FiTBench: Benchmark for Scene Graph Anticipation with Fine-grained Text Cues

Yao Liu
Wuhan Textile University
Wuhan, China
2415283050@wtu.edu.cn

Yangjun Ou
Wuhan Textile University
Wuhan, China
yjou@wtu.edu.cn

Li Mi
ETH Zürich
Zürich, Switzerland
li.mi@ethz.ch

Abstract

Scene graph anticipation (SGA) aims to predict future pair-wise relations among objects, represented as a scene graph, based only on past visual observations. However, predicting coherent and plausible future relations remains an open challenge due to the lack of fine-grained narrative reasoning capabilities to understand complex object interactions and their dynamics over time. Such reasoning requires modeling various contextual drivers (e.g., person emotions, object states) for video scene evolution. In this paper, we present FiTBench, a benchmark for scene graph anticipation with comprehensive semantics and diverse scenarios. Unlike the existing datasets that only provide scene graph annotations, FiTBench details cues of scene and narrative evolution by providing a systematic annotation pipeline and thus supports further fine-grained scene understanding. Moreover, a model-agnostic Text-Augmented Visual Semantics (TAVS) module is proposed to reason about narrative evolution by incorporating fine-grained text cues. Experimental results suggest that integrating fine-grained contextual cues is the key for improving the anticipation performance of SGA methods: six representative models achieve average gains of 4.33% and 3.12% on two FiTBench datasets, respectively. We release our data and code to the community, our project page: https://fitbench.github.io/page/.

Keywords: Scene Graph Anticipation, Future Prediction, Video Understanding, Scene Understanding

1. Introduction

Scene graph anticipation (24; 21) aims to predict the future structured representation of objects and their relationships (*e.g.*, scene graph) based only on past visual observations. By providing a basis for video scene evolution, SGA supports various video understanding tasks, such as action prediction (7; 20; 5; 35) and video question answering (44; 37).

Anticipating future relationships is inherently challenging, as it requires fine-grained narrative reasoning about key intrinsic states and semantic cues that drive relationship evolution in video narratives but are not directly reflected in visual appearances, so as to understand complex object interactions and model their dynamic processes over time. The modeling necessitates structured annotation of contextual drivers (*e.g.*, person emotions, object states, captions) of video scene evolution. More specifically, the **person emotions** imply their intentions for subsequent actions, thereby triggering changes in their relationships. As shown in Fig. 1 (a), the child's emotion changes from 'happy' to 'crying', which directly prompts the adult to approach and offer comfort, thereby driving the evolution of the relationship from '*walking on*' to '*hugging*'. The **object states** largely determine whether it will be used in the future and how it will be used. As shown in Fig. 1 (b), the bottle's state switches between '*inactive*' and '*active*'; it can be inferred that the person had used the bottle and is now returning it to the fridge, supporting the prediction of the <person-holding-fridge> relationship. Ignoring these fine-grained cues makes it difficult to understand the motivations behind relationship transitions. The **captions** provide overarching activity context that frames the narrative, enabling anticipation of relation sequences by complementing the above fine-grained cues with holistic scene semantics. For example, the phrase '*preparing dinner*' in a caption can suggest a sequence of upcoming relations, such as the <adult-stirring-simmering> relationship. These textual cues are closely interconnected, yet each contributes a distinct perspective on the scene: captions denote global activities, object states indicate functionality affordances, and person emotions signal intent transitions. Together, they complementarily provide fine-grained contextual divers that reveal the scene evolution for relation anticipation.

However, such information is *widely absent in existing benchmark datasets* (8; 42; 30; 21), fundamentally constraining models' ability to reason about narrative dynamics and anticipate coherent and plausible future relations. Moreover, most existing approaches for SGA lack effec-

Figure 1: Overview of FiTBench. The proposed FiTBench supports scene graph anticipation with videos in two views: (a) Exo-centric (third-person) and (b) Ego-centric (first-person). By extending the current video scene graph dataset with a systematically designed annotation pipeline, FiTBench provides three types of fine-grained text annotations: the *emotion* of the person, the *state* of objects, and the *captions* of key frames. The fine-grained annotations detail relation transitions and provide semantic cues for scene graph anticipation.

tive mechanisms to incorporate fine-grained text information and therefore exhibit limited narrative reasoning ability in complex video scenes.

To address these challenges, we propose a scene graph anticipation with **Fi**ne-grained **T**ext cues **Bench**mark, **FiT-Bench**, a new benchmark supports fine-grained narrative reasoning by detailing contextual drivers and scene evolution cues, including person emotions, object states, and captions, as shown in Fig. 1. FiTBench contains 10k videos sampled from two datasets: FiT-AG and FiT-PVSG, with both ego-centric and exo-centric viewpoints. The exo-centric views emphasize global visual relationship (e.g., <adult-in front of-floor> as shown in Fig. 1 (a)), while ego-centric views are dominated by hand-operated (e.g., <person-stirring-simmering> as shown in Fig. 1 (b)). Furthermore, to equip the current SGA model with fine-grained narrative reasoning ability, we propose a model-agnostic Text-Augmented Visual Semantics (TAVS) module. The module decomposes contextual text cues by integrating them separately. It takes contextual text cues as input and supports the relation evolution understanding of several state-of-the-art SGA models.

The main contributions of this work are:

- We propose FiTBench, which introduces fine-grained contextual drivers to facilitate the understanding of relation evolution. In particular, each video sample is enhanced with three annotations of subject emotions, object states, and frame-level captions.
- We propose TAVS, a model-agnostic plug-and-play module for SGA. Experimental results show consistent improvements over six representative SGA models on FiTBench by integrating fine-grained text cues with TAVS.

2. Background And Related Work

2.1. Video Visual Relationship Research

Video visual relationship research (34; 28; 32; 10; 1) aims to detect interactions between objects from video sequences, establishing a foundation for fine-grained scene understanding. Scene graph generation (18; 12; 16; 13) further integrates relational triples into a hierarchical graph structure. Starting from dynamic scene graph generation (SGG) (17; 11; 38; 40; 15; 23; 22; 42; 25; 39; 2; 6; 3), several methods (3; 25; 6) and datasets (42; 21; 22; 30; 8) were proposed. Building on the progress in SGG, scene graph anticipation (SGA) (24; 21) requires a deeper understanding of video scene evolution as it predicts the scene graph for the unseen future.

Dataset	Annotation				View		Task		Video	Video Hours	Avg. Len	Frames	Year
	Emotion	State	Caption	Bbox	Ego-Centric	Exo-Centric	SGG	SGA					
ImageNet-VidVRD (31)	✗	✗	✗	✓	✗	✓	✓	✗	1,000		-	-	2017
VidOR (30)	✗	✗	✗	✓	✗	✓	✓	✗	10,000	99h	720s	-	2019
Action Genome (8)	✗	✗	✓	✓	✗	✓	✓	✓	9,848	99h	35s	234k	2020
OpenPVSG (42)	✗	✗	✓	✗	✓	✓	✓	✗	400	8h	77s	153k	2023
VSGR (21)	✗	✗	✓	✓	✗	✓	✓	✓	3,748	-	-	2M	2025
FiTBench	✓	✓	✓	✓	✓	✓	✓	✓	10,248	107h	56s	387k	2025

Table 1: Comparison of the proposed FiTBench with existing datasets for video scene graph generation and anticipation in terms of annotation information and data volume. FiTBench is more modest in size and contains more fine-grained annotations.

