Situational Scene Graph for Structured Human-centric Situation Understanding

Chinthani Sugandhika[1,2,3] Chen Li[2,3] Deepu Rajan[1] Basura Fernando[1,2,3]
[1]Nanyang Technological University, Singapore.
[2]Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore.
[3]Institute of High-Performance Computing, Agency for Science, Technology and Research, Singapore.

Abstract

Graph based representation has been widely used in modelling spatio-temporal relationships in video understanding. Although effective, existing graph-based approaches focus on capturing the human-object relationships while ignoring fine-grained semantic properties of the action components. These semantic properties are crucial for understanding the current situation, such as where does the action takes place , what tools are used and functional properties of the objects. In this work, we propose a graph-based representation called Situational Scene Graph (SSG) to encode both human-object relationships and the corresponding semantic properties. The semantic details are represented as predefined roles and values inspired by situation frame, which is originally designed to represent a single action. Based on our proposed representation, we introduce the task of situational scene graph generation and propose a multi-stage pipeline Interactive and Complementary Network (InComNet) to address the task. Given that the existing datasets are not applicable to the task, we further introduce a SSG dataset whose annotations consist of semantic role-value frames for human, objects and verb predicates of human-object relations. Finally, we demonstrate the effectiveness of our proposed SSG representation by testing on different downstream tasks. Experimental results show that the unified representation can not only benefit predicate classification and semantic role-value classification, but also benefit reasoning tasks on human-centric situation understanding.

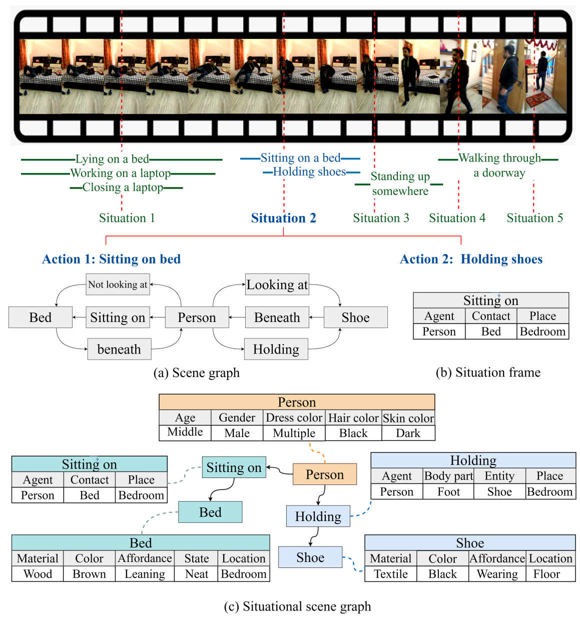

Figure 1. This video frame depicts a human-centric situation of the two concurrent actions "sitting on bed" and "holding shoes". Different structured action representation methods include, (a) Scene graph, (b) situation frame, **(c) Situational scene graph (ours)**: encompasses the person, objects, and verb predicate of human-object relations and their semantic role-values, providing a detailed schema with precisely defined structures to elaborate the components of one or more concurrent actions.

1. Introduction

Human-centric visual understanding has traditionally focused on identifying actions occurring in an image or video input [2, 17, 44]. Although superior performance has been achieved [24,42,50], these studies fail to capture the relationships between human and objects, which may be essential for human behavior understanding. More recently, graph based representations such as spatio-temporal scene graph (ST scene graph) [19] has been proposed to represent action in a structured manner. The graph based representation encodes human and objects as nodes connected together by pairwise relationships (as shown in 1a), and can support high-level inference tasks such as visual question answering (VQA) [12, 16, 18, 22, 38, 46, 53], image captioning [60, 60], referring expressions [59] image grounded dialog [11] and image retrieval [23] etc. However, the ST scene graph falls short of capturing the fine-grained properties of the relationship components, such as where the action takes place, functional properties of the objects (e.g. affordance) or any tools used by the person to perform the actions etc. Such fine-grain properties of human, objects and their pair-wise relationships are crucial in real-world applications, such as in robotics where the robot needs to understand not only the human behavior but also the properties of the objects (e.g. affordance) in order to interact with them.

Another line of research [61,62] based on situation frames offer a well-structured schema to represent these semantic properties in terms of predefined roles and values, providing richer context (as shown in Fig. 1b). However, they are limited to representing a single action, which is not applicable to scenarios with multiple simultaneous actions. In view of the limitations of existing approaches, we propose **Situational Scene Graph (SSG)** to represent both human-object relations and the corresponding semantic properties in a unified representation. As shown in Fig. 1c, we leverage the well-structured semantic role-value (SRV) pairs to encode the properties of the components within verb relationships (i.e. person, object and verb predicate) of a spatio-temporal scene graph. Our SSG is beneficial for **Human-centric situation understanding** where we aim to understand the actions, human-object relations as well as semantic properties of each entity involved in the current situation. Moreover, another advantage of SSG is its ability to jointly model the human-object relations and the semantic properties which could benefit both semantic role-value prediction and predicate prediction. Take the scene of a person holding a cup as an example. The relationship between the person and the cup (i.e., 'holding') implies that the person is likely engaging in an action related to the cup such as drinking. Additionally, knowing the semantic properties of the cup (e.g., affordance) narrows down the possible actions or relations, and the relationship 'holding' also provides cues for the semantic properties of the cup. This mutual reinforcement enhances the performance of both semantic role-value prediction and predicate prediction.

Based on our proposed representation, we introduce the task of SSG generation and introduce a new framework named **In**teractive and **Com**plementary **Net**work (InComNet). Our InComNet can be factorized into four stages: (I) object SRV classification (II) verb predicate classification (III) verb predicate SRV classification and (IV) person SRV classification. Stage (II) classifies the verb predicates while stage (I), (III) and (IV) classify the semantic role-value of each semantic role of object, verb predicate of the relation instances and person respectively. Given that existing datasets are not applicable to this task, we further introduce a SSG dataset based on the Action Genome [19] annotations by manually collecting 562K+ situation frames.

