UNO: Unifying One-stage Video Scene Graph Generation via Object-Centric Visual Representation Learning

Huy Le[1*] Nhat Chung[1] Tung Kieu[2,3] Jingkang Yang[4] Ngan Le[5]
[1]FPT Software AI Center, Vietnam [2]Aalborg University, Denmark [3]Pioneer Centre for AI, Denmark
[4]S-Lab, Nanyang Technological University, Singapore [5]AICV Lab, University of Arkansas, USA
{lehuy2316, nhatchung14}@gmail.com tungkvt@cs.aau.dk jingkang001@ntu.edu.sg thile@uark.edu

Abstract

Video Scene Graph Generation (VidSGG) aims to represent dynamic visual content by detecting objects and modeling their temporal interactions as structured graphs. Prior studies typically target either coarse-grained box-level or fine-grained panoptic pixel-level VidSGG, often requiring task-specific architectures and multi-stage training pipelines. In this paper, we present UNO (UNified Object-centric VidSGG), a single-stage, unified framework that jointly addresses both tasks within an end-to-end architecture. UNO is designed to minimize task-specific modifications and maximize parameter sharing, enabling generalization across different levels of visual granularity. The core of UNO is an extended slot attention mechanism that decomposes visual features into object and relation slots. To ensure robust temporal modeling, we introduce object temporal consistency learning, which enforces consistent object representations across frames without relying on explicit tracking modules. Additionally, a dynamic triplet prediction module links relation slots to corresponding object pairs, capturing evolving interactions over time. We evaluate UNO on standard box-level and pixel-level VidSGG benchmarks. Results demonstrate that UNO not only achieves competitive performance across both tasks but also offers improved efficiency through a unified, object-centric design.

1. Introduction

Video Scene Graph Generation (VidSGG) aims to extract structured, dynamic representations from videos by modeling objects as nodes and their pairwise interactions as edges in spatio-temporal graphs. These structured representations offer both interpretability and compositionality, making VidSGG a critical component in various downstream tasks such as video understanding [19, 43, 51, 60], video reasoning [38, 65] and robotic reasoning [57].

*Corresponding author

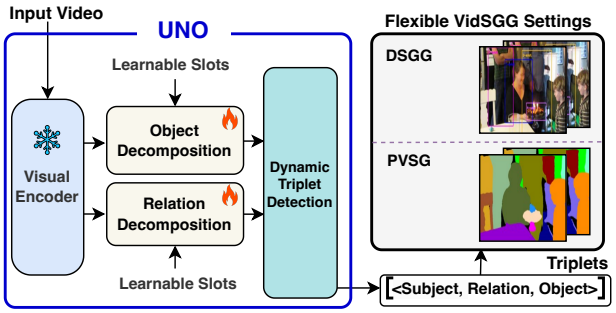

Figure 1. UNO. We introduce UNO, a unified framework for box-level VidSGG (DSGG) and pixel-level VidSGG (PVSG) settings.

Current VidSGG research primarily follows two directions, distinguished by the level of visual granularity: box-level VidSGG (also referred to as Dynamic Scene Graph Generation or DSGG) [19, 27, 34, 47, 49], and panoptic pixel-level VidSGG (also known as Panoptic Video Scene Graph Generation or PVSG) [52, 59, 60]. The former focuses on coarse-grained object representations using bounding boxes and typically models relationships at the frame level. The latter provides fine-grained, pixel-level representations using panoptic segmentation masks, where object trajectories are treated as graph nodes and interactions—including object-object and object-background—are captured throughout the video. Importantly, PVSG emphasizes the temporal consistency of object identities across frames.

In scenarios requiring multi-level scene understanding, a unified VidSGG model capable of handling both tasks is highly desirable. Such a model could flexibly adapt to diverse visual representations and support a wider range of applications without task-specific architectural redesign. However, achieving this unification is non-trivial due to the differing structural assumptions, temporal modeling requirements, and visual encoding strategies inherent to each task. Prior attempts have relied on multi-stage pipelines involving either box-level or pixel-level representation, followed by a tracking module [27, 36, 50, 60, 62], which introduce significant computational overhead and often result in sub-

Table 1. Comparison of different models on *VidSGG tasks*.

Methods	One-stage	Granularity-level		Temporal-level	
		Box	Pixel	Frame	Tracklet
STTran [8] ICCV'21	✗	✓	✗	✓	✗
APT [27] CVPR'22	✗	✓	✗	✓	✗
TEMPURA [34] CVPR'23	✗	✓	✗	✓	✗
PVSG [60] CVPR'23	✗	✗	✓	✓	✓
OED [49] CVPR'24	✓	✓	✗	✓	✗
MCL [36] AAAI'25	✗	✗	✓	✓	✓
DIFFVSGG [3] CVPR'25	✗	✓	✗	✓	✗
VISA [28] CVPR'25	✗	✓	✗	✓	✗
UNO (Ours)	✓	✓	✓	✓	✓

optimal performance due to decoupled learning and limited parameter sharing. On the other hand, designing a unified, end-to-end solution encounters the core challenge of learning a semantically consistent spatio-temporal representation that generalizes across varying levels of granularity while remaining aligned with the underlying video dynamics.

Beyond spatial granularity, VidSGG approaches also differ in their temporal-level representations, which can be broadly categorized into frame-level [8, 27, 34, 49] and tracklet-level [36, 60] methods. Frame-level methods construct scene graphs independently at each frame, aligning naturally with box-level VidSGG, but often fall short in modeling long-term interactions and maintaining temporal consistency. In contrast, tracklet-level methods are more accurate because they link object instances across frames to capture temporal dynamics explicitly, a strategy more common in pixel-level VidSGG. This lack of temporal generality hinders their ability to capture coherent and continuous scene dynamics—especially in applications requiring both fine-grained relationship modeling and long-range temporal reasoning. A detailed comparison of UNO with prior works is presented in Tab. 1.

To address this challenge, we propose UNO, a unified, object-centric, single-stage VidSGG framework that effectively supports both box-level and pixel-level tasks. Fig. 1 illustrates the overall concept of UNO. Our central hypothesis is that despite their differences, DSGG and PVSG share a common semantic context in object-centric representation which making it well-suited for such modeling. At the heart of UNO is an extended slot attention mechanism [31] that decomposes visual feature maps into compact object and relation slots. These slots serve as modular building blocks and form a shared latent representation space across both tasks. To ensure temporal consistency, we introduce *object temporal consistency learning*, which enforces the alignment of object slots across frames without the need for explicit tracking. Furthermore, we propose a *dynamic triplet prediction module* that efficiently associates relation slots with subject–object pairs while reducing redundancy in the predicted triplets.

We validate UNO on two standard benchmarks: Action Genome [19] for DSGG task and PVSG [60] for PVSG task. Experimental results show that UNO consistently outperforms state-of-the-art (SOTA) methods in both accuracy and computational efficiency. While UNO builds on prior concepts such as slot attention and temporal contrastive consistency, our key contribution lies in recontextualizing and integrating them into a unified, one-stage framework specifically designed for VidSGG. To the best of our knowledge, UNO is among the first approaches to jointly address both box-level and pixel-level VidSGG within an object-centric paradigm.

