sensors

MDPI

Article

HAtt-Flow: Hierarchical Attention-Flow Mechanism for Group-Activity Scene Graph Generation in Videos

Naga Venkata Sai Raviteja Chappa [1], Pha Nguyen [1], Thi Hoang Ngan Le [1], Page Daniel Dobbs [2] and Khoa Luu [1,*]

[1] Department of EECS, University of Arkansas, Fayetteville, AR 72701, USA; nchappa@uark.edu (N.V.S.R.C.); thile@uark.edu (T.H.N.L.)
[2] Department of Health, Human Performance and Recreation, University of Arkansas, Fayetteville, AR 72701, USA; pdobbs@uark.edu
* Corresponding author: khoaluu@uark.edu

Abstract: Group-activity scene graph (GASG) generation is a challenging task in computer vision, aiming to anticipate and describe relationships between subjects and objects in video sequences. Traditional video scene graph generation (VidSGG) methods focus on retrospective analysis, limiting their predictive capabilities. To enrich the scene-understanding capabilities, we introduced a GASG dataset extending the JRDB dataset with nuanced annotations involving *appearance, interaction, position, relationship, and situation* attributes. This work also introduces an innovative approach, a **H**ierarchical **Att**ention–**Flow** (HAtt-Flow) mechanism, rooted in flow network theory to enhance GASG performance. Flow–attention incorporates flow conservation principles, fostering competition for sources and allocation for sinks, effectively preventing the generation of trivial attention. Our proposed approach offers a unique perspective on attention mechanisms, where conventional "values" and "keys" are transformed into sources and sinks, respectively, creating a novel framework for attention-based models. Through extensive experiments, we demonstrate the effectiveness of our Hatt-Flow model and the superiority of our proposed flow–attention mechanism. This work represents a significant advancement in predictive video scene understanding, providing valuable insights and techniques for applications that require real-time relationship prediction in video data.

Keywords: group activity recognition; flow-attention; video scene graph

check for updates

Citation: Chappa, N.V.S.R.; Nguyen, P.; Le, T.H.N.; Dobbs, P.D.; Luu, K. HAtt-Flow: Hierarchical Attention-Flow Mechanism for Group-Activity Scene Graph Generation in Videos. *Sensors* **2024**, *24*, 3372. https://doi.org/10.3390/s24113372

Academic Editor: Alicja Wieczorkowska

Received: 11 April 2024
Revised: 14 May 2024
Accepted: 22 May 2024
Published: 24 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Visual scene understanding is a foundational challenge in computer vision, encompassing the interpretation of complex scenes, objects, and their relationships within images and videos. This task is particularly intricate for video, where temporal dynamics and multi-modal information introduce unique complexities. Group-activity video scene graph (GAVSG) generation, which involves predicting relationships between objects in a video across multiple frames, stands at the forefront of this endeavor.

In recent years, significant progress has been made in video understanding. Techniques such as video scene graph generation (VidSGG) have allowed us to extract high-level semantic representations from video content. However, VidSGG typically operates in a static, retrospective manner, constraining its predictive capabilities. The GAVSG dataset, on the other hand, extends the scope of visual scene understanding to anticipate and describe subject-and-object relationships and their temporal evolution.

In the closely linked domain of human–object interaction (HOI) [1], transferable techniques have proven effective for scene graph generation (SGG) tasks [2–18], some of which inspired the foundation of PSG dataset baselines [19,20].

In the context of SGG, diverse methodologies have been explored, from probabilistic graphical models and AND-OR grammar approaches [21–28] to knowledge graph embeddings like VTransE [29] and UVTransE [30]. Recent endeavors delve into challenges

such as the long-tailed distribution of predicates [31,32], visually irrelevant predicates [33], and precise bounding box localization [34]. As shown in Figure 1, we can observe that the previous methods can detect the subjects and objects in a scene. However, they need to generate a well-defined scene graph, whereas our method can learn all the nuanced relationships among the subjects and objects in the scene to produce a fine scene graph.

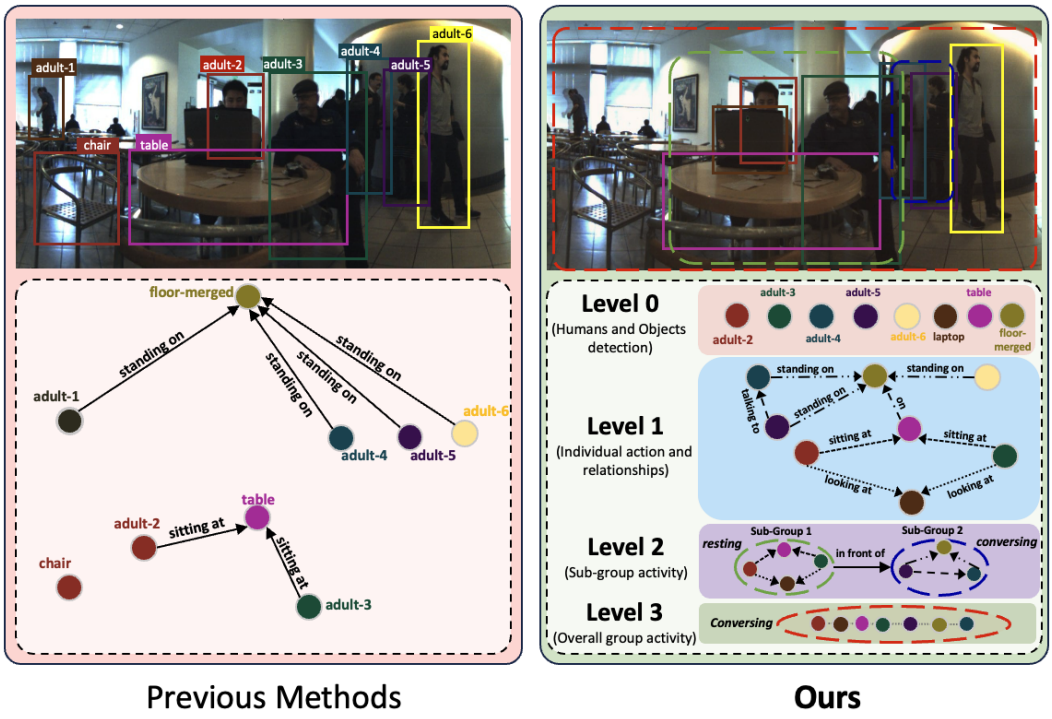


Figure 1. Comparison of HAtt-Flow results with other scene graph generation methods. **Best viewed in color and zoomed in.**

To address the limitation of the enriched learning of group activities in a scene, we introduced the GASG dataset, which includes nuanced annotations in the form of five different attributes. This can help set a better scene graph generation benchmark than the existing datasets in this domain. In this work, we propose a novel approach for GAVSG that draws inspiration from flow network theory, introducing flow–attention. This mechanism leverages flow conservation principles in both the source and sink aspects, introducing a competitive mechanism for sources and an allocation mechanism for sinks. This innovative approach mitigates the generation of trivial attention and enhances the predictive power of GAVSG. We build upon a new perspective on attention mechanisms, rooted in flow network theory, to design our GAVSG framework. The conventional attention mechanism aggregates information from "values" and "keys" based on the similarity between "queries". By framing attention in terms of flow networks, we transform values into sources and keys into endpoints, thus creating a fresh perspective on the attention mechanism.

Contributions: The main contributions of our work are threefold. **First**, we introduce a novel dataset (the dataset is available for verification at https://uark-cviu.github.io/GASG/) with nuanced attributes that aid the scene graph generation task in the group activity setting. **Second**, our work advances the state of the art in predictive video scene understanding by introducing flow–attention and redefining attention mechanisms via incorporating hierarchy awareness. **Third**, we demonstrate the effectiveness of our approach through extensive experiments and achieve state-of-the-art performance over the existing approaches.

2. Related Work

Group action recognition (GAR). Group action recognition (GAR) has witnessed a shift towards deep learning methodologies, notably convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [35–45]. Attention-based models and graph convolution networks are crucial in capturing spatial–temporal relations in group activities. Transformer-based encoders, often coupled with diverse backbone networks, excel in extracting features for discerning actor interactions in multimodal data [46]. Recent innovations, such as MAC-Loss, introduce dual spatial and temporal transformers for enhanced actor interaction learning [47]. The field continues to evolve with heuristic-free approaches like those by Tamura et al., simplifying the process of social group activity recognition and member identification [48].

Scene graph generation (SGG). In scene graph generation (SGG), the traditional two-stage paradigm involves object detection and pairwise predicate estimation [49–60]. Recent advancements include knowledge graph embeddings, graph-based architectures, energy-based models, and linguistic supervision [56,61–69]. To address challenges like long-tailed distribution and visually irrelevant predicates, the field has seen a pivot towards panoptic segmentation-based SGG, inspired by the simultaneous generation of scene graphs and semantic segmentation masks [34]. Notably, insights from the closely linked domain of human–object interaction (HOI) have influenced SGG techniques [2–18,70,71].