In summary, our key contributions can be listed as follows. Firstly, we introduce a unified SSG representation to encode both human-object relations and the corresponding semantic properties. Secondly, we introduce the challenging task of SSG generation and propose the InComNet baseline. Thirdly, we introduce the SSG dataset, which consists of 2.5K video clips with human-generated annotations for 25.5K human-centric situations encompassing **25.5K** person SRV frames, **61K** object SRV frames, **52K** verb predicate SRV frames. Lastly, we demonstrate the effectiveness of our proposed SSG representation by testing on different downstream tasks. Experimental results show that the unified representation can not only benefit semantic role-value and predicate prediction, but also benefit reasoning tasks on human-centric situation understanding.

2. Previous work

Structured visual representation methods: Classical structured scene representation methods include scene graphs [23] and situation frames [62]. Recent advancements such as spatio-temporal scene graphs [19] and spatio-temporal situations [43] leveraged the principles laid down by these classical methods to enhance video understanding. Other methods that build on them include STAR [53], scene graphs fusion [54], panoptic scene graph [56] and Panoptic Video Scene Graph (PVSG) [58]. Some notable advancements brought by the above methods include, improving the pixel level accuracy in localization [56, 58] and sub-scene graph fusion [54]. In contrast, situational scene graph takes a unique approach by seeking to leverage the strengths of both scene graph and situation frame to construct a more detailed schema with properly defined structures for representing and encoding components within action while also enabling to capture situations with multiple concurrent actions. Prior work have introduced problem tasks paired with the above representations such as, spatio-temporal/panoptic scene graph generation [19, 56], situation recognition [62] and video semantic role-labelling [43]. Similarly, based on our representation we also introduce a challenging problem task called situational scene graph generation.

Existing datatsets: Table 1 provides some popular video understanding benchmarks that are related to us. One major trend in the early stage is the provision of a large number of video clips with single action labels. Around year 2020, the community started focusing on action decomposition to delve into dynamics within actions. Although these decompositions offer somewhat holistic schemas for representing actions—such as person, object, and relationships, they often fall short in providing structured methods for encoding the properties of those decomposed elements (refer to Table 1 in the supplementary material for a comparison of existing action representations and their required annotations). Following this trajectory, we introduce the SSG benchmark, which further decomposes these action elements into their sub-semantic properties by utilizing semantic roles and values. The benchmark includes a total of **562K** human-generated semantic role-value annotations.

Visual representational models: To generate scene graphs

Dataset	# videos	# hours	# actions	Objects		Verb predicates		# persons	Object SR		Verb predicate SR		Person SR	
				# categories	# instances	# categories	# instances		# roles	# instances	# roles	# instances	# roles	# instances
ActivityNet (2015) [2]	28K	648	200	-	-	-	-	-	-	-	-	-	-	-
DALY (2016) [52]	8k	31	10	41	3.6K	-	-	-	-	-	-	-	-	-
Charades (2016) [44]	10K	82	157	37	-	-	-	-	-	-	-	-	-	-
AVA (2018) [17]	504K	108	80	-	-	49	-	-	-	-	-	-	-	-
EPIC-Kitchen (2018) [10]	-	55	125	331	-	-	-	-	-	-	-	-	-	-
HACS Clips (2019) [66]	0.4K	833	200	-	-	-	-	-	-	-	-	-	-	-
Kinetics-700 (2019) [3]	650K	1794	700	-	-	-	-	-	-	-	-	-	-	-
CAD120++ (2019) [67]	0.5K	0.57	10	13	64K	6	32K	-	-	-	-	-	-	-
Action Genome (2020) [19]	10k	82	157	35	0.4M	25	1.7M	220K	-	-	-	-	-	-
VidSitu (2021) [43]	29.2K	83	-	5.6K	-	1.5K	397K	-	-	-	5	**202K**	-	-
STAR (2021) [53]	22K	80	-	37	-	24	-	-	-	-	-	-	-	-
PVSG (2023) [58]	0.4K	9	-	257	7.5K	57	4.1K	1.1K	-	-	-	-	-	-
SSG (ours)	2.5K	22	157	35	61K	16	52K	25.5K	**16**	**271K**	**12**	164K	**5**	**127K**

Table 1. An analysis of SSG in comparison to other publicly available video understanding benchmarks. SR refers to semantic roles.

and situation frames from images, the approaches explored include Conditional Random Field models [23, 61, 62], Recurrent Neural Networks [35, 37, 55], Graph Neural Networks [25, 26, 39, 57], few-shot learning [4] and transformers [5, 6, 8, 51]. For video data, transformers have been widely used to generate the spatio-temporal scene graphs and spatio-temporal situation frames [7, 28, 34, 43, 47, 58]. Meanwhile, Large Vision Language Models (LVLMs) like CLIP [40], ALIGN [20], LLaVa [32] and VILA [29] have been instrumental in developing powerful joint semantic representations of vision and language. Different vision tasks try to leverage these LVLMs in various ways. For example, recent methods have employed the CLIP embeddings for situation recognition [41] and action recognition [21] attaining excellent results in respective problems. Inspired by this line of research, we also utilize CLIP embedding to solve our new problem situational scene graph generation.

3. Situational Scene Graph

A Situational Scene Graph (SSG) is formed by a set of semantic role-value pairs that elaborates several important semantic entities in a situation(s). As shown in Fig. 2 (tables, employing color codes green, blue and red depict the semantic roles and their associated values for the semantic entities person, object and verb predicate of the relation instance), these semantic entities include the actor or the person, the object instances the person interacts with, and the verb predicate of human-object relation instances. Furthermore, the overall situation may comprise the execution of one or more high-level human actions.