2. Related Works

Video Scene Graph Generation (VidSGG). VidSGG is an extension of Scene Graph Generation [2] that analyses videos to identify objects and their relationships, representing this information as a structured graph to support high-level video understanding tasks [38, 41, 43, 51, 57, 65]. Researchers have explored how VidSGG can be leveraged on different granularities of video content, from coarse bounding boxes [8, 9, 19, 29, 47] to fine-grained panoptic masks [52, 59, 60], to represent dynamic interactions among objects with varying levels of precision. In fact, the literature has largely diverged into two directions:

Dynamic Scene Graph Generation–DSGG [8, 9, 19, 29, 47] adopts the box-level approach to VidSGG that involves detection and tracking of object instances, capturing both spatial and temporal relationships to form graphs. In particular, Action Genome [19] supports DSGG with bounding boxes, relationship labels, and actions of human-object interactions. Various strategies have been proposed to predict objects and classify their pair-wise relationships [8, 9, 29]. Recently, OED [49] reformulates DSGG as a set prediction problem on object boxes, and leverages pair-wise features to represent each subject-object pair within the scene graph. We refer to this as box-level VidSGG.

Video Panoptic Scene Graph Generation–PVSG [52, 59, 60] requires nodes in scene graphs to be grounded by precise, pixel-level segmentation masks to facilitate fine-grained scene understanding. In the PVSG benchmark [60], frames in a video are assigned with panoptic masks that provide pixel-level detail of object and background boundaries, where methods [52, 60] have had to capture both evolving object-background interactions and object-object interactions to create a cohesive, dynamic scene graph representation. We refer to this as pixel-level VidSGG.

Interestingly, VidSGG strategies are often multi-stage pipelines [8, 9, 29, 47, 52, 59], while OED [49] recently pioneered a one-stage VidSGG approach for DSGG, eliminating the need for external tracking or multi-stage optimization. Although OED could not be directly adopted into PVSG, it highlights an important design consideration for stable learning and efficient end-to-end VidSGG modeling. *Building on this insight, our research introduces a novel one-stage framework that unifies dynamic scene graph generation for*

both DSGG and PVSG in videos, setting it apart from prior approaches.

Object-centric Representation Learning. Object-centric representation learning has been adopted to focus on entities [23, 35, 38, 44, 58] that are directly meaningful for study predictions. It is employed to uncover modular structures and independent mechanisms, such as objects and their relationships, from multi-object visual inputs [21, 40, 63]. Conventional models rely on object-/region-specific priors to facilitate reasoning and comprehension [35, 44, 58, 64]. Meanwhile, recent works have leveraged slot attention [23, 31, 40] to facilitate object-centric representations from raw scene features or for embodied and robotic features [7, 15]. In particular, slot attention has used GRU [6] and competitive attention mechanisms to bind to modular structures in the input [21, 31], potentially maintaining them through time for video understanding [40]. However, only a few studies have been done to consider an object-centric perspective for VidSGG besides straightforward visual detection, despite how it can consistently disentangle object semantics from general scene details. *In this work, our research aims to extend their utility to capture objects' modular structures and relationships for a unified VidSGG.*

3. Methodology

3.1. Preliminary and Motivation

In this subsection, we first introduce the definition and notation for existing VidSGG tasks, covering both DSGG and PVSG. We then explain how these tasks can be unified through our proposed one-stage framework, `UNO`.

Current VidSGG tasks. Given a video $\mathbf{V} = \langle I_1, \ldots, I_T \rangle$ of T frames, where each frame $I_t \in \mathbb{R}^{H_{\text{in}} \times W_{\text{in}} \times 3}$, VidSGG aims to produce a sequence of scene graphs $\mathbf{G} = \langle G_1, \ldots, G_T \rangle$, where each G_t represents the scene graph for frame I_t. Each G_t consists of triplets and is defined as $G_t = \{\textit{subject}, \textit{relation}, \textit{object}\}$. The goal is to model the conditional probability $\mathbb{P}(\mathbf{G} \mid \mathbf{V})$. To simplify, existing methods [27] reformulate this as predicting detected objects and their pairwise relations.

$$\mathbb{P}(\mathbf{G} \mid \mathbf{V}) = \mathbb{P}(\mathbf{B}, \mathbf{O}, \mathbf{R} \mid \mathbf{V}). \tag{1}$$

Here, $\mathbf{B} = \langle \mathbf{B}_1, \ldots, \mathbf{B}_T \rangle$ is the set of bounding boxes for entire video $\mathbf{V}$, where $\mathbf{B}_t = \{B_1, \ldots, B_{M_t}\}$ is the set of bounding boxes of M_t objects in the t-frame. Similarly, $\mathbf{O} = \langle \mathbf{O}_1, \ldots, \mathbf{O}_T \rangle$ is the set of object labels for entire video $\mathbf{V}$, where $\mathbf{O}_t = \{O_1, \ldots, O_{M_t}\}$ is the set of object labels for the t-frame. $\mathbf{R} = \langle \mathbf{R}_1, \ldots, \mathbf{R}_T \rangle$ is the set of relation for entire video $\mathbf{V}$, where $\mathbf{R}_t = \{R_1, \ldots, R_L\}$ is the set of relations for the t-frame. This approach is termed as coarse-grained box-level VidSGG task (DSGG [19]), and the conditional probability is factorized as follows:

$$\begin{aligned} \mathbb{P}(\mathbf{G} \mid \mathbf{V}) &= \mathbb{P}(\mathbf{G}_{\text{DSGG}} \mid \mathbf{V}) \\ &= \mathbb{P}(\mathbf{B} \mid \mathbf{V})\, \mathbb{P}(\mathbf{O} \mid \mathbf{B}, \mathbf{V})\, \mathbb{P}(\mathbf{R} \mid \mathbf{O}, \mathbf{B}, \mathbf{V}). \end{aligned} \tag{2}$$

A prior study [60] extended this formulation to handle the fine-grained pixel-level VidSGG by replacing the bounding boxes $\mathbf{B}$ with mask tubes $\mathbf{M}$, which represent each object in the entire video. This leads to the formulation: $\mathbb{P}(\mathbf{G} \mid \mathbf{V}) = \mathbb{P}(\mathbf{M}, \mathbf{O}, \mathbf{R} \mid \mathbf{V})$. We refer to this as the fine-grained pixel-level VidSGG task (PVSG [60]), and the conditional probability is factorized as follows.

$$\begin{aligned} \mathbb{P}(\mathbf{G} \mid \mathbf{V}) &= \mathbb{P}(\mathbf{G}_{\text{PVSG}} \mid \mathbf{V}) \\ &= \mathbb{P}(\mathbf{M} \mid \mathbf{V})\, \mathbb{P}(\mathbf{O} \mid \mathbf{M}, \mathbf{V})\, \mathbb{P}(\mathbf{R} \mid \mathbf{O}, \mathbf{M}, \mathbf{V}). \end{aligned} \tag{3}$$

where $\mathbf{M} = \langle \mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_{M_{\texttt{obj}} + M_{\texttt{bg}}} \rangle$ refers to the list of non-overlapping binary mask tubes of each object, where $M_{\texttt{obj}}$ and $M_{\texttt{bg}}$ is the number of objects and background appear in the video. For object i, the mask tube $\mathbf{m}_i \in \{0, 1\}^{T \times H_{\text{in}} \times W_{\text{in}}}$ collects all tracked masks in each frame.

