HIG: Hierarchical Interlacement Graph Approach to Scene Graph Generation in Video Understanding

Trong-Thuan Nguyen, Pha Nguyen, Khoa Luu
CVIU Lab, University of Arkansas
{thuann, panguyen, khoaluu}@uark.edu

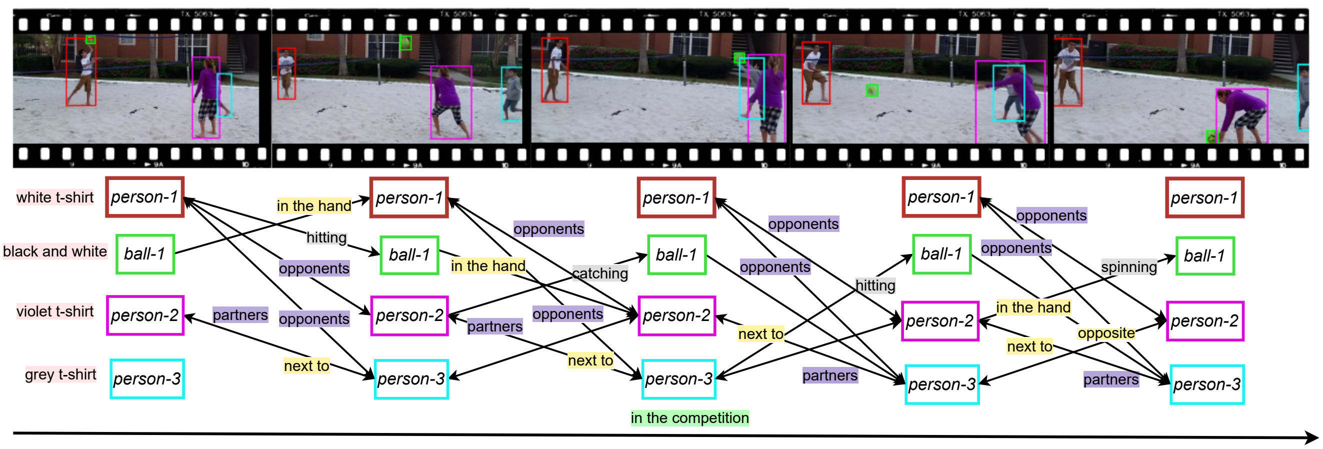


Figure 1. An example from our *ASPIRe* dataset for Visual Interactivity Understanding. The top row shows keyframes with the bounding boxes. **Appearance**, **Situation**, **Position**, **Interaction**, and **Relation** are attributes presented in the dataset. **Best viewed in color**.

Abstract

Visual interactivity understanding within visual scenes presents a significant challenge in computer vision. Existing methods focus on complex interactivities while leveraging a simple relationship model. These methods, however, struggle with a diversity of appearance, situation, position, interaction, and relation in videos. This limitation hinders the ability to fully comprehend the interplay within the complex visual dynamics of subjects. In this paper, we delve into interactivities understanding within visual content by deriving scene graph representations from dense interactivities among humans and objects. To achieve this goal, we first present a new dataset containing Appearance-Situation-Position-Interaction-Relation predicates, named ASPIRe, offering an extensive collection of videos marked by a wide range of interactivities. Then, we propose a new approach named Hierarchical Interlacement Graph (HIG), which leverages a unified layer and graph within a hierarchical structure to provide deep insights into scene changes across five distinct tasks. Our approach demonstrates superior performance to other methods through extensive experiments conducted in various scenarios.

1. Introduction

Visual interaction and relationship understanding have witnessed significant advancements in computer vision in recent years. Various methods, including deep learning, have been introduced, particularly in achieving advanced comprehension of diverse relationships for a holistic visual understanding. Traditional methods span from action recognition and localization to intricate processes like video captioning [19, 41, 46], spatio-temporal detection [35, 51] and video grounding [17, 22, 29]. However, these tasks often interpret visual temporal sequences in a constrained, uni-dimensional way. In addition, relation modeling techniques, including scene graph generation [13, 42, 44] and visual relationship detection [27, 49], adhere to predefined relation categories, limiting the scope for discovering more diverse relationships.

Delving into the Visual Interactivity Understanding problem [13, 27, 44], we introduce a new dataset, characterized by 5× larger interactivity types, including **A**ppearance-**S**ituation-**P**osition-**I**nteraction-**Re**lation, named *ASPIRe*. To this end, we introduce the Hierarchical Interlacement Graph (HIG), a novel approach to the Interactivity Understanding problem. The proposed HIG framework integrates the evolution of interactivities over time. It presents an intuitive

Table 1. Comparison of available datasets. # denotes the number of the corresponding item. The top sub-block of the table is the summary of image datasets, and the bottom is video datasets. **Single** and **Double** are the attribute types as defined in Subsec. 4.1. **H-H**, **H-O**, **O-O** indicate the interactivity between *Human and Human*, *Human and Object*, *Object and Object*.

Datasets	#Videos	#Frames	#Subjects	#RelCls	#Settings	Annotations			Attributes			
									Single	Double		
						BBox	Mask	#Annotations		H-H	H-O	O-O
Visual Genome [15]	-	**108K**	**33K**	**42K**	1	✓	✗	**3.8M**	✗	✗	✓	✓
PSG [42]	-	49K	80	56	1	✓	✓	538.2K	✗	✓	✓	✓
VidOR [27]	**10K**	-	80	50	1	✓	✗	50K	✗	✓	✓	✓
Action Genome [13]	**10K**	234K	25	25	1	✓	✗	**476.3K**	✗	✗	✓	✗
VidSTG [49]	**10K**	-	80	50	1	✓	✗	50K	✗	✓	✓	✓
EPIC-KITCHENS [6]	700	11.5K	21	13	1	✓	✗	454.3K	✗	✗	✓	✗
PVSG [44]	400	153K	126	57	1	✓	✓	-	✗	✓	✓	✓
ASPIRe (Ours)	1.5K	**1.6M**	**833**	**4.5K**	**5**	✓	✓	167.8K	✓	✓	✓	✓

modeling technique and lays the groundwork for enriched comprehension of visual activities and complex interactivities. HIG operates with a unique *unified layer* at every level to jointly process interactivities. This strategy simplifies operations and eliminates the intricacies of multilayers. Instead of perceiving video content as a monolithic block, HIG models an input video with a *hierarchical structure*, promoting a holistic grasp of object interplays. Each level delves into essence insights, benefiting from the features of different levels, thereby capturing scene changes over time.

