one-stage end-to-end generation. Beyond spatio-temporal modeling, TEMPURA [18] and FloCoDe [9] address the long-tailed predicate distribution through memory-guided training and correlation debiasing, respectively. STTran and DSG-DETR serve as the backbone relation prediction models in our experiments.

2.2 Weakly-Supervised Video Scene Graph Generation

To alleviate the heavy annotation burden of fully-supervised methods, weakly-supervised approaches have been proposed. This direction was first explored in the image domain: [24] grounds unlocalized relation labels to image regions, and [33] learns scene graphs directly from natural language supervision without bounding box or relation annotations. Extending this paradigm to the temporal domain, PLA [1] is the first to formulate a weakly-supervised setting for video scene graph generation, annotating only the middle frame of each video clip with unlocalized triplets. Since bounding boxes are unavailable, an off-the-shelf detector generates object proposals for all frames, which are then matched to the triplet annotations via class identity to produce pseudo-localized labels. Although annotations are provided only for the middle frame, PLA propagates

supervision to neighboring frames by treating detections in those frames that share the same class as the annotated entities as additional training samples, enabling frame-level scene graph generation across the entire clip. Building upon this pipeline, TRKT [30] observes that off-the-shelf detectors trained on static images produce suboptimal proposals in dynamic scenes and refines the detector with temporal information. More recently, NL-VSGG [10] further reduces annotation costs by leveraging video captions instead of explicit triplet annotations. Notably, these methods commonly adopt two-stage architectures [3,4] without adapting the relation prediction model to the weakly-supervised regime. Moreover, the sparse, single-frame supervision shared by these video methods makes the detection distribution shift discussed in Section 1 particularly acute—a challenge specific to the video setting that prior work leaves unaddressed. Our work fills this gap by equipping the relation prediction model with pair affinity-aware mechanisms tailored for the weakly-supervised video setting.

3 Method

3.1 Problem Formulation

In FS-VSGG, a model is trained to generate a scene graph $\mathcal{G}_t = \{(s_i, p_i, o_i)\}_{i=1}^{N_t}$ for each frame t in a video clip, where s_i, p_i, and o_i denote the subject, predicate, and object of the i-th triplet, respectively. Each subject and object is associated with a bounding box and a class label, and N_t is the number of annotated triplets in frame t. Training requires localized scene graph annotations on every frame, consisting of bounding boxes, object categories, and pairwise relation labels.

In the weakly-supervised setting we consider, following PLA [1], only unlocalized triplets $\mathcal{G}^u = \{(c_i^s, p_i, c_i^o)\}_{i=1}^{M}$ are provided for the middle frame of each video clip, where c_i^s and c_i^o are the subject and object class labels without bounding boxes, p_i is the predicate, and M is the number of annotated triplets. An off-the-shelf object detector [32] is employed to generate object proposals $\mathcal{D}_t = \{(b_j, c_j, \text{conf}_j)\}_{j=1}^{K_t}$ for each frame t, where b_j, c_j, and conf_j denote the bounding box, predicted class, and confidence score of the j-th detection, respectively. These detections are then matched with the unlocalized triplets to produce pseudo-localized scene graphs for training.

3.2 Overview

As discussed in Section 1, directly transferring a fully-supervised two-stage pipeline to the weakly-supervised setting leaves the relation prediction model unable to distinguish interactive pairs from non-interactive ones. Our method addresses this by equipping the model with a learnable pair affinity, through a unified pipeline consisting of three components. Since pair affinity learning is supervised by the matched/unmatched partition, we first describe **Relation-Aware Matching (RAM)** (Section 3.3), a preprocessing step that produces this partition by leveraging cross-modal grounding to select the most relation-consistent

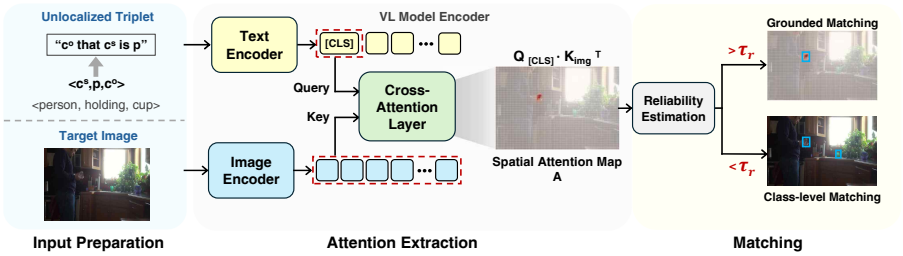

Fig. 3: **Overview of Relation-Aware Matching (RAM).** For each entity in a triplet, a relation-aware query is constructed and fed into a VL grounding model. The `[CLS]` cross-attention map localizes the described relation, and reliability estimation determines whether to perform grounded matching via Grounding Score (GS) or fall back to class-level matching. The figure illustrates the process for c^o; the same procedure is applied independently to c^s.

object among same-class candidates. **Pair Affinity Learning and Scoring (PALS)** (Section 3.4) then trains a binary interaction score on both matched and unmatched pairs and incorporates it into inference-time triplet ranking. Finally, **Pair Affinity Modulation (PAM)** (Section 3.5) uses pair affinity to modulate spatial and temporal attention, enabling effective contextual reasoning even when non-interactive pairs dominate the input.