Unifying VidSGG tasks. Given the current formulations of DSGG in Eq.2 and PVSG in Eq.3, the primary distinction lies in their temporal granularity. DSGG operates at the frame level, where $\mathbf{O}_t$ denotes the set of objects detected within each individual frame, and thus does not enforce temporal consistency across frames. In contrast, PVSG leverages mask tubes, which inherently depend on maintaining object consistency throughout the video. A naïve approach to unifying VidSGG would result in a multi-stage pipeline of suboptimal performance and computational cost due to different modeling strategies in each task and stage. To address this challenge, in this work, we consider two VidSGG tasks through a unified perspective and propose a framework to directly model $\mathbb{P}(\mathbf{G} \mid \mathbf{V})$ such that $\mathbf{G} = \{\mathbf{G}_{\text{DSGG}}$ or $\mathbf{G}_{\text{PVSG}}\}$ in an end-to-end training and inference manner. The primary challenge of our research lies in maintaining unified, spatio-temporal representations that align with video dynamics across both tasks, which we propose to address through an object-centric design, equipped with an object temporal consistency learning mechanism, ensuring a consistent and structured representation of objects and their relationships throughout the video sequence.

Design Principles. `UNO` follows three key principles. First, it is a one-stage unified framework that minimizes task-specific modifications and multi-stage processing while maximizing parameter sharing through object-centric representation. Second, it establishes a strong baseline for diverse VidSGG tasks by reducing computational cost compared to multi-stage methods without sacrificing performance. Finally, instead of fusing bounding boxes and masks at the output level, `UNO` employs a unified object-centric representation using Slot Attention, where object and relation slots serve as shared latent features and feed into task-specific heads for bounding box or panoptic mask prediction.

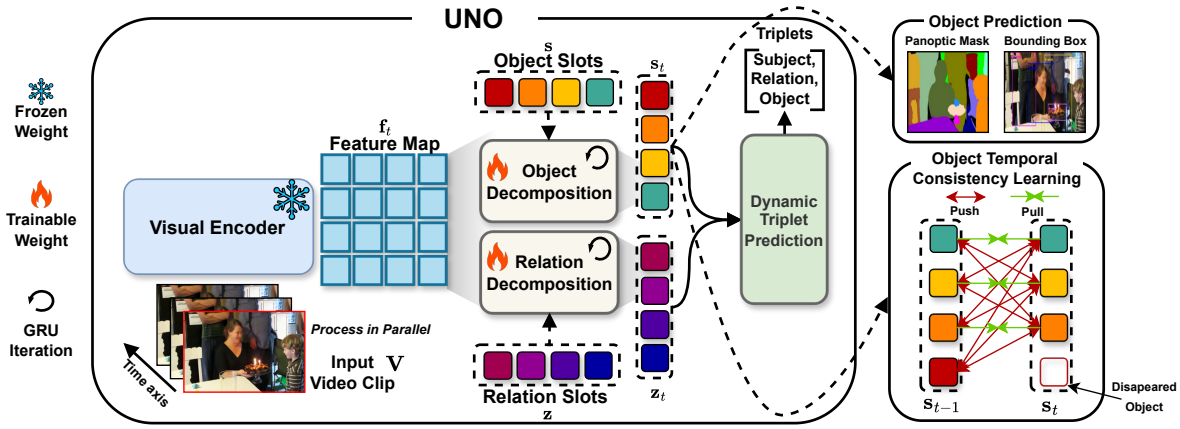

Figure 2. UNO Framework. Our architecture is powered with slot attention to efficiently decompose visual features into object and relation slots. The slots are also enabled with object temporal consistency learning to encourage their tracking through time. Finally, a dynamic triplet prediction module is integrated to align relation slots with their corresponding object slots, thereby obtaining the triplets of interest.

3.2. UNO Architecture

An overview of UNO is in Fig. 2. First, we utilize a frozen pre-trained visual encoder to extract feature maps from the last layer of each video frame. Next, we apply Slot Attention [31] to decompose these feature maps into modular slots, effectively capturing both object and relation representations. These slots serve as shared latent features and are passed through a task-specific prediction head–either for bounding boxes or for panoptic masks. To ensure spatio-temporal consistency, we introduce object temporal consistency learning, reinforcing stable slot features across the video. Finally, we propose a dynamic triplet prediction mechanism that associates relation slots with their corresponding object pairs. Integrated into our end-to-end framework, this mechanism minimizes redundancy in triplet prediction while enhancing model efficiency.

3.2.1. Visual Encoding

First, we employ a frozen, pre-trained vision model as the Visual Encoder, where the last-layer feature map encodes rich object cues [1, 10, 37, 45, 48], providing a spatial prior for learning object positions at varying granularities. Given a frame I_t, we extract its feature map $\mathbf{f}_t \in \mathbb{R}^{H_{\text{enc}} \times W_{\text{enc}} \times D_{\text{enc}}}$, where H_{enc} and W_{enc} denote the spatial dimensions, and D_{enc} represents the feature channel size.

Although pre-trained feature maps capture rich object information, they often entangle semantics, grouping similar objects [10, 48]. To address this, a decomposition module is essential for disentangling these features into distinct semantics, enabling precise object and relation predictions for triplet construction.

3.2.2. Object Decomposition

We employ Slot Attention [31], inspired by recent advances in object discovery [23, 24, 63], as a clustering mechanism to group semantically meaningful patches from $\mathbf{f}_t$ into predefined slots, where each slot corresponds to a distinct object region. Unlike previous studies [23, 39] that sample from a prior distribution, we initialize N object slots (denoted as $\mathbf{s}$) as learnable tokens. We then decompose the feature map $\mathbf{f}_t$ of t-th frame into N frame-wise object slot features $\mathbf{s}_t = \{\mathbf{s}_t^1, \ldots, \mathbf{s}_t^N\}$, $\mathbf{s}_t \in \mathbb{R}^{N \times D\texttt{slot}}$. This design encourages each slot to capture modular semantics [20, 56], enabling object-consistent decompositions through its unique *competition mechanism* that supports maintaining coherent slot representation across frames.

