DIFFVSGG: Diffusion-Driven Online Video Scene Graph Generation

Mu Chen[1], Liulei Li[2], Wenguan Wang[1]†, Yi Yang[1]
[1] ReLER, CCAI, Zhejiang University [2] ReLER, AAII, University of Technology Sydney
https://github.com/kagawa588/DiffVsgg

Abstract

Top-leading solutions for Video Scene Graph Generation (VSGG) typically adopt an offline pipeline. Though demonstrating promising performance, they remain unable to handle real-time video streams and consume large GPU memory. Moreover, these approaches fall short in temporal reasoning, merely aggregating frame-level predictions over a temporal context. In response, we introduce DIFFVSGG*, an online VSGG solution that frames this task as an iterative scene graph update problem. Drawing inspiration from Latent Diffusion Models (LDMs) which generate images via denoising a latent feature embedding, we unify the decoding of object classification, bounding box regression, and graph generation three tasks using one shared feature embedding. Then, given an embedding containing unified features of object pairs, we conduct a step-wise Denoising on it within LDMs, so as to deliver a clean embedding which clearly indicates the relationships between objects. This embedding then serves as the input to task-specific heads for object classification, scene graph generation, etc.* DIFFVSGG *further facilitates continuous temporal reasoning, where predictions for subsequent frames leverage results of past frames as the conditional inputs of LDMs, to guide the reverse diffusion process for current frames. Extensive experiments on three setups of Action Genome demonstrate the superiority of* DIFFVSGG.

1. Introduction

Video Scene Graph Generation (VSGG) is receiving growing attention as it benefits a wide range of downstream tasks (*e.g.*, video caption [34, 91], video retrieval [31, 50], and visual question answering [49, 87]). To deliver a holistic understanding of the underlying spatial-temporal dynamics within scenes, this task aims to construct a sequence of directed graphs where nodes represent objects and edges describe inter-object relationships (*a.k.a.*, *predicate*).

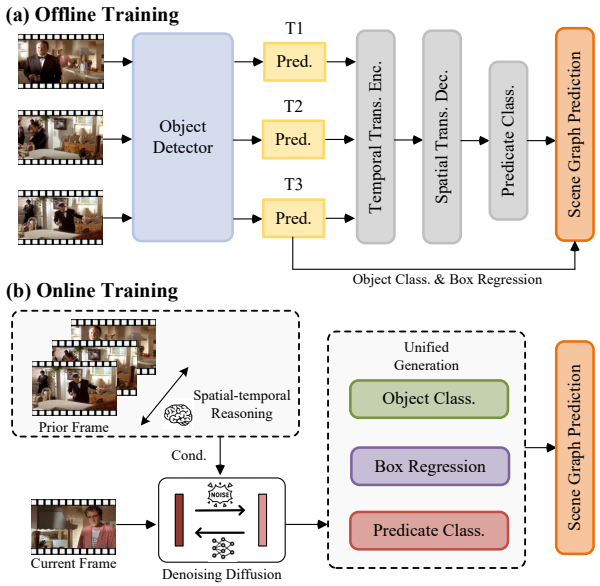

Figure 1. (a) Existing VSGG solutions typically adopt an offline training pipeline, dividing the problem into various components: object detection, temporal association, and contextual aggregation. (b) DIFFVSGG introduces a new paradigm that performs spatial-temporal reasoning directly as each frame is processed sequentially, enabling progressive, online updates to the scene graph.

Top-leading VSGG approaches typically adopt an offline pipeline [30, 69, 94], where scene graphs are generated independently for each frame and then aggregated along the temporal dimension (*i.e.*, Fig. 1(a)). Though demonstrating strong performance, they require full video sequences as inputs, which faces challenges in processing long videos containing hundreds of frames due to GPU memory constraints, and is unable to handle real-time video streams for applications like autonomous driving and augmented reality [100]. Moreover, the dealing of temporal cues focuses solely on the global aggregation of frame-level predictions via Transform blocks. This falls short in the reasoning over temporal space, which is essential for modeling the dynamic changes of interactions between subjects and objects, and

†Corresponding author.

potentially benefits predicate prediction. Though a surge of early work has sought to facilitate explicit reasoning over the temporal domain via messaging passing [90] or spatial-temporal graph [74], a large performance gap remains when compared to these offline temporal aggregation approaches.

To address these challenges, we propose DIFFVSGG, a high performance online approach for VSGG in leverage of Latent Diffusion Models (LDMs) [77, 81] (*i.e.*, Fig. 1(b)). In VSGG, the scene graph is dynamically evolved throughout the progression of video frames, with nodes and edges being continuously updated to precisely reflect the latest video content. Such an iterative update process shares a similar spirit of LDMs, which progressively remove noise to generate new samples from data distributions. Naturally, our motivation is to develop a VSGG model that aligns with the denoising principle of LDMs to iterative refine scene graphs along the temporal dimension, facilitating online temporal reasoning within this reverse diffusion process.

To achieve this, we tackle VSGG from a unified perspective, organizing the decoding of object classification, bounding box regression, and graph generation three tasks from a shared feature embedding, which serves as the input and output of the Denoising U-Net. Concretely, from the *spatial perspective*, given the object detection results from each frame, multiple object pairs can be constructed. Then, for each object pair, we integrate **i)** the visual feature of two objects, **ii)** the union feature between, and **iii)** the locations of two objects, into a unified embedding. This embedding serves as the input to the denoising U-Net, after steps of denoising, a clean embedding clearly describing the objects as well as their inter-object relationships can be delivered. From the *temporal perspective*, DIFFVSGG conducts frame-by-frame reasoning where the result for each subsequent frame is delivered by iteratively refining the predictions of previous frames via reverse diffusion. Specifically, object positions and contextual information from prior frames are leveraged as conditions to guide the denoising of the shared feature embeddings for current frames. This encourages continuous temporal reasoning as the video progresses, allowing DIFFVSGG to effectively capture long-term spatiotemporal dependencies and adapt to complex motion patterns in an online manner. To further unlock the potential of LDMs, we build a memory bank to store positions for each object. In this way, motion information including acceleration and deceleration of object can be explicitly calculated. These motion cues are integrated into conditional inputs for the reverse diffusion step, which can help infer relationships such as `following` or `approaching` between objects.

DIFFVSGG distinguishes itself in several aspects: **First**, it tackles VSGG in an online manner to continuously address an unlimited number of frames, while being friendly to devices with limited GPU memory. **Second**, the prediction of each video is implemented as a reverse diffusion process, which elegantly encodes spatial and temporal reasoning into the Denoising step. **Third**, the temporal cues (*i.e.*, historic predictions) are propagated into subsequent frames as conditions and participating in the prediction, while prior work simply aggregates predictions of all frames along the temporal dimension. **Fourth**, the learning of object classification, bounding box regression, and graph generation three tasks are jointly optimized with a shared feature embedding. Such a unified learning paradigm allows for the solving of VSGG from a global view, and avoids error made in one task propagating to subsequent tasks. **Fifth**, this unified decoding simplifies the VSGG pipeline and eliminates cumbersome handcrafted modules such as non-maximum suppression (NMS), and entity matching across frames.

To the best of our knowledge, DIFFVSGG is the first work that treats the VSGG task as an iterative denoising problem along the temporal dimension. Extensive experiments on Action Genome (AG) [40] demonstrate the superiority of our proposed method. Notably, DIFFVSGG achieves SOTA results across all three setups and surpasses the top-leading solutions (*e.g.*, DSG-DETR [30]) by **3.3** in terms of R@10, highlighting the great potential of utilizing diffusion models for visual relation understanding in videos.