Video scene graph generation (VidSGG). VidSGG, initiated by Shang et al. [72], explores spatio-temporal relations in videos. Research has delved into spatio-temporal conditional bias, the domain shift between image and video scene graphs, and embodied semantic approaches using intelligent agents. Notable methods include TRACE [73], which separates relation prediction and context modeling, and embodied semantic SGG, employing reinforcement learning for path generation via intelligent agents [74,75]. Yang et al. [76] proposed a transformer–encoder-based baseline model to evaluate their proposed panoptic video scene graph dataset, which included fusing the extracted features of the subjects in the scene.

Limitation of Prior Datasets

We present a detailed comparison of existing datasets in Table 1. However, upon further examination, it becomes apparent that the limitations of prior datasets are particularly pronounced when considering the intricate nature of group activities within visual content. Many existing datasets have primarily focused on specific types of individual actions, often overlooking the complex dynamics of group activities in real-world interactions. This narrow focus has hindered the development of models capable of addressing a diverse range of action classifications, thereby limiting their adaptability to varied real-world scenarios.

Moreover, earlier datasets have often provided sparse annotations and underscored relationships within isolated fragments of the relational graph, neglecting the complexities of broader and more intricate scenes. This sparse annotation may lead to a lack of comprehensive relationship modeling and potential biases in the developed models.

Furthermore, there exists a significant gap in the representation of scenes featuring dense crowds of people. These settings present formidable challenges related to occlusion management and require a nuanced understanding of complex interactions within such contexts.

Table 1. Comparison of existing datasets. **GA** is the group activity label, **H-H**, **H-O**, and **O-O** represent interactions between *human and human, human and object, and object and object*.

Datasets	Settings	Annotations			Attributes		
		BBox	IDs	GA	H-H	H-O	O-O
ActivityNet [77]	1	✓	✗	✗	✗	✓	✗
MSRVTT [78]	1	✓	✗	✗	✗	✓	✗
MSVD [79]	1	✓	✗	✗	✗	✓	✓
PSG [80]	1	✓	✓	✗	✓	✓	✓
PVSG [76]	1	✓	✓	✗	✓	✓	✓
GASG (Ours)	**5**	✓	✓	✓	✓	✓	✓

In response to these identified limitations, we introduce the *group activity scene graph* (GASG) dataset. This dataset directly addresses the aforementioned issues by featuring various scenes and settings (represented as attributes in the annotations), spanning distinct scenarios and effectively differentiating it from prior datasets. The GASG dataset excels in capturing five critical features of group activities: appearance, situation, position, interaction, and relations. Additionally, it comprehensively tracks the movements and interactions of individuals and sub-groups, facilitating a profound understanding of their dynamics and activities over time. With its rich dataset of these aspects, the GASG dataset lays the groundwork for a new paradigm in understanding complex group activities within scenarios characterized by dense populations, thereby pushing the boundaries of group activity recognition.

3. Dataset Overview

The GASG dataset offers a diverse array of sub-group and group activities, enriching the landscape of research in group activity recognition. It encompasses 27 categories of individual actions, such as walking and talking, aligning with the categories present in the JRDB-Act dataset. Additionally, our dataset presents 11 distinct categories of sub-group activities, ranging from standing closely and chatting to complex interactions like group evolution and collaborative work. These sub-group activities meticulously capture the nuances of interpersonal dynamics within smaller groups, providing a nuanced portrayal of real-world interactions.

Furthermore, the dataset encapsulates seven categories of group activities, including walking, conversing, commuting, resting, office working, waiting, and sitting. These activities encapsulate commonplace scenarios encountered in real-world settings, serving as a robust foundation for analyzing collective behaviors and interactions within larger groups. By incorporating a comprehensive spectrum of activities, the GASG dataset empowers researchers to delve deeply into the complexities of group dynamics and advance the frontiers of group activity recognition.

3.1. Data Collection and Annotation

This GASG dataset comprises a rich collection of videos from the JRDB dataset, offering a unique perspective on sub-group and overall group activities. This dataset provides comprehensive coverage of various activities and introduces essential tracking information.

The tracking information within the GASG dataset (as shown in Figure 2) facilitates a detailed understanding of individuals and sub-groups within each frame. This information includes the trajectory, position, and interactions of each actor. The dataset defines five key aspects for comprehensive scene understanding:

- *Interaction*: This aspect characterizes the dynamic interactions between subjects and objects, shedding light on how individuals and sub-groups engage with each other.
- *Position*: The dataset includes precise data on the location and orientation of subjects and objects, enhancing the analysis of their spatial relationships during activities.

- *Appearance*: Visual traits of subjects and objects are meticulously captured, allowing for detailed examinations of their attributes and characteristics.
- *Relationship*: Understanding the associations and connections between subjects and objects is essential for deciphering the complex interplay within group activities. This aspect provides insight into the underlying dynamics of relationships within the scenes.
- *Situation*: To provide environmental context, the GASG dataset offers descriptors highlighting the contextual information surrounding subjects and objects, enabling researchers to consider the broader setting in their analyses.

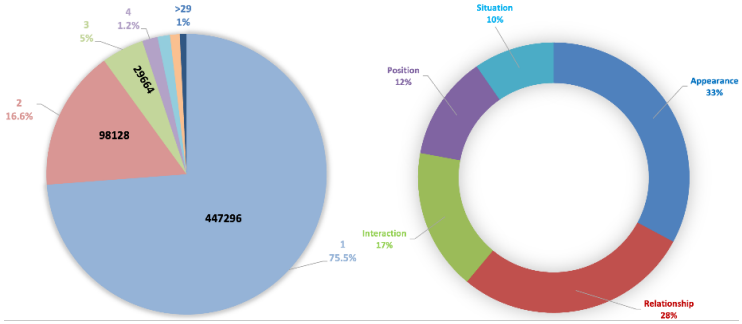


Figure 2. A sample video from our group activity scene graph (GASG) dataset. The top row displays keyframes featuring overlaid bounding boxes, each annotated with a unique ID for consistency. Below, the timeline tubes provide a comprehensive temporal representation of scene graph annotations for distinct attributes, including ***appearance, interaction, position, relationship, and situation***. These annotations offer nuanced details, enhancing scene understanding and contributing to a more refined video content analysis. **Best viewed in color and zoomed in.**

These five key aspects—interaction, position, appearance, relationship, and situation—form the backbone of the dataset's annotation structure, providing a holistic view of the diverse activities and interactions within the sub-group and overall group scenarios. This level of detail sets GASG apart, making it a valuable resource for research in scene comprehension, action recognition, and group activity analysis. The annotation process is detailed in the Appendix B.

3.2. Dataset Statistics

The accompanying pie chart in Figure 3b delves into the complexity of attributes in our dataset, revealing that "Appearance" consists of 33% of the total annotations, which is the highest, "Relationship" consists of 28%, "Interaction" consists of 17%, "Position" consists of 12%, and "Situation" consists of 10%, which is the least. We explore the distribution of social activity labels in Figure 3a, focusing on the sizes of social groups. The chart in the figure provides a nuanced view of social group sizes in the dataset. Specifically, 75.5%, 16.6%, 5%, and 1.2% of social groups consist of one, two, three, and four members, respectively. Interestingly, only 1% of the dataset includes groups with five or more members, with the maximum observed group size being 29 members.



a.) Number of Social Groups b.) Distribution of Attributes

Figure 3. Statistics of the GASG dataset, number of social groups, and attributes in the dataset. **Best viewed in color and zoomed in.**

4. Methodology

In this section, we present the methodology of our proposed HAtt-Flow approach for robust group activity recognition. We introduce three key modules: input preparation, hierarchical awareness induction, and feature flow–attention mechanism, which are overviewed below:

Input preparation: We prepare input node and edge embeddings for the graph transformer layer. This module uniquely incorporates both textual and visual features, facilitating a holistic representation of the underlying data. The novel aspect here is the integration of both modalities into a unified representation, enabling the model to leverage complementary information from both sources for improved recognition accuracy.

Hierarchical awareness induction: We propose enriching the vision and language branches through a novel hierarchy-aware attention mechanism. This module introduces hierarchical aggregation priors to guide the model in capturing complex relationships within the data. The novelty lies in integrating hierarchical information into the attention mechanism, allowing the model to capture multi-level dependencies and semantic hierarchies within group activities.

Feature flow-attention mechanism: Inspired by flow network theory, we introduce a feature flow–attention mechanism to prevent the generation of trivial attention s. This module incorporates competitive and allocation principles to enhance the model's ability to capture relevant features within group activities. The innovation here is the introduction of a flow-based attention mechanism, which enables the model to dynamically allocate attention based on feature importance, leading to more robust and interpretable representations of group activities.

Additionally, we present our training loss formulation tailored for the HAtt-Flow architecture, which encourages the joint learning of textual and visual features for improved group activity recognition.