In addition, the proposed HIG framework promotes dynamic *adaptability* and *flexibility*, empowering the model to adjust its structure and functions to capture the interactivities throughout video sequences. This adaptability is further showcased as the HIG framework proficiently tackles five distinct tasks, demonstrating its extensive flexibility in decoding various interactivity nuances. The proposed HIG framework is not confined to specific tasks or domains, emphasizing its broad applicability and potential.

The Contributions of this Work. There are three main contributions to this work. First, we develop a new dataset named *ASPIRe* for the Visual Interactivity Understanding problem, augmented with a high number of predicate types to capture the complex interplay in the real world. Second, we propose the Hierarchical Interlacement Graph (HIG), standing out with its hierarchical graph structure and unified layer to ensure scalability and flexibility, comprehensively capturing intricate interactivities within video content. Finally, through comprehensive experiments, including the evaluation of other methods on our *APSIRe* dataset and HIG model on both video and image datasets, we prove the advantages of the proposed approach that achieves the State-of-the-Art (SOTA) results.

2. Related Work

2.1. Dataset and Benchmarks

Dataset. Action Genome [13] introduces a comprehensive video database with action and spatiotemporal scene graph annotations. VidOR [27] and EPIC-KITCHENS [6] focus on object and relationship detection and egocentric action recognition. Ego4D [10], VidSTG [49], and PVSG [44] further enrich scene understanding and video scene graph resources. These datasets provide crucial benchmarks for evaluating scene understanding, detailed in Table 1.

Benchmarks. Current benchmarks primarily rely on relation classification for identifying inter-object associations. Action Genome [13] integrates spatiotemporal to Visual Genome [15] to establish scene graphs with action recognition using SGFB. VidOR [27] provides 10K videos for benchmarking video object detection and visual relation detection. EPIC-KITCHENS-100 [6] offers a varied dataset with 100 hours of video, 20M frames, and 90K actions. Ego4D [10] focuses on first-person video data, addressing past, present, and future aspects across nearly 3.6K videos. VidSTG [49] introduces the Video Grounding for Multi-Form Sentences (STVG) task, augmenting VidOR with additional sentence annotations. Recently, PVSG [44] expanded PSG [42], advancing video graph generation.

2.2. Interactivity Modeling Approaches

Video Situation Recognition. The VidSitu [26] benchmark provides a collection of events and situations for evaluation, covering verb prediction, semantic role prediction, and event relations prediction. In a related approach within this benchmark, VideoWhisperer [14] adopts a global perspective for video comprehension, utilizing self-attention across all video clips. Furthermore, the LVU [36] benchmark is tailored for self-supervised video representation learning, with a strong focus on hierarchical methodologies.

Video Understanding. This contains a wide range of tasks and research efforts. Action recognition has advanced significantly through graph-based [38], few-shot learning [31, 34], and transformer-based [4] approaches. Another area of interest is Object Retrieval [24, 45] and Spatiotemporal Detection [23, 35, 51], which involves object detection/segmentation, relation detection and moment retrieval in video content. Additionally, there are challenges such as Visual Question Answering [33, 37, 37] and Video Captioning [19, 41, 46]. Recently, Video Grounding [17, 22, 29, 40] has provided

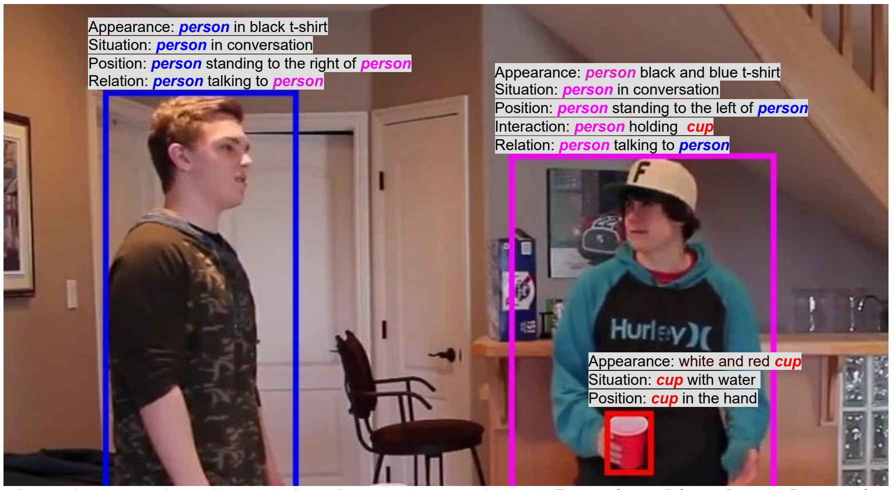
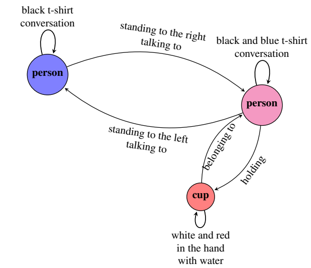


Figure 2. Example and annotations in our *ASPIRe* dataset. **Best viewed in color and zoom in.**



(a) A graph representation of the attributes in Fig. 2.

S_j \ S_i		Person	Object
Person	Position	✓	✓
	Interaction	✗	✗
	Relation	✓	✓
Object	Position	✗	✗
	Interaction	✓	✗
	Relation	✗	✗

(b) Summary of annotated *double-actor* attributes between two actors in our *ASPIRe* dataset. *appearance* and *situation* are *single-actor* attributes as in 4.1.

activities through natural language in visual content.

Scene Graph Generation. Biswas et al. [1] introduce a Bayesian strategy for debiasing scene graphs in images, enhancing recall without retraining. PE-Net [52] leveraging prototype alignment to improve entity-predicate matching in a unified embedding space, incorporating novel learning and regularization to reduce semantic ambiguity. PSGTR [42] and PSGFormer [42] introduce recent innovations in scene graph generation, which utilizes a transformer encoder-decoder to implicitly model scene graph triplets. Recently, PSG4DFormer [43] has been proposed to predict segmentation masks and then track them to create associated scene graphs through a relational component.

For dynamic scenes, TEMPURA [25] utilizes temporal consistency and memory-guided training to enhance the detection of infrequent visual relationships in videos. Cho et al. [5] introduce the Davidsonian Scene Graph (DSG) for assessing text-to-image alignment, operating a VQA module to process atomic propositions from text prompts and quantifying the alignment between text and image. Further, advancements by [9, 16, 21, 42] have adapted scene graph techniques to video, focusing on temporal relationships and advancing comprehensive scene understanding.