2. Related Work

Scene Graph Generation (SGG). SGG involves detecting object instances and classifying their pairwise visual relations in an image, which is essential for comprehensive visual tasks [7, 8, 11, 53, 97, 98, 117, 118]. Recent SGG approaches seek to comprehend visual context by aggregating spatial context through various strategies, including explicit message passing [14, 20, 56, 57, 103, 112], graph structure modeling [83, 85, 99, 106, 110], external knowledge integration [3, 14, 33, 35, 109, 111], and transformer-based networks [12, 18, 23, 24, 43, 48, 55, 59, 64, 72, 79].

Extended from SGG, Video Scene Graph Generation (VSGG) aims to ground visual relationships jointly in spatial and temporal dimensions. Prior research primarily concentrates on addressing the long-tail distribution problem observed in prominent benchmarks [40]. Numerous unbiased approaches [16, 21, 46, 51, 54, 68, 86, 105] have been devised to handle infrequent predicate classes arising from this distribution. Furthermore, recent VSGG methods underscore the importance of spatial-temporal learning, and make use of the sequence-processing ability of Transformer [2, 70, 84] to capture temporal continuity. Nonetheless, these approaches often depend on complex post-processing that decouples spatial and temporal learning into two independent steps, and simply emphasizes on the consistent matching between frames.

In contrast, DIFFVSGG handles both spatial and temporal reasoning from a unified perspective, where the reasoning on temporal dimension is naturally achieved via utiliz-

ing the spatial reasoning results of prior frames as the conditions to guide predictions for the current frames.

Diffusion Models Beyond Image Generation. Diffusion Models (DMs) have surpassed many other generative models in image generation [6, 22, 37, 38, 92, 104] by successively denoising images. Beyond image generation, DMs inherently perform implicit discriminative reasoning while generating data, which proves highly effective in visual tasks that require complex relationship modeling and spatiotemporal reasoning. Therefore, a surge of work has adapted generative diffusion models for tasks including image segmentation [1, 4, 5, 15, 32, 44, 93], object detection [13, 76, 102], object tracking [65, 66, 101], and monocular depth estimation [25, 67, 78, 116]. Recent research has also utilizes DMs for more complex tasks that require high-level visual understanding abilities such as visual-linguistic understanding [47], scene generation [39], and human-object interaction detection [41, 52]. The potential of DMs in processing graph data has also been explored, encompassing a range of advanced tasks. This includes early work employing score-based methods for generating permutation-invariant graphs [42, 71], as well as recent approaches focused on enhancing graph neural networks [114, 115] and advancing graph generation techniques [107].

In this work, we borrow the step-wise denoising ability of latent diffusion models to enable spatial reasoning within a single frame. This is achieved by iteratively refining the union embeddings of object pairs, which serve to inform the prediction for bounding boxes and predicate classes, *etc*. On the other hand, temporal reasoning across frames is facilitated via conditional prompting, *i.e*., using predictions of prior frames which contain rich location and inter-object relationship cues to guide the denoising of current frames.

Temporal Reasoning in Videos. Reasoning within the temporal dimension poses significant challenges for achieving high-level comprehension in video-related tasks [9, 10, 27–29, 61, 73, 108]. Recent advancements in video question answering seek to tackle temporal reasoning by employing attention mechanisms [26, 82] or incorporating external memory modules [45, 75] across video frames. In addition, to accurately capture the start, progression, and end of an action, action recognition approaches [89, 95] model complex motion patterns and dependencies in sequence. Video object detection (VOD) [19, 62] focuses on accurately capturing object trajectories, even in the presence of occlusions and abrupt movements. The top-leading VSGG solutions [30, 69, 94] instead aggregate predictions across frames to maintain the temporal consistency of objects.

In summary, although existing work facilitates temporal reasoning from different perspectives and demonstrates it effectiveness, the potential of diffusion models remains largely underexplored. In fact, the sequential denoising in time, where each step is informed by previous one, offers a suitable tool to reconstruct the states for current frames based on observations from past frames. This insight motivates the proposal of DIFFVSGG, which tackles the VSGG task via graph denoising over the temporal dimension.

3. Methodology

In this section, we first give a brief introduction to the general background of latent Diffusion Models (§3.1), then elaborate on the overall design of our proposed DIFFVSGG (§3.2), and finally present the detailed information on the network architecture and training objectives (§3.3).

3.1. Preliminary: Latent Diffusion Models

To begin, we give an illustration on how diffusion processes [81] are used to model data distributions and generate high-quality samples. Specifically, we consider a continuous-time Markov chain with $t \in [0, T]$:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \boldsymbol{I}), \tag{1}$$

where β_t is a small noise variance term at timestep t, and $\mathcal{N}$ denotes a Gaussian distribution. The corrupted data at any intermediate time step t can be derived recursively as:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}}\epsilon, \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative product of α_t over the timesteps, $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ represents the sampled noise. The reverse process essentially involves learning to reverse the noise addition at each step. The cleaned data at time step t can be written recursively as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}}}\epsilon_\theta(x_t, t), \tag{3}$$

where a time-conditioned denoising neural network $\epsilon_\theta(x, t)$ is trained to minimize the mean squared error between the true and predicted noise at each timestep:

$$\mathcal{L}_{\text{DM}} := \mathbb{E}_{x_0,\epsilon,t}\Big[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\Big]. \tag{4}$$

Building on the diffusion process described above, latent diffusion models (LDMs) [77] first encode samples into a low-dimensional latent space $z = \mathcal{E}(x)$ using an encoder $\mathcal{E}$, and then apply the diffusion process within this compressed space. Moreover, LDMs introduce the conditioning mechanisms which allow for control over the generated output based on additional input y such as text, labels, or images. Consequently, the training objective for LDMs is given as:

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{\mathcal{E}(x_0),y,\epsilon,t}\Big[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t, c_\theta(y))\|_2^2\Big], \tag{5}$$

where c_θ is a conditioning model to encode y.

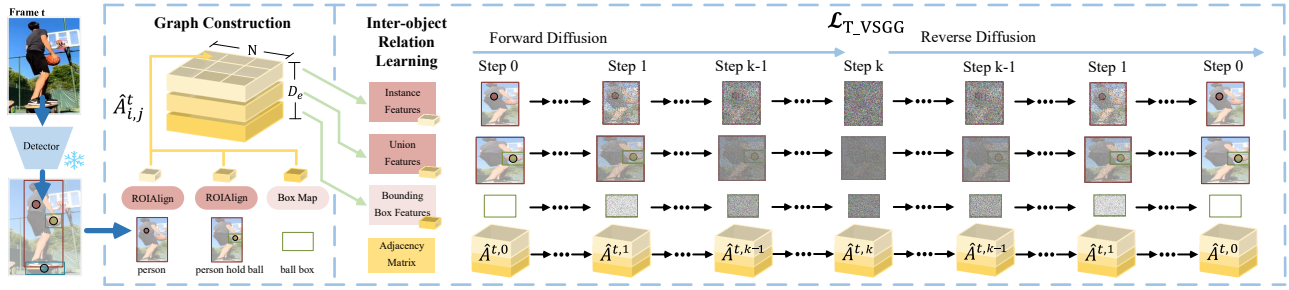

Figure 2. Overview of our proposed inter-object relationship learning strategy using latent diffusion models.