Formally, let us define the set of object classes as $\mathcal{O} = \{O_1, \cdots, O_n\}$, and the person category by P. Person instance in a video frame forms relations with object instances where each relation is represented by a triplet of the form $\langle$*Person, Predicate, Object*$\rangle$, e.g., $\langle$person, holding, dish$\rangle$ in Fig. 2. In the SSG, verb predicates like "holding," "drinking from," and "sitting on" are used to describe human-object interactions because they provide more detailed information about actions than spatial predicates like "in," "under," or "beneath" [15]. The set of verb predicate classes is denoted by $\mathcal{R} = \{R_1, \cdots, R_m\}$. The novelty in the SSG constructs lies in the fact that we annotate the semantic role-value pairs for each entity category and instance (person, object and verb predicates) in a scene graph of a person performing one or more high-level actions. The semantic role-value pair structure, known as the *Frame Structure*, is predefined based on the literature [37, 62]. The set of semantic roles for person class is predefined and is denoted by $S(P) = \{S_1(P), S_2(P), \cdots\}$. Similarly, the semantic roles of each object category are also predefined and denoted by $S(O_k) = \{S_1(O_k), S_2(O_k), S_3(O_k), \cdots\}$ for each object category O_k. Finally, the semantic roles of each verb predicate category are denoted by $S(R_k) = \{S_1(R_k), S_2(R_k), S_3(R_k), \cdots\}$ for each verb predicate category R_k. The associated values v of each semantic role (i.e. $S(P), S(O), S(R)$) depend on the specifics of the person, object and verb predicate of the relation instances in a given frame (e.g. $S_{Place}(Holding) = Kitchen$). The set of all distinct semantic-role values is denoted by $\mathcal{V} = \{v_1, v_2, \cdots\}$. In each video frame, there is one person instance executing one or more high-level human actions forming relations with object instances. Each person, object, or relation instance's semantic role is associated with a value

$$S(A) = v. \tag{1}$$

where $A \in \{P, \mathcal{O}, \mathcal{R}\}$ and $v \in \mathcal{V}$.

3.1. Situational scene graph generation

For a given video frame, the SSG generation task requires the classification of the (1) verb predicate classes and (2) semantic role-values of the person, object instances, and verb predicate of the relation instances as follows.

(1) Verb predicate (VP) classification: Given a video frame consisting of a person and class labels of object instances with their bounding boxes as inputs, a model is required to recognise the verb predicate class $\mathcal{R}$ of the relation instances between person and object instances.

(2) Semantic role-value (SRV) classification: Here the model is required to recognize the role-value v of each semantic role of person, object, and verb predicate of relation

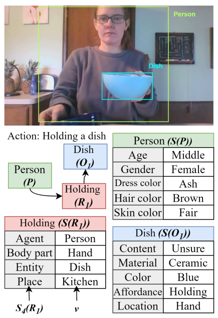

Figure 2. This video frame illustrates a situation of the action 'holding a dish'.

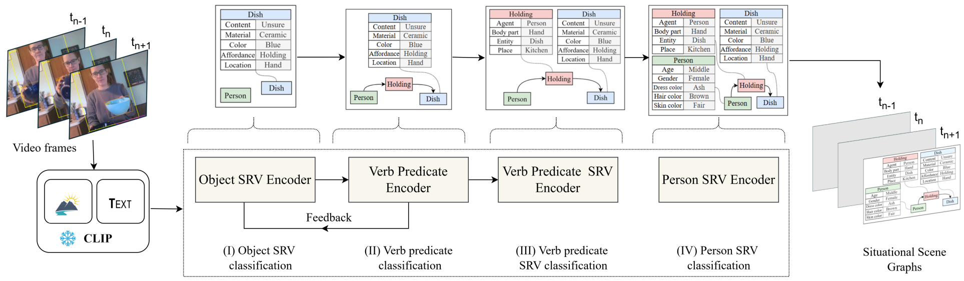

Figure 3. The pipeline of our proposed InComNet. Given a set of video frames, our model uses CLIP to extract necessary feature embeddings from each frame and then classifies SRV of objects, verb predicates, SRV of verb predicates and SRV of person. Finally, the situational scene graph is obtained on the right side. The InComNet stage (II) correspond to the SSG sub-task (1) and stages (I), (III) and (IV) correspond to the SSG sub-task (2).

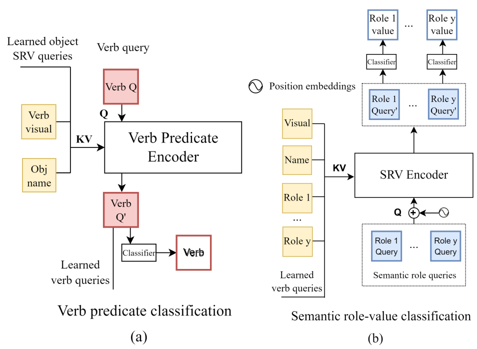

Figure 4. Architectures of verb predicate and SRV encoders.

instances in a given frame as shown in Eq. 1. As mentioned earlier, while semantic roles are shared across a category, the values depend on the specific instance of that category.

3.2. Proposed SSG model: InComNet

We hypothesize that semantic role-value classification helps predicate classification (in scene graph) and vice versa. Based on this, we introduce the InComNet, **In**teractive and **Com**plementary **Net**work, a framework designed for Situational Scene Graph (SSG) generation, which also demonstrates the dataset's utility for this task. As shown in Fig. 3, InComNet has four stages: (I) object SRV classification, (II) verb predicate classification (III), verb predicate SRV classification, and (IV) person SRV classification. The stage (II) classifies the verb predicates in the relation instances between person and object instances while stages (I), (III) and (IV) classify the semantic role-value of each semantic role of object, verb predicate of the relation instances and person respectively as in Eq. 1. In InComNet, we use transformer encoders with cross attention, where learnable query vectors serve as queries, and a sequence of input features act as keys and values. During training, these learnable queries interact with the features in the input sequence and construct useful learned query representations that can be used for the classifications. Accordingly, we employ four such encoders, namely, object SRV encoder, verb predicate encoder, verb predicate SRV encoder and persons SRV encoder corresponding to the InComNet stages (I), (II), (III) and (IV), respectively. The verb predicate encoder's detailed architecture is presented in Fig. 4a. Object, verb predicate and person SRV encoders follow the same architecture as shown in Fig. 4b with the exception that person SRV encoder does not include an intermediate input for learned verb queries.