Formally, following the standard slot attention procedure [31], we employ three linear transformation heads to map the object slots $\mathbf{s}$ into $\texttt{Query}\ \mathbf{q} \in \mathbb{R}^{N \times D_{\text{enc}}}$, while frame-wise feature maps $\mathbf{f}_t$ into $\texttt{Key}\ \mathbf{k} \in \mathbb{R}^{H_{\text{enc}} W_{\text{enc}} \times D_{\text{enc}}}$, and $\texttt{Value}\ \mathbf{v} \in \mathbb{R}^{H_{\text{enc}} W_{\text{enc}} \times D_{\text{enc}}}$. We iteratively calculate attention score and update slot representations via Gated Recurrent Unit (GRU) [5]. Mathematically, we formulate each iteration as:

$$\begin{aligned} \tilde{\mathbf{a}}_{i,j} &= \frac{e^{\mathbf{a}_{i,j}}}{\sum_{l=1}^{N} e^{\mathbf{a}_{l,j}}}, \quad \text{where} \quad \mathbf{a} = \frac{1}{\sqrt{D}} \mathbf{q}\mathbf{k}^{\top}. \\ \mathbf{w}_{i,j} &= \frac{\tilde{\mathbf{a}}_{i,j}}{\sum_{l=1}^{H_{\text{enc}} W_{\text{enc}}} \tilde{\mathbf{a}}_{i,l}}, \\ \mathbf{s}_t &= \text{GRU}(\texttt{inputs} = \mathbf{w}\mathbf{v}, \texttt{states} = \mathbf{s}_t) \end{aligned} \tag{4}$$

Here, the attention weights $\tilde{\mathbf{a}}$ are normalized with $softmax$ along the slot dimension, and the weighted mean coefficient $\mathbf{w}$ aggregates the $\texttt{Value}\ \mathbf{v}$ to update the slots. This mechanism encourages competition among slots, ensuring each slot captures distinct object features. We use the object slots $\mathbf{s}_t$ at the final iteration as the distilled object tokens from the feature maps $\mathbf{f}_t$.

Object Prediction. Since each object slot captures multi-level granularity information, using a lightweight prediction head is sufficient for object class, bounding box, and mask prediction. Thus, we employ three lightweight prediction heads, each consisting of a feed-forward network (FFN) with two linear layers–a functional layer followed by a task-specific layer for each output. The classification head

outputs object classes $\hat{\mathbf{O}}_t = \text{FFN}_{\texttt{cls}}(\mathbf{s}_t)$, while the box head predicts object coordinates $\hat{\mathbf{B}}_t = \text{FFN}_{\texttt{box}}(\mathbf{s}_t)$. For mask prediction, a lightweight decoder with four transpose convolutions [32] upsamples the feature map $\mathbf{f}_t$ to the original frame size, resulting in $\mathbf{f'}_t = \text{Dec}(\mathbf{f}_t)$. The panoptic mask is then obtained by applying a matrix multiplication between the object slots $\mathbf{s}_t$ and the upsampled feature map $\mathbf{f'}_t$ in the mask head, resulting in a binary mask of each object slot $\hat{\mathbf{m}}_t = \text{FFN}_{\texttt{mask}}(\mathbf{s}_t \cdot \mathbf{f'}_t)$.

Object Temporal Consistency Learning. Maintaining consistent spatio-temporal representations for object slots is crucial for tracking coherent object features over time. Slot Attention provides a strong basis for object-centric video understanding. However, prior studies have noted that slot-based representations struggle to maintain temporal consistency across consecutive frames in a video [61]. To address this limitation, and inspired by prior works [14, 25], we incorporate object temporal consistency learning using a contrastive loss. Specifically, slots matching the same ground truth index across frames are treated as positive samples, while all others are negatives. Formally, given the i-th object slot $\mathbf{s}_t^i$ at frame t, we define $\mathbf{s}_{t-1}^i$ as the positive target and $\mathbf{s}_{t-1}^j$ as negatives ($j \neq i$). The right side of Fig. 2 illustrates the concept of object temporal consistency learning. The corresponding loss function is formulated as:

$$\mathcal{L}_{\text{consistency}} = -\sum_{\mathbf{s}_{t-1}^i} \log \frac{e^{(\mathbf{s}_t^i \cdot \mathbf{s}_{t-1}^i)}}{e^{(\mathbf{s}_t^i \cdot \mathbf{s}_{t-1}^i)} + \sum_{\mathbf{s}_{t-1}^j} e^{(\mathbf{s}_t^i \cdot \mathbf{s}_{t-1}^j)}}. \quad (5)$$

This loss encourages positive slots to stay close while pushing negative slots apart. By focusing only on matched slots, our method mitigates suboptimal updates from noisy, unmatched slots, ensuring that identical object slots remain aligned across frames. This alignment enables slots to refine each other's features, enhancing spatio-temporal consistency.

3.2.3. Relation Decomposition

Existing VidSGG approaches [49] predict relations sequentially after object prediction, leading to high computational complexity and potential inaccuracy if objects are missed. In contrast, UNO predicts objects and relations simultaneously, enabling a parallelized execution that significantly reduces complexity. Using Eq. 4, we apply an enhanced slot attention mechanism to decompose the feature map $\mathbf{f}_t$ into relation slots, where each slot captures potential interaction regions. This mechanism enables relation slots to capture the entire spatial context of a frame, rather than being restricted to object pair intersections. This broader coverage significantly enhances relation prediction performance and supports our dynamic triplet mechanism (see Sec. 3.2.4). Specifically, we initialize K relation slots (denoted as $\mathbf{z}$) as learnable tokens. Then, for the t-th frame, we obtain frame-wise relation slot features $\mathbf{z}_t = \{\mathbf{z}_t^1, \ldots, \mathbf{z}_t^K\}$, where $\mathbf{z}_t \in \mathbb{R}^{K \times D\texttt{slot}}$. Each frame-wise relation slot $\mathbf{z}_t$ is then passed through a classification FFN head to predict relations: $\hat{\mathbf{R}}_t = \text{FFN}_{\texttt{rel}}(\mathbf{z}_t)$. Consequently, both relation slots $\mathbf{z}_t$ together with its relation class $\hat{\mathbf{R}}_t$ and object slots $\mathbf{s}_t$ are jointly extracted from $\mathbf{f}_t$ at time step t.

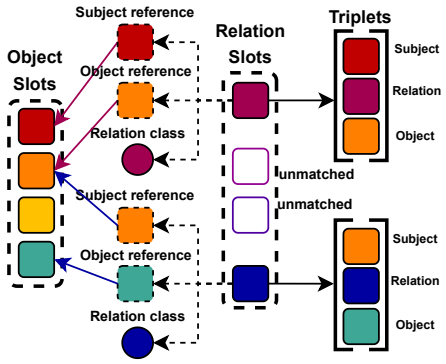

Figure 3. Dynamic triplet prediction module that predicts subject and object references to object slots from relation slots.

3.2.4. Dynamic Triplet Prediction

Existing methods either learn an adjacency matrix [52] to identify subject-object pairs or directly predict pairwise subject-object embeddings [49], followed by sequential relation prediction. In contrast, our approach introduces a dynamic triplet prediction mechanism that directly associates N objects with K relations, eliminating the need to construct an $N \times N$ object pair matrix. In theory, K can be as large as N^2, since each object may interact with every other object. However, in practice, $K \ll N$ due to the inherent sparsity of real-world interactions—only a small subset of object pairs exhibit meaningful relationships. This reduces redundancy and duplication in triplet prediction while maintaining high performance. Fig. 3 illustrates the proposed module.