We utilized pre-trained visual and textual backbones to extract the corresponding subject features in the video, $\mathbf{v}$, and textual features, $\mathbf{t}$. In the input preparation section, these are h to represent the nodes and e to represent the edges. However, we denote visual nodes as $\mathbf{v}$, textual nodes as $\mathbf{t}$, and textual edges as $\mathbf{t}_e$ in Figure 4. We used the graph transformer layer and the graph transformer layer with edge features to extract the corresponding feature representations. The former is tailored to graphs lacking explicit edge attributes, while the latter incorporates a dedicated edge feature pipeline to integrate available edge information, maintaining abstract representations at each layer.

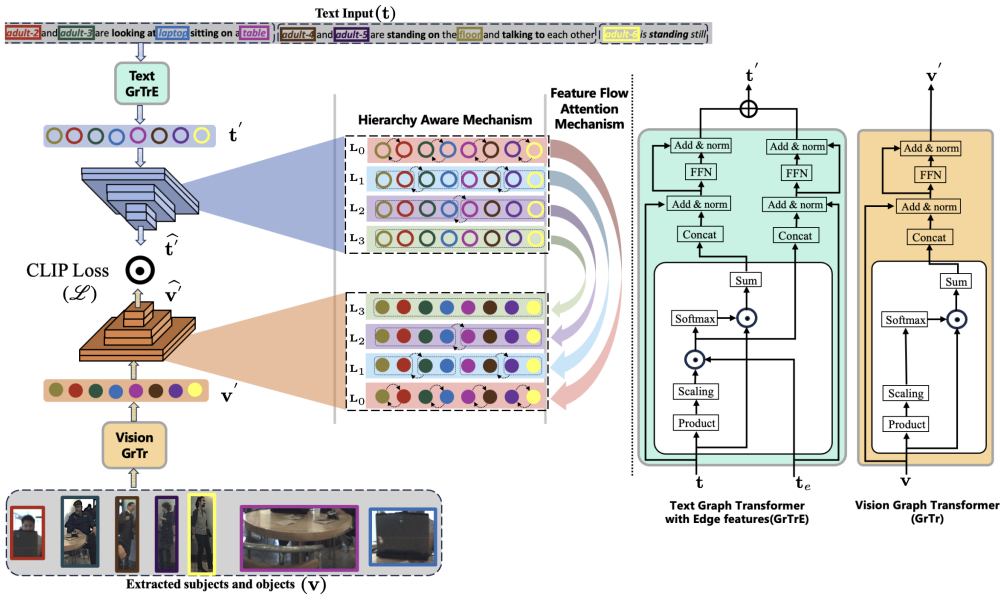


Figure 4. Overall architecture of the proposed HAtt-Flow network. The extracted visual and textual features are passed through their respective graph transformers to obtain corresponding node features. These nodes are passed through the hierarchy-aware-based transformer encoder models to have enriched features, including a feature flow–attention mechanism to enhance cross-modality learning. Finally, we use CLIP loss to optimize the learned features. Please refer to Figure 1 for the details of levels $\mathbf{L}_0$, $\mathbf{L}_1$, $\mathbf{L}_2$, and $\mathbf{L}_3$.

Now, let us proceed to detail each module in the subsequent subsections.

4.1. Input Preparation

Initially, we prepared input node and edge embeddings for the graph transformer layer. In the context of our model, text features are employed to generate both nodes and edges, whereas vision features are exclusively utilized for generating nodes. Consider a graph, $\mathcal{G}$, with node features represented as text features, $\alpha_i \in \mathbb{R}^{d_n \times 1}$, for each node, i, and edge features, also derived from text, denoted as $\beta_{ij} \in \mathbb{R}^{d_e \times 1}$ for edges between nodes i and j. The input node features, α_i, and edge features, β_{ij}, undergo a linear projection to be embedded into d-dimensional hidden features, h_i^0 and e_{ij}^0.

$$\hat{h}i^0 = A^0\alpha_i + a^0; eij^0 = B^0\beta_{ij} + b^0, \tag{1}$$

Here, $A^0 \in \mathbb{R}^{d \times d_n}$, $B^0 \in \mathbb{R}^{d \times d_e}$ and $a^0, b^0 \in \mathbb{R}^d$ are parameters of the linear projection layers. The pre-computed node positional encodings of dimension k are linearly projected and added to the node features $\hat{h}_i^0$.

$$\lambda_i^0 = C^0\lambda_i + c^0; h_i^0 = \hat{h}_i^0 + \lambda_i^0, \tag{2}$$

Here, $C^0 \in \mathbb{R}^{d \times k}$, and $c^0 \in \mathbb{R}^d$. Notably, positional encodings are only added to the node features at the input layer and not during intermediate graph transformer layers. Detailed information about the graph transformer layers is presented in the Appendix C.

4.2. Hierarchical Awareness Induction

We propose enriching the vision and language branches through a hierarchy-aware attention mechanism. In line with the conventional transformer architecture, we divide modality inputs into low-level video patches and text tokens. These are recursively merged based on semantic and spatial similarities, gradually forming more semantically concentrated clusters, such as video objects and text phrases. We define hierarchy aggregation priors with the following aspects:

Tendency to merge. Patches and tokens are recursively merged into higher-level clusters that are *spatially and semantically* similar. If two nearby video patches share similar appearances, merging them is a natural step to convey the same semantic information.

Non-splittable. Once patches or tokens are merged, they will *never be split* in later layers. This constraint ensures that hierarchical information aggregation never degrades, preserving the complete process of hierarchy evolution layer by layer.

We incorporate these hierarchy aggregation priors into an attention mask, C, serving as an extra inductive bias to help the conventional attention mechanism in transformers better explore hierarchical structures adapted to each modality format—a 2D grid for videos and a 1D sequence for texts. Thus, the proposed hierarchy-aware attention is defined as follows:

$$Hierarchy_Attention = \left(C \odot \text{softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)\right)V \tag{3}$$

Note that C is shared among all heads and progressively updated bottom-up across transformer layers. We elaborate on the formulations of the hierarchy-aware mask, C, for each modality as follows.

Hierarchy Induction for Language Branch

In this section, we reconsider the tree-transformer method from the perspective of the proposed hierarchy-aware attention, explaining how to impose hierarchy aggregation priors on C in three steps.

Generate neighboring attention score. The merging tendency of adjacent word tokens is described through neighboring attention scores. Two learnable key and query matrices, $W_Q^{'}$ and $W_K^{'}$, transfer any adjacent word tokens, (t_i, t_{i+1}). The neighboring attention score, $s_{i,i+1}$, is defined as their inner product:

$$s_{i,i+1} = \frac{(t_i W_Q^{'}) \cdot (t_{i+1} W_K^{'})}{\sigma_t} \tag{4}$$

Here, σ_t is a hyperparameter controlling the scale of the generated scores. A softmax function for each token, t_i, is employed to normalize its merging tendency with two neighbors:

$$p_{i,i+1}, p_{i,i-1} = \text{softmax}(s_{i,i+1}, s_{i,i-1}) \tag{5}$$

For neighbor pairs (t_i, t_{i+1}), the neighboring affinity score $\hat{a}_{i,i+1}$ is the geometric mean of $p_{i,i+1}$ and $p_{i+1,i}$: $\hat{a}_{i,i+1} = \sqrt{p_{i,i+1} \cdot p_{i+1,i}}$. From a graph perspective, it describes the strength of edge $e_{i,i+1}$ by comparing it with edges $e_{i-1,i}$ ($p_{i,i+1}$ vs. $p_{i,i-1}$) and $e_{i+1,i+2}$ ($p_{i+1,i}$ vs. $p_{i+1,i+2}$).

Enforcing non-splittable property. A higher neighboring affinity score indicates that two neighbor tokens are more closely bonded. To ensure that merged tokens will not be split, layer-wise affinity scores, $a_{i,i+1}^l$, should increase as the network goes deeper, i.e., $a_{i,i+1}^l \geq a_{i,i+1}^{l-1}$ for all l. It helps to gradually generate a desired hierarchy structure:

$$a_{i,i+1}^l = a_{i,i+1}^{l-1} + \left(1 - a_{i,i+1}^{l-1}\right)\hat{a}_{i,i+1}^l \tag{6}$$

Similarly, we formulate the *hierarchy induction for visual branch*, detailed in the Appendix D.

4.3. Feature Flow–Attention Mechanism

In the following representation, we use the corresponding nodes, h, from the respective branches of language and visual graph transformers as the queries (**Q**), keys (**K**), and values (**V**).

Inspired by flow network theory, the flow–attention mechanism introduces a competitive mechanism for sources and an allocation mechanism for sinks, preventing the generation of trivial attention. In a flow network framework, attention is viewed as the flow of information from sources to sinks. The results (**R**) act as endpoints receiving the

inbound information flow, and the values ($\mathbf{V}$) serve as sources providing the outgoing information flow.

Flow capacity calculation. For a scenario with n sinks and m sources, incoming flow, I_i, for the i-th sink and outgoing flow, O_j, for the j-th source are calculated as follows:

$$\begin{aligned} I_i &= \phi(\mathbf{Q}_i) \sum_{j=1}^{m} \phi(\mathbf{K}_j)^{\mathsf{T}}, \\ O_j &= \phi(\mathbf{K}_j) \sum i = 1^n \phi(\mathbf{Q}_i)^{\mathsf{T}}, \end{aligned} \tag{7}$$

where $\phi(\cdot)$ is a non-negative function.