3.2. DIFFVSGG: Diffusion Models for VSGG

Problem Definition. Considering a video sequence represented as $\mathcal{I} = \{I^1, \cdots, I^T\}$ containing T frames, the objective of VSGG is to generate a sequence of scene graph $\mathcal{G} = \{G^1, \cdots, G^T\}$ over time, where each element $\mathcal{G}^t$ represents the corresponding scene graph for frame I^t. Each graph G^t is defined as $G^t = (\mathcal{V}^t, \mathcal{E}^t)$, with $\mathcal{V}^t$ and $\mathcal{E}^t$ denoting the sets of graph nodes and edges, respectively. Each node $v_i^t \in \mathcal{V}^t$ includes attributes such as category and location for object i, while each edge $e_{i,j}^t \in \mathcal{E}^t$ describes the inter-object relationship (*a.k.a.*predicate) between subject i and object j. In this manner, $\mathcal{G}$ captures all objects (*i.e.*, $\mathcal{V}$), their interactions (*i.e.*, $\mathcal{E}$), and the dynamics as they evolve over time (*i.e.*, from $t = 1$ to $t = T$) within the scene.

Graph Construction. Given a video $\mathcal{I} = \{I^1, \cdots, I^T\}$, we first employ an on-the-shelf object detector $\mathcal{F}_{\text{det}}$ for each frame:

$$\{\boldsymbol{F}^t, \mathcal{B}^t, \mathcal{O}^t\} = \mathcal{F}_{\text{det}}(I^t), \tag{6}$$

where $\boldsymbol{F}^t$ is the feature extracted by backbone, $\mathcal{B}^t = \{b_1^t, \cdots, b_{N_t}^t\}$ and $\mathcal{O}^t = \{o_1^t, \cdots, o_{N_t}^t\}$ are bounding box and class predictions for N_t objects detected from frame I^t, respectively. This serves to initialize an adjacency matrix $\boldsymbol{A}^t \in \mathbb{R}^{N_t \times N_t \times D_e}$, of which each element aims to represent the inter-object relationships between object i and j:

$$\begin{aligned} \boldsymbol{A}_{i,j}^t &= [\boldsymbol{F}_{o_i}^t; \boldsymbol{F}_{o_i,o_j}^t; \boldsymbol{F}_{b_i}^t] \in \mathbb{R}^{D_e}, \\ \boldsymbol{A}_{j,i}^t &= [\boldsymbol{F}_{o_j}^t; \boldsymbol{F}_{o_j,o_i}^t; \boldsymbol{F}_{b_j}^t] \in \mathbb{R}^{D_e}. \end{aligned} \tag{7}$$

Here [] refers to the concatenation of features, implemented as `torch.cat()`. $\boldsymbol{F}_{o_i}^t = \mathcal{F}_{\text{ROI}}(\boldsymbol{F}^t, b_i^t)$ is the instance-level feature for object i extracted via ROIAlign [36] (*i.e.*, $\mathcal{F}_{\text{ROI}}$), $\boldsymbol{F}_{o_i,o_j}^t = \mathcal{F}_{\text{ROI}}(\boldsymbol{F}^t, b_i^t \cup b_j^t)$ is the feature mapped from the union box of object i and j (*i.e.*, $b_i^t \cup b_j^t$), and $\boldsymbol{F}_{b_i}^t$ describes bounding box b_i^t using a box-to-feature mapping function identical to that in [112]. $\boldsymbol{A}_{i,j}^t$ represents the relationship predicted from subject i, which differs from $\boldsymbol{A}_{j,i}^t$ where j is considered the subject. In this way, the matrix $\boldsymbol{A}^t$ encodes both node features of each object and the edge features between any pair of them. Note that such subject-based encoding does not incorporate $\boldsymbol{F}o_j$ and $\boldsymbol{F}b_j$ into $\boldsymbol{A}_{i,j}$ (*i.e.*, $[\boldsymbol{F}_{o_i}; \boldsymbol{F}_{b_i}; \boldsymbol{F}_{o_i,o_j}; \boldsymbol{F}_{o_j}; \boldsymbol{F}_{b_j}]$), which hinders the neural network to distinguish between the subject (i) and object (j). Additionally, as the number of instances (*i.e.*, N_t) typically varies across frames due to the emergence or disappearance of some instances, we pad $\boldsymbol{A}^t$ to a fixed size of $N \times N$ where $N > \max(N_1, \cdots, N_T)$, using randomly generated feature embedding from Gaussian distributions.

Inter-object Relationship Learning via LDMs. Next we aim to facilitate the learning of inter-object relationships within LDMs using ground truth scene graph (*i.e.*, Fig. 2). Since we use an off-the-self object detector, the extracted feature $\boldsymbol{F}^t$ remains static for each frame. Therefore, with access to the ground truth bounding box annotations, we can compute precise edge features via Eq. 7 from $\boldsymbol{F}^t$. The adjacency matrix $\hat{\boldsymbol{A}}^t$ which exactly encodes the inter-object features can also be obtained. Here $\hat{\boldsymbol{A}}^t$ is the ground truth of inter-object relationships, initialized using features derived from ground-truth bounding boxes, where object pairs with no relations are represented as empty entries. $\hat{\boldsymbol{A}}^t$ instructs the learning of a LDM which is responsible for recovering object features and their relationships from random noise.

• ***Forward Process.*** Let $\hat{\boldsymbol{A}}^{t,0}$ represent the clean adjacency matrix. The noise injection process to progressively perturb this matrix is defined following Eq. 2, and expressed as:

$$\hat{\boldsymbol{A}}^{t,k} = \sqrt{\overline{\alpha}_t} x \hat{\boldsymbol{A}}^{t,0} + \sqrt{1 - \overline{\alpha}} \epsilon, \tag{8}$$

where $\hat{\boldsymbol{A}}^{t,k}$ denotes the noisy adjacency matrix at step k. This enables the model to learn robust feature representations, through the exposure to degraded versions of $\hat{\boldsymbol{A}}^{t,0}$.

• ***Reverse Process.*** To recover the original adjacency matrix $\hat{\boldsymbol{A}}^{t,0}$, we employ a denoising U-Net ϵ_θ which is trained to iteratively remove noise starting from the initial noisy matrix $\hat{\boldsymbol{A}}^{t,K}$. This process follows the denoising step defined in Eq. 3 and proceeds as:

$$\hat{\boldsymbol{A}}^{t,k-1} = \frac{1}{\sqrt{\alpha_t}} (\hat{\boldsymbol{A}}^{t,k} - \frac{\beta_t}{\sqrt{1 - \overline{\alpha}}} \epsilon_\theta(\hat{\boldsymbol{A}}^{t,k}, k), \tag{9}$$

where ϵ_θ is optimized to recover meaningful inter-object relationships. Following Eq. 5, we define the training objective $\mathcal{L}_{\text{VSGG}}$ by minimizing the spatial structure loss:

$$\mathcal{L}_{\text{VSGG}} := \mathbb{E}_{\hat{\boldsymbol{A}}^{t,k}, \epsilon, k} \left[\|\epsilon - \epsilon_\theta(\hat{\boldsymbol{A}}^{t,k}, k)\|_2^2 \right]. \tag{10}$$

$\mathcal{L}_{\text{VSGG}}$ encourages ϵ_θ to accurately predict and remove the noise at each step k, so as to restore $\hat{\boldsymbol{A}}^t$ structured in the

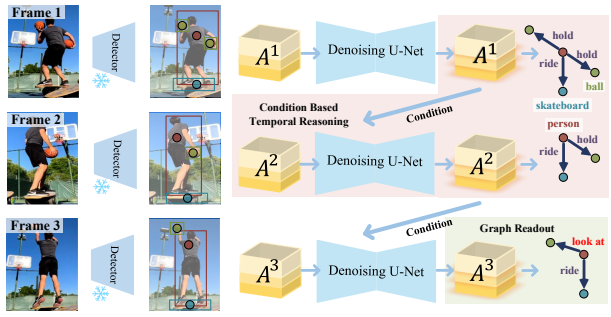

Figure 3. Overview of our proposed temporal prompting strategy.

way that preserves spatial and relational coherence. During inference, given the adjacency matrix $\boldsymbol{A}^t$ from an arbitrary frame t, we consider it noisy and utilize the well-trained denoising U-Net above to refine it, so as to deliver an updated version that indicates the interaction between objects, as well as the location and category of objects: $\boldsymbol{A}^{t,k-1} = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{A}^{t,k} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}}}\epsilon_\theta(\boldsymbol{A}^{t,k}, k)$. Unlike the training process which utilizes $\hat{\boldsymbol{A}}$ derived from ground truth bounding boxes, $\boldsymbol{A}$ here is obtained from the predicated ones. Note that the main challenge of VSGG lies in modeling of the dynamic changes of relation between object and subject, and the diffusion models primarily focus on recovering object features and their relationships. Thus, the slight bias in predicated bounding boxes, especially after ROIAlign, causes negligible impacts to relation predictions.