3.3 Relation-Aware Matching

In WS-VSGG, where bounding box annotations are unavailable, pseudo-localized supervision must be constructed by matching unlocalized triplet annotations with detected objects. Since each unlocalized triplet (c^s, p, c^o) provides only class-level identity without spatial information, existing methods assign all detections whose predicted class matches a given entity as positive pseudo-labels. Formally, for each entity $c \in \{c^s, c^o\}$, the set of class-matched detections is $\mathcal{D}_c = \{d_j \in \mathcal{D}_t : c_j = c\}$, which inevitably includes false matches when multiple instances of the same class coexist in a frame (e.g., cup 2 in Figure 1(b)). These false matches negatively affect both objectives in our relation prediction model: predicate classification receives incorrect relation labels from non-participating instances, and pair affinity learning (Section 3.4) suffers from corrupted matched/unmatched partitions where non-interactive pairs are incorrectly treated as positive.

To provide cleaner supervision for both objectives, we refine this partition by leveraging vision-language grounding to identify the most relation-consistent instance among same-class candidates. The key requirement is a mechanism that can associate a relation-aware textual description with a specific object instance, rather than merely identifying objects by category. Vision-language grounding models [13,15] are well suited for this role, as they can localize image regions corresponding to free-form text descriptions. We adopt GroundingDINO [15] in particular, as its encoder architecture provides readily accessible dense cross-modal

attention maps. For each entity in the triplet, we construct a query that embeds the target class within its relational context (e.g., for ⟨person, holding, cup⟩, "cup that the person is holding"). This query is fed into the model along with the input frame, and we extract the `[CLS]` cross-attention map $\mathbf{A} \in \mathbb{R}^{H_a \times W_a}$ from the final feature enhancer layer. Since the query encodes not just the target class but also its relational context, $\mathbf{A}$ naturally concentrates on the specific instance engaged in the described relation.

While attention maps often provide useful spatial signals, they are not always reliable. Since our goal is to identify a single object instance that best matches the described relation, spatially concentrated attention indicates that the model has found a clear visual referent, while dispersed attention suggests either failure to distinguish among candidates or absence of the target object. We leverage this observation by using spatial concentration as a proxy for localization reliability. Given $\mathbf{A}$, we normalize it into a probability distribution $\mathbf{W}$ and define the reliability score r as:

$$\sigma_{\text{spat}} = \sqrt{\sum_{i,j} \mathbf{W}_{ij} \left[(x_j - \mu_x)^2 + (y_i - \mu_y)^2\right]}, \quad r = \exp\left(-\frac{\sigma_{\text{spat}}}{\sqrt{H_a^2 + W_a^2}}\right), \quad (1)$$

where (x_j, y_i) are spatial coordinates in the attention map of size $H_a \times W_a$, (μ_x, μ_y) is the weighted centroid of $\mathbf{W}$, and σ_{spat} measures spatial dispersion. If r exceeds a threshold τ_r, we perform grounded matching; otherwise, we fall back to class-level matching. For reliable attention maps, each candidate detection $d_k \in \mathcal{D}_c$ is scored using a Grounding Score (GS) that combines two complementary measures: concentration C, the fraction of total attention mass captured by the detection, and density ρ, the attention mass per unit area:

$$C(b_k) = \frac{\sum_{i,j \in b_k} \mathbf{A}_{ij}}{\sum_{i,j} \mathbf{A}_{ij}}, \quad \rho(b_k) = \frac{\sum_{i,j \in b_k} \mathbf{A}_{ij}}{|b_k|_a}, \quad (2)$$

where $|b_k|_a$ denotes the box area in attention-map coordinates. The final grounding score is $\text{GS}(b_k) = C(b_k) \cdot \sigma(\rho(b_k))$, with $\sigma(\cdot)$ denoting the sigmoid function. We select the detection with the highest GS as the match, provided it exceeds a threshold τ_{gs}. If neither condition is met, the triplet is discarded.