Flow conservation. We establish the preservation of incoming flow capacity for each sink, maintaining the default value at 1, effectively "locking in" the information forwarded to the next layer. This conservation strategy ensures that the outgoing flow capacities of sources engage in competition, with their collective sum strictly constrained to 1. Likewise, by conserving the outgoing flow capacity for each source at the default value of 1, essentially "fixing" the information acquired from the previous layer, the conservation of incoming and outgoing flow capacities is enforced via normalizing operations:

$$\frac{\phi(\mathbf{K})}{\mathbf{O}}, \frac{\phi(\mathbf{Q})}{\mathbf{I}}, \tag{8}$$

In this context, the ratio denotes element-wise division, with $\frac{\phi(\mathbf{K})}{\mathbf{O}}$ dedicated to source conservation and $\frac{\phi(\mathbf{Q})}{\mathbf{I}}$ assigned to sink conservation.

This normalization process ensures the preservation of flow capacity for each source and sink token, as evidenced by the following equations:

$$\begin{aligned} \text{source-}j\text{:}\ \frac{\phi(\mathbf{K}_j)^{\mathsf{T}}}{O_j} \sum_{i=1}^{n} \phi(\mathbf{Q}_i) &= \frac{\sum_{i=1}^{n} \phi(\mathbf{Q}_i)\phi(\mathbf{K}_j)^{\mathsf{T}}}{O_j} = 1 \\ \text{sink-}i\text{:}\ \frac{\phi(\mathbf{Q}_i)^{\mathsf{T}}}{I_i} \sum_{j=1}^{m} \phi(\mathbf{K}_j) &= \frac{\sum_{j=1}^{m} \phi(\mathbf{K}_j)\phi(\mathbf{Q}_i)^{\mathsf{T}}}{I_i} = 1 \end{aligned} \tag{9}$$

These equations replicate the same computations as Equation (7). The initial equation concerns the outgoing flow capacity of the j-th source after the normalization process $\frac{\phi(\mathbf{K})}{\mathbf{O}}$. In contrast, the second equation corresponds to the incoming flow capacity of the i-th sink after the normalization process $\frac{\phi(\mathbf{Q})}{\mathbf{I}}$. In both instances, the capacities are identical to the default value of 1.

The conserved incoming flow, $\widehat{\mathbf{I}}$, and outgoing flow, $\widehat{\mathbf{O}}$, are represented as follows:

$$\begin{aligned} \widehat{\mathbf{I}} &= \phi(\mathbf{Q}) \sum_{j=1}^{m} \frac{\phi(\mathbf{K}j)^{\mathsf{T}}}{Oj}, \\ \widehat{\mathbf{O}} &= \phi(\mathbf{K}) \sum_{i=1}^{n} \frac{\phi(\mathbf{Q}i)^{\mathsf{T}}}{Ii} \end{aligned} \tag{10}$$

Flow–attention mechanism. We introduce the flow–attention mechanism, leveraging competition induced via incoming flow conservation for sinks. In $\widehat{\mathbf{O}}$, sources compete while maintaining a fixed flow-capacity sum, revealing source significance. $\widehat{\mathbf{I}}$ represents the

sink information when the source outgoing capacity is 1, reflecting aggregated information allocation to each sink. The flow–attention equations are as follows:

$$\begin{aligned} \text{Competition: } \widehat{\mathbf{V}} &= \text{Softmax}(\widehat{\mathbf{O}}) \odot \mathbf{V} \\ \text{Aggregation: } \mathbf{A} &= \frac{\phi(\mathbf{Q})}{\mathbf{I}} \left(\phi(\mathbf{K})^{\mathsf{T}} \widehat{\mathbf{V}}\right) \\ \text{Allocation: } \mathbf{R} &= \text{Sigmoid}(\widehat{\mathbf{I}}) \odot \mathbf{A}. \end{aligned} \tag{11}$$

In the "Competition" stage, $\widehat{\mathbf{V}}$ is determined through the application of the Softmax function to $\widehat{\mathbf{O}}$, followed by element-wise multiplication with $\mathbf{V}$. The "Aggregation" step, denoted as $\mathbf{A}$, is computed using the presented equation. Lastly, the "Allocation" phase calculates $\mathbf{R}$ by employing the Sigmoid function on $\widehat{\mathbf{I}}$, which is then element-wise multiplied with $\mathbf{A}$.

4.4. Training Loss

To adapt the contrastive pretraining objective for video and text features in the HAtt-Flow architecture, the objective function can be expressed as follows:

$$\mathscr{L} = -\frac{1}{M}\sum_{i}^{M} \log \frac{\exp\left(\mathbf{v}_i^{\top} \mathbf{u}_i / \tau\right)}{\sum_{j=1}^{M} \exp\left(\mathbf{v}_i^{\top} \mathbf{u}_j / \tau\right)} - \frac{1}{M}\sum_{i}^{M} \log \frac{\exp\left(\mathbf{u}_i^{\top} \mathbf{v}_i / \tau\right)}{\sum_{j=1}^{M} \exp\left(\mathbf{u}_i^{\top} \mathbf{v}_j / \tau\right)} \tag{12}$$

Here, $\mathbf{v}$ and $\mathbf{u}$ represent the video and text feature vectors, τ is the learnable temperature parameter, and M is the total number of video–text pairs, i.e., the total number of labels.

5. Experimental Results

5.1. Experiment Settings

Dataset details. Our dataset adopts a division strategy from JRDB [81], where videos are segregated at the sequence level, ensuring the entirety of a video sequence is allocated to a specific split. The 54 video sequences are distributed, with 20 for training, seven for validation, and 27 for testing. To align with the evaluation practices of analogous datasets, our evaluation is centered on keyframes sampled at one-second intervals, resulting in 1419 training samples, 404 validation samples, and 1802 test samples.

Implementation details. Our framework, implemented in PyTorch, undergoes training on a machine featuring four NVIDIA Quadro RTX 6000 GPUs. During training, we adopt a batch size of 2 and leverage the Adam Optimizer, commencing the training process with an initial learning rate set at 0.0001.

Evaluation metrics. We evaluate the model using two tasks: (1.) predicate classification (PredCls) and (2.) video scene graph generation (VSGG). The video scene graph generation (VSGG) task aims to generate descriptive triplets for an input video. Each triplet, denoted as $(r_i, t_1, t_2, o_s, {m_s}^{(t_1,t_2)}, o_o, {m_o}^{(t_1,t_2)})$, consists of a relation, r_i, occurring between time points t_1 and t_2, connecting a subject, o_s, (class category) with mask tube ${m_s}^{(t_1,t_2)}$ and an object, o_o, with mask tube ${m_o}^{(t_1,t_2)}$. Evaluation metrics for PredCls and VSGG adhere to scene graph generation (SGG) standards, utilizing Recall@K (R@K) and mean Recall@K (mR@K). Successful recall for a ground-truth triplet $(\hat{o}_s, \hat{m}_s^{(\hat{t}_1,\hat{t}_2)}, \hat{o}_o, \hat{m}_o^{(\hat{t}_1,\hat{t}_2)}, \hat{r}_i^{(\hat{t}_1,\hat{t}_2)})$ requires accurate category labels and IOU volumes between predicted and ground-truth mask tubes above 0.5. The soft recall is recorded when these criteria are met, considering the time IOU between predicted and ground-truth intervals.

5.2. Comparison with the State of the Art

We present our comparisons with state-of-the-art (SOTA) methods in Tables 2 and 3 for our dataset and PSG dataset. In direct comparison with the methods above, the HAtt-Flow model exhibits a notable performance advantage, establishing itself as the current state of the art. This superiority is attributed to its proficiency in capturing intricate

social activities among subjects across spatial and temporal dimensions. On the GASG dataset, our proposed method outperforms existing SGG methods by a significant margin on all metrics except for the R/mR@20 of the VSGG task. On the PSG dataset, it is evident that the proposed method dominated the other methods to demonstrate state-of-the-art performance.

Table 2. Comparison with SOTA methods on **GASG dataset**. Bold numbers indicate the best results.

Method	Modality (X)	R/mR@20		R/mR@50		R/mR@100	
		PredCls	VSGG	PredCls	VSGG	PredCls	VSGG
IMP [54]	Image	31.9/9.55	16.5/6.52	36.8/10.9	18.2/7.05	38.9/11.6	18.6/7.23
	Video	-	-	-	-	-	-
MOTIFS [55]	Image	44.9/20.2	20.0/9.10	50.4/20.1	21.7/9.57	52.4/22.9	22.0/9.67
	Video	-	-	-	-	-	-
VCTree [56]	Image	45.3/20.7	20.6/9.70	50.8/22.6	22.1/10.2	52.7/23.3	22.5/10.2
	Video	-	-	-	-	-	-
GPSNet [58]	Image	31.5/13.2	17.8/2.03	39.9/16.4	19.6/7.49	44.7/18.3	20.1/7.67
	Video	-	-	-	-	-	-
PSG [80]	Image	-	31.4/16.2	-	32.9/21.5	-	36.1/22.7
	Video	-	-	-	-	-	-
PVSG [76]	Image	-	38.3/18.1	-	41.7/20.8	-	43.2/23.7
	Video	-	**13.6/10.2**	-	19.2/11.1	-	26.5/14.7
Ours	Image	**57.4/35.2**	**42.2/21.4**	**60.2/36.1**	**44.5/23.1**	**63.7/39.5**	**48.1/26.9**
	Video	**27.1/14.3**	11.2/9.1	**29.5/17.71**	**19.6/12.3**	**41.7/24.2**	**30.2/18.1**

Table 3. Comparison with SOTA methods on **PSG dataset**. Bold numbers indicate the best results.