Leveraging the effectiveness of CLIP [40] embeddings, we derive visual and semantic representations using CLIP's vision ($\psi_v()$) and text ($\psi_t()$) encoders. Additionally, we use a visually prompted frame embedding, employing a new visual prompt named a **translucent background prompt**. In this prompt, we applied a translucent pink overlay to the non-interested regions in the video frame, allowing the model to pay more attention to the interested region. Here the interested region refers to the bounding box region (e.g. in object SRV classification, the prompt is applied to the non-bounding box region of the object while retaining the bounding box region in its original state. Refer to supplementary material Section 6 for more details). Let us denote the prompted frame by f^{pr} and the corresponding visual feature by $U_{f^{pr}} = \psi_v(f^{pr})$. Similarly, text embedding of the object class name U_{O_i} where $O_i \in \mathcal{O}$, and text embeddings of the object semantic role names $\{U_{S_1(O_i)}, U_{S_2(O_i)}, \cdots\}$ are obtained from the CLIP text encoder ψ_t.

Object SRV classification: In the first iteration, the object SRV encoder receives a sequence of key/values $KV = \{U_{f^{pr}}, U_{O_i}, U_{S_1(O_i)}, \cdots, U_{S_x(O_i)}\}$, where x is the maximum number of semantic roles for an object in the SSG dataset. For objects with fewer than x roles, the remaining embeddings are zero-padded. Queries are formed by x learnable query vectors, producing x role query embeddings $Q^O_r = \{q^O_{r_1}, \cdots, q^O_{r_x}\}$. A linear classifier predicts the correct role-values using Q^O_r. In our semantic frame definition, the positions of roles for objects, verb predicates and person remain consistent across the dataset. Hence, only the output embedding corresponding to the appropriate role is utilized during loss calculation.

Verb predicate classification: We hypothesize that object semantic roles and values enhance the classification of verb predicates in the relation between a person and object. Thus, the output of the object SRV encoder from stage (I) is used as additional input for classifying verb predicates R_i in stage (II). The input sequence for stage (II) includes the prompted frame $U_{f^{pr}}$, object class embedding U_{O_i}, and the learned object role query embeddings Q_r^O forming the keys and values $KV = \{U_{f^{pr}}, U_{O_i}, Q_r^O\}$. A learnable verb predicate query vector is then used to generate the verb predicate query embedding Q^R.

Verb predicate SRV classification: The verb predicate SRV encoder receives an input sequence comprising the prompted frame (verb visual features) $U_{f^{pr}}$, the learned verb predicate query embedding Q^R, the object class name embedding U_{O_i} associated with the verb predicate, and the role name embeddings $\{U_{S_1(R_i)}, \cdots, U_{S_y(R_i)}\}$. This sequence forms the keys and values $KV = \{U_{f^{pr}}, Q^R, U_{O_i}, U_{S_1(R_i)}, \cdots, U_{S_y(R_i)}\}$, while y learnable verb predicate role query vectors generate the corresponding role query embeddings Q_r^R. From the second iteration onward, the learned verb predicate query embedding Q^R from the predicted verb predicate instances is used as an additional input to stage (I), complementing the object SRV classification, based on the hypothesis that verb predicates assist in classifying object semantic role-values. Consequently, in iteration 2, the input sequence for stage (I) becomes $KV = \{U_{f^{pr}}, U_{O_i}, Q^R, U_{S_1(O_i)}, \cdots, U_{S_x(O_i)}\}$.

Person SRV classification: The input sequence for stage (IV) includes the prompted frame (person visual features) $U_{f^{pr}}$, the text embedding of the person name[1] U_P, and role name embeddings $\{U_{S_1(P)}, U_{S_2(P)}, \cdots, U_{S_z(P)}\}$, forming the keys and values $KV = \{U_{f^{pr}}, U_P, U_{S_1(P)}, \cdots, U_{S_z(P)}\}$ for the person SRV encoder. Queries are generated by z learnable person role query vectors, producing z learned person role query embeddings Q_r^P. The learned query embeddings Q_r^O, Q^R, Q_r^R and Q_r^P of stages (I), (II), (III) and (IV) are passed through four linear classifiers to generate the final situational scene graph by classifying object semantic role-values, verb predicates, verb predicate semantic role-values and person semantic role-values respectively.

Stages (I), (II), and (III) are iterated d times, with their losses calculated by summing cross-entropy losses across all iterations. The loss for stage (IV) is computed using cross-entropy loss between predicted and ground truth role-values.

[1] in all cases we use term "person".

3.3. SSG dataset

We introduce **S**ituational **S**cene **G**raph (SSG) dataset to facilitate the proposed human-action representation of situational scene graphs. SSG dataset is built on Action Genome's [19] spatio-temporal scene graphs and Action Genome is built on Charades [44]. However, unlike the Action Genome whose goal is to decompose Charades' video-level actions by annotating person, objects and their pairwise relationships, SSG's goal is to provide precise structures to further elaborate these action partonomies using semantic roles and role-values. Moreover, unlike VidSitu [43] which annotates a single action with semantic roles for every 2-second clip, SSG dataset provides frame-level situations in a more granular way than VidSitu. An analysis of it in comparison to other publicly available video understanding datasets is given in Table 1. Among the main components of situational scene graphs, persons, object instances and verb predicates of the relation instances are already annotated in the Action Genome. We provide additional annotations for semantic roles and values for the above semantic entities.