2.2. Future Prediction Task and Benchmarks

Future prediction in video can be roughly divided into two themes: 1) *generating future frames or trajectories* (*e.g.* trajectory forecasting (19; 27; 29), location forecasting, and video forecasting (36; 35).) and 2) *predicting future labels or states* (*e.g.*, future human behavior (5), action (7; 20) or object state forecasting (43; 4; 26)). Among them, SGA (24; 45) aims to predict feature relation triples, which can be categorized into a second theme and connects relation-level and object-level predictions. Compared to previous SGA datasets in Table 1, FiTBench is more modest in size, coverage of videos from different views, and, more importantly, contains comprehensive fine-grained contextual drivers that reveal the scene evolution and support fine-grained narrative reasoning.

3. Benchmark Construction

3.1. Fine-grained Text Cues

Emotion annotation describes the emotional state of a person in each video frame, providing critical clues for relational evolution in specific scenarios like social interactions. For example, at moments T2 and T5 in Fig. 1 (a), when the child looks '*happy*', the relationship between the child and the cake is '*looking at*'. When the child starts '*crying*', the relationship changes to '*not looking at*' the cake.

State annotation is used to capture changes in the functional and physical properties of objects in a video. Actions often trigger these changes and can lead to relational transitions. For example, from T1 to T2 in the Fig. 1 (a), the camera changes from '*still*' to '*shaky*', which then triggers <child-not looking at-cake>.

Frame-level caption annotation provides diverse contextual information by describing each video frame with text. For example, in Fig. 1 (b), the caption says '*I put something into the fridge*' which suggests that the person is about to open the refrigerator, leading to the <person-holding-fridge> relationship.

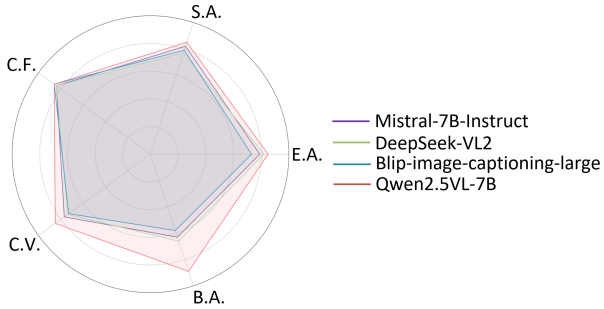

Figure 2: Pilot Study. We manually evaluated the performance of four VLM models (*i.e.*, Mistral-7B-Instruct (9), DeepSeek-VL2 (41), Blip-image-captioning-large (14), Qwen2.5VL-7B (33)) across five metrics using 100 samples: 1) Emotion annotation Accuracy (E.A.); 2) State annotation Accuracy (S.A.); 3) Caption annotation accuracy for describing individual Frames (C.F.); 4) Caption annotation accuracy for describing Video content (C.V.); and 5) Bounding box annotation Accuracy (B.A.).

3.2. Annotation Pipeline

To obtain fine-grained text cues, we develop a systematic annotation pipeline using the large-scale Vision-Language Model (VLM) Qwen2.5VL-7B (33) (Qwen2.5VL).

Pilot study. We manually evaluated the performance of four VLM models (*i.e.*, Mistral-7B-Instruct (9), DeepSeek-VL2 (41), Blip-image-captioning-large (14), Qwen2.5VL-7B (33)) across five metrics using 100 samples: 1) Emotion annotation Accuracy (E.A.); 2) State annotation Accuracy (S.A.); 3) Caption annotation accuracy for describing individual Frames (C.F.); 4) Caption annotation accuracy for describing Video content (C.V.); and 5) Bounding box annotation Accuracy (B.A.). The results, as shown in the Fig. 2, indicate that all VLM models perform exceptionally well in describing frame images. However, the captions generated by the Qwen2.5VL more accurately reflect the content of the entire video. Furthermore, the Qwen2.5VL demonstrates significantly superior performance in generating bounding boxes.