Methods	R/mR@20		R/mR@50		R/mR@100	
	PredCls	VSGG	PredCls	VSGG	PredCls	VSGG
IMP [54]	30.5/8.97	17.9/7.35	35.9/10.5	19.5/7.88	38.3/11.3	20.1/8.02
MOTIFS [55]	45.1/19.9	20.9/9.60	50.5/21.5	22.5/10.1	52.5/22.2	23.1/10.3
VCTree [56]	45.9/21.4	21.7/9.68	51.2/23.1	23.3/10.2	53.1/23.8	23.7/10.3
GPSNet [58]	38.8/17.1	18.4/6.52	46.6/20.2	20.0/6.97	50.0/21.3	20.6/7.2
PSG [80]	-	28.2/15.4	-	32.1/20.3	-	35.3/21.5
Ours	**52.4/25.6**	**32.9/18.4**	**56.1/28.3**	**35.3/21.6**	**62.7/32.12**	**41.34/23.1**

5.3. Ablation Study

Flow-attention direction. We introduced a novel flow–attention mechanism between the hierarchical transformers handling text and vision. To explore the impact of the flow direction between these networks, we conducted experiments as detailed in Table 4. Our findings validate that optimal results are achieved when attention flows from the text to the vision transformer. Conversely, performance declines in the opposite direction, notably when no attention flows. This observation suggests that the cross-attention mechanism enhances the model's contextual learning capacity primarily when the flow is from text to vision because text often provides high-level semantic information and context that can guide the understanding of visual content. However, it is worth noting that further investigation is warranted to fully understand the reasons behind the decline in performance when attention flows in the opposite direction. Potential improvements could involve refining the interaction between the text and vision transformers to better leverage complementary information from both modalities.

Importance of hierarchy awareness. We incorporated hierarchy awareness into the transformer framework to bolster the model's scene graph generation capabilities. The experimental results, detailed in Table 5, affirm that hierarchical awareness optimally

enhances scene graph generation. Conversely, performance declines in the absence of this design. This is likely because, when the model is aware of the hierarchy during the generation of video scene graphs, it accurately predicts all relevant nodes and their relationships (edges). However, future work could explore alternative methods for incorporating hierarchy awareness to further improve performance, such as fine-tuning the hierarchical structure or exploring different aggregation techniques.

Attribute analysis. In Table 6, we evaluate the impact of different attributes in the dataset on the model's performance in scene graph generation. Our findings affirm that including all attributes results in optimal scene graph generation. Conversely, the performance exhibits a decline when we consider individual attributes one by one. We can clearly observe that performance is proportional to the attribute distribution, as shown in Table 3, i.e., the more the number of attributes, the better performance. This underscores the significance of leveraging all attributes, indicating that they collectively enhance the model's capacity to grasp intricate contexts, enabling accurate scene graph generation. However, further investigation into the interplay between different attributes and their impact on performance could provide valuable insights for refining the model architecture and training process.

Table 4. Ablation study for **flow–attention direction**. Bold numbers indicate the best results.

Flow–Attention Direction	Modality (X)	R/mR@20		R/mR@50		R/mR@100	
		PredCls	VSGG	PredCls	VSGG	PredCls	VSGG
T $\nrightarrow$ X	Image	32.1/17.4	18.3/9.2	37.15/19.2	21.4/10.5	41.6/21.8	23.5/14.7
	Video	10.2/5.3	4.7/2.4	12.7/7.5	7.1/4.2	16.2/10.8	9.8/7.3
X $\rightarrow$ T	Image	45.3/23.84	20.4/10.8	48.1/25.6	23.5/12.4	52.8/27.3	29.2/16.1
	Video	15.4/7.1	7.2/4.7	18.3/9.4	9.7/6.2	24.7/12.1	14.1/9.4
T $\rightarrow$ X	Image	**57.4/35.2**	**42.2/21.4**	**60.2/36.1**	**44.5/23.1**	**63.7/39.5**	**48.1/26.9**
	Video	**27.1/14.3**	**11.2/9.1**	**29.5/17.71**	**19.6/12.3**	**41.7/24.2**	**30.2/18.1**

Table 5. Ablation study for **hierarchy awareness**. Bold numbers indicate the best results.

Hierarchy Awareness	Modality (X)	R/mR@20		R/mR@50		R/mR@100	
		PredCls	VSGG	PredCls	VSGG	PredCls	VSGG
✗	Image	46.1/22.3	23.7/12.5	47.2/24.1	24.5/13.4	52.6/28.3	29.8/16.1
	Video	17.5/11.3	10.4/8.7	22.1/14.1	14.1/9.61	27.4/15.2	25.7/16.3
✓	Image	**57.4/35.2**	**42.2/21.4**	**60.2/36.1**	**44.5/23.1**	**63.7/39.5**	**48.1/26.9**
	Video	**27.1/14.3**	**11.2/9.1**	**29.5/17.71**	**19.6/12.3**	**41.7/24.2**	**30.2/18.1**

Table 6. Ablation study for **attributes** in our dataset. Bold numbers indicate the best results.

Attributes	R/mR@20		R/mR@50		R/mR@100	
	PredCls	VSGG	PredCls	VSGG	PredCls	VSGG
Appearance	14.5/2.3	1.4/0.3	17.2/6.1	2.3/0.5	21.1/9.1	7.8/2.1
Relationship	13.2/1.8	1.3/0.5	16.4/5.2	2.1/0.4	20.8/8.7	7.5/2.4
Interaction	11.4/1.4	0.9/0.4	14.7/4.1	1.7/0.7	17.7/6.4	6.2/1.6
Position	8.1/0.7	1.5/0.51	10.6/3.7	1.4/0.4	14.2/4.7	4.1/0.9
Situation	5.7/0.2	0.7/0.2	7.2/1.7	0.9/0.3	10.1/2.8	2.8/0.4
All together	**27.1/14.3**	**11.2/9.1**	**29.5/17.71**	**19.6/12.3**	**41.7/24.2**	**30.2/18.1**

6. Qualitative Analysis

To gain deeper insights into the performance of HAtt-Flow, we employed visualization to illustrate its scene graph generation predictions using our dataset. As depicted in

Figure 5, our model substantially improves overall scene graph generation compared to PSGFormer. The effectiveness of our hierarchy-aware attention–flow mechanism contributes significantly to this enhancement, providing our model with superior context modeling capabilities for visual representations guided by textual inputs.

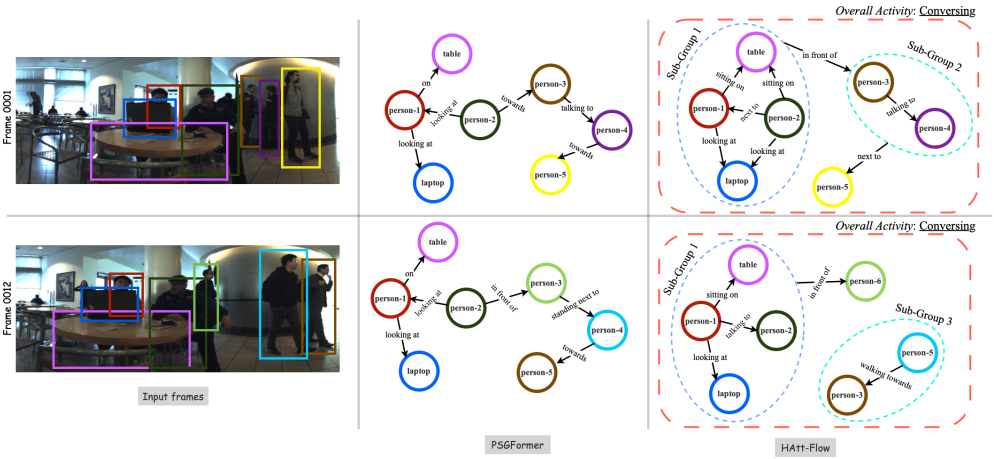


Figure 5. The visualization of the scene graphs generated via PSGFormer [80] and our approach. We can observe that [80] could only detect the subjects, but not accurate groups and their interactions. In contrast, the HAtt-Flow is accurate in graph generation and overall group activity prediction. **Best viewed in color and zoomed in.**

7. Conclusions

In this work, we introduced a pioneering dataset designed with nuanced attributes, specifically tailored to enhance the scene graph generation task within the context of group activities. Our contributions extend to advancing predictive video scene understanding, propelled by the introduction of flow–attention and a paradigm shift in attention mechanisms through hierarchy awareness. Via rigorous experimentation, we demonstrated the efficacy of our approach, showcasing significant improvements over the previously existing methods.