3.4 Pair Affinity Learning and Scoring

Given the refined matched/unmatched partition from RAM, the relation prediction model receives all detected pairs. We introduce two parallel embeddings for each pair: a relation embedding $\mathbf{R}$ that encodes which predicate the pair is performing, and a pair affinity embedding $\mathbf{P}$ that encodes whether the pair actively participates in a meaningful interaction. For each subject-object pair, the initial relation embedding $\mathbf{R}_0$ is constructed from the concatenation of subject and object detection features, their spatial union feature, and class embeddings, following the pair representation used in the prior works [3,4]. The pair affinity embedding $\mathbf{P}_0$ is then derived from $\mathbf{R}_0$ via a projection MLP: $\mathbf{P}_0 = \text{MLP}_{\text{init}}(\mathbf{R}_0)$,

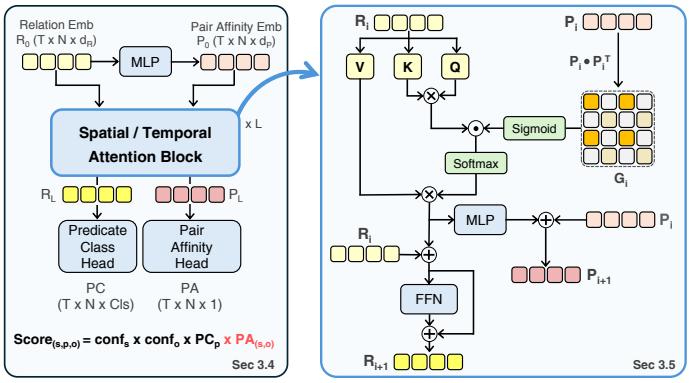

Fig. 4: Overview of PALS and PAM. **(a)** The relation embedding $\mathbf{R}_0$ and pair affinity embedding $\mathbf{P}_0$ are jointly updated through L spatial and temporal attention blocks, then decoded into predicate classification scores PC and pair affinity scores PA for inference. **(b)** Inside each attention block, the affinity matrix $\mathbf{G}_i = \mathbf{P}_i\mathbf{P}_i^\top$ gates the attention logits, and the attention output updates both $\mathbf{R}$ and $\mathbf{P}$ via residual connections. The figure illustrates spatial attention for clarity; temporal attention follows the same mechanism with backbone-specific sequence grouping across frames.

as the visual and semantic cues in $\mathbf{R}_0$ already provide a reasonable prior for estimating initial interaction likelihood.

As illustrated in Figure 4(a), both embeddings are jointly refined through L spatial and temporal attention layers, progressively diverging toward their respective objectives. The two embeddings are decoupled since predicate classification and interaction estimation have fundamentally different characteristics: the former is a fine-grained multi-class problem defined over interacting pairs, while the latter is a binary decision that must also handle the far more numerous non-interactive pairs. This distinction is particularly important under weakly-supervised settings, where limited and noisy supervision makes it difficult for a single embedding to implicitly learn both objectives. Explicit separation allows each embedding to be optimized with its own supervision signal and data distribution, reducing the burden on the model under constrained supervision.

After the final layer, the pair affinity score $\mathrm{PA}_{(s,o)}$ is obtained by projecting the pair affinity embedding through an MLP followed by a sigmoid function:

$$\mathrm{PA}_{(s,o)} = \sigma\Big(\mathrm{MLP}_{\mathrm{PA}}(\mathbf{P}_L^{(s,o)})\Big) \in [0,1], \tag{3}$$

$$\mathcal{L}_{PA} = -\frac{1}{2}\left(\frac{1}{|\mathcal{P}^+|}\sum_{(s,o)\in\mathcal{P}^+}\log \mathrm{PA}_{(s,o)} + \frac{1}{|\mathcal{P}^-|}\sum_{(s,o)\in\mathcal{P}^-}\log(1-\mathrm{PA}_{(s,o)})\right), \tag{4}$$

where $\mathcal{P}^+$ and $\mathcal{P}^-$ denote matched and unmatched pairs after RAM, respectively. Since unmatched pairs vastly outnumber matched ones in WS-VSGG, we adopt

a class-balanced formulation that averages the loss separately over each set to prevent the majority class from dominating training.

At inference, predicted triplets are ranked by a composite score to determine the final scene graph. In conventional WS-VSGG pipelines [1,30], this score relies solely on detection confidence scores conf_s, conf_o and predicate classification score PC_p: $\text{Score}(s, p, o) = \text{conf}_s \cdot \text{conf}_o \cdot \text{PC}_p$. However, this scoring function can penalize genuinely interactive pairs. Objects involved in interactions are often partially occluded or motion-blurred, resulting in lower detection confidence that allows clearly visible but non-interactive background objects to outscore them. To explicitly account for interaction likelihood in the ranking, we incorporate the pair affinity score $\text{PA}_{(s,o)}$ as an additional multiplicative factor:

$$\text{Score}(s, p, o) = \text{conf}_s \cdot \text{conf}_o \cdot \text{PC}_p \cdot \text{PA}_{(s,o)}, \tag{5}$$

This ensures that non-interactive pairs are suppressed in the final ranking regardless of their detection confidence or predicate scores, directly addressing the core limitation of existing pipelines.