Emotion and state annotation are generated by pro-

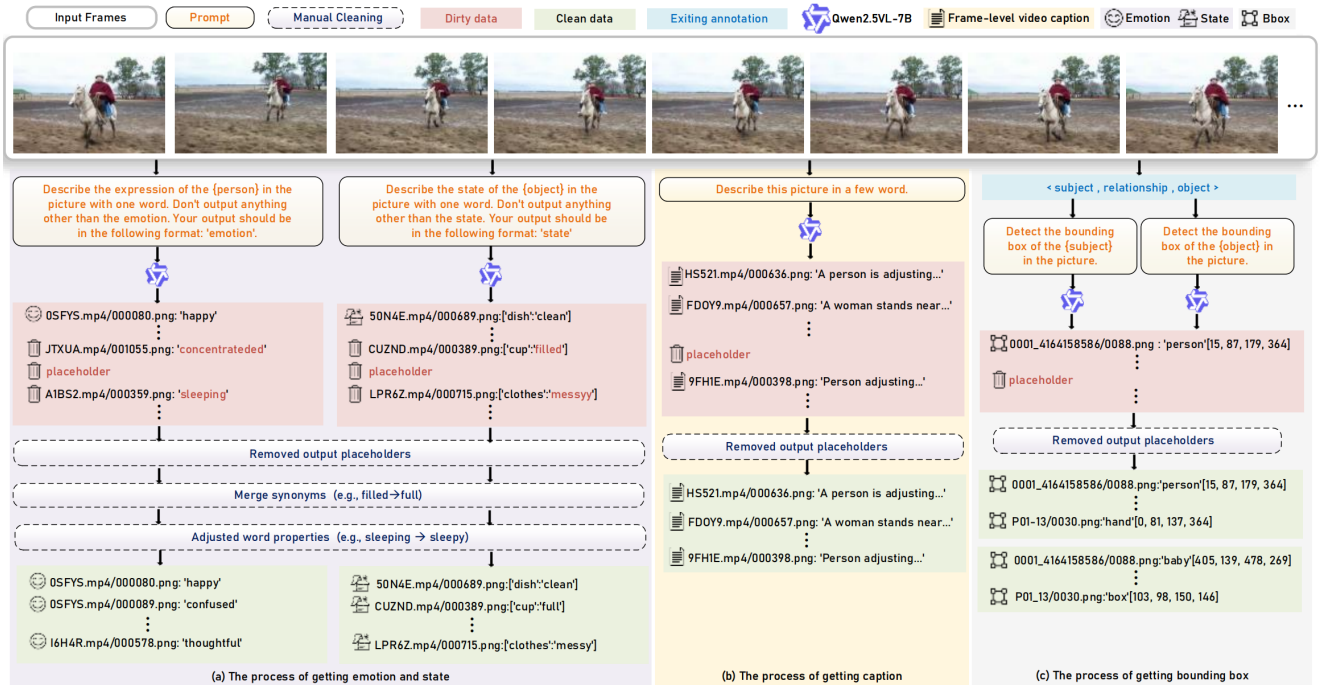

Figure 3: Annotation pipeline. (a) The process of annotating emotion and state. We employ a prompt-based visual language reasoning method to obtain emotion and state annotations for each image frame. We manually removed output placeholders, merged synonyms, and adjusted word properties. (b) The process of annotating frame-level captions. Leveraging the powerful image description capabilities of Qwen2.5VL-7B (33), high-quality video frame captions were obtained after manual cleaning and removing output placeholders. (c) The process of annotating bounding box. To ensure bounding box annotations align with scene graph anticipation targets, we extract subject-object pairs from existing <subject-relationship-object> triples. These pairs, along with video frames, are input into Qwen2.5VL-7B (33) for accurate localization, and invalid and noisy boxes are manually removed.

cessing each video frame with Qwen2.5VL, and restricting its output to word semantic labeling with these two prompts: 'Describe the expression of the person in the picture with one word. Don't output anything other than the emotion. Your output should be in the following format: 'emotion'.' and 'Describe the state of the object in the picture with one word. Don't output anything other than the state. Your output should be in the following format: 'state'.' To ensure annotation quality, we manually removed output placeholders, merged synonyms (e.g., merge '*full*' and '*filled*'), and adjusted the word properties. (*e.g.*, '*sleeping*' to '*sleepy*').

Frame-level caption annotation for the FiT-AG dataset is created by using the VLM to generate a text description for each frame. Qwen2.5VL has powerful image description capabilities, and we input the video frame by frame into VLM. Most of the results obtained are compliant video subtitles, and very few placeholders are output. We manually removed output placeholders and a few abnormal samples. The FiT-PVSG dataset comes from aligning the original segment-level captions to each frame.

Bounding-box annotation for the FiT-PVSG dataset as shown in the Fig 3 (c). We extract all involved subject-object pairs based on the existing <subject-relationship-object> triples to ensure that the bounding box annotations are consistent with the predicted goals of the scene graph. The subject-object pairs of interest are entered into Qwen2.5VL along with the video frames, which can effectively localize the bounding boxes of the objects we are interested in.

3.3. Datasets Statistic

FiTBench contains two datasets, FiT-AG and FiT-PVSG, which select videos from the Action Genome dataset (8) (9848 videos, 25 relationship classes) and the OpenPVSG dataset (42) (400 videos, 72 relationship classes), respectively. Among them, the FiT-PVSG dataset contains 111 ego-centric videos. Table 2 reports the number of frames we annotated and the percentage of total frames that were labeled. Our annotations reach a high annotation density, with an average coverage of 89.3%.

Emotion annotation. We annotated 333k frames in the FiTBench with 36 emotion categories. In Fig. 4 (a), we

	Emotion	State	Caption	BBox
Number	333,821	381,516	409,889	401,743
Percentage	90.6%	85.2%	91.5%	89.7%

Table 2: Annotation statistics in FiTBench. The number of annotated frames and their proportion of the total frames for each annotation type in FiTBench.

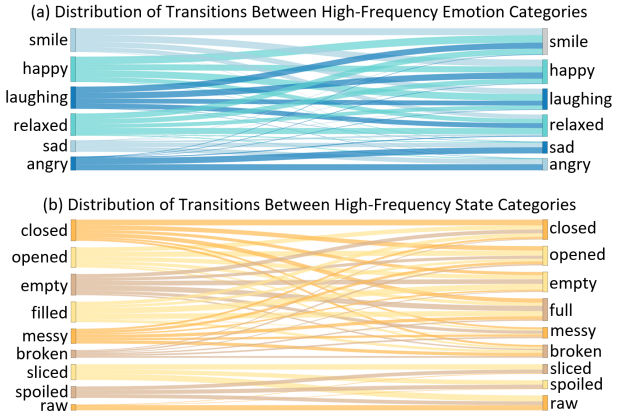

Figure 4: Distribution of transitions between high-frequency (a) emotion and (b) state categories. Emotion transitions are reasonable (*e.g.*, '*smile*' rarely changes to '*sad*'). State transitions are influenced by object properties, container-type objects (*e.g.*, cups) transition from '*full*' to '*empty*', while food-type objects (*e.g.*, meat) transition from '*raw*' to '*spoiled*'. Generally, states belonging to different object properties do not transition into one another (*e.g.*, '*empty*' does not transition to '*raw*')

show the distribution of emotion transitions for the high-frequency emotions, thereby confirming that the transitions between the emotions we annotated are reasonable. For instance, '*smile*' often changes to '*content*' or '*happy*', and rarely to '*sleepy*' or '*sad*'.

State annotation. We annotated 381k frames in the FiTBench with 164 state categories. In Fig. 4 (b), we illustrate the distribution of state transitions for some high-frequency state categories. It can be observed that state transitions do not occur randomly but are influenced by the object's properties. For instance, container-type objects (*e.g.*, cups) transition from '*full*' to '*empty*', while food-type objects (*e.g.*, meat) transition from '*raw*' to '*spoiled*'. Generally, states belonging to different object properties do not transition into one another (*e.g.*, '*empty*' does not transition to '*raw*').

Frame-level caption annotation. We annotated 409k frames in the FiTBench, averaging one caption per 1.1 frames—ensuring dense and comprehensive coverage. As shown in Fig. 5 (a), the length of caption characters in the FiT-AG dataset and FiT-PVSG dataset is concentrated in the 70-80 and 50-60 range, respectively. As shown in Fig. 5 (b), while captions provide general scene descriptions, they often lack details about emotions and object states. The proposed annotations of emotion and state serve as a complement to the captions.

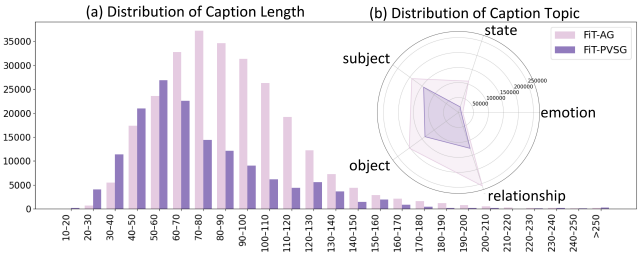

Figure 5: Distribution of (a) caption lengths measured in characters and (b) caption topics in FiTBench. The proposed annotations can effectively complement the topic of emotion, as state in the caption.