Pairwise Index Matching. Each relation is defined as an interaction between two objects, with one as the subject and the other as the object. Thus, each relation slot inherently encodes information about the subject-object pair. We reformulate the problem as mapping relation slots to specialized representations that store the corresponding pair of object slots, enabling direct matching without constructing an $N \times N$ matrix. Formally, given K relation slots at the t-frame, $\mathbf{z}_t = \langle \mathbf{z}_t^1, \ldots, \mathbf{z}_t^K \rangle$, for each $\mathbf{z}_t^j$, we use two FFNs to generate subject-object pair of reference embeddings: $\mathbf{p}_j^s$ and $\mathbf{p}_j^o$.

$$\mathbf{p}_j^s = \text{FFN}_s(\mathbf{z}_t^j), \quad \mathbf{p}_j^o = \text{FFN}_o(\mathbf{z}_t^j) \quad (6)$$

Next, we find the indices of the subject and object that correspond to the object slots by matching the reference embeddings with the object slots $\mathbf{s}_t$ using a similarity function. Specifically, we aim to identify the most relevant subject $\mathbf{s}_t^i \in \mathbf{s}_t$ and object $\mathbf{s}_t^{i'} \in \mathbf{s}_t$ given a specific subject reference embedding $\mathbf{p}_j^s$ and object reference embedding $\mathbf{p}_j^o$, respectively. This process is formulated as follows.

$$\hat{i}_j^s = \underset{i \le N}{\arg\max}\ \text{sim}(\mathbf{p}_j^s, \mathbf{s}_t^i),\ \hat{i}_j^o = \underset{i' \le N}{\arg\max}\ \text{sim}(\mathbf{p}_j^o, \mathbf{s}_t^{i'}) \quad (7)$$

Here, $\text{sim}(u, v) = \frac{u \cdot v}{||u|| \cdot |v||}$ is the similarity function. The predicted indices $\hat{i}_j^s$ and $\hat{i}_j^o$ correspond to the subject and object pair of the relation slot $\mathbf{z}_t^j$. This dynamic matching process effectively links relation slots with their most relevant pair of object slots, forming coherent triplets without relying on predefined adjacency matrices. This adaptive strategy enhances generalization across diverse object relations in videos. The final triplet prediction of t-th frame is the set of K triplets, $\{\langle \mathbf{s}_t^{\hat{i}_j^s}, \mathbf{z}_t^i, \mathbf{s}_t^{\hat{i}_j^o} \rangle\}_{i=1}^K$, which can be replaced with corresponding bounding box and mask depends on the task.

Triplet Duplication Reduction. Unlike prior work [49], which predicts triplets as a set and often results in duplicate detections, our method leverages slot attention with a built-in competitive mechanism to mitigate redundancy. This design ensures distinct object slots and relation slots focus on appropriate regions, producing unique object and relation predictions for triplets without requiring post-processing steps like Non-Max Suppression (NMS) [18].

3.3. Training Objectives

During training, we first perform Hungarian matching [12] between the predicted and ground-truth object boxes/masks to assign object slots, followed by supervision for detection and classification. The loss function for the DSGG task is defined as:

$$\begin{aligned}\mathcal{L}_{\text{DSGG}} =& \lambda_{\text{obj\_cls}} \mathcal{L}_{\text{obj\_cls}}(\hat{\mathbf{O}}_t, \mathbf{O}_t) + \lambda_{\text{box}} \mathcal{L}_{\text{box}}(\hat{\mathbf{B}}_t, \mathbf{B}_t) \\ &+ \lambda_{\text{GIoU}} \mathcal{L}_{\text{GIoU}}(\hat{\mathbf{B}}_t, \mathbf{B}_t),\end{aligned} \quad (8)$$

where $\mathcal{L}_{\text{obj\_cls}}$ is the Cross Entropy (CE) loss, $\mathcal{L}_{\text{box}}$ is the ℓ_1 loss, and $\mathcal{L}_{\text{GIoU}}$ is the GIoU loss [42].

For the PVSG task, the loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{PVSG}} =& \lambda_{\text{obj\_cls}} \mathcal{L}_{\text{obj\_cls}}(\hat{\mathbf{O}}_t, \mathbf{O}_t) + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}}(\hat{\mathbf{m}}_t, \mathbf{m}_t) \\ &+ \lambda_{\text{dice}} \mathcal{L}_{\text{dice}}(\hat{\mathbf{m}}_t, \mathbf{m}_t),\end{aligned} \quad (9)$$

where $\mathcal{L}_{\text{mask}}$ is the CE loss, and $\mathcal{L}_{\text{dice}}$ is the Dice loss [46, 54].

Finally, we re-apply Hungarian matching between the predicted and ground-truth relations to align relation slots while supervising both relation classification and index prediction. The matching between the indices of the predicted subjects/objects and their ground-truth indices is formulated as a classification problem, where indices are converted to one-hot vectors. The relation loss is formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{Rel}} =& \lambda_{\text{rel\_cls}} \mathcal{L}_{\text{rel\_cls}}(\hat{\mathbf{R}}_t, \mathbf{R}_t) + \lambda_{\text{sidx}} \mathcal{L}_{\text{sidx}}(\hat{i}^s, i^s) \\ &+ \lambda_{\text{oidx}} \mathcal{L}_{\text{oidx}}(\hat{i}^o, i^o),\end{aligned} \quad (10)$$

where $\mathcal{L}_{\text{rel\_cls}}$, $\mathcal{L}_{\text{sidx}}$, and $\mathcal{L}_{\text{oidx}}$ are all CE losses.

4. Experiment Results

4.1. Experimental Settings

Datasets & Evaluation Metrics. We conduct experiments on the Action Genome [19] for the DSGG task and the PVSG [60] for the PVSG task. We evaluate the DSGG task following the setting from [49] and the PVSG task following the setting from [60]. We train and evaluate the model strictly separately for each benchmark/task, without any data mixing or augmentation strategies.

Implementation Details. We adopt both Vision Transformer (ViT) and Convolutional Neural Networks (CNNs) as the backbone for the Visual Encoder to perform frame-wise feature extraction. More specifically, for ViT, we use ViT-S/14, ViT-B/14, ViT-L/14 [11] with pre-trained weights from DINO [37]; for CNNs, we use ResNet-50 [16] pre-trained weights from MoCo [17]. The number of slots are empirically chosen and can be observed that the optimal number of slots is influenced by both the number of object classes in the dataset and the objects present in the video. Therefore, for DSGG, we set $N = 40$ object slots and $K = 24$ relation slots; for PVSG, we use $N = 96$ object slots and $K = 40$ relation slots.

4.2. Comparison with State of the Arts

Results on Box-Level VidSGG (DSGG). Tab. 2 compares `UNO` against SOTA methods on the Action Genome dataset. `UNO` not only surpasses the second-best one-stage OED by a clear margin but also outperforms multi-stage methods with tracking mechanisms such as APT, TR^2, TPT, and TEMPURA. The results on the SGDET task highlight `UNO`'s strong VidSGG capabilities in DSGG. Our approach simultaneously localizes objects and predicts relations, achieving 45.2% R@20 (↑4.3% over the second best) under *With Constraint* and 49.7% R@20 (↑5.7%) under *No Constraint*. Similarly, in the PredCLS task, where oracle object tracks from ground truth are provided, `UNO` surpasses all other methods across various metrics, reaching 80.3% R@20 under *With Constraint* and 98.1% R@20 under *No Constraint*. Despite multi-stage methods benefiting from oracle tracks and directly aggregating accurate spatio-temporal context, `UNO` still outperforms them. However, our results indicate room for improvement in object localization.