Composing the dataset: In the Action Genome, each person-object instance is characterized by three relationship types: attention, spatial, and contact. We define semantic role-values for verb relationships, termed verb predicates, which capture how a person interacts with an object to perform higher-level actions (2). Among the 17 verb predicate classes, we excluded *other relationship* due to its vagueness and difficulty in decomposition into specific semantic roles and values. The remaining 16 verb predicates, combined with semantic roles and values, effectively represent ongoing actions. We retained all 35 object classes and the person class from the Action Genome without additional filtering. Next, following the concept of semantic roles in FrameNet [14] and previous work [62], we assigned semantic roles for the person, object and verb predicate classes. The raw semantic roles, particularly those related to objects and person, were further refined to ensure their applicability in visually describing a scene. We developed a web-based application software called Semantic Role Labelling Tool to facilitate the annotation process. Annotations were carried out over a period of three months by students with a Computer Science background. Refer to the supplementary material Section 1 for more details about the annotation process, the labelling tool, distribution of the dataset and further dataset statistics.

4. Experiments

Experimental settings: We use our SSG dataset for the SSG generation task. We use the translucent background prompt in all the visual prompt-related experiments and trained the InComNet with a learning rate of 0.001, and ExponentialLR scheduler with Adamax optimizer, on a 49 GB NVIDIA RTX 6000 GPU.

Evaluation metrics: For the verb predicate classification task, we use the metric **accuracy** (Acc) [62] under "with constraint" predicate classification (PREDcls) [19] setting. Further, since semantic role-value classification is similar to the semantic role labelling in conventional verb-based situation frames, we adapt existing metrics **value** and **value-all** [37] along with a new metric called **value-two**. However, unlike in [37, 62], we have only single annotation per record. Therefore, we assess the value by determining whether the model accurately predicts at least one role-value for a given role out of all the roles associated with that particular semantic entity. The value-all metric evaluates if the model can accurately predict all the semantic role-values, while the value-two metric checks if at least two semantic role-values are correctly predicted out of all the roles. We use top-1 setting for verb predicate SRV classification where semantic role-values are considered incorrect if the predicted verb is incorrect. Since we do not perform object detection, we use the ground truth setting for object and person SRV classifications assuming ground truth object/person to be known [37, 62]. For SRV classification, we also introduce a new metric called **role-based accuracy** (role-based acc.) to understand the role-based performance. It is calculated by first determining the accuracy for each role separately, and then averaging across all roles.

4.1. Situational Scene Graph Generation

Baselines: We compare the performance of InComNet over six baselines under three categories: (1) CLIP zero-shot inference [40], (2) CLIP linear probing [40] and (3) VILA inference [29]. In CLIP linear probing [40], we attach a linear classifier on top of the frozen CLIP vision encoder for each object SRV, verb predicate, verb predicate SRV and person SRV classifications. In VILA inference, we prompt VILA-1.5-3B model in interleaved image-text VQA MCQ style with MCQ questions generated using the SSG annotations. Under this category we showcase both zero-shot and impact of fully fine-tuning an LVLM like VILA on the SSG task. The option space for verb predicate and SRV MCQ questions includes the set of verb predicate classes and unique set of role-values associated with each semantic role in the SSG dataset respectively. More details about fine-tuning settings for VILA, including the prompts used are given in the supplementary material Section 2.

InComNet models: We have three InComNet variants as (1) InComNet-224 (base model), (2) InComNet-336-Frozen and (3) InComNet-336-FT. We use ViT-B-32 and ViT-L-14-336 CLIP vision and text encoders to extract image and text feature embeddings in (1) and (2) respectively, while in (3), we fine-tune ViT-L-14-336 CLIP model on SSG dataset and then use this model to extract the required embeddings. We perform all ablation studies and hyper-parameter analysis etc. using the InComNet-224 base model. When fine-tuning the Clip ViT-L-14-336 model on our SSG dataset, we created a detailed description about each video frame with the SSG annotations. Refer supplementary material Section 3 for more details about CLIP fine-tuning on SSG dataset. We refer InComNet's individual tasks as the scenario where its sub-tasks- object SRV, verb predicate and verb predicate SRV are trained independently without inter-task communication.

Table 2 shows the performance comparison of InComNet with baselines over the SSG dataset. Our analysis reveals following key insights: **(1) Superior performance of InComNet-336-FT**: InComNet-336-FT outperforms VILA1.5-3b-FT by an average margin of 15.5% and CLIP linear probing-336-FT model by an average margin of 34.3%; **(2) Enhances performance through joint and iterative training**: Overall performance of object SRV, verb predicate and verb predicate SRV classifications improves by 4% when they are iteratively and jointly trained with complementary exchange of input-output information from other sub-tasks than training each sub-task individually. This demonstrates the synergistic benefits of our approach, which validates our hypothesis that semantic role-value classification and predicate classification mutually enhance each other; **(3) Impact of higher image resolution**: Increasing the image resolution from 224 to 336 pixels in the InComNet-336-Frozen model significantly enhances its capability to capture fine-grained details, resulting in an average performance improvement of 5.6% over the InComNet-224. This underscores the importance of higher resolution for improved model performance in detailed image analysis; **(4) Benefits of finetuning CLIP features:** Finetuning CLIP vision and text encoders for specific tasks like ours would lead to notable performance gains. Specifically, the ViT-L-14-336-FT model shows a 7.2% improvement over the ViT-L-14-336-Frozen model; **(5) Challenges and opportunities with SSG:** SSG task presents a notable challenge to foundational models like VILA when prompted with large option space. While the value and value-two metrics suggest that the model is capable of identifying one or two dominant properties, the lower performance in the value-all metric highlights that it struggles to accurately identify the majority of the properties associated with people, objects, and verb relationships. This underscores the potential for continued research, positioning SSG as a promising area for exploration by LVLMs.