Limitations

While the novel flow–attention mechanism introduced in this work draws inspiration from flow network theory and offers promising advancements, several limitations warrant consideration. Primarily, the implementation of flow–attention imposes heightened computational demands, which may pose challenges for deployment in resource-constrained settings. This computational complexity underscores the need for efficient algorithms and hardware acceleration techniques to enable practical use in real-world applications. Furthermore, the performance of the flow–attention model is greatly influenced by the quality and quantity of available training data. Variations in data quality or insufficient sample sizes can impact the model's robustness and generalization capabilities, highlighting the importance of extensive and diverse datasets for achieving optimal results. Addressing these computational and data-related limitations is crucial to maximizing the practicality and effectiveness of the proposed flow–attention mechanism. Future research endeavors should focus on developing strategies to mitigate computational demands while maintaining performance levels and exploring methods for enhancing the robustness of the model to variations in training data. By overcoming these challenges, the flow–attention mechanism can realize its full potential as a valuable tool in a wide range of real-world applications.

Author Contributions: Conceptualization, N.V.S.R.C., P.N. and K.L.; methodology, N.V.S.R.C. and P.N.; programming, N.V.S.R.C.; validation, N.V.S.R.C., P.N. and K.L.; formal analysis, N.V.S.R.C.; investigation, K.L.; resources, K.L.; data curation, N.V.S.R.C.; writing—original draft preparation, N.V.S.R.C.; writing—review and editing, P.N., T.H.N.L., P.D.D. and K.L.; visualization, N.V.S.R.C.; supervision, T.H.N.L., P.D.D. and K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The proposed dataset is available at https://uark-cviu.github.io/GASG/ under the CC-BY 4.0 license to help the research community. Please reach out to the corresponding author for further information.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Sensor Application's Correlation with Our Article

Our research leveraged a diverse array of sensor data to enable a comprehensive analysis and understanding. Within the JRDB dataset [81], the authors integrated data streams from various sensors, including RGB cameras and lidars, to capture rich environmental information and facilitate sensor-based applications.

Our annotations primarily focus on pedestrian detection and tracking, utilizing data from multiple sensors to enhance accuracy and robustness. Specifically, we include data streams from RGB cameras positioned in the upper row of the cylindrical stereo camera suite, which capture images for annotation purposes. Additionally, the previous authors utilized point clouds obtained from both the upper and lower 16-channel lidars (Velodynes) for 3D bounding box annotations around pedestrians.

This multi-sensor approach enables us to integrate spatial information from different sources, facilitating the precise localization and tracking of pedestrians in complex environments. By harnessing the complementary strengths of RGB cameras and Lidars, our dataset provides a rich source of annotated sensor data for applications such as pedestrian detection, tracking, and environmental perception.

Appendix B. GASG Dataset Annotation Pipeline

GPT4RoI. Our initial step involves using GPT4RoI to generate textual descriptions corresponding to input bounding boxes. GPT4RoI integrates visual and linguistic data and adeptly handles spatial instructions. During processing, GPT4RoI replaces <$region_i$> tags in these instructions with results from RoIAlign, derived directly from the image's features. This process creates a unique fusion of region-specific data with language embeddings. For an enhanced multimodal understanding, this combination of embeddings is then interpreted via the Vicuna [82] model, a specialized instance of LLaMA [83]. This allows us to input bounding boxes around objects and prompt the system for detailed descriptions, covering aspects like appearance, situation, positioning, interactions, and relationships. For instance, when we input bounding boxes around objects and ask the system questions, such as determining the relationship between individuals in <$region_1$> and <$region_2$>, the system responds with detailed, context-rich descriptions.

Post-processing with Spacy. After generating text with GPT-4RoI, we utilize Spacy v3.0 (https://spacy.io/, accessed on 20 October 2023), a Python library for natural language processing, to refine the text further. We specifically use Spacy to add grammatical tags to each word in the text. This tagging involves identifying the grammatical role of each word and determining whether it is a noun, verb, or adjective, among others. This process is essential for understanding the sentence structure and ensuring that the text is accurate in its content and grammatically coherent.

Human curation and filtering. For the final step, we rely on human expertise to ensure the highest quality of our output. Our team carefully reviews the Spacy-processed text using a specially designed filter that helps categorize interactivity types. This human oversight is essential for maintaining the highest standards of accuracy and relevance. It enables us to meticulously confirm and refine the interaction types identified by the LLM, ensuring that our final label is precise.

Appendix C. Data Format

```
data[{
    'file_name': str,
    'height': int,
    'width': int,
    'image_id': int,
    'frame_index': int,
    'video_id': int,
    'segments_info':[{
        'id': int,
        'track_id': int,
        'category_id': int,
        'iscrowd': 0 or 1,
        'isthing': 0 or 1,
        'area': int,
        'sub-group_id': int,
        'group_activity':int
    }]
    'appearances': [{
        'segment_id': int,
        'app_id': int
    }]
    'situations': [{
        'segment_id': int,
        'sit_id': int
    }]
    'positions': [{
        'segment_id': int,
        'segment_id': int,
        'pos_id': int
    }]
    'interactions': [{
        'segment_id': int,
        'segment_id': int,
        'inter_id': int
    }]
    'relations': [[
        'segment_id': int,
        'segment_id': int,
        'rel_id': int
    ]]
}]
```

Appendix C.1. Basic Image Information

This section details the fundamental attributes of each image:

- `file_name`: The name of the image file.
- `height`: The height of the image in pixels.
- `width`: The width of the image in pixels.
- `image_id`: A unique identifier for the image.
- `frame_index`: The index of the frame within the video sequence.
- `video_id`: An identifier for the video or image collection to which this image belongs.

Appendix C.2. Segment Information

This section includes the `segments_info` key, which is a list of segments within the image. Each segment contains the following:

- `id`: A unique identifier for the segment.
- `track_id`: An identifier to track the segment across different frames.
- `category_id`: An identifier for the category of the object in the segment.
- `iscrowd`: A binary value indicating whether the segment represents a crowd.
- `isthing`: A binary value indicating whether the segment represents a "thing" (as opposed to "stuff" like a banner, blanket, curtain, pillow, or towel).
- `area`: The area covered by the segment in the image.

Appendix C.3. Interactivity Attributes

This section encompasses lists of `predicate_appearances`, `predicate_situations`, `predicate_interactions`, and `predicate_relations` for each segment. For single-actor attributes (i.e., appearances and situations), the structure is as follows:

- `segment_id`: An identifier for the segment.
- `id`: An identifier for the interactivity type.

For double-actor attributes (i.e., positions, interactions, and relations), the structure includes two different `segment_id`s to represent the interactivity between two segments:

- `segment_id_1`: An identifier for the first segment.
- `segment_id_2`: An identifier for the second segment.
- `id`: An identifier for the interactivity type.

These descriptors represent lists of integers, specifying various aspects of the subject, object, individual, and group activities for each bounding box within the annotations and `segments_info`. Please find the complete train and test annotations in the dataset link provided (under Data Availability Statement).

Appendix D. More Implementation Details

We present the pseudo-code for hierarchy induction in Algorithm A1 and our flow–attention mechanism in Algorithm A2.

Algorithm A1 Hierarchy induction.

Input: All neighboring affinity scores, $a^{l}_{(i,j),(i',j')}$, for $l = \{1, \ldots, N\}$ layers, a list of "break" threshold values $\{\theta_1, \ldots, \theta_N\}$ for every layer.

$l \leftarrow N$ ▷ Start from the highest layer
Initialize a nested list, $\mathcal{B} = \{B_1, \ldots, B_N\}$ ▷ Store break edges of each layer
while $l > 0$ **do**
  for each edge $(i,j),(i',j')$ in the patch graph **do**
    if $a^{l}_{(i,j),(i',j')} < \theta_l$ **then**
      if $l = N$ **then**
        Append the edge $(i,j),(i',j')$ to B_l ▷ Break the edge $(i,j),(i',j')$ in the top layer N
      else
        if edge $(i,j),(i',j')$ not in B_{l+1} **then**
          Append the edge $(i,j),(i',j')$ to B_l ▷ Break the edge $(i,j),(i',j')$ in layer l
        end if
      end if
    end if
  end for
  $l \leftarrow l - 1$ ▷ Move to the next lower layer
end while
Draw visual hierarchy based on $\mathcal{B}$; then remove redundant edges by finding connected components.

Algorithm A2 Multi-head flow–attention mechanism (normal version).