2.3. Limitations of Prior Datasets

Existing datasets exhibit notable limitations that hinder a comprehensive understanding of interactivity within visual content. Many of these datasets primarily focus on *a limited set of interactivity types*, overlooking the complexity of real-world interactions. This restricted scope has impeded the development of models capable of handling a variety of interactivities, thereby limiting their applicability to diverse scenarios. Moreover, previous datasets predominantly emphasize relationships within *single connected components of the relational graph*, neglecting complex scenes. Sparse annotations in some datasets further constrain relationship modeling, often failing to provide comprehensive coverage and potentially leading to model bias.

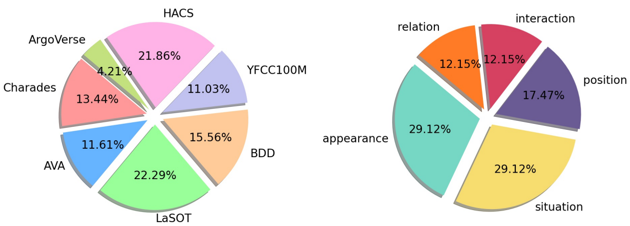
To address these limitations, we introduce the new *ASPIRe* dataset to Visual Interactivity Understanding. The diversity of the *ASPIRe* dataset is showcased through its wide range of scenes and settings, distributed in seven scenarios. Therefore, *ASPIRe* distinguishes itself from earlier datasets, including five types of interactivity, as in Fig. 2.

3. Dataset Overview

3.1. Dataset Collection and Annotation

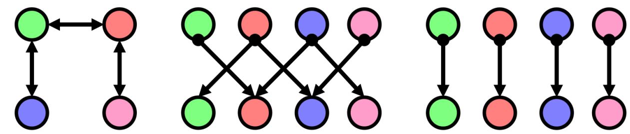
We introduce a dataset compiled from seven distinct sources, each contributing unique perspectives to our collection. The ArgoVerse [3] and BDD [47] datasets focus on outdoor driving scenes, providing valuable insights into real-world traffic scenarios. In contrast, the LaSOT [7] and YFCC100M [32] datasets consist of in-the-wild videos, capturing a diverse spectrum of human experiences and online interactions. Additionally, our dataset incorporates content from the AVA [11], Charades [28], and HACS [50] datasets, encompassing videos that depict various human interactions, including interactions between humans and objects. This compilation results in a diverse scene featuring 833 objects. Therefore, the *ASPIRe* dataset enhances the understanding of activities, surpassing traditional image datasets like Visual Genome [15] and PSG [42] by integrating video data. This crucial integration brings a dynamic dimension to scene analysis that is conspicuously absent in static datasets. *ASPIRe* stands out for its exceptional detail, demonstrating the dynamic interactivities over time. *ASPIRe* has a depth of interactivities context that is notably comprehensive of other datasets while only presenting the relationship of humans, including VidOR [27], Action Genome [13] and PVSG [44], marking a considerable stride in the scene understanding.

To this end, we introduce a structured annotation file



(a) Video sources. (b) Interactivity types.

Figure 3. Statistics from the proposed *ASPIRe* dataset.



(i) double-actor (ii) single-actor

(a) Interactivities **(b) Interlacements**

Figure 4. The terminologies used in our proposed *ASPIRe* dataset and *Hierarchical Interlacement Graph*.

anchored by a primary key named `data`. This file assembles dictionaries associated with a particular frame and detailed annotations. Each dictionary contains two crucial lists: `segments_info` and `annotations`. The `segments_info` list is a collection of dictionaries that describe the individual segments of the image, and the `annotations` list consists of dictionaries that offer bounding boxes and masking details for each segment. Additionally, objects identified within these segments and annotations are assigned the `track_id` to maintain the identity within a video. In particular, the annotations within the *ASPIRe* dataset are distinguished by five interactivity descriptors: (i) `appearances` details visual traits of subjects or objects; (ii) `situations` describes the environmental context; (iii) `positions` identifies the location and orientation; (iv) `interactions` captures the dynamic actions between *Human-Object*; (v) `relations` define the connections and associations between *Human-Human*.

3.2. Dataset Statistics

The *ASPIRe* dataset is quantitatively analyzed in Table 1 and visually represented in Fig. 3. *ASPIRe* contains 1,488 videos covering 833 object categories and 4,549 interactivities, including appearances, situations, positions, interactions, and relationships. The dataset is especially remarkable for its videos that depict a comprehensive and intricate variety of interactivities among subjects, with the number of appearances recorded at 722, situations at 2,902, positions at 130, interactions at 565, and relations at 230. Furthermore, the dataset features objects annotated with boxes and masks, amounting to 167,751 detail annotations.

We provide a detailed analysis of average occurrences within each video of the *ASPIRe* dataset. On average, subjects are featured at 4.5 per video, showcasing diversity in the presence of objects. Both the frequency of appearances and situations remain steady at an average of 4.5 occurrences per video, suggesting a uniform representation of visual elements and their contextual narratives. Positions have a marginally lower average of 4.3 per video. Interactions and relationships averaged around 4.0 instances per video.

4. Methodology

4.1. Terminologies

Fig. 4 illustrates our definitions for analyzing interactivities temporally. Fig. 4a shows the original definition of interactivities within the subjects as annotated in our proposed *ASPIRe* dataset. Interactivities refer to the relationship between subjects. Fig. 4b illustrates a new term *Interlacements*, which are interactivities that span across two or sets of nodes in time or frames. *Interlacements* is our novel design representing how the interactivities evolve in our proposed HIG model, which will be present in the next Section 5. Fig. 4b has two parts, including *double-actor* and *single-actor* attribute interlacements. Fig. 4b(i) defines double-actor attributes. double-actor attributess include *position*, *interaction*, and *relation*, which are attributes that involve two subjects. Fig. 4b(ii) defines single-actor attributes. Single-actor attributes include *appearance* and *situation*, attributes of individual subjects.