CLIP inference results are sub-optimal, likely due to its training focus on tasks like image classification, which may introduce a bias towards salient objects within the image. Furthermore, the CLIP model is not tailored for structured prediction similar to our InComNet. In contrast, SSG requires classifying *multiple objects, verb predicates, the person and their semantic role-values in a single video frame*, making the accurate mapping of these elements more challenging. Further ablation studies about InComNet, its it-

Method	Object SRV classification				Verb predicate classification acc.	Verb predicate SRV classification				Person SRV classification			
	Role-based acc.	Value	Value-two	Value-all		Role-based acc.	Value	Value-two	Value-all	Role-based acc.	Value	Value-two	Value-all
CLIP-ViT-B-32 zero-shot inference [40]	1.5	6.9	0.2	0.0	8.9	0.4	0.1	0.0	0.0	0.1	0.2	0.0	0.0
CLIP linear probing-224 [40]	11.0	52.7	2.9	1.1	40.4	3.3	39.0	0.0	0.0	17.3	83.2	3.3	0.0
CLIP linear probing-336-Frozen [40]	11.2	54.4	3.5	1.1	40.1	3.4	38.5	0.0	0.0	19.1	89.7	3.6	0.0
CLIP linear probing-336-FT [40]	11.2	54.5	3.6	1.8	47.5	3.8	46.1	0.0	0.0	19.2	90.3	4.9	0.0
VILA1.5-3b-z-shot (MCQ all options) [29]	11.2	40.4	8.8	0.0	19.6	14.1	19.1	13.0	1.7	35.0	88.2	57.2	0.0
VILA1.5-3b-FT (Instruction-tuned MCQ all options) [29]	21.0	69.2	26.4	0.2	61.3	35.3	59.6	35.4	3.2	44.4	95.7	74.6	1.0
InComNet's individual tasks (no iterations/feedback)	45.5	86.2	54.0	5.7	66.9	26.9	65.3	63.4	28.3	-	-	-	
InComNet-224	46.8	86.6	55.0	5.9	67.8	29.4	66.2	64.7	32.0	39.8	97.4	77.4	1.6
InComNet-336-Frozen	47.4	87.6	58.3	6.8	68.5	36.3	66.7	64.8	32.6	42.2	98.0	77.7	1.8
InComNet-336-FT	**49.6**	**89.5**	**62.1**	**7.8**	**70.8**	**38.4**	**69.0**	**67.4**	**38.7**	**49.3**	**98.5**	**85.8**	**1.9**

Table 2. Performance of InComNet for SSG generation task on SSG dataset. The bold and underlined font show the best and the second best result respectively.

erative refinement of the results including visualizations, per-class and per-role performances of verb predicate and SRV classification can be found in the Supplementary material Section 4. Implementation details and hyper-parameter analysis of the InComNet are given in supplementary material Section 5. Translucent background prompt proved to be the most effective visual prompt for the SSG task compared to existing visual prompts [1, 64] and more details about prompt evaluation can be found in the supplementary material Section 6.

4.2. Applications of situational scene graphs

We now demonstrate how our unified SSG representation can benefit both situation recognition and predicate classification tasks in complementary ways. These tasks are paired with the key building blocks of our representation i.e. situation frames [62] and ST scene graphs [19] respectively.

(1) Situation recognition on SSG dataset: Since situation frame is a key foundational component of our SSG representation, situation recognition task could also be tested on the SSG dataset. The only difference between typical situation recognition problem on datasets such as imSitu [62] and situation recognition on SSG dataset is that, in imSitu there is only one activity verb and its associated semantic roles and values whereas, the SSG dataset includes multiple activity verbs, each paired with semantic roles and values corresponding to multiple concurrent actions within a single video frame. Therefore, in SSG situation recognition problem, the models should be able to classify all the verb predicates and their associated semantic role-values in a video frame which adds another layer of complexity. Accordingly, we purpose the SOTA situation recognition models i.e. Clipsitu also to predict all the verb predicates and their semantic role-values.

Thanks to our SSG representation, as shown in Table 3, our InComNet models consistently outperform clipsitu models and VILA models, achieving average performance improvements of 31.1% and 18.2%, respectively, across all metrics. Furthermore, it should be noted that, our unified representation enables the simultaneous prediction of verbs and their semantic role-values within a single pipeline, fostering a complementary exchange of information and inter-task

Method	Top-1 verb accuracy	Top-1 verb predicate SRV classification			GT verb predicate SRV classification		
		Value	Value-two	Value-all	Value	Value-two	Value-all
Clipsitu MLP [41]		40.1	30.7	0.8	97.1	55.3	9.0
Clipsitu TF [41]	41.5	40.0	33.0	2.5	97.1	57.0	8.5
Clipsitu XTF [41]		39.9	33.0	2.2	97.1	54.9	6.2
VILA1.5-3b-z-shot [29]	19.6	19.1	13.0	1.7	96.8	54.3	5.5
VILA1.5-3b-FT [29]	61.3	59.6	35.4	3.2	**98.2**	82.0	30.4
InComNet-224	67.8	66.2	64.7	34.0	97.2	94.0	51.9
InComNet-336-Frozen	68.5	66.7	64.8	32.6	97.3	92.7	38.5
InComNet-336-FT	**69.8**	**68.2**	**66.9**	**41.2**	**98.2**	**94.8**	**58.1**

Table 3. Performance of SOTA task-specific situation recognition models and foundation models on SSG dataset for situation recognition. The bold and underlined font show the best and the second best result respectively.

communication on verb predicate and SRV classification. In contrast, both the Clipsitu and VILA models depend on a two-stage prediction approach, where the two tasks are handled separately.