Results on Pixel-Level VidSGG (PVSG). The PVSG dataset presents highly dynamic videos and frequent substantial changes in camera angles. Tab. 3 demonstrates that both the Image Panoptic Segmentation + Tracking [4, 55] (IPS+T) model and Video Panoptic Segmentation [4, 26] (VPS) baselines fall short compared to our end-to-end `UNO`. It is essential to focus on R/mR@20, as it serves as a primary performance metric [60]. Notably, our model substantially outperforms both IPS+T and VPS at R/mR@20 across mask

Table 2. Performance comparison with SOTA DSGG methods on Action Genome dataset. The best results are in **bold**.

Method	Backbone	With Constraint						No Constraint					
		SGDET			PredCLS			SGDET			PredCLS		
		R@10↑	R@20↑	R@50↑	R@10↑	R@20↑	R@50↑	R@10↑	R@20↑	R@50↑	R@10↑	R@20↑	R@50↑
Multi-stage Method													
STTran [8] ICCV'21	ResNet-101	25.2	34.1	37.0	68.6	71.8	71.8	24.6	36.2	48.8	77.9	94.2	99.1
APT [27] CVPR'22	ResNet-101	26.3	36.1	38.3	69.4	73.8	73.8	25.7	37.9	50.1	78.5	95.1	99.2
STTran-TPI [53] ACM MM'22	ResNet-101	26.2	34.6	37.4	69.7	72.6	72.6	-	-	-	-	-	-
TR^2 [50] ICRA'23	ResNet-101	26.8	35.5	38.3	70.9	73.8	73.8	27.8	39.2	50.0	83.1	96.6	99.9
VsCGG [33] ACM MM'23	ResNet-101	27.4	35.8	38.2	70.1	73.4	73.5	29.3	40.2	48.9	78.8	94.9	99.2
TEMPURA [34] CVPR'23	ResNet-101	28.1	33.4	34.9	68.8	71.5	71.5	29.8	38.1	46.4	80.4	94.2	99.4
DSG-DETR [13] WACV'23	ResNet-101	30.3	34.8	36.1	-	-	-	32.1	40.9	48.3	-	-	-
TPT [62] TMM'23	ResNet-101	-	-	-	-	-	-	32.0	39.6	51.5	85.6	97.4	99.9
TD^2-Net [30] AAAI'24	ResNet-101	28.7	-	37.1	70.1	-	73.1	30.5	-	49.3	81.7	-	99.8
One-stage Method													
OED [49] CVPR'24	ResNet-50	33.5	40.9	48.9	73.0	76.1	76.1	35.3	44.0	51.8	83.3	95.3	99.2
UNO (Ours)	ResNet-50	35.4	42.2	49.5	73.7	76.9	78.1	36.6	46.1	53.9	84.7	96.1	99.9
	ViT-S/14	36.7	43.1	50.2	74.2	78.5	79.6	37.5	47.5	54.5	85.9	96.6	**100.0**
	ViT-B/14	38.2	44.7	51.9	75.6	79.4	80.4	39.9	48.2	56.3	87.4	97.2	**100.0**
	ViT-L/14	**39.3**	**45.2**	**53.8**	**76.8**	**80.3**	**82.5**	**40.8**	**49.7**	**57.1**	**88.3**	**98.1**	**100.0**

Table 3. Performance comparison with SOTA PVSG methods on PVSG dataset. The best results are in **bold**. Next, ♢ and ♡ stands for the relation predictor [60]: 1D Convolution and Transformer Encoder, respectively.

Method	Backbone	vIOU Threshold = 0.5						vIOU Threshold = 0.1					
		R@20↑	R@50↑	R@100↑	mR@20↑	mR@50↑	mR@100↑	R@20↑	R@50↑	R@100↑	mR@20↑	mR@50↑	mR@100↑
Multi-stage Method													
Image Panoptic Segmentation + Tracking [4, 55]													
♢PVSG [60]	ResNet-50	3.88	5.24	6.71	2.55	3.29	5.36	10.06	14.99	18.13	8.98	12.21	15.47
♡PVSG [60]	ResNet-50	3.88	5.66	6.18	2.81	4.12	4.44	9.01	14.88	17.51	6.69	11.28	13.20
♡MCL [36]	ResNet-50	3.98	5.97	7.44	2.98	4.20	5.15	10.59	16.98	22.33	9.56	12.39	17.47
♢MCL [36]	ResNet-50	4.51	6.08	7.76	3.56	4.38	5.86	11.43	17.30	22.85	9.57	13.13	17.48
Video Panoptic Segmentation [4, 26]													
♢PVSG [60]	ResNet-50	0.42	0.63	0.63	0.25	0.67	0.67	8.07	11.01	12.89	7.84	9.78	10.77
♡PVSG [60]	ResNet-50	0.42	0.73	1.05	0.61	0.76	0.92	6.50	9.64	12.26	5.75	8.25	9.51
♡MCL [36]	ResNet-50	0.63	1.05	1.05	0.83	0.76	0.76	6.71	10.27	13.42	6.94	8.68	12.09
♢MCL [36]	ResNet-50	0.84	1.26	1.26	0.98	1.22	1.22	8.18	12.90	14.22	8.00	11.47	13.59
One-stage Method													
UNO (Ours)	ResNet-50	6.23	7.37	8.65	5.60	6.84	8.21	13.83	19.27	24.63	11.65	15.94	19.99
	ViT-S/14	7.45	8.46	9.69	6.83	7.50	9.26	14.11	20.71	25.11	12.40	16.82	20.78
	ViT-B/14	8.71	9.19	10.46	7.56	8.99	9.83	15.76	21.87	26.58	13.12	17.25	21.81
	ViT-L/14	**9.44**	**10.83**	**11.59**	**8.25**	**9.72**	**10.86**	**17.54**	**23.82**	**27.32**	**14.81**	**18.31**	**22.14**

Table 4. Multi-task training ablation on PVSG dataset.

Training setting	AP50↑	PQ↑	R@20↑
With bounding box only	28.5	-	-
With mask only	-	47.4	8.03
With both bounding box & mask	**30.6**	**48.9**	**9.44**

Table 5. Temporal consistency learning ablation.

Method	DSGG task			PVSG task		
	R@10↑	R@20↑	R@50↑	R@20↑	R@50↑	R@100↑
w/o $\mathcal{L}_{\text{consistency}}$	38.1	41.3	49.4	7.16	9.23	12.17
w/ $\mathcal{L}_{\text{consistency}}$	**39.3**	**45.2**	**53.8**	**9.44**	**10.83**	**11.59**

overlaps of 0.5 and 0.1. For instance, UNO holds the current peak for R@20 at 9.44% in the vIOU threshold of 0.5, indicating that, on average, 01 in every 13 ground-truth triplets is successfully recalled, compared to IPS+T's best performance of 01 in roughly every 25 triplets (R@20 of 4.51%). However, lowering the threshold to a more lenient 0.1 raises UNO's score to approximately 17.54%, allowing the model to recall 02 out of every 13 triplets. This suggests that while the model shows higher efficacy than others for recognizing key video content, there remains considerable room for improvement.