4.2. Problem Formulation

Given a video input $\in \mathbb{R}^{T\times H\times W\times 3}$ consisting of T frames and frame size of $H \times W$, we identify a set of distinct subjects, represented as vertices in our graph, $V_t = \{S_i\}_t$ at a particular time t and an interactivity set I as in Eqn. (1).

$$I(S_i, S_j) = \Big\{\mathcal{A}(S_i), \mathcal{S}(S_i), \mathcal{PO}(S_i, S_j), \mathcal{IN}(S_i, S_j), \mathcal{RE}(S_i, S_j)\Big\} \tag{1}$$

It encapsulates all possible interactivities between subjects. Each element in I provides a fine-grain classification of the interactivity types. These interactivities are appearance $\mathcal{A}(S_i)$, situation $\mathcal{S}(S_i)$ to express the single-actor attributes, and position $\mathcal{PO}(S_i, S_j)$, interaction $\mathcal{IN}(S_i, S_j)$ and relation $\mathcal{RE}(S_i, S_j)$ give the double-actor attributes, respectively. The primary objective is to construct a function f. For each pair of subjects and each frame in the video, f identifies the most fitting interactivities from the set I. This function is represented in Eqn. (2).

$$f : V_t \times V_t \rightarrow I \tag{2}$$

For every pair of objects drawn from V_t, the function f learns to predict an interactivity set I, defining the Visual Interactivity Understanding task.

5. Our Proposed Method

Eqn. (2) is the primary objective in this problem. Our design of the graph structure, as in Fig. 5, will be described below.

5.1. Hierarchical Interlacement Graph (HIG)

HIG model is designed to capture the complex dynamics of object interactivity across both spatial and temporal dimensions. It represents a video as a sequence of graphs $\{G_t(V_t, E_t)\}_{t=1}^{T}$ at the first layer, where each graph G_t corresponds to a pair of frames. Here, V_t denotes the set of nodes, and E_t represents the set of edges at time t. As the model progresses through subsequent layers, it combines graphs from the previous layer to form new, more comprehensive graphs, culminating in a single graph cell at the highest level L, representing the entire video interlacement.

HIG Blocks. The HIG model consists of HIG blocks, each representing a distinct level of interactivity within the hierarchical structure. These blocks function consistently across all levels $l \in \{1, \dots, L\}$. At each level l, the model integrates graphs from the previous level to enhance the understanding of interactivity across spatial and temporal dimensions, as detailed in Algorithm 1.

The feature representation $\mathcal{F}_t^{(l)}(S_i)$ is dynamically updated for every node S_i at each level l and time frame t. This update involves transformations and aggregations of information from the neighboring nodes of S_i. Each node S_i in the graph encapsulates a feature set that evolves through the hierarchical levels, progressing horizontally across levels and vertically across time frames, starting from $t = 1$ to $T_l = T - l + 1$ at each level. Specifically, at each level, the model transitions from processing a larger number of simpler graphs to fewer, more complex graphs. The feature representation $\mathcal{F}_t^{(l)}(S_i)$ at level l, with $l > 1$, is derived by aggregating transformed features of neighboring nodes from the previous level $l-1$ as shown in Eqn. (3).

$$\mathcal{F}_t^{(l)}(S_i) = \sum_{S_j \in \mathcal{N}(S_i)} \mathcal{F}_t^{(l-1)}(S_j) \tag{3}$$

In Eqn. (3), the feature representation of a node at level l is the sum of the transformed features of its neighboring nodes from the previous level. For each node S_i, the function $\mathcal{N}$ identifies a set of neighboring nodes that share similar attributes based on similarity scores. This procedure enhances the comprehensiveness of each node feature set as it ascends through the hierarchical layers.

Message-Passing Mechanism. In our hierarchical design, nodes are interconnected through a message-passing mechanism. The message $m_t^{(l)}(S_i, S_j)$ at level l and time t is influenced by the weight matrix $\mathcal{W}_{ij}^{(l)}$ and the feature vector $\mathcal{F}_t^{(l-1)}(S_j)$ transmitted from S_j to S_i. The message from node S_j to S_i is represented as in Eqn. (4).

$$m_t^{(l)}(S_i, S_j) = \mathcal{W}_{ij}^{(l)} \cdot \mathcal{F}_t^{(l-1)}(S_j) \tag{4}$$

Algorithm 1 HIG Construction and Feature Embedding

- **Input:** Frames as graphs $\{G_t(V_t, E_t)\}_{t=1}^{T}$; initial features $\mathcal{F}_t^{(0)}(S_i)$ for each node S_i; number of hierarchical levels L; weight matrices $\mathcal{W}_{ij}^{(l)}$ for all levels $l \in \{1, \dots, L\}$ and node pairs $S_i, S_j \in V_t$.
- **Output:** $I(S_i, S_j)$

for $l = 1$ to L **do**
  $T_l \leftarrow T - l + 1$
  for $t = 1$ to T_l **do**
    $G_{l,t}(V_{l,t}, E_{l,t}) \leftarrow \text{ConstructGraph}(G_t, l)$
    for $S_i \in V_{l,t}$ **do**
      $m_t^{(l)}(S_i, S_j) \leftarrow \mathcal{W}_{ij}^{(l)} \cdot \mathcal{F}_t^{(l-1)}(S_j), \forall S_j \in \mathcal{N}(S_i)$
      $\mathcal{F}_t^{(l)}(S_i) \leftarrow \sum_{t=1}^{T_l} \mathcal{F}_t^{(l-1)}(S_j), \forall S_j \in \mathcal{N}(S_i)$
    end for
  end for
end for
$(V'_t, E'_t) \leftarrow (V'_{L,T_L}, E'_{L,T_L})$
$\{\mathcal{F}'_t(S_i)\}_{S_i \in V'_t} \leftarrow \{\mathcal{F}'_{L,T_L}(S_i)\}_{S_i \in V'_{L,T_L}}$
for $(S_i, S_j) \in V'_t \times V'_t$ **do**
  $I(S_i, S_j) \leftarrow \mathcal{C}\left(m_1^{(L)}(S_i, S_j), \mathcal{F}_1^{(L)}(S_i)\right)$
end for

In Eqn. (4), the message is a product of the weight matrix specific to that level and the feature vector of the sending node. The message $m_t^{(l)}(S_i, S_j)$ is transmitted from node S_j to node S_i shaped by the dimensions of the weight matrix $\mathcal{W}_{ij}^{(l)}$ and the feature vector $\mathcal{F}_t^{(l-1)}(S_j)$. The weight matrix $\mathcal{W}_{ij}^{(l)}$, critical at level l, typically has a shape of $(D_l \times D_{l-1})$, where D_l denotes the feature dimension at level l and D_{l-1} represents the dimension at the preceding level $l-1$. Concurrently, $\mathcal{F}_t^{(l-1)}(S_j)$ is the feature vector of the node S_j from the previous layer, formatted as a column vector with dimensions $(D_{l-1} \times 1)$.