(2) Predicate classification with situational scene graphs on Action Genome dataset: In our SSG dataset, we provided annotations for only 8% of the Action Genome dataset. In this experiment, we trained the InComNet model for predicate classification (under PredCLS setting and with-constraint strategy) on this limited 8% and used transfer learning to infer object and verb predicate semantic role-value annotations for the entire Action Genome test set during inference. This approach enabled effective evaluation of the predicate classification problem across the entire Action Genome dataset with the help of its 8% SSG annotations. It is worth to mention that, while others have used the entire Action Genome dataset during training, our model has used only 8% Action Genome frames for which we have provided the SSG annotations. We attach three linear classifiers for the InComNet's verb predicate encoder for the three relationship types in Action Genome. The results in Table 4 demonstrate that we can achieve comparable performance with SOTA methods with just 8% data.

4.3. Reasoning on human-centric situations

Next, we evaluate the usability of our SSG representation for human-centric situation understanding by leveraging the zero-shot image-text interleaved MCQ inference capabilities of the VILA 1.5-3B model [29]. For this experiment, we generate 139K+ MCQ questions using the annotations in the

Method	R@10	R@20	R@50
VRD [33]	51.7	54.7	54.7
M-FREQ [63]	62.4	65.1	65.1
MSDN [27]	65.5	68.5	68.5
VCTree [45]	66.0	69.3	69.3
RelDN [65]	66.3	69.5	69.5
GBS-Net [30]	66.8	69.9	69.9
STTran TPI [49]	<u>69.7</u>	72.6	72.6
APT [28]	69.4	**73.8**	**73.8**
TD^2-Net(P) [31]	70.1	-	73.1
STTran [7]	68.6	71.8	71.8
DSG-DETR [13]	68.4	71.7	71.7
TR^2 [48]	**70.9**	**73.8**	**73.8**
TEMPURA [36]	68.8	71.5	71.5
CLIP zero-shot [40]	19.4	20.9	20.9
InComNet-224	69.1	69.2	69.2
InComNet-336-Frozen	67.8	70.8	70.9
InComNet-336-FT	69.4	<u>72.7</u>	<u>72.7</u>

Table 4. Performance of SOTA predicate classification models on Action Genome dataset. **Our method uses only 8% AG frames for training**. The bold and underlined font show the best and the second best result respectively.

Input	Obj SRV questions	Verb predicate questions	Verb predicate SRV questions	Person SRV questions
(1) Question	51.1	54.1	60.3	25.1
(2) Image + question	53.9	60.1	70.6	46.2
(3) Image + predicted graph + question	<u>60.0</u>	<u>70.8</u>	<u>75.1</u>	<u>50.0</u>
(4) Image + GT graph + question	**76.9**	**85.5**	**78.5**	**71.8**

Table 5. SSG VQA on VILA [29]. The bold and underlined font show the best and the second best result respectively.

SSG test set each having **four** options. Thus, the test questions set comprises of 12K verb predicate questions, 38K verb predicate SRV questions, 64K object SRV questions and 25K person SRV questions. Performance is measured using accuracy. The results are shown in Table 5. In (3) and (4) the predicted/ GT SSG graph in text format is incorporated as an in-context reference. The results demonstrate that, incorporating the predicted SSG graph as an in-context yields an average of 10% improvement in accuracy while GT SSG graph leads to 24% marking a certain upper bound on performance when using graph data. This underscores the value of explicitly representing the elements in situations using the SSG graph structure for improved human-centric situated understanding and reasoning. Additionally, it highlights the potential avenues for enhancing future SSG models on this area.

5. Challenges and outlook

Challenges The SSG tasks are particularly challenging due to the need for (1) strong visual cues to identify the attributes of smaller objects (e.g., the material of a dish), subtle variations in human poses (e.g., sitting vs. leaning), precise object locations (e.g., a laptop on a table or on someone's lap), recognizing specific body parts involved in actions, and the tools used to perform those actions. Additionally, (2) situational common sense is essential for understanding semantic role-values for certain roles, such as object affordances (e.g., a doorknob's affordance when opening a door is unlocking, while it is locking when closing). Finally, (3) unlike problems such as typical situation recognition which requires recognising one salient activity verb and its semantic role-values, the SSG task particularly requires recognising multiple verb predicates, multiple objects, the person and their semantic role-values from a single video frame that correspond to multiple concurrent actions.

Outlook SSG dataset has shown potential to enhance performance in existing vision tasks such as situation recognition, predicate classification and human-centric situation reasoning by leveraging its detailed annotations. Beyond these tasks, we also foresee the potential of SSG in pushing the current VLMs towards a more granular level of video perception and reasoning specially in the following areas: **(1) Video QA**: While current video questions answering datasets such as STAR [53] and AGQA [16] lack fine-grain details about entities involved in the situations, the inclusion of semantic roles and values of those entities could be crucial for enhancing the situated reasoning capabilities of the VLMs, which we list as a future work; **(2) Dense video captioning**: SSG annotations can also be utilized for dense video captioning and serve as a benchmark for video-description pre-training and evaluation. We also list this as a potential future direction; **(3) Video generation**: recent work explores image/video generation from structured representations [9]. Given the fine-grained semantic properties embedded in situational scene graphs, we anticipate that future SSG models could enable the generation of videos from situational scene graphs, opening new avenues in video synthesis.

6. Conclusion

We introduce situational scene graphs, a unified graph representation which can represent both human-object relations and semantic properties of the entities involved in a human centric situation. In doing so, we also introduce a new task called situational scene graph generation accompanied by a new dataset, SSG and a new model to address this task. Finally we demonstrate the utility of our SSG representation for situation recognition, predicate classification and human-centric situation understanding. The need for extensive annotations to construct SSG representation may be seen as a limitation. Yet, the advancement of structured semi-supervised learning approaches may allow us to generate better annotations unleashing the full potential of SSGs for human-centric situation understanding and reasoning.