4.3. Ablation Studies

Effect of Multi-Task Learning. Tab. 4 highlights the impact of multi-task learning and cross-task synergy on PVSG, which provides both box-level and pixel-level VidSGG ground truths. AP50 denotes Average Precision@0.5, while PQ is the Panoptic Quality [22]. Using UNO, we observe that training with both bounding boxes and masks improves all metrics, with AP50 increasing from 28.5 to 29.6, PQ from 47.4 to 48.3, and R@20 from 8.03 to 9.44. These results suggest that box-level and pixel-level VidSGG data complement each other, enabling richer representations that enhance individual task performances.

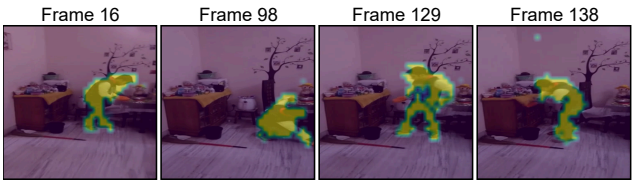

Figure 4. Spatio-temporal consistency of an object slot over time.

Figure 5. Visualization results of relation slots on Action Genome.

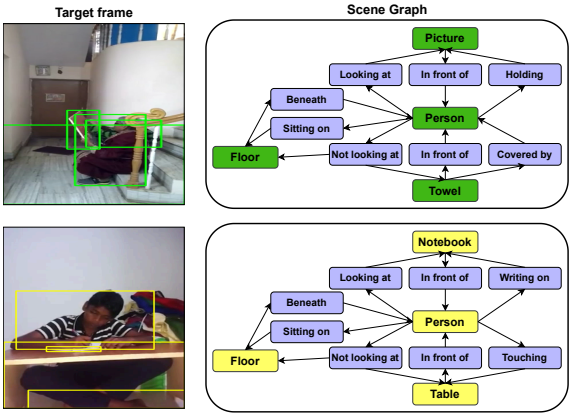

Figure 6. Visualization results on DSGG task. we illustrate a case of a "Person" sitting on the "Floor" while holding a "Picture", with a "Towel" wrapped around his back.

Multi-Object Spatio-Temporal Consistency. By enabling UNO with slot attention to capture modular object features from frozen visual representations, we detect a form of temporal consistency, where object slots retain stable representations as a video sequence evolves, which can be observed from Tab. 5 across both tasks even without $\mathcal{L}_{\text{consistency}}$. By explicitly incorporating $\mathcal{L}_{\text{consistency}}$ that aims to align slots through time, we observe a significant improvement. It is also supported from the results of PVSG in Tab. 3, where volume IOU is involved to consider mask consistency through time. Such finding results in what we term Multi-Object Spatio-Temporal Consistency, where spatial features (slots binding to visual features) and temporal features (object transitions over time) are cohesively integrated for improved accuracy via UNO, with one such case illustrated in Fig. 4.

4.4. Qualitative Results

DSGG. Fig. 4 illustrates a long time step (frame 16 to frame 138) of a test video sequence in Action Genome. It reveals that our method is able to distinguish object instances with semantic structure at the mask level through object slots, even without such labels in Action Genome. This indicates

Figure 7. Visualization results of UNO on PVSG. UNO is qualitatively shown to address a complex task with a video of two dogs (*i.e.* "Dog-4" and "Dog-9") playing around in a living room with a toy (*i.e.* "Toy-7") on the floor (*i.e.* "Floor-1") and the sofa (*i.e.* "Sofa-10"), cluttered in the background are miscellaneous objects (*i.e.* "Shelf-3", "Shelf-8").

that UNO is able to maintain a coherent spatio-temporal representation. Fig. 5 visualize relation slots of a test video sequence in Action Genome. It demonstrates the ability of UNO to capture meaningful relational semantics via slot attention, which are interestingly shown as highlighted areas between actors and objects, indicating that UNO can interpret structural semantics that correspond with relations in a spatial manner. Another example is shown in Fig. 6, where UNO can capture the person, objects, and their interactions.

PVSG. Fig. 7 visualizes the result of UNO on PVSG. On the top part, UNO demonstrates the extraction of semantic masks from an example video, and in the bottom part, it showcases the consistent prediction of object-relation semantics across time. This emphasizes UNO's ability to handle dynamic interactions and complex environments.

5. Conclusion

We propose UNO, a unified framework that effectively addresses both coarse-grained and fine-grained tasks. By incorporating an enhanced slot attention mechanism and object temporal consistency learning, UNO learns robust, modular representations that adapt dynamically to box-level and pixel-level visual granularity. Additionally, we integrate a dynamic triplet prediction module to establish precise, relation-specific associations between objects, improving efficiency while reducing redundancy. Our empirical results on Action Genome and PVSG convey that UNO offers an effective, streamlined solution for VidSGG, advancing the state-of-the-art in unified, object-centric spatio-temporal representation learning for VidSGG.

Limitations. UNO may face challenges in handling object disappearances or reappearances. It also operates with a fixed number of slots that limit its adaptability to complex dynamics, such as crowded environments, fast-motion scenarios, or when numerous small objects quickly maneuver.

Broader Impacts. UNO introduces a new paradigm for unified VidSGG, providing a computationally efficient and flexible framework that facilitates easy development and serves as a structured representation generation to enhance a wide range of video understanding and reasoning tasks.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9630–9640, 2021. 4

[2] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1–26, 2023. 2

[3] Mu Chen, Liulei Li, Wenguan Wang, and Yi Yang. DIFFVSGG: diffusion-driven online video scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1280–1289, 2022. 6, 7

[5] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014. 4

[6] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 3

[7] Nhat Chung, Taisei Hanyu, Toan Nguyen, Huy Le, Frederick Bumgarner, Duy Nguyen Ho Minh, Khoa Vo, Kashu Yamazaki, Chase Rainwater, Tung Kieu, Anh Nguyen, and Ngan Le. Rethinking progression of memory state in robotic manipulation: An object-centric perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. 3

[8] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16372–16382, 2021. 2, 7

[9] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7023–7032, 2019. 2

[10] Shuangrui Ding, Rui Qian, Haohang Xu, Dahua Lin, and Hongkai Xiong. Betrayed by attention: A simple yet effective approach for self-supervised video object segmentation. In *Proceedings of the European Conference on Computer Vision*, 2024. 4

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021. 6

[12] Jack Edmonds and Richard M. Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *J. ACM*, 19(2):248–264, 1972. 6

[13] Shengyu Feng, Hesham Mostafa, Marcel Nassar, Somdeb Majumdar, and Subarna Tripathi. Exploiting long-term dependencies for generating dynamic scene graphs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5130–5139, 2023. 7

[14] Tobias Fischer et al. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE TPAMI*, 45(12):15380–15393, 2023. 5