Hierarchical Aggregation. As the HIG model navigates its hierarchical structure, it systematically combines and refines node features from the initial to the ultimate level. This transition is marked by the aggregation and transformation of node features, ensuring that the nuanced details captured at lower levels are integrated into higher-level context. This process culminates at the highest level L, where the model consolidates all the refined features into a single graph cell at $t = 1$ as in Eqn. (5).

$$\mathcal{F}_1^{(L)}(S_i) = \sum_{S_j \in \mathcal{N}(S_i)} \mathcal{F}_1^{(L-1)}(S_j) \tag{5}$$

Eqn. (5) indicates the final feature representation $\mathcal{F}_1^{(L)}(S_i)$ at level L is an aggregation of the transformed features of its neighboring nodes from the previous level. This final representation encapsulates the comprehensive interactivity information from all hierarchical levels.

Interactivity Prediction. For every pair of nodes (S_i, S_j), the function $\mathcal{C}$ is employed to analyze their interactivity. This function considers both the message $m_1^{(L)}(S_i, S_j)$, which

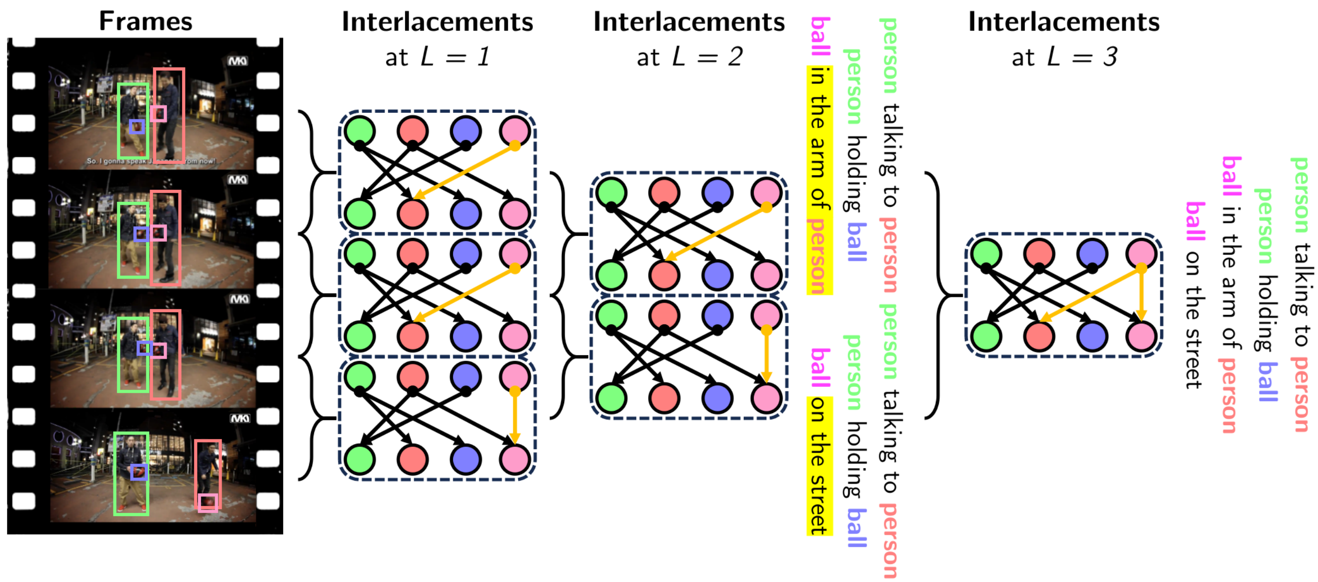


Figure 5. Our proposed *Hierarchical Interlacement Graph*. The highlighted attributes denote the temporal changes in the graph. Then, all predicted interactivities are accumulated into the next hierarchy level. *A higher-level graph cell covers a bigger portion of video frames.*

encapsulates the interactivity between the nodes, and the feature representation $\mathcal{F}_1^{(L)}(S_i)$, which reflects the features of the node S_i at the highest hierarchical level. The prediction function is formulated as in Eqn. (6).

$$I(S_i, S_j) = \mathcal{C}\left(m_1^{(L)}(S_i, S_j), \mathcal{F}_1^{(L)}(S_i)\right) \tag{6}$$

In Eqn. (6), $I(S_i, S_j)$ represents the predicted interactivities between nodes S_i and S_j. The classification function $\mathcal{C}$ operates on the features and messages at the highest hierarchical level to produce a fine-grained classification on the edge connecting these nodes. The output of this function is represented in the set I, where each element provides a detailed classification of the five interactivity types, including appearance ($\mathcal{A}$), situation ($\mathcal{S}$), position ($\mathcal{PO}$), interaction ($\mathcal{IN}$), and relation ($\mathcal{RE}$).

Designing a framework as our HIG model, involving data with varying subjects has distinct advantages. First, graphs are well-suited for the task, where the number of subjects can vary. Second, the message-passing mechanism allows interactivities to be exchanged between neighboring nodes. Finally, HIG allows for a contextual understanding of where and when information occurs in the video, which is essential for tasks that require precise timestamps of events or actions.

5.2. Training Loss

The training loss is an integral component of the HIG model, leveraging a hierarchical weight-sharing strategy and a sequential unfreezing technique. The following section outlines the training process in detail.

Sequential Training Strategy. The HIG framework employs a hierarchical weight-sharing strategy to enhance the efficiency of the training process. By sharing weights across different levels of the GNN hierarchy, the model takes advantage of a reduction in the total number of parameters, which operates as a regularizing mechanism to improve model generalization. In particular, training within the HIG framework is conducted through a sequential unfreezing strategy. Initially, the base level is activated, and subsequent levels are progressively unfrozen. This strategy allows the network to adapt to the feature embeddings $\mathcal{F}_t^{(l)}(S_i)$, which are refined at each level l and time step t.

At each level, the Focal Loss function [18] is employed for edge classification, following [13, 42, 44], as in Eqn. (7).

$$\mathcal{L}(\mathcal{F}_t^{(l)}(S_i)) = -\alpha_t (1 - p_t(\mathcal{F}_t^{(l)}(S_i)))^\gamma \log(p_t(\mathcal{F}_t^{(l)}(S_i))) \tag{7}$$

where p_t measures the probability for the class, α_t is a weighting factor, and γ is a parameter that adjusts the rate.