Condition Based Temporal Reasoning. For a given video stream, we can infer future frames based on the content of preceding frames. This observation serves as the motivation for our temporal prompting strategy (*i.e.*, Fig. 3), where we condition the denoising process of $\boldsymbol{A}^t$ on the denoised adjacency matrix of prior frames (*i.e.*, $\boldsymbol{A}^{t-1,0}$), as follows:

$$\boldsymbol{A}^{t,k-1} = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{A}^{t,k} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}}}\epsilon_\theta(\boldsymbol{A}^{t,k}, k, \boldsymbol{A}^{t-1,0}), \quad (11)$$

where the condition decoder c_θ in Eq. 5 is discarded, as $\boldsymbol{A}^{t-1,0}$ has already shared the same dimension as $\boldsymbol{A}^{t,k}$. With this condition-based temporal association for denoising, the training objective for ϵ_θ is updated as:

$$\mathcal{L}_{\text{T\_VSGG}} := \mathbb{E}_{\boldsymbol{A}^{t,k}, \boldsymbol{A}^{t-1,0}, \epsilon, k}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{A}^{t,k}, \boldsymbol{A}^{t-1,0}, k)\|_2^2\right]. \quad (12)$$

Here $\mathcal{L}_{\text{T\_VSGG}}$ encourages ϵ_θ to generate temporally consistent outputs, enabling a smooth and coherent denoising of adjacency matrix across frames. Moreover, since new instances keep emerging as the video progresses, we calculate similarities between ROIAligned features of the potential new object (*i.e.*, $\boldsymbol{F}^t_{o_p}$) and all other objects in the past frame (*i.e.*, $\boldsymbol{F}^{t-1}_{o_i}$), and consider o_p as new object if the similarities is smaller than 0.2 for all other objects. Then, the random padded features in $\boldsymbol{A}$ during graph construction are replaced with union and instance features of these new instances.

Motion Enhanced Denoising for VSGG. Motion cues serve as indicators to describe object movements in a given scene. They provide essential context for understanding the position and duration of events. Moreover, the speed and direction of object movements can reveal intentions for interactions (*e.g.*, approaching), which aids in temporal reasoning by inferring whether they might complete an action or reach a goal. Therefore, given a box prediction $b_i^t = \{x_i^t, y_i^t, w_i^t, h_i^t\}$ for object i, we explicitly calculate the approaching speed between object i and j as follows:

$$\begin{aligned} d_{i,j}^t &= \sqrt{(x_i^t - x_j^t)^2 + (y_i^t - y_j^t)^2}, \\ v_{i,j}^t &= (d_{i,j}^t - d_{i,j}^{t-\Delta t})/\Delta t, \end{aligned} \quad (13)$$

where Δt is the interval of frame. Given $v^t \in \mathbb{R}^{N^t \times N^t}$, we pad it into the same size with $\boldsymbol{A}^t$ (*i.e.*, $v^t \in \mathbb{R}^{N \times N}$) and inject it into $\boldsymbol{A}^{t,k}$ at each denoising step k: $\boldsymbol{A}^{t,k} = \boldsymbol{A}^{t,k} + v^t$.

Graph Readout. Next we investigate how to deliver SGG predictions for each frame based on the denoised adjacency matrix $\boldsymbol{A}^{t,0}$. Specifically, the prediction for predicate of each subject-object pair is obtained as follows:

$$r_{i,j}^t = \text{softmax}(\mathcal{F}_{\text{pred}}(\boldsymbol{A}_{i,j}^{t,0})) \in \mathbb{R}^{N_{\text{pred\_cls}}} \quad (14)$$

where $\mathcal{F}_{\text{pred}}$ is an MLP-based classifier and $N_{\text{pred\_cls}}$ represents the number of predicate classes. For object classification and bounding box regression, to render a global view, we utilize the entire row i of $\boldsymbol{A}^{t,0}$ (*i.e.*, $\boldsymbol{A}_i^{t,0}$) where all elements consider i as the subject, to deliver the prediction:

$$o_i^t = \text{softmax}(\mathcal{F}_{\text{obj}}(\boldsymbol{A}_i^{t,0})) \in \mathbb{R}^{N_{\text{obj\_cls}}}, \quad (15)$$

$$b_i^t = \text{sigmoid}(\mathcal{F}_{\text{box}}(\boldsymbol{A}_i^{t,0})) \in \mathbb{R}^4, \quad (16)$$

where $N_{\text{obj\_cls}}$ is the number of object classes, $\mathcal{F}_{\text{obj}}$ and $\mathcal{F}_{\text{box}}$ are the MLP-based classifier and projector, respectively.

3.3. Implementation Details

Network Configuration. DIFFVSGG is an online VSGG framework built upon the iterative diffusion diagram. It comprises three components: one off-the-shelf detector to deliver object detection results, the LDMs for inter-object relationship denosing, and projector heads for object classification, predicate prediction, and bounding box regression.

Training Objective. The training process of DIFFVSGG consists of two stages. In the first stage, we pre-train the denoising U-Net ϵ_θ in Eq. 9, using $\hat{\boldsymbol{A}}^t$ constructed from ground truth bounding box annotations for objects. In the second stage, we optimize the MLP classifier and projector in Eq. 14-15, which generate the final VSGG predictions.

Specifically, given relation prediction p_r and ground truth y_r, the relation classification loss is computed as:

$$\mathcal{L}_{\text{pred\_cls}} = -\sum_{i=1}^{C_r} y_{r,i} \log(p_{r,i}). \quad (17)$$

Method	PredCLS						SGCLS						SGDET					
	R@10	R@20	R@50	mR@10	mR@20	mR@50	R@10	R@20	R@50	mR@10	mR@20	mR@50	R@10	R@20	R@50	mR@10	mR@20	mR@50
RelDN [113]	20.3	20.3	20.3	6.2	6.2	6.2	11.0	11.0	11.0	3.4	3.4	3.4	9.1	9.1	9.1	3.3	3.3	3.3
TRACE [88]	27.5	27.5	27.5	15.2	15.2	15.2	14.8	14.8	14.8	8.9	8.9	8.9	13.9	14.5	14.5	8.2	8.2	8.2
VRD [63]	51.7	54.7	54.7	-	-	-	32.4	33.3	33.3	-	-	-	19.2	24.5	26.0	-	-	-
Motif Freq [112]	62.4	65.1	65.1	-	-	-	40.8	41.9	41.9	-	-	-	23.7	31.4	33.3	-	-	-
MSDN [57]	65.5	68.5	68.5	-	-	-	43.9	45.1	45.1	-	-	-	24.1	32.4	34.5	-	-	-
VCTREE [85]	66.0	69.3	69.3	-	-	-	44.1	45.3	45.3	-	-	-	24.4	32.6	34.7	-	-	-
GPS-Net [60]	66.8	69.9	69.9	-	-	-	45.3	46.5	46.5	-	-	-	24.7	33.1	35.1	-	-	-
STTran [17]	68.6	71.8	71.8	37.8	40.1	40.2	46.4	47.5	47.5	27.2	28.0	28.0	25.2	34.1	37.0	16.6	20.8	22.2
APT [58]	69.4	73.8	73.8	-	-	-	47.2	48.9	48.9	-	-	-	26.3	36.1	38.3	-	-	-
STTran-TPI [96]	69.7	72.6	72.6	37.3	40.6	40.6	47.2	48.3	48.3	28.3	29.3	29.3	26.2	34.6	37.4	15.6	20.2	21.8
TR2 [94]	70.9	73.8	73.8	-	-	-	47.7	48.7	48.7	-	-	-	26.8	35.5	38.3	-	-	-
TEMPURA [69]	68.8	71.5	71.5	42.9	46.3	46.3	47.2	48.3	48.3	34.0	35.2	35.2	28.1	33.4	34.9	18.5	22.6	23.7
DSG-DETR [30]	-	-	-	-	-	-	50.8	52.0	52.0	-	-	-	30.3	34.8	36.1	-	-	-
DIFFVSGG	**71.9**	**74.5**	**74.5**	**48.1**	**50.2**	**50.2**	**52.5**	**53.7**	**53.7**	**37.3**	**38.4**	**38.4**	**32.8**	**39.9**	**45.5**	**20.9**	**23.6**	**26.2**