[15] Taisei Hanyu, Nhat Chung, Huy Le, Toan Nguyen, Yuki Ikebe, Anthony Gunderman, Duy Nguyen Ho Minh, Khoa Vo, Tung Kieu, Kashu Yamazaki, Chase Rainwater, Anh Nguyen, and Ngan Le. Slotvla: Towards modeling of object-relation representations in robotic manipulation. *arXiv preprint arXiv:2511.06754*, 2025. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016. 6

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020. 6

[18] Jan Hendrik Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6477, 2017. 6

[19] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 1, 2, 3, 6

[20] Baoxiong Jia, Yu Liu, and Siyuan Huang. Improving object-centric learning with query optimization. In *Proceedings of the International Conference on Learning Representations*, 2023. 4

[21] Jindong Jiang, Fei Deng, Gautam Singh, Minseung Lee, and Sungjin Ahn. Slot state space models. *CoRR*, abs/2406.12272, 2024. 3

[22] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9404–9413, 2019. 7

[23] Chi-Hsi Kung, Shu-Wei Lu, Yi-Hsuan Tsai, and Yi-Ting Chen. Action-slot: Visual action-centric representations for multi-label atomic activity recognition in traffic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18451–18461, 2024. 3, 4

[24] Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. SCOUTER: slot attention-based classifier for explainable image recognition.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1026–1035, 2021. 4

[25] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video k-net: A simple, strong, and unified baseline for video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18825–18835, 2022. 5

[26] Xiangtai Li et al. Video k-net: A simple, strong, and unified baseline for video segmentation. In *CVPR*, 2022. 6, 7

[27] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13874–13883, 2022. 1, 2, 3, 7

[28] Yanjun Li, Zhaoyang Li, Honghui Chen, and Lizhi Xu. Unbiased video scene graph generation via visual and semantic dual debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[29] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020. 2

[30] Xin Lin, Chong Shi, Yibing Zhan, Zuopeng Yang, Yaqi Wu, and Dacheng Tao. Td^2-net: Toward denoising and debiasing for video scene graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3495–3503, 2024. 7

[31] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 11525–11538, 2020. 2, 3, 4

[32] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 5

[33] Jiale Lu, Lianggangxu Chen, Youqi Song, Shaohui Lin, Changbo Wang, and Gaoqi He. Prior knowledge-driven dynamic scene graph generation with causal inference. In *Proceedings of the ACM International Conference on Multimedia*, pages 4877–4885, 2023. 7

[34] Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K Roy-Chowdhury. Unbiased scene graph generation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22803–22813, 2023. 1, 2, 7

[35] Tushar Nagarajan and Kristen Grauman. Shaping embodied agent behavior with activity-context priors from egocentric video. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 29794–29805, 2021. 3

[36] Thong Thanh Nguyen, Xiaobao Wu, Yi Bin, Cong-Duy T. Nguyen, See-Kiong Ng, and Anh Tuan Luu. Motion-aware contrastive learning for temporal panoptic scene graph generation. In *AAAI Conference on Artificial Intelligence*, 2025. 1, 2, 7

[37] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. 4, 6

[38] Rohith Peddi, Saksham Singh, Saurabh, Parag Singla, and Vibhav Gogate. Towards scene graph anticipation. In *Proceedings of the European Conference on Computer Vision*, pages 159–175, 2025. 1, 2, 3

[39] Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16629–16641, 2023. 4

[40] Rui Qian, Shuangrui Ding, and Dahua Lin. Rethinking image-to-video adaptation: An object-centric perspective. In *Proceedings of the European Conference on Computer Vision*, pages 329–348, 2024. 3

[41] Gorjan Radevski, Marie-Francine Moens, and Tinne Tuytelaars. Revisiting spatio-temporal layouts for compositional action recognition. In *Proceedings of the British Machine Vision Conference*, pages 110–125, 2021. 2

[42] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019. 6

[43] Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. Action scene graphs for long-form understanding of egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2

[44] Junyao Shi, Jianing Qian, Yecheng Jason Ma, and Dinesh Jayaraman. Composing pre-trained object-centric representations for robotics from "what" and "where" foundation models. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 15424–15432, 2024. 3

[45] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *Proceedings of the British Machine Vision Conference*, pages 310–334, 2021. 4

[46] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Proceedings of the International Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248, 2017. 6

[47] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13688–13697, 2021. 1, 2

[48] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3554–3563, 2023. 4

[49] Guan Wang, Zhimin Li, Qingchao Chen, and Yang Liu. Oed: Towards one-stage end-to-end dynamic scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27938–27947, 2024. 1, 2, 5, 6, 7

[50] 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, 7

[51] Junyao Wang, Arnav Vaibhav Malawade, Junhong Zhou, Shih-Yuan Yu, and Mohammad Abdullah Al Faruque. RS2G: data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7478–7487, 2024. 1, 2

[52] Jinghao Wang, Zhengyu Wen, Xiangtai Li, Zujin Guo, Jingkang Yang, and Ziwei Liu. Pair then relation: Pair-net for panoptic scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10452–10465, 2024. 1, 2, 5

[53] Shuang Wang, Lianli Gao, Xinyu Lyu, Yuyu Guo, Pengpeng Zeng, and Jingkuan Song. Dynamic scene graph generation via temporal prior inference. In *Proceedings of the ACM International Conference on Multimedia*, pages 5793–5801, 2022. 7

[54] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: segmenting objects by locations. In *Proceedings of the European Conference on Computer Vision*, pages 649–665, 2020. 6

[55] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? In *Proceedings of the Advances in Neural Information Processing Systems*, pages 726–738, 2021. 6, 7

[56] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 4

[57] Zhenjia Xu, Zhanpeng He, Jiajun Wu, and Shuran Song. Learning 3d dynamic scene representations for robot manipulation. In *Proceedings of the Conference on Robot Learning*, pages 126–142, 2020. 1, 2

[58] Zihui Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In *Proceedings of the Advances in Neural Information Processing Systems*, 2023. 3

[59] Jingkang Yang, Jun Cen, Wenxuan Peng, Shuai Liu, Fangzhou Hong, Xiangtai Li, Kaiyang Zhou, Qifeng Chen, and Ziwei Liu. 4d panoptic scene graph generation. In *Proceedings of the Advances in Neural Information Processing Systems*, 2023. 1, 2

[60] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, and Ziwei Liu. Panoptic video scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18675–18685, 2023. 1, 2, 3, 6, 7

[61] Andrii Zadaianchuk et al. Object-centric learning for real-world videos by predicting temporal feature similarities. In *NeurIPS*, 2023. 5

[62] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. End-to-end video scene graph generation with temporal propagation transformer. *IEEE Transactions on Multimedia*, 2023. 1, 7

[63] Yi Zhou et al. Slot-vps: Object-centric representation learning for video panoptic segmentation. In *CVPR*, pages 3083–3093, 2022. 3, 4

[64] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. VIOLA: imitation learning for vision-based manipulation with object proposal priors. *CoRR*, abs/2210.11339, 2022. 3

[65] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan S. Kankanhalli. Explainable video action reasoning via prior knowledge and state transitions. In *Proceedings of the ACM International Conference on Multimedia*, pages 521–529, 2019. 1, 2