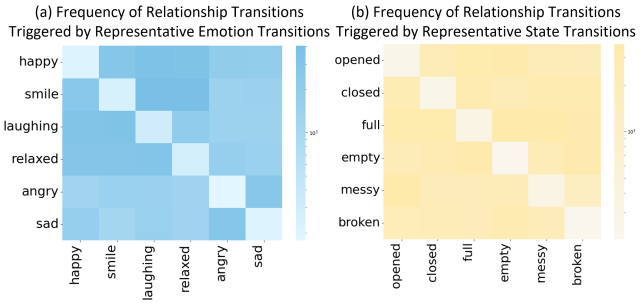

Figure 6: Quality Analysis. The frequency of relationship transitions triggered by certain representative (a) emotion transitions and (b) state transitions. The upper and lower triangular regions of the transition matrix represent the frequency of relationship transitions when emotions or states change, while the diagonal represents the frequency of relationship transitions when emotions or states remain unchanged. Values in the transition matrix are displayed on a logarithmic scale to more clearly highlight differences between categories. Both the upper and lower triangular areas exhibit high values, indicating that emotion transitions and state transitions influence relationship transitions.

3.4. Quality Analysis

Fig. 6 illustrates the frequency of relationship transitions triggered by certain representative (a) emotion transitions and (b) state transitions. The upper and lower triangular regions of the transition matrix represent the frequency of relationship transitions when emotions or states change, while the diagonal represents the frequency of relationship transitions when emotions or states remain unchanged. Values in the transition matrix are displayed on a logarithmic scale to more clearly highlight differences between categories.

Emotion applicability. Both the upper and lower triangular regions exhibit high values, indicating a significant association between emotional transitions and relationship

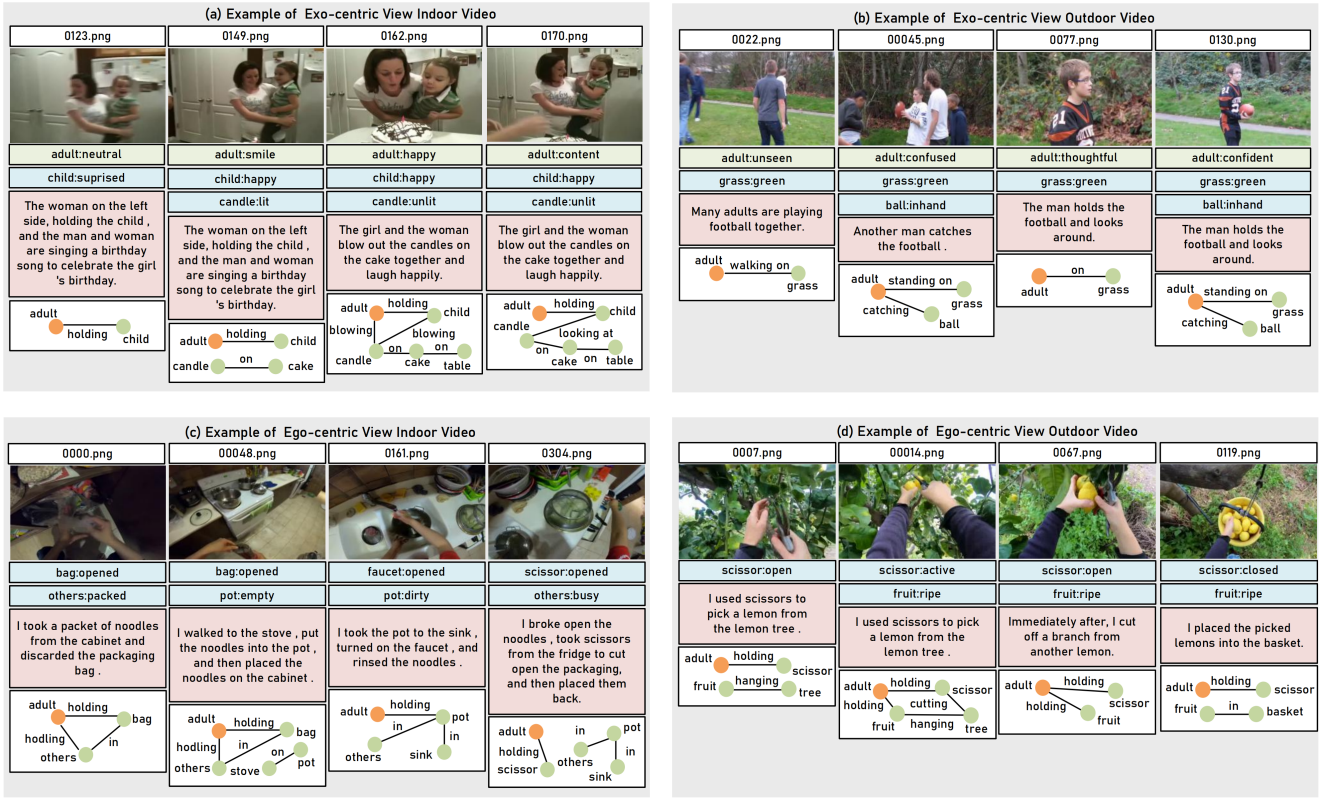

Figure 7: Annotation Reliability. (a) Examples of exo-centric view in indoor video. The VLM model determines that the emotion of an adult with a blurred face is most likely '*smile*' by analyzing the contextual information from the preceding and following video segments. This contextual understanding generates annotations for frames that human annotators are unable to recognize. (b) Examples of exo-centric view in outdoor video. The higher annotation accuracy observed in outdoor videos can be attributed to better lighting conditions and clearer visual cues. Examples of ego-centric view in (c) indoor video and (d) outdoor video. The ego-centric view is filmed from the subject's perspective, with the camera typically moving alongside the subject. The visual focus remains consistently centered on the subject's current behavioral scene (such as hand movements and object interactions). This perspective minimizes unnecessary background noise, aiding VLM models in capturing state-related cues.

transitions. For instance, when emotion transitions from '*relaxed*' to '*happy*', the probability of a relationship transition increases markedly. This contrasts sharply with diagonal values (*e.g.*, '*relaxed*' changes to '*relaxed*'), suggesting minimal relationship adjustments during stable emotional periods. In addition to the lower frequency of diagonal transitions, there are also instances of other low-frequency transitions (*e.g.*, '*sad*' changes' to '*happy*'). This is because fewer samples exhibit this type of emotional shift.

State applicability. Since state transitions are constrained by object categories (*e.g.*, container objects can shift between '*empty*' and '*full*'), we report the frequency of relation changes triggered by a subset of transferable state transitions. Both the upper and lower triangular regions exhibit high values, indicating a significant correlation between state changes and relationship transitions. For instance, when a bottle transitions from '*full*' to '*empty*', the probability of a relationship transition is high. This occurs because the subject likely performed an action like drinking water, and the occurrence of such actions is often followed by relationship transitions.

In summary, fluctuations in subject emotion and object state often reflect or trigger transitions in the relationship. Therefore, incorporating emotion and state transitions as contextual cues significantly enhances the explainability and accuracy of relationship modeling in video understanding systems.