Table 1. Comparison of state-of-the-art VSGG methods on Action Genome `test` [40] under the ***w* constraint** setting.

Similarly, given object class prediction p_o and ground truth y_o, the object classification loss is defined as:

$$\mathcal{L}_{\text{obj\_cls}} = -\sum_{j=1}^{C_o} y_{o,j} \log(p_{o,j}). \tag{18}$$

For bounding box regression, given the ground truth bounding box t_k (*e.g.*, (x, y, w, h)) and the predicted bounding box $\hat{t}_k$, the regression loss is formulated as:

$$\mathcal{L}_{\text{box\_reg}} = \text{Smooth L1}(t_k - \hat{t}_k). \tag{19}$$

The training objectives for each stage are formulated as:

$$\begin{aligned} &\text{Stage 1}: \quad \mathcal{L} = \mathcal{L}_{\text{T\_VSGG}}, \\ &\text{Stage 2}: \quad \mathcal{L} = \mathcal{L}_{\text{pred\_cls}} + \mathcal{L}_{\text{obj\_cls}} + 0.5\mathcal{L}_{\text{box\_reg}}, \end{aligned} \tag{20}$$

4. Experiment

4.1. Experimental setting

Dataset. We evaluate DIFFVSGG on Action Genome (AG) [40], the largest video scene graph generation dataset comprising over 10K videos extended from the Charades dataset [80]. This dataset includes 1,715,568 predicate instances across 25 predicate classes, spanning 234,253 video frames, and features 476,229 bounding boxes across 35 object categories. The 1,715,568 instances for 25 predicate classes are divided into attention, spatial and contacting, three different types. Following prior work [17, 58], we use the same training/testing split.

Training. In the first training stage which optimizes the LDMs with ground-truth bounding box annotations, we set the learning rate to 10^{-4} and use the Adam optimizer. The batch size is set to 2048, and the model is trained for 100 epochs. Each input clip consists of five frames sampled at random time intervals, enabling conditional temporal reasoning. The denoising U-Net ϵ_θ remains frozen once after pre-training. In the second training stage, the classifiers and projectors are trained for 10 epochs with a batch size of 8. We use the AdamW optimizer with a learning rate of 10^{-5}, which decays by a factor of 5 halfway through the training.

Evaluation Setup. In line with previous studies[30, 40, 58, 63], three standard evaluation protocols are adopted:

- **PredCLS:** With oracle-provided object labels and bounding boxes as well as grounding-truth subject-object pairs, PredCLS assesses the model capability to predict predicate labels for each subject-object pair.
- **SGCLS:** Building on PredCLS, SGCLS requires the simultaneous prediction of both predicate labels and the associated subject-object pairs for each predicate.
- **SGDET:** As the most challenging task, SGDET requires generating complete scene graphs from scratch, including object detection, subject-object pair selection, and predicate classification. Detection is considered accurate if the overlap between prediction and ground truth exceeds 0.5.

Evaluation Metric. We employ the Recall@k where $k \in 10, 20, 50$ as the evaluation metric, to measure the proportion of ground truth elements within the top-k predictions. Additionally, Mean-Recall@k is also adopted to ensure the evaluation is not biased toward high-frequency classes. The evaluation is performed under two different scenarios:

- ***w* constraint**: Each subject-object pair is restricted to a maximum of one predicate.
- ***w/o* constraints**: Each subject-object pair is allowed to have multiple predicates simultaneously.

4.2. Comparison with State-of-the-arts

Tables 1-2 present the main experimental results of DIFFVSGG against several top-leading approaches on Action Genome `test` [40]. Following existing VSGG works [17, 58], we firstly select a few representative image-level SGG methods such as RelDN [113] and GPS-Net [60]. Then, we compare DIFFVSGG with existing video-level solutions such as STTran[17], APT[58], *etc*. In general, DIFFVSGG outperform image-level SGG methods by a large

Method	PredCLS						SGCLS						SGDET					
	R@10	R@20	R@50	mR@10	mR@20	mR@50	R@10	R@20	R@50	mR@10	mR@20	mR@50	R@10	R@20	R@50	mR@10	mR@20	mR@50
RelDN [113]	44.2	75.4	89.2	31.2	63.1	75.5	25.0	41.9	47.9	18.6	36.9	42.6	13.6	23.0	36.6	7.5	18.8	33.7
VRD [63]	51.7	54.7	54.7	-	-	-	32.4	33.3	33.3	-	-	-	19.2	24.5	26.0	-	-	-
Motif Freq [112]	62.4	65.1	65.1	-	-	-	40.8	41.9	41.9	-	-	-	23.7	31.4	33.3	-	-	-
MSDN [57]	65.5	68.5	68.5	-	-	-	43.9	45.1	45.1	-	-	-	24.1	32.4	34.5	-	-	-
VCTREE [85]	66.0	69.3	69.3	-	-	-	44.1	45.3	45.3	-	-	-	24.4	32.6	34.7	-	-	-
TRACE [88]	72.6	91.6	96.4	50.9	73.6	82.7	37.1	46.7	50.5	31.9	42.7	46.3	26.5	35.6	45.3	22.8	31.3	41.8
GPS-Net [60]	76.0	93.6	99.5	-	-	-	-	-	-	-	-	-	24.5	35.7	47.3	-	-	-
STTran [17]	77.9	94.2	99.1	51.4	67.7	82.7	54.0	63.7	66.4	40.7	50.1	58.5	24.6	36.2	48.8	20.9	29.7	39.2
APT [58]	78.5	95.1	99.2	-	-	-	55.1	65.1	68.7	-	-	-	25.7	37.9	50.1	-	-	-
TR2 [94]	83.1	96.6	99.9	-	-	-	57.2	64.4	66.2	-	-	-	27.8	39.2	50.0	-	-	-
TEMPURA [69]	80.4	94.2	99.4	61.5	85.1	98.0	56.3	64.7	67.9	48.3	61.1	66.4	29.8	38.1	46.4	24.7	33.9	43.7
DSG-DETR [30]	-	-	-	-	-	-	59.2	69.1	72.4	-	-	-	32.1	40.9	48.3	-	-	-
DIFFVSGG	**83.1**	**94.5**	99.1	**66.3**	**90.5**	**98.4**	**60.5**	**70.5**	**74.4**	**51.0**	**64.2**	**68.8**	**35.4**	**42.5**	**51.0**	**27.2**	**37.0**	**45.6**

Table 2. Comparison of state-of-the-art VSGG methods on Action Genome `test` [40] under the ***w/o* constraint** setting.