Input: $\mathbf{Q} \in \mathbb{R}^{n\times d}, \mathbf{K} \in \mathbb{R}^{m\times d}, \mathbf{V} \in \mathbb{R}^{m\times d}$

$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathtt{Split}(\mathbf{Q}), \mathtt{Split}(\mathbf{K}), \mathtt{Split}(\mathbf{V})$ // $\mathbf{Q} \in \mathbb{R}^{n\times h\times \frac{d}{h}}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{m\times h\times \frac{d}{h}}$

$\mathbf{Q}, \mathbf{K} = \mathtt{Sigmoid}(\mathbf{Q}), \mathtt{Sigmoid}(\mathbf{K})$

$$\mathbf{I} = \mathtt{Sum}\left(\mathbf{Q} \odot \mathtt{Broadcast}\left(\mathtt{Sum}(\mathbf{K}, \mathtt{dim=0}), \mathtt{dim=0}\right), \mathtt{dim=2}\right) \quad // \ \mathbf{I} \in \mathbb{R}^{n\times h}$$

$$\mathbf{O} = \mathtt{Sum}\left(\mathbf{K} \odot \mathtt{Broadcast}\left(\mathtt{Sum}(\mathbf{Q}, \mathtt{dim=0}), \mathtt{dim=0}\right), \mathtt{dim=2}\right) \quad // \ \mathbf{O} \in \mathbb{R}^{m\times h}$$

$$\widehat{\mathbf{I}} = \mathtt{Sum}\left(\mathbf{Q} \odot \mathtt{Broadcast}\left(\mathtt{Sum}(\mathbf{K}/\mathbf{O}, \mathtt{dim=0}), \mathtt{dim=0}\right), \mathtt{dim=2}\right) \quad // \ \widehat{\mathbf{I}} \in \mathbb{R}^{n\times h}$$

$$\widehat{\mathbf{O}} = \mathtt{Sum}\left(\mathbf{K} \odot \mathtt{Broadcast}\left(\mathtt{Sum}(\mathbf{Q}/\mathbf{I}, \mathtt{dim=0}), \mathtt{dim=0}\right), \mathtt{dim=2}\right) \quad // \ \widehat{\mathbf{O}} \in \mathbb{R}^{m\times h}$$

$$\mathbf{R} = \mathtt{Matmul}\left(\mathbf{Q}/\mathbf{I}, \mathtt{Matmul}\left(\mathbf{K}, \mathbf{V} \odot \mathtt{Softmax}(\widehat{\mathbf{O}})\right)\right) \odot \mathtt{Sigmoid}(\widehat{\mathbf{I}}) \quad // \ \mathbf{R} \in \mathbb{R}^{n\times h\times \frac{d}{h}}$$

Return R

References

1. Gupta, S.; Malik, J. Visual semantic role labeling. *arXiv* **2015**, arXiv:1505.04474.
2. Gkioxari, G.; Girshick, R.; Dollár, P.; He, K. Detecting and recognizing human-object interactions. In Proceedings of the IEEE Conference on Computer Vision and Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
3. Kato, K.; Li, Y.; Gupta, A. Compositional learning for human object interaction. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
4. Chao, Y.W.; Liu, Y.; Liu, X.; Zeng, H.; Deng, J. Learning to detect human-object interactions. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018.
5. Wang, T.; Anwer, R.M.; Khan, M.H.; Khan, F.S.; Pang, Y.; Shao, L.; Laaksonen, J. Deep contextual attention for human-object interaction detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
6. Li, Y.L.; Zhou, S.; Huang, X.; Xu, L.; Ma, Z.; Fang, H.S.; Wang, Y.; Lu, C. Transferable interactiveness knowledge for human-object interaction detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 16–20 June 2019.
7. Zhou, T.; Wang, W.; Qi, S.; Ling, H.; Shen, J. Cascaded human-object interaction recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Long Beach, CA, USA, 16–20 June 2020.
8. Wang, T.; Yang, T.; Danelljan, M.; Khan, F.S.; Zhang, X.; Sun, J. Learning human-object interaction detection using interaction points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Seattle, WA, USA, 14–16 June 2020.
9. Hou, Z.; Peng, X.; Qiao, Y.; Tao, D. Visual compositional learning for human-object interaction detection. In Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020.
10. Li, Y.L.; Liu, X.; Lu, H.; Wang, S.; Liu, J.; Li, J.; Lu, C. Detailed 2d–3d joint representation for human-object interaction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Seattle, WA, USA, 14–19 June 2020.
11. Gao, C.; Xu, J.; Zou, Y.; Huang, J.B. Drg: Dual relation graph for human-object interaction detection. In Proceedings of the 16th European Conference ECCV, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020.
12. Kim, B.; Choi, T.; Kang, J.; Kim, H.J. Uniondet: Union-level detector towards real-time human-object interaction detection. In Proceedings of the 16th European Conference ECCV, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020.
13. Liu, Y.; Chen, Q.; Zisserman, A. Amplifying key cues for human-object-interaction detection. In Proceedings of the 16th European Conference ECCV, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020.
14. Tamura, M.; Ohashi, H.; Yoshinaga, T. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Virtual, 19–25 June 2021.
15. Hou, Z.; Yu, B.; Qiao, Y.; Peng, X.; Tao, D. Affordance transfer learning for human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Virtual, 19–25 June 2021.
16. Zhang, A.; Liao, Y.; Liu, S.; Lu, M.; Wang, Y.; Gao, C.; Li, X. Mining the benefits of two-stage and one-stage hoi detection. *NeurIPS* **2021**, *34*, 17209–17220.
17. Wang, S.; Duan, Y.; Ding, H.; Tan, Y.P.; Yap, K.H.; Yuan, J. Learning Transferable Human-Object Interaction Detector with Natural Language Supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, New Orleans, LA, USA, 19–24 June 2022.

18. Zhang, F.Z.; Campbell, D.; Gould, S. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, New Orleans, LA, USA, 19–24 June 2022.
19. Kim, B.; Lee, J.; Kang, J.; Kim, E.S.; Kim, H.J. Hotr: End-to-end human-object interaction detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Virtual, 19–25 June 2021.
20. Zou, C.; Wang, B.; Hu, Y.; Liu, J.; Wu, Q.; Zhao, Y.; Li, B.; Zhang, C.; Zhang, C.; Wei, Y.; et al. End-to-end human object interaction detection with hoi transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Virtual, 19–25 June 2021.
21. Amer, M.R.; Xie, D.; Zhao, M.; Todorovic, S.; Zhu, S.C. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In Proceedings of the ECCV: 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 187–200.
22. Amer, M.R.; Todorovic, S.; Fern, A.; Zhu, S.C. Monte carlo tree search for scheduling activity recognition. In Proceedings of the ICCV International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1353–1360.
23. Amer, M.R.; Lei, P.; Todorovic, S. Hirf: Hierarchical random field for collective activity recognition in videos. In Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 572–585.
24. Amer, M.R.; Todorovic, S. Sum product networks for activity recognition. *IEEE Trans. Anal. Mach. Intell.* **2015**, *38*, 800–813. [CrossRef] [PubMed]
25. Lan, T.; Wang, Y.; Yang, W.; Robinovitch, S.N.; Mori, G. Discriminative latent models for recognizing contextual group activities. *IEEE Trans. Anal. Mach. Intell.* **2011**, *34*, 1549–1562. [CrossRef] [PubMed]
26. Lan, T.; Sigal, L.; Mori, G. Social roles in hierarchical models for human activity recognition. In Proceedings of the 2012 IEEE Conference on Computer Vision and Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: New York, NY, USA, 2012; pp. 1354–1361.
27. Shu, T.; Xie, D.; Rothrock, B.; Todorovic, S.; Chun Zhu, S. Joint inference of groups, events and human roles in aerial videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4576–4584.
28. Wang, Z.; Shi, Q.; Shen, C.; Van Den Hengel, A. Bilinear programming for human activity recognition with unknown mrf graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Portland, OR, USA, 23–28 June 2013; pp. 1690–1697.
29. Zhang, H.; Kyaw, Z.; Chang, S.F.; Chua, T.S. Visual translation embedding network for visual relation detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Honolulu, HI, USA, 21–26 June 2017.
30. Hung, Z.S.; Mallya, A.; Lazebnik, S. Contextual translation embedding for visual relationship detection and scene graph generation. *IEEE Trans. Anal. Mach. Intell.* **2020**, *43*, 3820–3832. [CrossRef] [PubMed]
31. Tang, K.; Niu, Y.; Huang, J.; Shi, J.; Zhang, H. Unbiased scene graph generation from biased training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Virtual, 13–19 June 2020.
32. Desai, A.; Wu, T.Y.; Tripathi, S.; Vasconcelos, N. Learning of Visual Relations: The Devil is in the Tails. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021.
33. Liang, Y.; Bai, Y.; Zhang, W.; Qian, X.; Zhu, L.; Mei, T. Vrr-vg: Refocusing visually-relevant relationships. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
34. Khandelwal, S.; Suhail, M.; Sigal, L. Segmentation-grounded Scene Graph Generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021.
35. Bagautdinov, T.; Alahi, A.; Fleuret, F.; Fua, P.; Savarese, S. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 4315–4324.
36. Deng, Z.; Vahdat, A.; Hu, H.; Mori, G. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Las Vegas, NV, USA, 26–27 July 2016; pp. 4772–4781.
37. Ibrahim, M.S.; Muralidharan, S.; Deng, Z.; Vahdat, A.; Mori, G. A hierarchical deep temporal model for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Las Vegas, NV, USA, 26 June–1 July 2016: pp. 1971–1980.
38. Ibrahim, M.S.; Mori, G. Hierarchical relational networks for group activity recognition and retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 721–736.
39. Li, X.; Choo Chuah, M. Sbgar: Semantics based group activity recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2876–2885.
40. Qi, M.; Qin, J.; Li, A.; Wang, Y.; Luo, J.; Van Gool, L. Stagnet: An attentive semantic rnn for group activity recognition. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 101–117.
41. Shu, X.; Tang, J.; Qi, G.; Liu, W.; Yang, J. Hierarchical long short-term concurrent memory for human interaction recognition. *IEEE Trans. Anal. Mach. Intell.* **2019**, *43*, 1110–1118. [CrossRef] [PubMed]
42. Wang, M.; Ni, B.; Yang, X. Recurrent modeling of interaction context for collective activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 3048–3056.