3.5. Annotation Reliability

To ensure the quality of our annotations, we conducted a manual verification process with expert reviewers on a subset of the video dataset, which includes both indoor and outdoor videos under exo-centric and ego-centric viewpoints. We had 100 experts manually check the annotations

		Indoor			Outdoor		
		Emotion	State	Caption	Emotion	State	Caption
Exo-centric	Accept	88.1	90.3	96.0	93.4	91.0	97.1
	Disagree	9.2	5.4	4.4	6.5	8.8	2.4
Ego-centric	Accept	–	98.6	97.9	–	98.5	97.3
	Disagree	–	3.3	2.1	–	1.7	1.6

Table 3: Annotation acceptance (%) across different viewpoints and environments.

of 100 video samples in FiTBench. For each video sample, the experts first checked whether the labels were complete. Then, they check whether the emotion labels and state labels match the subject and object, whether the bounding box can fit the target, and whether the video caption reasonably describes the content of the image. Each sample is checked by two experts. The results of the review, shown in Table 3, demonstrate that most annotations are reliable across different perspectives and scenarios. For exo-centric indoor videos, the emotion annotations achieved an accept rate of 88.1% and a disagree rate of 9.2%. While slightly lower than outdoor videos (accept: 93.4%, disagree: 6.5%), this drop in agreement is mainly attributed to challenges such as poor lighting conditions or partial occlusions in indoor scenes, which can obscure facial expressions. Notably, we observed that large language-vision models tend to infer emotion from temporal context, even when facial cues are not visible. As shown in Fig. 7 (a), the face of the adult in the first frame is blurred, but the model still labels this frame with a '*smile*' emotion label. This is because, based on the preceding and following segments of the video, the model can determine that the adult's emotion is likely to continue as a '*smile*'. This ability to utilize context generates annotations for frames that appear ambiguous to human annotators. Although this introduces a small amount of error, it also reflects the model's reasoning ability in complex environments.

For exo-centric outdoor videos, the annotation quality remains high across all categories, with emotion, state, and caption labels showing acceptance rates above 91%. As shown in Fig. 7 (b), the higher annotation accuracy observed in outdoor videos can be attributed to better lighting conditions and clearer visual cues, which facilitate both model predictions and human verification, especially for action and context-related annotations.

In the ego-centric setting, where the subject's face is typically not visible, we intentionally excluded emotion annotations, ensuring consistency and avoiding unreliable labeling. For other annotations (state and caption), acceptance rates are consistently high—above 97.0%, with disagreement rates below 3.5%, and both state and caption annotations achieve high accuracy. The main reason is that the ego-centric views are shot from the subject's point of view, and the camera usually moves with the subject, with the visual focus steadily focused on the subject's current behavior scene (such as hand movements and object interaction). This perspective reduces unnecessary background interference, which helps the model capture clues related to states, such as '*empty*' in Fig. 7 (c) and '*ripe*' in Fig. 7(d). Thereby improving the clarity of state judgments.

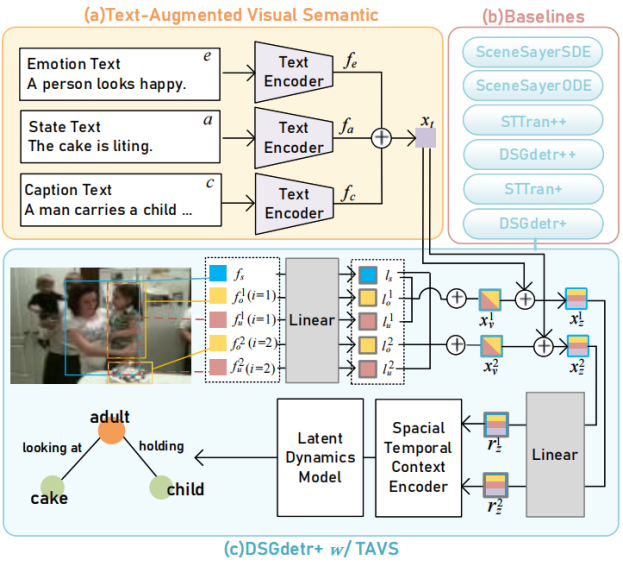

Figure 8: The proposed TAVS method with different baselines. (a) The Text-Augmented Visual Semantic (TAVS) module adds frame-level text features, which are encoded using the CLIP model. (b) We chose six methods as baselines for adaptation. (c) Specific fusion strategies for integrating the TAVS module into each baseline are described.

In summary, while a small portion of the emotion annotations in indoor settings may be affected by visual ambiguity, the overall annotation quality remains high and consistent. Furthermore, observed disagreements often stem from model-generated context-based predictions rather than random noise, which can still provide meaningful supervision for downstream learning tasks.

4. Method

In this section, we describe the proposed Text-Augmented Visual Semantic module (TAVS) (as shown in Fig. 8), a plug-and-play module that enhances the narrative reasoning ability of existing baselines by fusing fine-grained text cues of relation transition. The TAVS module is model-agnostic and can be easily integrated into different baseline models.

4.1. Task Definition

The SGA task aims to predict future relationships between subjects and objects as scene graphs. More specifically, let S and O denote the subject and object set, and the P denote the predicate set. The relationship set R can be

defined as $R = \{r(s,p,o)|s \in S, o \in O, p \in P\}$, where s, p and o are respectively the subject, predicate and object in a relationship triplet $r(s,p,o)$. SGA aims to compute the probability: $R = R(r^{0:T}|V^{-H:0}, r^{-H:0}, s \in S, o \in O, p \in P, r \in R)$, where V^t is the frame at time t, t=0 represents the present. $V^{-H:0}$ denotes the visual history of H previous key frames. $r^{0:T}$ and $r^{-H:0}$ represent the predicates in the future T key frames and the past H frames, respectively.

4.2. Preliminaries

The baseline model uses visual representations for scene graph anticipation. As shown in Fig. 8(c), three types of features are extracted: the subject and object features, and their union region features.

For each subject s and its associated set of i-th objects o, where $i \in \{1, 2, ...\}$, and their union region u, the model extract their visual region features f_s, f_o^i, f_u^i via vision backbone and project them via fully connected layers:

$$\begin{aligned} l_s &= \text{Linear}_s(f_s), \\ l_o^i &= \text{Linear}_o(f_o^i), \\ l_u^i &= \text{Linear}_u(f_u^i). \end{aligned} \tag{1}$$

Then all three are concatenated to form the final visual relationship features x_v^i:

$$x_v^i = \text{Concat}(l_s, l_o^i, l_u^i). \tag{2}$$

4.3. Text-Augmented Visual Semantic Module

In order to understand complex object interactions and their dynamics over time, we incorporate fine-grained text cues to reason about narrative evolution, including person emotion e, object state a, and frame-level caption c. These features are encoded using the CLIP model.