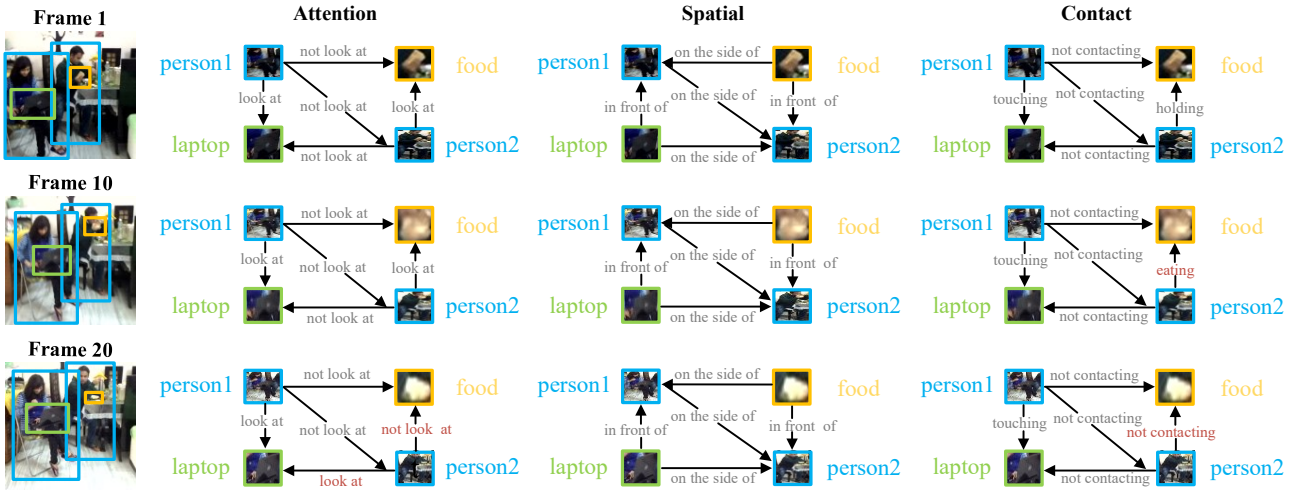

Figure 4. Visualization results on Action Genome `test` [40]. All results are given under the SGDET setup. Predicates in red indicate relationships are transformed to another one.

margin in all metrics. Concretely, under the ***w* constraint** setup, our proposed method achieves R@10/R@20/R50 scores of 32.8/39.9/45.5, surpassing GPS-Net which achieves 24.7/33.1/35.1 scores by 8.1/6.8/10.4 scores, respectively. Similar trends can be observed on the ***w/o* constraint** setup. All of the above firmly demonstrates the effectiveness of incorporating temporal cues to tackle relation understanding in dynamic scene.

Additionally, when compared to the video-level counterparts, DIFFVSGG can still deliver state-of-the-art performance. Specifically, though DSG-DETR[30] adopts a modern DETR-like architecture and uses a Transformer-based decoder with all frames as the inputs to aggregate temporal cues, our DIFFVSGG which conducts online inference still outperforms it by a solid gap (*i.e.*, 35.4/42.5/51.0 *vs*. 32.1/40.9/48.3 under the ***w/o* constraint**). More importantly, when it comes to the comparison of mR metric which prevents the bias toward high-frequency classes, DIFFVSGG obtains significantly higher performance compared to existing work. This suggests that using latent diffusion models (LDMs) to model the inter-object relationship is effective, and it successfully learned meaning distributions which can prevent overfitting to high-frequency classes. Similar conclusions can be drawn on PredCLS and SGCLS two evaluation protocols, where DIFFVSGG can still deliver SOTA performance under both ***w* constraint** and ***w/o* constraint** two setups.

4.3. Qualitative Results

We provide visualization results of UNIALIGN in Fig. 4. It can be seen that our proposed method is able to generate accurate scene graphs across various challenging scenarios, such as fast motion and occlusion.

4.4. Diagnostic Experiment

To assess the effectiveness of the detailed designs of DIFFVSGG and gain deeper insights, we conduct a series of experiments on the AG `test` [40]. All performance metrics are reported under the SGDET setup.

Key Component Analysis. We first examine the efficacy

#	$\mathcal{L}_{\text{VSGG}}$	$\mathcal{L}_{\text{T\_VSGG}}$	Motion	*w* constraints			*w/o* constraints		
				R@10	R@20	R@50	R@10	R@20	R@50
1				26.9	35.3	38.0	27.5	38.9	48.5
2	✓			29.7	37.7	41.5	30.3	40.1	49.9
3		✓		32.0	39.3	44.6	33.9	42.0	50.3
4		✓	✓	**32.8**	**39.9**	**45.5**	**35.4**	**42.5**	**51.0**

Table 3. Analysis on key components of DIFFVSGG.

#	Union Box Feature	Subject Feature	Subject Location	*w* constraints			No Constraints		
				R@10	R@20	R@50	R@10	R@20	R@50
1	✓			29.5	36.7	42.0	32.3	39.2	47.5
2	✓	✓		31.0	38.2	43.6	33.7	40.8	48.7
3	✓	✓	✓	**32.8**	**39.9**	**45.5**	**35.4**	**42.5**	**51.0**

Table 4. Analysis on elements to construct adjacency matrix $\boldsymbol{A}^t$.

Step T	*w* constraints			*w/o* constraints		
	R@10	R@20	R@50	R@10	R@20	R@50
10	31.7	38.5	42.9	34.0	41.1	49.5
20	**32.8**	**39.9**	**45.5**	**35.4**	**42.5**	**51.0**
50	33.2	40.5	46.4	35.7	43.1	51.4
100	33.1	40.3	45.9	35.5	42.8	51.1

Table 5. Analysis on forward and reverse diffusion steps.

# Layer	*w* constraints			*w/o* constraints		
	R@10	R@20	R@50	R@10	R@20	R@50
2	31.9	38.2	44.5	34.0	42.1	50.1
3	**32.8**	**39.9**	**45.5**	**35.4**	**42.5**	**51.0**
5	32.4	39.6	46.1	34.4	43.0	50.7
6	32.6	39.3	46.3	34.7	44.0	50.9

Table 6. Analysis of the layer configuration of the U-Net.

Element	*w* constraints			*w/o* constraints		
	R@10	R@20	R@50	R@10	R@20	R@50
$\boldsymbol{A}_{i}^{t,0}$	**32.8**	**39.9**	**45.5**	**35.4**	**42.5**	**51.0**
$\boldsymbol{A}_{i,j}^{t,0}$	31.7	37.7	43.5	33.6	41.2	49.6

Table 7. Analysis on the elements used for graph readout.

Model	Parameter Number	Inference Time	*w* constraints			*w/o* constraints		
			R@10	R@20	R@50	R@10	R@20	R@50
TRACE [88]	66.7 M	8.4 FPS	13.9	14.5	14.5	26.5	35.6	45.3
STTran [17]	51.1 M	10.9 FPS	25.2	34.1	37.0	24.6	36.2	48.8
TEMPURA [69]	53.5 M	9.6 FPS	28.1	33.4	34.9	29.8	38.1	46.4
DSG-DETR [30]	65.5 M	7.3 FPS	30.3	34.8	36.1	32.1	40.9	48.3
DIFFVSGG	**46.7** M	8.7 FPS	**32.8**	**39.9**	**45.5**	**35.4**	**42.5**	**51.0**

Table 8. Analysis on the trainable parameters and inference time.

of essential components of DIFFVSGG in Table 3, where the first row denotes the baseline model directly using bounding box predictions to construct scene graphs without denoising refinement. After integrating LDMs to model complex inter-object relationships within the scene (*i.e.*, *row* #2), DIFFVSGG enjoys considerable improvement on both w and w/o constraint setups. Next, applying condition based temporal reasoning to capture long-term temporal dependencies further boosts the performance to 32.0 and 33.9 on two setups. Finally, upon the incorporation of motion enhanced denoising (*i.e.*, *row* #4), DIFFVSGG obtains the best performance on both setups, suggesting that motion cues can effectively enhance the awareness of temporal cues through conditional promoting.