Loss Aggregation. The losses computed at each hierarchical level are aggregated to determine the total loss for the model as in Eqn. (8). This aggregation ensures that the training signal is comprehensive and encapsulates the learning objectives at each hierarchy level. The HIG framework promotes a nuanced training process, empowering the GNN to model the inherent hierarchical structures.

$$\mathcal{L}_{\text{total}} = \sum_l \mathcal{L}(\mathcal{F}_t^{(l)}(S_i)) \tag{8}$$

6. Experiment Results

6.1. Implementation Details

Dataset. The training set comprises 55K subjects and 197K interactivities across 500 videos, while the validation set,

Table 2. Comparison against baseline methods on single-actor attributes.

Method	Interlacement	R/mR@20	R/mR@50	R/mR@100
Vanilla	Appearance	10.88 / 0.09	12.19 / 0.09	14.16 / 0.08
	Situation	2.87 / 0.02	5.29 / 0.03	9.05 / 0.03
Handcrafted	Appearance	11.09 / 0.11	12.26 / 0.13	14.27 / 0.17
	Situation	3.08 / 0.04	5.36 / 0.07	9.16 / 0.12
Convolution	Appearance	11.32 / 0.11	12.28 / 0.25	14.32 / 0.22
	Situation	3.31 / 0.04	5.38 / 0.19	9.21 / 0.17
Transformer	Appearance	12.35 / **0.62**	13.89 / **0.64**	16.10 / **0.66**
	Situation	4.54 / 0.55	6.99 / **0.58**	10.99 / 0.61
HIG (Our)	Appearance	**15.02** / 0.60	**18.60 / 0.64**	**20.11**/ 0.65
	Situation	**5.01 / 0.56**	**7.02** / 0.55	**12.01 / 0.63**

Table 3. Comparison against baseline methods on double-actor attributes.

Method	Interlacement	R/mR@20	R/mR@50	R/mR@100
Vanilla	Position	10.52 / 0.50	21.97 / 0.55	38.05 / 0.62
	Interaction	10.16 / 0.12	22.35 / 0.13	39.91 / 0.14
	Relation	9.71 / 0.32	21.96 / 0.36	39.11 / 0.40
Handcrafted	Position	10.73 / 0.52	22.04 / 0.59	38.16 / 0.71
	Interaction	10.37 / 0.14	22.42 / 0.17	40.02 / 0.23
	Relation	9.92 / 0.34	22.03 / 0.40	39.22 / 0.49
Convolution	Position	10.96 / 0.52	22.06 / 0.71	38.21 / 0.76
	Interaction	10.60 / 0.14	22.44 / 0.29	40.07 / 0.28
	Relation	10.15 / 0.34	22.05 / 0.52	39.27 / 0.54
Transformer	Position	11.04 / **0.83**	22.52 / 0.90	38.84 / 1.02
	Interaction	10.68 / **0.45**	22.90 / **0.48**	40.70 / **0.52**
	Relation	10.23 / **0.65**	22.51 / **0.71**	39.90 / **0.96**
HIG (Ours)	Position	**13.02** / 0.09	**24.52 / 1.33**	**42.33 / 1.12**
	Interaction	**12.02** / 0.11	**24.65** / 0.12	**41.65** / 0.14
	Relation	**10.26** / 0.29	**23.72** / 0.34	**41.47** / 0.39

derived from a test split, includes 113K subjects and 400 interactivities across 988 videos. In addition, we use PSG [42] to evaluate our performance on the image data.

Model Configurations. This work uses the PyTorch framework and operates on 8$\times$ NVIDIA RTX A6000 GPUs. It utilizes a training batch size of 1 and employs the `AdamW` Optimizer, starting with an initial learning rate of 0.0001. We employ PyTorch Geometric [8] for constructing graphs where nodes represent detections and edges signify potential interactivities. It integrates a ResNet-50 [12] backbone trained with DETR [2]. Our framework involves edge pruning using `scatter_min` and `scatter_max` for aggregating node features such as bounding box coordinates and track identification. Then, the framework calculates cosine similarity and selects the *top-k* ($k = 12$) nearest neighbors.

Metrics. Inspired by [27, 42, 44], we calculate the recall metric for the Visual Interactivity Understanding task to predict a set of triplets that accurately describe the input video. The model predicts the category labels for the subject, object, and predicate within each triplet. Each triplet represents a distinct interactivity in the range time t_1 and t_2. Moreover, each triplet corresponds to a specific subject in single-actor scenarios and a pair of subjects in double-actor scenarios based on a predefined set. To this end, we leverage the standard metrics used in activity understanding, including $R@K$ and $mR@K$ utilized to evaluate the recall of top K categories and their mean recall, respectively.

Table 4. Comparison at different video sampling rates of our HIG.

Sampling Rate	Interlacement	R/mR@20	R/mR@50	R/mR@100	FPS
2 (Half)	Appearance	12.13 / 0.59	12.25 / 0.63	7.48 / 0.64	26.4
	Situation	2.12 / 0.55	5.67 / 0.54	8.62 / 0.62	
	Position	10.13 / 0.08	18.17 / 1.32	29.7 / 1.11	
	Interaction	9.13 / 0.10	18.30 / 0.11	29.02 / 0.13	
	Relation	7.37 / 0.28	17.37 / 0.33	28.84 / 0.38	
1 (Full)	Appearance	15.02 / 0.60	18.60 / 0.64	20.11 / 0.65	24.2
	Situation	5.01 / 0.56	7.02 / 0.55	12.01 / 0.63	
	Position	13.02 / 0.09	24.52 / 1.33	42.33 / 1.12	
	Interaction	12.02 / 0.11	24.65 / 0.12	41.65 / 0.14	
	Relation	10.26 / 0.29	23.72 / 0.34	41.47 / 0.39	

6.2. Ablation Study

Baseline Methods. We re-implemented four baseline methods introduced in [44] and presented in Table 2 and Table 3 since the official implementation is unavailable. Table 2 compares all baseline methods and the HIG along single-actor attributes, and Table 3 compares double-actor attributes.

Instead of evaluating each video frame separately, e.g. level $l = 1$, we prefer the *higher levels* prediction, where confident score ≥ 0.9. A higher hierarchy level covers a more significant portion of the video frame, as in Fig. 5. This approach *effectively reduces noise and produces a higher recall rate*. In particular, the HIG method is better at recognizing single-actor attributes than other baselines, including Transformer, Convolution, Handcrafted, and Vanilla. Specifically, the HIG model is 2.67% higher than the Transformer, the best method in baseline at $R@20$. HIG is also better for the double-actor attributes, especially in figuring out interactions and relations. It is 1.34% higher than Transformers at $R@20$ when identifying interactions. We visualize keyframe predictions in a video, as shown in Fig. 6.