As shown in Fig. 8(a), for each image frame, we construct a list of textual descriptions t, $t \in \{e, a, c\}$. These text messages are passed through the CLIP text encoder to obtain a feature representation:

$$f_t = \text{CLIP}(t), t \in \{e, a, c\}\,. \tag{3}$$

The resulting three text features are combined to form frame-level text features x_t. Frame-level text features are then obtained by aligning these features with the frames corresponding to the visual subject-object pair:

$$x_t = \text{Concat}_{t \in \{e,a,c\}}\, f_t. \tag{4}$$

Then, features from the baseline module (x_v^i) and the TAVS module (x_t) are fused as x_z^i after mapping each of them to a uniform dimension:

$$x_z^i = \text{Concat}(x_v^i, x_t). \tag{5}$$

Finally, x_z^i is further sent to a linear layer to obtain the final fused feature r^i for relation classification.

$$r^i = \text{Linear}_z(x_z^i). \tag{6}$$

5. Experiments

5.1. Settings

We evaluate six representative scene graph anticipation methods on FiTBench. Following the previous SGA literature (24), during inference, seen video frames, active object category labels, and accurate bounding box annotations are sent to the models. The models need to capture relation dynamics and predict future relations.

We evaluated methods under three different scenarios: including the separate train/test on 1) **FiT-AG** and 2) **FiT-PVSG** datasets, and 3) **transfer setting**, where the model is trained on FiT-AG with only exo-centric video and evaluated on the ego-centric subset in the FiT-PVSG. The unique transfer setting explores the models' ability to transfer knowledge learned in the exo-centric views to the ego-centric views.

To construct future scene graphs under these settings, we adopt two widely established strategies from the SGG (12; 6): **With constraint strategy** builds a simplified scene graph by allowing only one relational predicate between any pair of objects. **No constraint strategy** permits multiple relational predicates between the same object pairs, keeping all predicted relations.

5.2. Metrics

Following the previous methods (24), we use Recall@k and Mean Recall@k as the main evaluation metrics, with k chosen from $\{10, 20, 50\}$. To evaluate the model's ability to handle temporal dynamics, we vary the initial clip ratio $\mathcal{F}$ of the input video to values 0.3, 0.5, 0.7, and 0.9.

5.3. Baselines

We select six representative methods from SGA literature (24) as our adaptation baselines: STTran+, DSGDetr+, STTran++, DSGDetr++, SceneSayerODE, and SceneSayerSDE. The first four are extended versions of STTran (3) and DSGDetr (6), respectively.

STTran+ & DSGDetr+ use a basic transformer architecture to build an anticipatory transformer. They generate future relation representations by processing the observed relation context over time.

STTran++ & DSGDetr++ add a temporal encoder to handle relation representations. Their loss function is designed to decode both expected and observed relation representations at the same time.

SceneSayerODE is based on Neural Ordinary Differential Equations, which models the change of object relations as a nonlinear, deterministic process. It captures continuous space-time dynamics using time-invariant vector fields.

SceneSayerSDE uses Neural Stochastic Differential Equations to model real-world uncertainties, such as blurry images or short-term occlusions, as stochastic noise.

Dataset	Method	Recall						Mean Recall					
		With Constraint			No Constraint			With Constraint			No Constraint		
		10	20	50	10	20	50	10	20	50	10	20	50
FiT-AG	STTran+	23.0	25.1	25.2	31.7	47.3	63.5	8.4	9.3	9.3	12.5	23.6	48.6
	STTran+ *w*/TAVS	**26.4**	**29.4**	**29.5**	**38.2**	**54.2**	**64.8**	**11.4**	**13.4**	**13.5**	**17.8**	**28.3**	**53.8**
	DSGDetr+	**30.8**	**32.8**	**32.8**	30.5	47.3	62.8	7.1	7.8	7.8	9.5	19.6	46.0
	DSGDetr+ *w*/TAVS	30.0	32.6	32.7	**38.5**	**53.2**	**64.4**	**9.5**	**10.8**	**10.9**	**13.7**	**24.4**	**51.2**
	STTran++	24.7	26.9	26.9	34.1	48.9	63.7	9.4	10.5	10.5	14.0	25.2	48.6
	STTran++ *w*/TAVS	**33.9**	**37.3**	**37.3**	**43.5**	**57.6**	**65.0**	**13.9**	**16.3**	**16.3**	**20.2**	**32.2**	**57.4**
	DSGDetr++	24.3	26.5	26.5	33.2	49.2	64.0	9.2	10.4	10.5	13.9	25.6	49.2
	DSGDetr++ *w*/TAVS	**33.0**	**35.6**	**35.7**	**41.3**	**55.7**	**64.6**	**12.7**	**14.4**	**14.4**	**17.5**	**30.2**	**56.3**
	SceneSayerODE	37.1	39.5	39.5	43.1	56.2	64.4	17.1	18.7	18.7	**23.4**	**36.5**	**57.1**
	SceneSayerODE *w*/TAVS	**38.2**	**41.5**	**41.5**	**45.7**	**58.2**	**65.0**	**15.9**	**19.0**	**19.1**	22.3	34.7	56.7
	SceneSayerSDE	**38.9**	41.5	41.5	45.7	58.4	65.2	16.7	19.0	19.0	22.2	36.1	59.2
	SceneSayerSDE *w*/TAVS	38.7	**42.0**	**42.0**	**46.4**	**59.2**	**65.3**	**17.9**	**21.3**	**21.4**	**25.3**	**39.3**	**61.2**
FiT-PVSG	STTran+	10.2	**11.0**	11.0	26.3	41.8	49.4	1.0	**1.2**	**1.2**	**2.2**	6.0	18.8
	STTran+ *w*/TAVS	**10.3**	11.0	**11.1**	25.8	**43.9**	**50.7**	**1.1**	**1.2**	**1.2**	2.1	**7.2**	**20.3**
	DSGDetr+	14.1	14.9	15.0	30.4	**43.1**	50.7	1.1	1.3	1.3	**2.3**	**7.6**	**19.8**
	DSGDetr+ *w*/TAVS	**36.3**	**37.1**	**37.2**	**36.3**	**42.4**	**49.7**	**1.6**	**1.8**	**1.8**	**2.3**	7.4	19.5
	STTran++	**10.3**	11.0	11.1	26.3	42.2	49.3	1.0	1.2	1.2	2.2	4.9	17.3
	STTran++ *w*/TAVS	**10.3**	**11.1**	**11.2**	**26.4**	**42.7**	**49.8**	**1.1**	**1.2**	**1.2**	**2.2**	**5.8**	**19.4**
	DSGDetr++	13.1	13.9	14.0	29.3	41.9	49.2	1.1	1.2	1.2	2.2	**5.3**	**18.8**
	DSGDetr++ *w*/TAVS	**32.3**	**33.1**	**33.1**	**35.6**	**42.4**	48.6	**1.5**	**1.7**	**1.7**	**2.3**	5.2	16.2
	SceneSayerODE	37.0	37.6	37.6	23.9	36.8	45.1	**1.7**	1.7	1.7	1.5	3.6	**12.4**
	SceneSayerODE *w*/TAVS	**37.6**	**38.4**	**38.5**	**36.4**	**40.7**	**44.6**	**1.7**	**1.8**	**1.8**	**2.0**	**4.3**	11.5
	SceneSayerSDE	28.0	28.6	28.6	11.2	16.4	28.6	1.9	**2.2**	**2.2**	2.2	5.8	15.0
	SceneSayerSDE *w*/TAVS	**37.4**	**38.1**	**38.2**	**29.2**	**41.5**	**48.4**	**2.1**	**2.2**	**2.2**	**3.0**	**6.9**	**19.5**

Table 4: Compare the baseline method and the baseline method with the TAVS module (baseline *w*/TAVS) on the FiT-AG and FiT-PVSG datasets at $\mathcal{F} = 0.3$.