Graph Construction. We study the impact of using various features to construct the adjacency matrix $\boldsymbol{A}^t$ in Table 4. As shown, incorporating all three types of features, as described in Eq. 7, achieves the best performance.

Number of Diffusion Step. Next we investigate the effect of different number of diffusion steps. As shown in Table 5, the best performance is achieved at $T = 50$. However, to balance the performance and efficiency efficiency, we set $T = 20$, which yields a slight reduction in performance.

Number of Layer in Denoising U-Net. We further analyze the impact of using different number of layers in the encoder and decoder of denoising U-Net. As shown in Table 6, DIFFVSGG achieves similar performance when the layer number exceeds 2. For efficiency, we set it to 3.

Graph Readout. We explore different graph readout strategies in Table 7. Here $\boldsymbol{A}_{i}^{t,0}$ refers to utilize the entire row i of $\boldsymbol{A}^{t,0}$ to deliver the predictions (*i.e.*, Eq. 15-16), while $\boldsymbol{A}_{i,j}^{t,0}$ denoting utilize one element at (i, j) of $\boldsymbol{A}^{t,0}$ to make prediction. It can be seen that aggregating all elements in a row that contains the same subject delivers higher performance.

Running Efficiency. Finally we probe the running efficiency of DIFFVSGG. As seen in Table 8, our method requires the fewest trainable parameters among all competitors. Though inference speed is slightly slower than existing work due to step-wise denoising, the improvement in performance compensates for this limitation.

5. Conclusion

We presented DIFFVSGG, a diffusion-based VSGG solution that makes contributions in three aspects. First, as an online approach, it effectively addresses challenges such as GPU memory constraints and real-time processing while delivering high performance, highlighting its potential for real-world application. Second, we propose a learning strategy for LDMs to generate inter-object relationships from random subject-object pairs, with ground-truth object locations as the annotations. Third, in leverage of LDMs, DIFFVSGG provides a new approach to enable temporal reasoning via dynamic, iterative refinement of scene graphs across frames, conditioned on predictions of prior frame. We hope this work could provide a new perspective for video analysis through reverse diffusion along the temporal dimension.

Acknowledgment

This work was supported in part by the Natural Science Foundation of Zhejiang Province (LDT23F02023F02), the Fundamental Research Funds for the Central Universities (226-2024-00058), and the National Natural Science Foundation of China (No. 62372405). The authors would like to thank [optional: specific individuals or institutions] for their valuable discussions and feedback. We also acknowledge the use of [optional: datasets, computational resources, or tools] that contributed to this research.

References

[1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 3

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 2

[3] Stephan Baier, Yunpu Ma, and Volker Tresp. Improving visual relationship detection using semantic modeling of scene descriptions. In *ISWC*, 2017. 2

[4] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2021. 3

[5] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *CVPR*, 2022. 3

[6] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024. 3

[7] Mu Chen, Zhedong Zheng, Yi Yang, and Tat-Seng Chua. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. In *ACM MM*, 2023. 2

[8] Mu Chen, Minghan Chen, and Yi Yang. Uahoi: Uncertainty-aware robust interaction learning for hoi detection. *Computer Vision and Image Understanding*, 247: 104091, 2024. 2

[9] Mu Chen, Liulei Li, Wenguan Wang, Ruijie Quan, and Yi Yang. General and task-oriented video segmentation. In *ECCV*, 2024. 3

[10] Mu Chen, Zhedong Zheng, and Yi Yang. Pipa++: towards unification of domain adaptive semantic segmentation via self-supervised learning. *arXiv preprint arXiv:2407.17101*, 2024. 3

[11] Mu Chen, Zhedong Zheng, and Yi Yang. Transferring to real-world layouts: A depth-aware framework for scene adaptation. In *ACM MM*, 2024. 2

[12] Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Hydra-sgg: Hybrid relation assignment for one-stage scene graph generation. In *ICLR*, 2025. 2

[13] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *ICCV*, 2023. 3

[14] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *CVPR*, 2019. 2

[15] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. In *CVPR*, 2023. 3

[16] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *ACM MM*, 2021. 2

[17] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *ICCV*, 2021. 6, 7, 8

[18] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11169–11183, 2023. 2

[19] Yiming Cui. Feature aggregated queries for transformer-based video object detectors. In *CVPR*, 2023. 3

[20] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *CVPR*, 2017. 2

[21] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Learning of visual relations: The devil is in the tails. In *ICCV*, 2021. 2

[22] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[23] Qi Dong, Zhuowen Tu, Haofu Liao, Yuting Zhang, Vijay Mahadevan, and Stefano Soatto. Visual relationship detection using part-and-sum transformers with composite queries. In *ICCV*, 2021. 2

[24] Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In *CVPR*, 2022. 2

[25] Yiqun Duan, Xianda Guo, and Zheng Zhu. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023. 3

[26] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, 2019. 3

[27] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *ICCV*, 2019. 3

[28] Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. Navigation instruction generation with bev perception and large language models. In *ECCV*, 2024.

[29] Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. Scene map-based prompt tuning for navigation instruction generation. In *CVPR*, 2025. 3

[30] Shengyu Feng, Hesham Mostafa, Marcel Nassar, Somdeb Majumdar, and Subarna Tripathi. Exploiting long-term dependencies for generating dynamic scene graphs. In *WACV*, 2023. 1, 2, 3, 6, 7, 8

[31] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 1

[32] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. In *NeurIPS*, 2022. 3

[33] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *CVPR*, 2019. 2

[34] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013. 1

[35] Wenguan Wang Guikun Chen, Jin Li. Scene graph generation with role-playing large language models. In *NeurIPS*, 2025. 2

[36] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 4

[37] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[38] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 3

[39] Xiaolin Hong, Hongwei Yi, Fazhi He, and Qiong Cao. Human-aware 3d scene generation with spatially-constrained diffusion models. *arXiv preprint arXiv:2406.18159*, 2024. 3

[40] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, 2020. 2, 6, 7

[41] Jian-Yu Jiang-Lin, Kang-Yang Huang, Ling Lo, Yi-Ning Huang, Terence Lin, Jhih-Ciang Wu, Hong-Han Shuai, and Wen-Huang Cheng. Record: Reasoning and correcting diffusion for hoi generation. In *ACM MM*, 2024. 3

[42] Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *ICML*, 2022. 3

[43] Deunsol Jung, Sanghyun Kim, Won Hwa Kim, and Minsu Cho. Devil's on the edges: Selective quad attention for scene graph generation. In *CVPR*, 2023. 2

[44] Boah Kim, Yujin Oh, and Jong Chul Ye. Diffusion adversarial representation learning for self-supervised vessel segmentation. In *ICLR*, 2023. 3

[45] Junyeong Kim, Minuk Ma, Trung Pham, Kyungsu Kim, and Chang D Yoo. Modality shifting improvedion network for multi-modal video question answering. In *CVPR*, 2020. 3

[46] Boris Knyazev, Harm De Vries, Cătălina Cangea, Graham W Taylor, Aaron Courville, and Eugene Belilovsky. Graph density-aware losses for novel compositions in scene graph generation. *arXiv preprint arXiv:2005.08230*, 2020. 2

[47] Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. Are diffusion models vision-and-language reasoners? In *NeurIPS*, 2023. 3