3.5 Pair Affinity Modulation

While PALS addresses the structural limitation of the training pipeline, the dominance of non-interactive pairs in the detection space also affects the internal reasoning of the relation prediction model. Modern VSGG architectures [3,4,18,9] perform contextual reasoning through attention layers, where each pair attends to other pairs in the sequence to refine its representation. Attending to other interactive pairs provides complementary relational cues beneficial for predicate determination, but the current detection distribution severely limits such interactions in both spatial and temporal attention. In temporal attention, this effect is further compounded as non-interactive pairs from adjacent frames propagate uninformative context across time. To this end, we propose Pair Affinity Modulation (PAM), which modulates attention based on pair affinity embeddings.

Within each attention block at layer i, we construct a pair-wise affinity matrix $\mathbf{G}_i = \mathbf{P}_i \mathbf{P}_i^\top$ over the attention sequence, where $\mathbf{P}_i$ follows the same sequence composition as $\mathbf{R}_i$. The composition depends on the attention type: spatial attention operates over pairs within a single frame, while temporal attention operates over pairs across multiple frames, with the specific grouping determined by the backbone architecture [3,4]. Since $\mathbf{G}_i$ and the attention logits share the same sequence structure, each element of $\mathbf{G}_i$ directly corresponds to an attention weight between two pairs, enabling element-wise gating. We use $\mathbf{G}_i$ to gate the attention logits via element-wise multiplication with the sigmoid $\sigma(\mathbf{G}_i)$, where $\mathbf{Q}_i$, $\mathbf{K}_i$, $\mathbf{V}_i$ are the query, key, and value projections of the relation embedding $\mathbf{R}_i$:

$$\text{Attn}_i = \text{Softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^\top}{\sqrt{d_k}} \odot \sigma(\mathbf{G}_i)\right) \mathbf{V}_i, \tag{6}$$

suppressing information flow from low-affinity pairs and redistributing attention mass toward interactive pairs. By filtering noise before it propagates temporally,

PAM helps maintain consistent relation predictions across frames. The attention output updates the relation embedding $\mathbf{R}$ via FFN and the pair affinity embedding $\mathbf{P}$ via projection MLP, both with residual connections. While the gating mechanism modulates attention at every layer, $\mathbf{G}_i$ itself receives no direct supervision. To ensure that the final affinity matrix preserves a clear distinction between interactive and non-interactive pairs, we apply triplet ranking loss to $\mathbf{G}_L$:

$$\mathcal{L}_{\text{PAM}} = \sum_{(a,b^+,b^-)} \max\Big(0,\ \mathbf{G}_L^{(a,b^-)} - \mathbf{G}_L^{(a,b^+)} + m\Big)\,, \tag{7}$$

where (a, b^+) are distinct pairs both matched to annotated relations, b^- is unmatched, and m is a margin. This encourages interactive pairs to attend more strongly to each other than to non-interactive ones, enforcing relative ordering without constraining absolute affinity values.

The overall training objective combines relation classification with pair affinity losses: $\mathcal{L} = \mathcal{L}_{\text{rel}} + \lambda_{\text{PA}}\mathcal{L}_{\text{PA}} + \lambda_{\text{PAM}}\mathcal{L}_{\text{PAM}}$, where $\mathcal{L}_{\text{rel}}$ is the standard relation classification loss applied only to $\mathcal{P}^+$, following the backbone architecture [3,4]. Since PAM operates solely by gating attention logits, it is applicable as a plug-and-play module to any attention-based relation prediction architecture, as we verify with both STTran [3] and DSG-DETR [4] in our experiments.

4 Experiments

4.1 Dataset and Implementation Details

We evaluate on Action Genome (AG) [7], which provides frame-level scene graph annotations for 234,253 frames from 9,201 videos, covering 36 object categories and 25 predicate categories. Following prior works [1,30,10,19,28], we use 7,464 videos for training and 1,737 for testing, with only the unlocalized triplets of the middle frame as supervision. We adopt the Scene Graph Detection (SGDet) protocol, which jointly detects objects and predicts their relations from raw frames without ground-truth bounding boxes. We report Recall@K ($K \in \{10, 20, 50\}$) under two settings: *With Constraint*, where only the top-scoring predicate is retained for each subject-object pair, and *No Constraint*, where multiple predicates per pair are permitted.