Dataset	Method	Recall						Mean Recall					
		With Constraint			No Constraint			With Constraint			No Constraint		
		10	20	50	10	20	50	10	20	50	10	20	50
FiT-AG	STTran+	26.7	28.9	29.0	36.0	53.9	71.7	9.7	10.8	10.8	13.9	26.7	52.4
	STTran+ *w*/TAVS	**30.3**	**33.6**	**33.6**	**43.1**	**61.1**	**73.0**	**12.6**	**14.8**	**14.8**	**19.1**	**30.8**	**58.4**
	DSGDetr+	**35.0**	37.1	37.1	34.4	53.2	70.8	8.0	8.7	8.8	10.5	21.4	48.9
	DSGDetr+ *w*/TAVS	34.1	**37.2**	**37.2**	**43.3**	**60.3**	**72.6**	**11.3**	**13.1**	**13.1**	**15.9**	**27.5**	**55.2**
	STTran++	27.8	30.1	30.1	38.2	55.3	71.8	10.3	11.5	11.5	14.8	27.5	52.2
	STTran++ *w*/TAVS	**38.8**	**42.5**	**42.6**	**49.4**	**65.1**	**73.3**	**15.7**	**18.5**	**18.5**	**22.7**	**36.0**	**61.8**
	DSGDetr++	28.1	30.5	30.6	37.4	55.8	72.0	10.7	12.0	12.0	15.3	28.6	52.6
	DSGDetr++ *w*/TAVS	**38.3**	**41.3**	**41.3**	**47.4**	**63.5**	**73.0**	**14.7**	**16.8**	**16.9**	**19.9**	**34.2**	**61.0**
	SceneSayerODE	43.1	45.8	45.8	49.6	64.1	72.8	**20.0**	**21.9**	**21.9**	**26.0**	**39.7**	**62.9**
	SceneSayerODE *w*/TAVS	**43.3**	**46.9**	**46.9**	**51.4**	**65.6**	**73.2**	17.9	21.3	21.5	25.2	38.3	61.7
	SceneSayerSDE	43.8	46.6	46.6	51.4	65.7	73.4	18.9	21.1	21.2	24.4	40.0	63.8
	SceneSayerSDE *w*/TAVS	**43.9**	**47.5**	**47.6**	**52.1**	**66.5**	**73.5**	**20.2**	**24.1**	**24.3**	**28.0**	**43.2**	**65.2**
FiT-PVSG	STTran+	9.8	10.2	10.3	**27.9**	41.2	46.5	**1.0**	1.0	**1.1**	**2.3**	7.4	17.8
	STTran+ *w*/TAVS	**10.0**	**10.4**	**10.5**	27.0	**41.6**	**46.8**	**1.0**	**1.1**	**1.1**	1.7	**7.8**	**18.0**
	DSGDetr+	13.6	14.0	14.1	30.7	**42.4**	**47.6**	1.1	1.1	1.1	2.2	**9.6**	**18.9**
	DSGDetr+ *w*/TAVS	**33.3**	**33.7**	**33.8**	**34.5**	41.4	46.3	**1.5**	**1.6**	**1.6**	**2.3**	8.9	18.2
	STTran++	9.8	10.3	10.3	27.5	40.9	46.4	**1.0**	1.0	**1.1**	2.0	6.1	17.7
	STTran++ *w*/TAVS	**9.9**	**10.4**	**10.5**	**28.1**	**41.7**	**47.4**	**1.0**	**1.1**	**1.1**	**2.4**	**6.9**	**19.2**
	DSGDetr++	13.3	13.8	13.9	29.5	40.6	**46.8**	1.1	1.1	1.1	2.1	**6.4**	**17.7**
	DSGDetr++ *w*/TAVS	**31.7**	**32.2**	**32.3**	**34.6**	**40.8**	46.3	**1.5**	**1.5**	**1.6**	**2.4**	6.3	17.0
	SceneSayerODE	34.9	35.2	35.3	26.0	36.7	**43.2**	**1.6**	**1.7**	**1.7**	1.6	3.7	12.5
	SceneSayerODE *w*/TAVS	**35.1**	**35.5**	**35.6**	**34.7**	**39.3**	**43.2**	**1.6**	1.6	1.6	**2.0**	**5.6**	**13.0**
	SceneSayerSDE	26.3	26.8	26.8	13.0	19.3	31.2	1.4	**1.8**	**1.8**	3.0	5.9	**14.3**
	SceneSayerSDE *w*/TAVS	**35.0**	**35.4**	**35.5**	**29.9**	**38.8**	**44.0**	**1.6**	1.6	1.6	**4.4**	**7.0**	12.5

Table 5: Compare the baseline method and the baseline method with the TAVS module (baseline *w*/TAVS) on the FiT-AG and FiT-PVSG datasets at $\mathcal{F} = 0.5$.

5.4. Experimental Results

Table 4, Table 5, Table 6 and Table 7 sequentially present the test results of FiTBench at $\mathcal{F} = 0.3$, $\mathcal{F} = 0.5$, $\mathcal{F} = 0.7$ and $\mathcal{F} = 0.9$. These results strongly demonstrate the effectiveness of leveraging fine-grained text cues for narrative reasoning in SGA.

1) Standard Experimental Results Analysis

Performance on FiT-AG. After applying the TAVS module to STTran+, DSGDetr+, STTran++, DSGDetr++, and SceneSayerSDE methods, improvements were observed across all $\mathcal{F}$ values. For instance, STTran++ *w*/TAVS achieved an average 9.53% gain at $\mathcal{F} = 0.7$. Although these methods differ in model architecture and loss func-