Video Sampling Rates. Table 4 explores the influence of frame sampling rates on the performance of the HIG model in deployment. Our analysis focuses on evaluating the performance under a reduced number of frames. In the ASPIRe dataset, the testing set includes 988 videos, totaling 10,456,48 frames. We address the efficiency of the HIG model by halving the number of frames in each video. In particular, we discard one frame out of every two successive frames. Our experiment demonstrates a trade-off between recall score and inference time. Significantly, the performance of the HIG model decreases in terms of recall, but the FPS increases by 2.2 FPS.

6.3. Comparison with State-of-the-Arts

Performance on *ASPIRe*. We provide the comparative analysis with SOTAs in Table 5, including IMP [39], MOTIFS [48], VCTree [30], and GPSNet [20]. In the ASPIRe dataset, the HIG method shows impressive results in identifying the position on recall at different top K. In addition, the HIG model performs well on identifying relations when it is higher than 1.13% at $R@20$ compared to GPSNet.

Scene Graph Generation (SGG). We extend the capability of the HIG model while incorporating image-based scene graph generation into the training process presented in Ta-

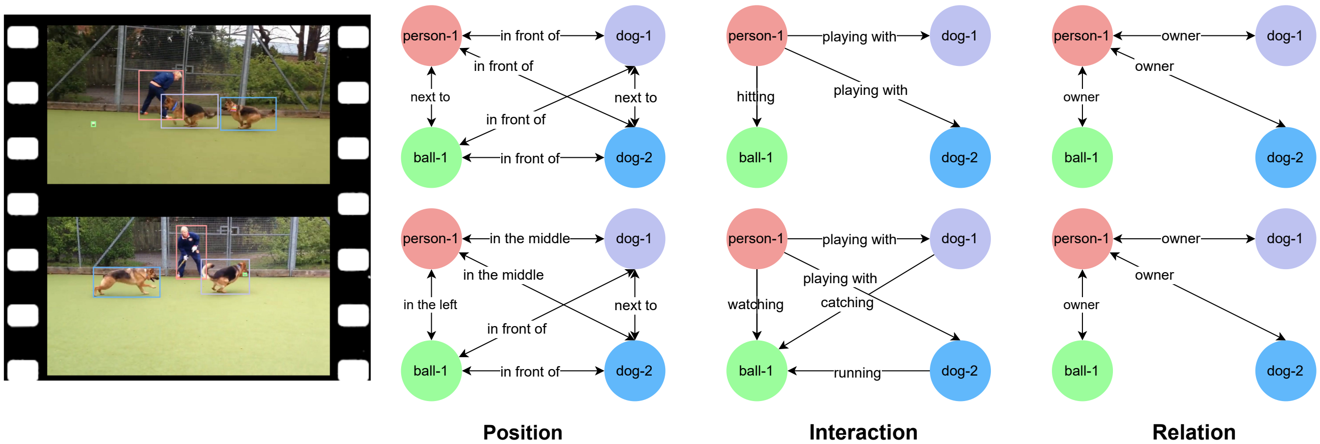


Figure 6. Qualitative results of position, interaction, and relation from scene graphs generated from the HIG model.

Table 5. Comparison against previous methods on *ASPIRe*.

Method	Interlacement	R/mR@20	R/mR@50	R/mR@100
IMP [39]	Position	9.70 / 0.49	9.70 / 0.49	9.70 / 0.49
	Interaction	12.79 / 0.08	12.79 / 0.08	12.79 / 0.08
	Relation	11.51 / 0.32	11.51 / 0.32	11.51 / 0.32
MOTIFS [48]	Position	6.89 / 0.48	8.49 / 0.38	8.70 / 0.40
	Interaction	8.83 / 0.12	10.33 / 0.12	10.57 / 0.12
	Relation	8.72 / 0.32	10.26 / 0.32	10.55 / 0.32
VCTree [30]	Position	4.18 / 0.39	6.75 / 0.40	8.59 / 0.42
	Interaction	6.23 / 0.10	9.58 / 0.10	11.63 / 0.10
	Relation	6.51 / 0.27	9.82 / 0.28	11.51 / 0.28
GPSNet [20]	Position	12.89 / **1.26**	12.89 / 1.26	12.89 / **1.26**
	Interaction	10.89 / **0.11**	10.89 / **0.12**	10.89 / 0.12
	Relation	9.87 / **0.35**	9.87 / **0.35**	9.87 / 0.35
HIG (Ours)	Position	**13.02** / 0.09	**24.52 / 1.33**	**42.33** / 1.12
	Interaction	**12.02 / 0.11**	**24.65 / 0.12**	**41.65 / 0.14**
	Relation	**10.26** / 0.29	**23.72** / 0.34	**41.47 / 0.39**

Table 6. Comparison against previous methods on SGG task.

Method	Interlacement	R/mR@20	R/mR@50	R/mR@100
IMP [39]	Position	0.25 / 0.36	0.29 / 0.35	0.30 / 0.33
	Interaction	0.71 / 0.13	0.98 / 0.12	1.15 / 0.13
	Relation	0.80 / 0.26	0.81 / 0.25	0.84 / 0.24
MOTIFS [48]	Position	0.23 / 0.43	0.23 / 0.43	0.31 /0.38
	Interaction	0.39 / 0.11	0.94 / 0.11	1.17 / 0.10
	Relation	0.31 / 0.30	0.32 / 0.28	0.53 / 0.32
VCTree [30]	Position	0.13 / 0.23	0.14 / 0.22	0.14 / 0.21
	Interaction	0.55 / 0.07	0.65 / 0.08	0.72 / 0.08
	Relation	0.39 / 0.18	0.39 / 0.20	0.43 / 0.21
GPSNet [20]	Position	0.09 / 0.46	1.17 / 0.37	1.32 / 0.46
	Interaction	0.99 / 0.09	1.02 / 0.09	1.11 / 0.09
	Relation	0.14 / 0.23	0.16 / 0.13	0.29 / 0.23
HIG (Ours)	Position	**1.00 / 0.42**	**2.40 / 0.44**	**4.87 / 0.47**
	Interaction	**1.30 / 0.09**	**3.45 / 0.11**	**6.93 / 0.12**
	Relation	**1.26 / 0.27**	**3.43 / 0.30**	**7.02 / 0.32**

ble 6. We only compare the double-actor attributes since the prior method was designed for interactivities in a pair of subjects. The superior performance of the HIG method across all interlacement and metrics in the SGG task underscores its advanced proficiency in attribute recognition within frame-based scene graph generation scenarios. The HIG model is higher than 3.55%, 5.82%, and 6.73% at $R@100$ compared to the best-performing previous method, GPSNet, regarding position, interaction, and relation.