We follow the two-stage training pipeline of PLA [1] and use VinVL [32], pretrained on four public object detection datasets (COCO [14], OpenImages [12], Objects365 [23], and Visual Genome [11]), as the off-the-shelf detector. For the relation prediction backbone, we experiment with both STTran [3] and DSG-DETR [4]. For RAM, we use GroundingDINO [15] with a Swin-B [16] backbone, with reliability threshold $\tau_r = 0.3$ and grounding score threshold $\tau_{gs} = 0.2$. All experiments are conducted on a single NVIDIA RTX 3090 GPU. Further hyperparameter settings and training configurations are provided in the Supplementary Material, **Sec. A.3**.

Table 1: Comparison with state-of-the-art methods on the AG dataset (SGDet). Methods are grouped by backbone and supervision level: *Zero-shot* denotes methods without any task-specific training, and *Weak* denotes training with unlocalized scene graph annotations. † indicates results reproduced with official code under identical settings. Best and second-best results among weakly-supervised methods within each backbone are in **bold** and underlined, respectively.

Backbone	Supervision	Method	With Constraint			No Constraint		
			R@10	R@20	R@50	R@10	R@20	R@50
RLIPv2 [31]	Zero-shot	Vanilla	5.06	8.37	10.05	5.98	14.60	21.42
STTran [3]	Full	Vanilla	25.20	34.10	37.00	24.60	36.20	48.80
	Weak	PLA [1]	15.39	21.44	26.24	15.83	22.83	31.74
		PLA + **Ours**	**22.24**	**26.48**	**28.00**	**23.20**	**30.24**	**37.47**
		TRKT [30]	17.56	22.33	27.45	18.76	24.49	33.92
		TRKT†	15.11	20.63	26.02	15.73	22.64	31.23
		TRKT† + **Ours**	19.63	24.29	27.55	21.13	27.95	35.89
DSG-DETR [4]	Full	Vanilla	30.30	34.80	36.10	32.10	40.90	48.30
	Weak	PLA	15.47	21.33	25.86	15.66	22.71	31.90
		PLA + **Ours**	**21.88**	**25.92**	27.41	**23.13**	**30.34**	**37.51**
		TRKT†	15.23	20.10	25.92	15.60	22.09	31.18
		TRKT† + **Ours**	19.32	24.81	**27.86**	21.08	28.14	35.96

4.2 Comparison with State-of-the-Art

Table 1 compares our method with existing approaches on the AG dataset under the SGDet protocol. We apply our full pipeline on top of two WS-VSGG baselines, PLA [1] and TRKT [30], using two relation prediction backbones, STTran [3] and DSG-DETR [4]. We additionally include RLIPv2 [31], an image-level relation detection model applied in a zero-shot manner, and the fully-supervised vanilla models as reference points, where the latter serves as an upper bound for each backbone.

Our method yields consistent improvements across all four baseline–backbone combinations, with average gains of +5.47 in R@10 under With Constraint and +6.43 under No Constraint. Improvements are consistent across all K values, as shown in Table 1. Notably, our best configuration (PLA + Ours on STTran) achieves 88.3% and 94.3% of the fully-supervised upper bound at R@10 under the With-Constraint and No-Constraint protocols, respectively, while relying only on single-frame unlocalized annotations without bounding box supervision. These results substantially narrow the gap between weak and full supervision, highlighting the importance of explicitly modeling pairwise interactivity in the weakly-supervised regime. Beyond AG, we further validate our framework on VidHOI [2], a larger-vocabulary dataset (78 object and 50 predicate categories),

Table 2: Ablation study on the AG dataset (SGDet). Best results are in **bold** and second best are underlined.

	RAM	PALS	PAM	With Constraint			No Constraint		
				R@10	R@20	R@50	R@10	R@20	R@50
(a)				15.39	21.44	26.24	15.83	22.83	31.74
(b)		✓		19.79	24.00	26.34	21.03	26.94	33.63
(c)		✓	✓	20.22	24.48	26.83	21.63	27.84	34.73
(d)	✓			15.90	21.87	26.63	16.46	23.33	32.36
(e)	✓	✓		22.14	25.74	27.07	23.12	29.94	36.35
(f)	✓	✓	✓	**22.24**	**26.48**	**28.00**	**23.20**	**30.24**	**37.47**

where it exhibits the same consistent gains; we defer the full results to the Supplementary Material, **Sec. C**.