Acknowledgment This research/project is supported by the National Research Foundation, Singapore, under its NRF Fellowship (Award# NRF-NRFF14-2022-0001) and funding allocation to B.F. by A*STAR under its SERC Central Research Fund (CRF).

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 7

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1, 3

[3] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 3

[4] Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. Scene graph prediction with limited labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2580–2590, 2019. 3

[5] Junhyeong Cho, Youngseok Yoon, and Suha Kwak. Collaborative transformers for grounded situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19659–19668, 2022. 3

[6] Junhyeong Cho, Youngseok Yoon, Hyeonjun Lee, and Suha Kwak. Grounded situation recognition with transformers. *arXiv preprint arXiv:2111.10135*, 2021. 3

[7] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16372–16382, 2021. 3, 8

[8] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[9] Yuren Cong, Jinhui Yi, Bodo Rosenhahn, and Michael Ying Yang. Ssgvs: Semantic scene graph-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2555–2565, 2023. 8

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 3

[11] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 1

[12] Bruno Cesar de Oliveira Souza, Marius Aasan, Helio Pedrini, and Adin Ramirez Rivera. Selfgraphvqa: a self-supervised graph neural network for scene-based question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4640–4645, 2023. 1

[13] Shengyu Feng, Hesham Mostafa, Marcel Nassar, Somdeb Majumdar, and Subarna Tripathi. Exploiting long-term dependencies for generating dynamic scene graphs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5130–5139, 2023. 8

[14] Charles J Fillmore, Christopher R Johnson, and Miriam RL Petruck. Background to framenet. *International journal of lexicography*, 16(3):235–250, 2003. 5

[15] 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

[16] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 1, 8

[17] 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. 1, 3

[18] Marcel Hildebrandt, Hang Li, Rajat Koner, Volker Tresp, and Stephan Günnemann. Scene graph reasoning for visual question answering. *arXiv preprint arXiv:2007.01072*, 2020. 1

[19] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 1, 2, 3, 5, 6, 7

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[21] Tianyu Jiang and Ellen Riloff. Exploiting commonsense knowledge about objects for visual activity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7277–7285, 2023. 3

[22] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 1

[23] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 1, 2, 3

[24] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16030, 2021. 1

[25] Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. Situation recognition with graph neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4173–4182, 2017. 3

[26] Shijie Li, Jinhui Yi, Yazan Abu Farha, and Juergen Gall. Pose refinement graph convolutional network for skeleton-based action recognition. *IEEE Robotics and Automation Letters*, 6(2):1028–1035, 2021. 3

[27] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*, pages 1261–1270, 2017. 8

[28] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13874–13883, 2022. 3, 8

[29] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 3, 6, 7, 8

[30] 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. 8

[31] Xin Lin, Chong Shi, Yibing Zhan, Zuopeng Yang, Yaqi Wu, and Dacheng Tao. Tdˆ 2-net: Toward denoising and debiasing for dynamic scene graph generation. *arXiv preprint arXiv:2401.12479*, 2024. 8

[32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3

[33] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 852–869. Springer, 2016. 8

[34] Yichao Lu, Himanshu Rai, Jason Chang, Boris Knyazev, Guangwei Yu, Shashank Shekhar, Graham W Taylor, and Maksims Volkovs. Context-aware scene graph generation with seq2seq transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15931–15941, 2021. 3

[35] Arun Mallya and Svetlana Lazebnik. Recurrent models for situation recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 455–463, 2017. 3

[36] 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. 8

[37] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 314–332. Springer, 2020. 3, 6

[38] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4542–4550, 2024. 1

[39] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In *Proceedings of the 27th ACM international conference on multimedia*, pages 84–93, 2019. 3

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4, 6, 7, 8

[41] Debaditya Roy, Dhruv Verma, and Basura Fernando. Clipsitu: Effectively leveraging clip for conditional predictions in situation recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 444–453, 2024. 3, 7

[42] Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: What can 8 learned tokens do for images and videos? *arXiv preprint arXiv:2106.11297*, 2021. 1

[43] 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, 3, 5

[44] 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. 1, 3, 5

[45] 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. 8

[46] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2017. 1

[47] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13688–13697, 2021. 3

[48] Jingyi Wang, Jinfa Huang, Can Zhang, and Zhidong Deng. Cross-modality time-variant relation learning for generating dynamic scene graphs. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8231–8238. IEEE, 2023. 8

[49] Shuang Wang, Lianli Gao, Xinyu Lyu, Yuyu Guo, Pengpeng Zeng, and Jingkuan Song. Dynamic scene graph generation

via temporal prior inference. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5793–5801, 2022. 8

[50] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 1

[51] Meng Wei, Long Chen, Wei Ji, Xiaoyu Yue, and Tat-Seng Chua. Rethinking the two-stage framework for grounded situation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2651–2658, 2022. 3

[52] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Human action localization with sparse spatial supervision. *arXiv preprint arXiv:1605.05197*, 2016. 3

[53] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 1, 2, 3, 8

[54] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021. 2

[55] 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. 3

[56] 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. 2

[57] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018. 3

[58] 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. 2, 3

[59] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4145–4154, 2019. 1

[60] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019. 1

[61] Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, and Ali Farhadi. Commonly uncommon: Semantic sparsity in situation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7196–7205, 2017. 2, 3

[62] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542, 2016. 2, 3, 5, 6, 7

[63] 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. 8

[64] Hao Zhang and Basura Fernando. Fine-grained regional prompt tuning for visual abductive reasoning. *arXiv preprint arXiv:2303.10428*, 2023. 7

[65] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11535–11543, 2019. 8

[66] 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

[67] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. Explainable video action reasoning via prior knowledge and state transitions. In *Proceedings of the 27th acm international conference on multimedia*, pages 521–529, 2019. 3