[48] Sanjoy Kundu and Sathyanarayanan N Aakur. Is-ggt: Iterative scene graph generation with generative transformers. In *CVPR*, 2023. 2

[49] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 1

[50] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. 1

[51] Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, and Jun Xiao. The devil is in the labels: Noisy label correction for robust scene graph generation. In *CVPR*, 2022. 2

[52] Liulei Li, Wenguan Wang, and Yi Yang. Human-object interaction detection collaborated with large relation-driven diffusion models. In *NeurIPS*, 2024. 3

[53] Liulei Li, Jianan Wei, Wenguan Wang, and Yi Yang. Neural-logic human-object interaction detection. In *NeurIPS*, 2024. 2

[54] Lin Li, Jun Xiao, Hanrong Shi, Hanwang Zhang, Yi Yang, Wei Liu, and Long Chen. Nicest: Noisy label correction and training for robust scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[55] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. In *CVPR*, 2022. 2

[56] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *CVPR*, 2017. 2

[57] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *CVPR*, 2017. 2, 6, 7

[58] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *CVPR*, 2022. 6, 7

[59] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Vrr-vg: Refocusing visually-relevant relationships. In *CVPR*, 2019. 2

[60] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, 2020. 6, 7

[61] Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *ICCV*, 2023. 3

[62] Xin Liu, Fatemeh Karimi Nejadasl, Jan C van Gemert, Olaf Booij, and Silvia L Pintea. Objects do not disappear: Video object detection by single-frame object location anticipation. In *CVPR*, 2023. 3

[63] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 6, 7

[64] Yichao Lu, Himanshu Rai, Jason Chang, Boris Knyazev, Guangwei Yu, Shashank Shekhar, Graham W Taylor, and Maksims Volkovs. Context-aware scene graph generation with seq2seq transformers. In *ICCV*, 2021. 2

[65] Run Luo, Zikai Song, Lintao Ma, Jinlin Wei, Wei Yang, and Min Yang. Diffusiontrack: Diffusion model for multi-object tracking. In *AAAI*, 2024. 3

[66] Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. Diffmot: A real-time diffusion-based multiple object tracker with non-linear prediction. In *CVPR*, 2024. 3

[67] Yifan Mao, Jian Liu, and Xianming Liu. Stealing stable diffusion prior for robust monocular depth estimation. *arXiv preprint arXiv:2403.05056*, 2024. 3

[68] Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *CVPR*, 2016. 2

[69] Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K Roy-Chowdhury. Unbiased scene graph generation in videos. In *CVPR*, 2023. 1, 3, 6, 7, 8

[70] Megha Nawhal and Greg Mori. Activity graph transformer for temporal action localization. *arXiv preprint arXiv:2101.08540*, 2021. 2

[71] Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon. Permutation invariant graph generation via score-based generative modeling. In *ICAIS*, 2020. 3

[72] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. In *CVPR*, 2019. 2

[73] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 3

[74] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In *ACM MM*, 2019. 2

[75] Tanzila Rahman, Shih-Han Chou, Leonid Sigal, and Giuseppe Carenini. An improved attention for visual question answering. In *CVPR*, 2021. 3

[76] Yasiru Ranasinghe, Deepti Hegde, and Vishal M Patel. Monodiff: Monocular 3d object detection and pose estimation with diffusion models. In *CVPR*, 2024. 3

[77] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[78] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 3

[79] Suprosanna Shit, Rajat Koner, Bastian Wittmann, Johannes Paetzold, Ivan Ezhov, Hongwei Li, Jiazhen Pan, Sahand Sharifzadeh, Georgios Kaissis, Volker Tresp, et al. Relationformer: A unified framework for image-to-graph generation. In *ECCV*, 2022. 2

[80] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 6

[81] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2, 3

[82] Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. Learning visual knowledge memory networks for visual question answering. In *CVPR*, 2018. 3

[83] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *CVPR*, 2021. 2

[84] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019. 2

[85] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 2019. 2, 6, 7

[86] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020. 2

[87] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016. 1

[88] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *ICCV*, 2021. 6, 7, 8

[89] Thanh-Dat Truong, Quoc-Huy Bui, Chi Nhan Duong, Han-Seok Seo, Son Lam Phung, Xin Li, and Khoa Luu. Direcformer: A directed attention in transformer approach to robust action recognition. In *CVPR*, 2022. 3

[90] Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi. Video relationship reasoning using gated spatio-temporal energy graph. In *CVPR*, 2019. 2

[91] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *ICCV*, 2015. 1

[92] Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. In *ICLR*, 2023. 3

[93] Hefeng Wang, Jiale Cao, Rao Muhammad Anwer, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Dformer: Diffusion-guided transformer for universal image segmentation. *arXiv preprint arXiv:2306.03437*, 2023. 3

[94] Jingyi Wang, Jinfa Huang, Can Zhang, and Zhidong Deng. Cross-modality time-variant relation learning for generating dynamic scene graphs. *arXiv preprint arXiv:2305.08522*, 2023. 1, 3, 6, 7

[95] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *CVPR*, 2021. 3

[96] Shuang Wang, Lianli Gao, Xinyu Lyu, Yuyu Guo, Pengpeng Zeng, and Jingkuan Song. Dynamic scene graph generation via temporal prior inference. In *ACM MM*, 2022. 6

[97] Wenguan Wang, Yi Yang, and Yunhe Pan. Visual knowledge in the big model era: Retrospect and prospect. *Frontiers of Information Technology & Electronic Engineering*, 26(1):1–19, 2025. 2

[98] Jianan Wei, Tianfei Zhou, Yi Yang, and Wenguan Wang. Nonverbal interaction detection. In *ECCV*, 2024. 2

[99] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: relational embedding for scene graph. In *NeurIPS*, 2018. 2

[100] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, 2022. 1

[101] Fei Xie, Zhongdao Wang, and Chao Ma. Diffusiontrack: Point set diffusion model for visual object tracking. In *CVPR*, 2024. 3

[102] Chenfeng Xu, Huan Ling, Sanja Fidler, and Or Litany. 3difftection: 3d object detection with geometry-aware diffusion features. In *CVPR*, 2024. 3

[103] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017. 2

[104] Minkai Xu, Alexander S Powers, Ron O Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. In *ICML*, 2023. 3

[105] Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In *ACM MM*, 2020. 2

[106] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *ECCV*, 2018. 2

[107] Ling Yang, Zhilin Huang, Zhilong Zhang, Zhongyi Liu, Shenda Hong, Wentao Zhang, Wenming Yang, Bin Cui, and Luxia Zhang. Graphusion: Latent diffusion for graph generation. *IEEE Transactions on Knowledge and Data Engineering*, 2024. 3

[108] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). In *ICML*, 2024. 3

[109] Yuan Yao, Ao Zhang, Xu Han, Mengdi Li, Cornelius Weber, Zhiyuan Liu, Stefan Wermter, and Maosong Sun. Visual distant supervision for scene graph generation. In *ICCV*, 2021. 2

[110] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *ECCV*, 2018. 2

[111] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *ICCV*, 2017. 2

[112] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018. 2, 4, 6, 7

[113] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *CVPR*, 2019. 6, 7

[114] Jialin Zhao, Yuxiao Dong, Ming Ding, Evgeny Kharlamov, and Jie Tang. Adaptive diffusion in graph neural networks. In *NeurIPS*, 2021. 3

[115] Songwei Zhao. Flexible diffusion for graph neural networks. *arXiv preprint arXiv:2010.04159*, 2020. 3

[116] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023. 3

[117] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen. Cascaded human-object interaction recognition. In *CVPR*, 2020. 2

[118] Tianfei Zhou, Siyuan Qi, Wenguan Wang, Jianbing Shen, and Song-Chun Zhu. Cascaded parsing of human-object interaction recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2827–2840, 2021. 2