4.3 Ablation Study

To analyze the contribution of each component, we conduct an ablation study by incrementally adding RAM, PALS, and PAM to the baseline. We adopt PLA [1] as our baseline (row (a)), which uses VinVL [32] as the off-the-shelf detector and STTran [3] as the relation prediction model. Since PAM modulates attention using the pair affinity embedding that PALS introduces, PAM requires PALS as a prerequisite. When PALS is applied without PAM, the pair affinity score is learned and used only for inference-time ranking, while the attention modulation and the $\mathcal{L}_{\text{PAM}}$ loss are disabled. Table 2 summarizes the results.

Effect of RAM. RAM alone yields only a modest improvement (row (a)→(d)); although it reduces label noise, it simultaneously reduces the number of matched pairs while the model still trains exclusively on them. However, RAM substantially amplifies PALS (row (b)→(e)), demonstrating that its primary value lies in providing cleaner supervision for pair affinity learning.

Effect of PALS. Adding PALS yields the largest individual gain (row (a)→(b)), confirming that the inability to filter non-interactive pairs is the primary bottleneck in WS-VSGG. When combined with RAM (row (d)→(e)), PALS achieves an even larger gain, as the cleaner matched/unmatched partition directly improves the supervision quality for pair affinity learning.

Effect of PAM. PAM provides consistent improvements over PALS alone (row (b)→(c) and row (e)→(f)), with gains more pronounced at higher K values, indicating that attention modulation improves predicate classification itself by suppressing noisy context from non-interactive pairs. We further verify this through a controlled experiment in the Supplementary Material, **Sec. F**.

Overall. The full model (row (f)) achieves the best performance across all metrics. All three components are synergistic, each yielding progressively larger improvements when the others are present. This interdependence validates our pipeline design, where each stage addresses a distinct weakness of WS-VSGG while reinforcing the others.

Table 3: Effect of RAM on pseudo-label quality. "–" denotes the original matching without RAM. Relative changes are shown in parentheses.

Detection	VL Model	Match	TP	Precision	Recall	F1
PLA	–	16,137	6,480	0.4016	**0.4423**	0.4209
	GLIP	9,109	5,363	0.5888 (+46.6%)	0.3660 (−17.2%)	0.4514 (+7.2%)
	GDINO	8,067	5,853	**0.7255** (+80.6%)	0.3995 (−9.7%)	**0.5153** (+22.4%)
TRKT	–	21,976	7,012	0.3191	**0.4786**	0.3829
	GLIP	11,945	5,472	0.4581 (+43.6%)	0.3735 (−22%)	0.4115 (+7.5%)
	GDINO	10,444	5,846	**0.5597** (+75%)	0.3990 (−17%)	**0.4659** (+22%)

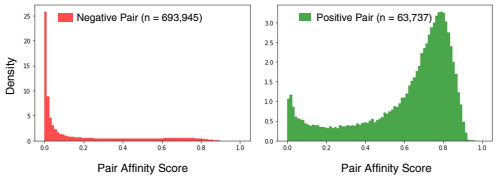

Fig. 5: Pair affinity score distributions on the AG test set. Negative pairs cluster near zero (left); positive pairs peak around 0.7–0.9 (right).

PA	With Constraint R@10	R@20	R@50	No Constraint R@10	R@20	R@50
w/o	17.08	23.04	27.57	16.26	23.10	32.41
w/	**22.24**	**26.48**	**28.00**	**23.20**	**30.24**	**37.47**

Table 4: Effect of pair affinity scoring at inference. Same weights; only scoring is toggled.

4.4 Component Analysis

Pseudo-label refinement via RAM. Table 3 reports the number of generated pseudo-labels (*Match*), ground-truth true positives (*TP*), and the resulting *Precision*, *Recall*, and *F1* across different detection–VL model combinations. While our main pipeline uses GDINO [15], we additionally evaluate with GLIP [13] to verify that RAM is not tied to a specific VL model. RAM consistently improves precision by a large margin regardless of the VL model: GDINO yields +80.6% for PLA and +75% for TRKT, while GLIP yields +46.6% and +43.6%, respectively, with only modest recall trade-offs. Notably, even after refinement, TRKT's best precision (0.56) remains well below PLA's (0.73), which explains the performance gap between TRKT + Ours and PLA + Ours in Table 1. Further analysis on RAM is provided in the Supplementary Material, **Sec. D**.

Effect of pair affinity scoring. Fig. 5 visualizes the learned pair affinity score distributions on the test set, separated by whether each object pair has a ground-truth relation (Positive, IoU > 0.5) or not (Negative). The two distributions are clearly separated, confirming that the model effectively learns to distinguish interactive from non-interactive pairs. Table 4 further quantifies this effect on end-task performance. Using the trained weights of PLA + Ours with the STTran backbone, toggling pair affinity scoring at inference yields consistent gains across all metrics, with the largest improvements at strict budgets: +5.16 in (W)R@10 and +6.94 in (N)R@10.