43. Yan, R.; Tang, J.; Shu, X.; Li, Z.; Tian, Q. Participation-contributed temporal dynamic model for group activity recognition. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1292–1300.
44. Chappa, N.V.; Nguyen, P.; Nelson, A.H.; Seo, H.S.; Li, X.; Dobbs, P.D.; Luu, K. SoGAR: Self-supervised Spatiotemporal Attention-based Social Group Activity Recognition. *arXiv* **2023**, arXiv:2305.06310.
45. Chappa, N.V.; Nguyen, P.; Nelson, A.H.; Seo, H.S.; Li, X.; Dobbs, P.D.; Luu, K. Spartan: Self-supervised spatiotemporal transformers approach to group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 5157–5167.
46. Gavrilyuk, K.; Sanford, R.; Javan, M.; Snoek, C.G. Actor-transformers for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 839–848.
47. Han, M.; Zhang, D.J.; Wang, Y.; Yan, R.; Yao, L.; Chang, X.; Qiao, Y. Dual-AI: Dual-path Actor Interaction Learning for Group Activity Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, New Orleans, LO, USA, 19–24 June 2022; pp. 2990–2999.
48. Tamura, M.; Vishwakarma, R.; Vennelakanti, R. Hunting Group Clues with Transformers for Social Group Activity Recognition. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part IV; Springer: Berlin/Heidelberg, Germany, 2022; pp. 19–35.
49. Johnson, J.; Krishna, R.; Stark, M.; Li, L.J.; Shamma, D.; Bernstein, M.; Fei-Fei, L. Image retrieval using scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Boston, MA, USA, 7–12 June 2015.
50. Dai, B.; Zhang, Y.; Lin, D. Detecting visual relationships with deep relational networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 June 2017.
51. Zhang, J.; Elhoseiny, M.; Cohen, S.; Chang, W.; Elgammal, A. Relationship proposal networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 June 2017.
52. Kolesnikov, A.; Kuznetsova, A.; Lampert, C.; Ferrari, V. Detecting visual relationships using box attention. In Proceedings of the International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
53. Qi, M.; Li, W.; Yang, Z.; Wang, Y.; Luo, J. Attentive relational networks for mapping images to scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition (CVPR 2019), Long Beach, CA, USA, 16–20 June 2019.
54. Xu, D.; Zhu, Y.; Choy, C.B.; Fei-Fei, L. Scene graph generation by iterative message passing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 June 2017.
55. Zellers, R.; Yatskar, M.; Thomson, S.; Choi, Y. Neural motifs: Scene graph parsing with global context. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018.
56. Tang, K.; Zhang, H.; Wu, B.; Luo, W.; Liu, W. Learning to Compose Dynamic Tree Structures for Visual Contexts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition (CVPR 2019), Long Beach, CA, USA, 16–20 June 2019.
57. Yang, J.; Lu, J.; Lee, S.; Batra, D.; Parikh, D. Graph r-cnn for scene graph generation. In Proceedings of the European Conference on Computer Vision (ECCV 20108), Minich, Germany, 8–14 September 2018.
58. Lin, X.; Ding, C.; Zeng, J.; Tao, D. Gps-net: Graph property sensing network for scene graph generation. In Proceedings of the IEEE/CVF European Conference on Computer Vision (CVPR 2020), Minich, Germany, 8–14 September 2020.
59. Chen, T.; Yu, W.; Chen, R.; Lin, L. Knowledge-embedded routing network for scene graph generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition (CVPR 2019), Long Beach, CA, USA, 16–20 June 2019.
60. Li, Y.; Ouyang, W.; Zhou, B.; Shi, J.; Zhang, C.; Wang, X. Factorizable net: An efficient subgraph-based framework for scene graph generation. In Proceedings of the IEEE/CVF European Conference on Computer Vision, (ECCV 2018), Minich, Germany, 8–14 September 2018.
61. Suhail, M.; Mittal, A.; Siddiquie, B.; Broaddus, C.; Eledath, J.; Medioni, G.; Sigal, L. Energy-Based Learning for Scene Graph Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition (CVPR 2021), Virtual, 19–25 June 2021.
62. Gu, J.; Zhao, H.; Lin, Z.; Li, S.; Cai, J.; Ling, M. Scene graph generation with external knowledge and image reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition (CVPR 2019), Long Beach, CA, USA, 16–20 June 2019.
63. Zareian, A.; Karaman, S.; Chang, S.F. Bridging knowledge graphs to generate scene graphs. In Proceedings of the ECCV: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020.
64. Zareian, A.; Wang, Z.; You, H.; Chang, S. Learning Visual Commonsense for Robust Scene Graph Generation. In Proceedings of the ECCV: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020.
65. Lu, C.; Krishna, R.; Bernstein, M.; Fei-Fei, L. Visual relationship detection with language priors. In Proceedings of the European Conference on Computer Vision (ECCV): 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016: pp. 852–869.
66. Zhong, Y.; Shi, J.; Yang, J.; Xu, C.; Li, Y. Learning to generate scene graph from natural language supervision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, (ICCV 2021), Virtual, 11–17 October 2021.
67. Ye, K.; Kovashka, A. Linguistic Structures as Weak Supervision for Visual Scene Graph Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition (CVPR 2021), Virtual, 19–25 June 2021.
68. Nguyen, P.; Quach, K.G.; Kitani, K.; Luu, K. Type-to-track: Retrieve any object via prompt-based tracking. In Proceedings of the NeurIPS 2023: 37th Annual Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.

69. Nguyen, P.; Truong, T.D.; Huang, M.; Liang, Y.; Le, N.; Luu, K. Self-supervised domain adaptation in crowd counting. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; IEEE: New York, NY, USA, 2022; pp. 2786–2790.
70. Nguyen, T.T.; Nguyen, P.; Luu, K. HIG: Hierarchical Interlacement Graph Approach to Scene Graph Generation in Video Understanding. *arXiv* **2023**, arXiv:2312.03050.
71. Quach, K.G.; Le, N.; Duong, C.N.; Jalata, I.; Roy, K.; Luu, K. Non-volume preserving-based fusion to group-level emotion recognition on crowd videos. *Pattern Recognit.* **2022**, *128*, 108646. [CrossRef]
72. Shang, X.; Ren, T.; Guo, J.; Zhang, H.; Chua, T.S. Video visual relation detection. In Proceedings of the 25th ACM international conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1300–1308.
73. Teng, Y.; Wang, L.; Li, Z.; Wu, G. Target adaptive context aggregation for video scene graph generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 13688–13697.
74. Dong, L.; Gao, G.; Zhang, X.; Chen, L.; Wen, Y. Baconian: A Unified Open-source Framework for Model-Based Reinforcement Learning. *arXiv* **2019**, arXiv:1904.10762.
75. Li, X.; Guo, D.; Liu, H.; Sun, F. Embodied semantic scene graph generation. In Proceedings of the Conference on Robot Learning, London, UK, 8–11 November 2022; pp. 1585–1594.
76. Yang, J.; Peng, W.; Li, X.; Guo, Z.; Chen, L.; Li, B.; Ma, Z.; Zhou, K.; Zhang, W.; Loy, C.C.; et al. Panoptic video scene graph generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 18675–18685.
77. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, J. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970.
78. Xu, J.; Mei, T.; Yao, T.; Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE Conference on Computer Vision and Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5288–5296.
79. Pan, Y.; Yao, T.; Li, H.; Mei, T. Video captioning with transferred semantic attributes. In Proceedings of the IEEE Conference on Computer Vision and Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6504–6512.
80. Yang, J.; Ang, Y.Z.; Guo, Z.; Zhou, K.; Zhang, W.; Liu, Z. Panoptic scene graph generation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 178–196.
81. Martin-Martin, R.; Patel, M.; Rezatofighi, H.; Shenoi, A.; Gwak, J.; Frankel, E.; Sadeghian, A.; Savarese, S. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE Trans. Anal. Mach. Intell.* **2021**, *45*, 6748–6765. [CrossRef] [PubMed]
82. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv* **2023**, arXiv:2306.05685.
83. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.