Performance on PSG. In addition, to demonstrate the effectiveness of our method in both image and video datasets,

Table 7. Comparison against previous methods on PSG [42].

Method	R/mR@20	R/mR@50	R/mR@100
IMP [39]	16.5 / 6.52	18.2 / 7.05	18.6 / 7.23
MOTIFS [48]	20.0 / 9.10	21.7 / 9.57	22.0 / 9.69
VCTree [30]	**20.6 / 9.70**	22.1 / 10.2	22.5 / 10.2
GPSNet [20]	17.8 / 7.03	19.6 / 7.49	20.1 / 7.67
PSGFormer [42]	18.6 / 16.7	20.4 / **19.3**	20.7 / **19.7**
HIG (Ours)	19.4 / 6.42	**22.3** / 8.13	**26.3** / 9.70

we present a comparison of the PSG dataset presented in Table 7. Since our model is well-tailored to the video datasets, it performs slightly lower than state-of-the-art methods on the image dataset at $R@20$. However, the graph representation still performs comparable results. Significantly, the HIG model is 3.8% higher than VCTree in terms of $R@100$.

7. Conclusion

We addressed the Visual Interactivity Understanding problem by introducing the *ASPIRe* dataset and the *Hierarchical Interlacement Graph*. The *ASPIRe* dataset, with its extensive range of predicate types, offers a nuanced view of interactivities, setting a new benchmark in the field. Meanwhile, HIG provides a unique hierarchical structure and unified layer, demonstrating exceptional scalability and flexibility in capturing complex interlacement in video content to handle five interactivity types. In addition, we provided extensive experiments showcasing the efficiency of HIG and achieving state-of-the-art results in both video and image datasets.

Limitations. While showcasing significant advancement in understanding interactivities, the HIG approach had limitations, particularly in challenging scenarios. Computing possible interlacements became a computational bottleneck, potentially hindering real-time application feasibility. Furthermore, the framework had to work on remembering in long-duration videos, where the continual learning of new interactivities could lead to the decay of previously acquired knowledge. The HIG model was tailored to video datasets. Therefore, image-based performance might not be optimal.

References

[1] Bashirul Azam Biswas and Qiang Ji. Probabilistic debiasing of scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10429–10438, 2023. 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 7

[3] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019. 3

[4] Jiawei Chen and Chiu Man Ho. Mm-vit multi-modal video transformer for compressed video action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1910–1921, 2022. 2

[5] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023. 3

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 2

[7] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5374–5383, 2019. 3

[8] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. 7

[9] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Not all relations are equal: Mining informative labels for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15596–15606, 2022. 3

[10] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[11] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[13] 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

[14] Zeeshan Khan, CV Jawahar, and Makarand Tapaswi. Grounded video situation recognition. *Advances in Neural Information Processing Systems*, 35:8199–8210, 2022. 2

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2, 3

[16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18869–18878, 2022. 3

[17] Mengze Li, Han Wang, Wenqiao Zhang, Jiaxu Miao, Zhou Zhao, Shengyu Zhang, Wei Ji, and Fei Wu. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23090–23099, 2023. 1, 2

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[19] Wang Lin, Tao Jin, Ye Wang, Wenwen Pan, Linjun Li, Xize Cheng, and Zhou Zhao. Exploring group video captioning with efficient relational approximation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15281–15290, 2023. 1, 2

[20] 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. 7, 8

[21] Xin Lin, Changxing Ding, Yibing Zhan, Zijian Li, and Dacheng Tao. Hl-net: Heterophily learning network for scene graph generation. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19476–19485, 2022. 3

[22] Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23100–23109, 2023. 1, 2

[23] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10840–10849, 2020. 2

[24] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23023–23033, 2023. 2

[25] 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. 3

[26] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5600, 2021. 2

[27] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287, 2019. 1, 2, 3, 7

[28] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 3

[29] Chaolei Tan, Zihang Lin, Jian-Fang Hu, Wei-Shi Zheng, and Jianhuang Lai. Hierarchical semantic correspondence networks for video paragraph grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18973–18982, 2023. 1, 2

[30] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019. 7, 8

[31] Anirudh Thatipelli, Sanath Narayan, Salman Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Bernard Ghanem. Spatio-temporal relation modeling for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19958–19967, 2022. 2

[32] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 3

[33] Aisha Urooj, Hilde Kuehne, Bo Wu, Kim Chheu, Walid Bousselham, Chuang Gan, Niels Lobo, and Mubarak Shah. Learning situation hyper-graphs for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14879–14889, 2023. 2

[34] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19948–19957, 2022. 2

[35] Haoqian Wu, Keyu Chen, Haozhe Liu, Mingchen Zhuge, Bing Li, Ruizhi Qiao, Xiujun Shu, Bei Gan, Liangsheng Xu, Bo Ren, et al. Newsnet: A novel dataset for hierarchical temporal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10669–10680, 2023. 1, 2

[36] Fanyi Xiao, Kaustav Kundu, Joseph Tighe, and Davide Modolo. Hierarchical self-supervised representation learning for movie understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9727–9736, 2022. 2

[37] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2804–2812, 2022. 2

[38] Jiazheng Xing, Mengmeng Wang, Yudi Ruan, Bofan Chen, Yaowei Guo, Boyu Mu, Guang Dai, Jingdong Wang, and Yong Liu. Boosting few-shot action recognition with graph-guided hybrid matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1740–1750, 2023. 2

[39] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 7, 8

[40] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022. 2

[41] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023. 1, 2

[42] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 1, 2, 3, 6, 7, 8

[43] Jingkang Yang, CEN Jun, Wenxuan Peng, Shuai Liu, Fangzhou Hong, Xiangtai Li, Kaiyang Zhou, Qifeng Chen, and Ziwei Liu. 4d panoptic scene graph generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[44] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, et al. 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

[45] Xitong Yang, Fu-Jen Chu, Matt Feiszli, Raghav Goyal, Lorenzo Torresani, and Du Tran. Relational space-time query in long-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6408, 2023. 2

[46] Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, and Ming-Hsuan Yang. Hierarchical modular network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition, pages 17939–17948, 2022. 1, 2

[47] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[48] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018. 7, 8

[49] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020. 1, 2

[50] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019. 3

[51] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, et al. Tuber: Tubelet transformer for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13598–13607, 2022. 1, 2

[52] Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. Prototype-based embedding network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22783–22792, 2023. 3