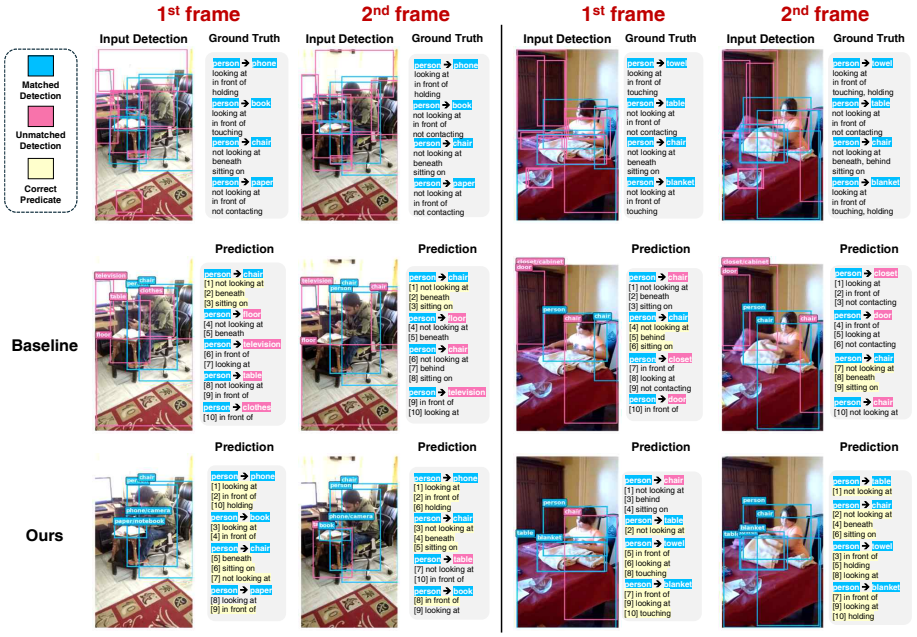

Fig. 6: Qualitative comparison of top-10 predictions on the AG test set. Two video clips are shown, each with two consecutive frames. For each frame, input detections are overlaid, where ground-truth matched detections are shown in blue and unmatched detections in pink. In the prediction lists, predicates are ranked from [1] (highest score) to [10], and correctly predicted predicates are highlighted in yellow. Compared to the baseline (PLA), our method ranks more interaction-relevant object pairs within the top-10, while the baseline frequently assigns high ranks to non-interactive objects (e.g., floor, clothes, closet, door).

4.5 Qualitative Results

Fig. 6 presents qualitative comparisons between PLA (baseline) and PLA + Ours on two consecutive frames from the AG test set. For each frame, we overlay the input detections and visualize the top-10 predicted triplets.

Compared to the baseline, our method consistently ranks triplets associated with matched detections within the top-10 predictions. Even in the presence of numerous noisy detections, our model prioritizes objects that participate in annotated interactions. In contrast, the baseline frequently assigns high ranks to unmatched detections, such as *television* and *clothes* in the left example, or *door* and *closet* in the right example. Moreover, our predictions contain more correctly predicted predicates, indicating improved relational accuracy alongside better object prioritization. These qualitative results demonstrate that our framework effectively suppresses spurious detections and promotes interaction-relevant object pairs, leading to more meaningful scene graph predictions.

5 Conclusion

We presented a model-agnostic framework for weakly-supervised video scene graph generation that addresses the core challenge of distinguishing interactive from non-interactive object pairs. RAM refines pseudo-labels via vision-language grounding, PALS learns a pair affinity score to re-rank predictions at inference, and PAM modulates backbone attention to suppress noisy context from non-interactive pairs. Experiments on Action Genome demonstrate consistent improvements across multiple baselines and backbones, narrowing the gap to full supervision to within 88% and 94% of the upper bound in (W)R@10 and (N)R@10, respectively. We believe the principle of explicitly modeling interactivity between detected objects can benefit broader weakly-supervised settings beyond scene graph generation.

Limitations and future work. The quality of pair affinity supervision is inherently bounded by the matched/unmatched partition. RAM relies on a pretrained VL model, and while the reliability estimation mitigates uncertain localizations, it remains a heuristic proxy. Incorrect grounding can still introduce false positives, while annotation incompleteness may leave false negatives. Developing more principled approaches to grounding reliability remains an important direction for future work.

Acknowledgements. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00340745), and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2026-25510490, Development of an AI-Linked Multimodal Media Authoring Platform Based on Real-Time Collaboration and Lightweight Technologies).

References

1. Chen, S., Xiao, J., Chen, L.: Video scene graph generation from single-frame weak supervision. In: The Eleventh International Conference on Learning Representations (2023)
2. Chiou, M.J., Liao, C.Y., Wang, L.W., Zimmermann, R., Feng, J.: St-hoi: A spatial-temporal baseline for human-object interaction detection in videos. In: Proceedings of the 2021 ACM workshop on intelligent cross-data analysis and retrieval. pp. 9–17 (2021)
3. Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., Yang, M.Y.: Spatial-temporal transformer for dynamic scene graph generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16372–16382 (2021)
4. Feng, S., Mostafa, H., Nassar, M., Majumdar, S., Tripathi, S.: Exploiting long-term dependencies for generating dynamic scene graphs. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5130–5139 (2023)

5. Grunde-McLaughlin, M., Krishna, R., Agrawala, M.: Agqa: A benchmark for compositional spatio-temporal reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11287–11297 (2021)
6. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
7. Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C.: Action genome: Actions as compositions of spatio-temporal scene graphs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10236–10247 (2020)
8. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: CVPR. pp. 1219–1228 (2018)
9. Khandelwal, A.: Flocode: Unbiased dynamic scene graph generation with temporal consistency and correlation debiasing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2516–2526 (2024)
10. Kim, K., Yoon, K., In, Y., Jeon, J., Moon, J., Kim, D., Park, C.: Weakly supervised video scene graph generation via natural language supervision. In: International Conference on Learning Representations (ICLR). pp. 39754–39776 (2025)
11. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**, 32–73 (2017)
12. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**, 1956–1981 (2020)
13. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10975 (2022)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014)
15. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55. Springer (2024)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
17. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European conference on computer vision. pp. 852–869. Springer (2016)
18. Nag, S., Min, K., Tripathi, S., Roy-Chowdhury, A.K.: Unbiased scene graph generation in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22803–22813 (2023)
19. Peddi, R., Dharmavarapu, S.R., Kancheti, S.S., Nakka, K.K., Anandhalli, M., Sridhar, S., Nicholson, A.: Towards open-vocabulary scene graph generation with prompt-based finetuning. In: European Conference on Computer Vision (2024)

20. Rodin, I., Furnari, A., Min, K., Tripathi, S., Farinella, G.M.: Action scene graphs for long-form understanding of egocentric videos. In: CVPR. pp. 18622–18632 (2024)
21. Shang, X., Li, Y., Xiao, J., Ji, W., Chua, T.S.: Video visual relation detection via iterative inference. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3654–3663 (2021)
22. Shang, X., Ren, T., Guo, J., Zhang, H., Chua, T.S.: Video visual relation detection. In: Proceedings of the 25th ACM International Conference on Multimedia. pp. 1300–1308 (2017)
23. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8430–8439 (2019)
24. Shi, J., Zhong, Y., Xu, N., Li, Y., Xu, C.: A simple baseline for weakly-supervised scene graph generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16393–16402 (2021)
25. Teng, Y., Wang, L., Li, Z., Wu, G.: Target adaptive context aggregation for video scene graph generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13688–13697 (2021)
26. Tsai, Y.H.H., Divvala, S., Morency, L.P., Salakhutdinov, R., Farhadi, A.: Video relationship reasoning using gated spatio-temporal energy graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10424–10433 (2019)
27. Wang, G., Li, Z., Chen, Q., Liu, Y.: Oed: Towards one-stage end-to-end dynamic scene graph generation. In: CVPR. pp. 27938–27947 (2024)
28. Wang, S., Gao, L., Lyu, X., Guo, Y., Zeng, P., Song, J.: Dynamic scene graph generation via temporal prior inference. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5793–5801 (2022)
29. Xiao, J., Zhou, P., Chua, T.S., Yan, S.: Video graph transformer for video question answering. In: European Conference on Computer Vision. pp. 39–58. Springer (2022)
30. Xu, Z., Lei, T., Li, Z., Wang, G., Chen, Q., Peng, Y., Liu, Y.: Trkt: Weakly supervised dynamic scene graph generation with temporal-enhanced relation-aware knowledge transferring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15812–15821 (2025)
31. Yuan, H., Zhang, S., Wang, X., Albanie, S., Pan, Y., Feng, T., Jiang, J., Ni, D., Zhang, Y., Zhao, D.: Rlipv2: Fast scaling of relational language-image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21649–21661 (2023)
32. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5579–5588 (2021)
33. Zhong, Y., Shi, J., Yang, J., Xu, C., Li, Y.: Learning to generate scene graph from natural language supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1823–1834 (2021)