Dataset	Method	Recall						Mean Recall					
		With Constraint			No Constraint			With Constraint			No Constraint		
		10	20	50	10	20	50	10	20	50	10	20	50
FiT-AG	STTran+	31.4	33.4	33.4	43.8	63.9	81.6	11.9	13.1	13.1	17.8	34.4	61.6
	STTran+ *w*/TAVS	**35.1**	**38.1**	**38.1**	**51.3**	**70.9**	**82.9**	**16.7**	**16.7**	**16.7**	**22.6**	**36.9**	**69.0**
	DSGDetr+	40.0	41.8	41.8	41.0	62.1	80.5	9.1	9.8	9.8	12.6	26.1	57.8
	DSGDetr+ *w*/TAVS	**40.1**	**42.9**	**42.9**	**51.9**	**70.3**	**82.7**	**13.9**	**15.7**	**15.7**	**20.3**	**34.5**	**67.1**
	STTran++	32.0	34.1	34.1	45.8	64.9	81.8	12.4	13.7	13.7	18.2	34.1	62.0
	STTran++ *w*/TAVS	**44.9**	**48.1**	**48.1**	**58.0**	**75.2**	**83.2**	**18.5**	**21.1**	**21.1**	**27.4**	**43.2**	**72.3**
	DSGDetr++	34.6	37.0	37.0	46.0	66.0	81.9	12.4	13.7	13.7	18.2	34.1	62.0
	DSGDetr++ *w*/TAVS	**45.3**	**48.3**	**48.3**	**56.9**	**73.9**	**83.2**	**18.3**	**20.8**	**20.8**	**25.5**	**41.4**	**72.6**
	SceneSayerODE	50.9	53.3	53.3	59.6	75.0	83.0	**23.3**	25.0	25.0	**31.9**	**48.7**	**73.0**
	SceneSayerODE *w*/TAVS	**51.0**	**54.3**	**54.3**	**61.2**	**76.1**	**83.2**	21.5	**25.1**	**25.1**	29.6	46.1	71.6
	SceneSayerSDE	51.2	53.7	53.7	61.0	75.8	83.3	22.4	24.5	24.5	29.7	46.1	74.0
	SceneSayerSDE *w*/TAVS	**51.5**	**54.7**	**54.8**	**61.9**	**77.0**	**83.5**	**23.5**	**27.6**	**27.6**	**34.1**	**51.6**	**76.6**
FiT-PVSG	STTran+	**10.9**	**11.9**	**12.0**	**29.0**	43.0	50.8	**1.2**	**1.3**	**1.3**	**2.0**	**6.7**	**18.6**
	STTran+ *w*/TAVS	**10.9**	**11.9**	**12.0**	28.9	**43.5**	**51.2**	**1.2**	**1.3**	**1.3**	2.0	5.5	18.5
	DSGDetr+	15.3	16.3	16.4	32.4	**45.1**	**52.5**	1.3	1.4	1.4	2.0	**7.0**	**20.8**
	DSGDetr+ *w*/TAVS	**37.0**	**37.9**	**38.0**	**37.3**	43.7	50.6	**1.7**	**1.9**	**1.9**	**2.1**	6.3	18.1
	STTran++	10.9	11.9	12.0	29.0	42.1	49.9	**1.2**	**1.3**	**1.3**	**2.0**	4.8	17.4
	STTran++ *w*/TAVS	**11.6**	**12.6**	**12.7**	**29.6**	**43.6**	**51.3**	**1.2**	**1.3**	**1.3**	**2.0**	**6.2**	**18.6**
	DSGDetr++	15.3	16.3	16.4	31.9	42.7	**50.3**	1.3	1.4	1.4	2.1	5.4	16.6
	DSGDetr++ *w*/TAVS	**36.6**	**37.6**	**37.7**	**37.6**	**43.7**	49.6	**1.7**	**1.9**	**1.9**	**2.2**	**5.5**	**16.7**
	SceneSayerODE	38.0	38.9	39.0	31.4	41.0	47.1	**1.8**	1.9	**1.9**	**1.9**	**6.0**	15.8
	SceneSayerODE *w*/TAVS	**38.8**	**39.6**	**39.7**	**38.3**	**43.5**	**49.4**	**1.8**	**1.9**	**1.9**	**2.2**	5.4	**17.1**
	SceneSayerSDE	31.9	32.4	32.5	19.1	25.8	39.0	**1.8**	**2.0**	**2.0**	2.2	4.6	**18.0**
	SceneSayerSDE *w*/TAVS	**38.8**	**39.6**	**39.7**	**33.8**	**42.0**	**48.8**	**1.8**	1.9	1.9	**2.4**	**5.4**	15.6

Table 6: Compare the baseline method and the baseline method with the TAVS module (baseline *w*/TAVS) on the FiT-AG and FiT-PVSG datasets at $\mathcal{F} = 0.7$.

Dataset	Method	Recall						Mean Recall					
		With Constraint			No Constraint			With Constraint			No Constraint		
		10	20	50	10	20	50	10	20	50	10	20	50
FiT-AG	STTran+	37.1	38.7	38.7	56.2	77.6	92.0	15.8	17.0	17.0	26.3	46.8	75.4
	STTran+ *w*/TAVS	**40.1**	**42.1**	**42.1**	**63.4**	**83.4**	**93.1**	**16.8**	**18.4**	**18.4**	**28.3**	**46.5**	**78.2**
	DSGDetr+	44.7	45.9	45.9	50.9	74.7	91.0	10.3	10.8	10.8	16.3	22.7	71.5
	DSGDetr+ *w*/TAVS	**46.4**	**48.7**	**48.7**	**64.9**	**83.5**	**93.0**	**17.0**	**18.5**	**18.5**	**27.9**	**46.9**	**78.4**
	STTran++	38.5	40.1	40.1	58.3	78.9	92.1	17.7	19.1	19.1	26.5	47.1	75.6
	STTran++ *w*/TAVS	**52.3**	**54.6**	**54.6**	**70.9**	**87.0**	**93.4**	**23.4**	**25.4**	**25.4**	**36.7**	**54.9**	**82.6**
	DSGDetr++	41.6	43.5	43.5	58.2	79.6	92.1	17.3	18.7	18.7	27.6	48.5	76.2
	DSGDetr++ *w*/TAVS	**53.3**	**55.5**	**55.5**	**70.6**	**86.7**	**93.4**	**23.9**	**26.1**	**26.1**	**35.8**	**54.2**	**82.4**
	SceneSayerODE	**59.7**	61.3	61.3	73.2	87.2	93.3	**27.9**	28.9	28.9	**39.5**	**57.5**	83.4
	SceneSayerODE *w*/TAVS	59.6	**61.9**	**61.9**	**74.0**	**87.7**	**93.4**	26.8	**29.4**	**29.4**	38.6	56.5	**86.6**
	SceneSayerSDE	58.5	60.4	60.4	73.6	87.4	93.5	26.4	27.7	27.7	37.6	56.1	87.1
	SceneSayerSDE *w*/TAVS	**60.5**	**62.8**	**62.8**	**74.7**	**88.4**	**93.6**	**28.7**	**31.8**	**31.8**	**42.2**	**60.4**	**89.0**
FiT-PVSG	STTran+	11.1	11.4	11.4	**34.5**	44.9	51.1	**1.2**	**1.3**	**1.3**	**2.6**	**7.7**	**22.1**
	STTran+ *w*/TAVS	**11.3**	**11.6**	**11.7**	33.4	**45.1**	**51.7**	**1.2**	**1.3**	**1.3**	2.4	7.2	20.4
	DSGDetr+	13.2	13.5	13.6	36.6	44.2	**52.7**	1.2	1.3	1.3	**2.6**	7.0	**24.1**
	DSGDetr+ *w*/TAVS	**38.5**	**38.8**	**38.9**	**39.6**	**44.9**	**51.3**	**1.8**	**1.9**	**1.9**	2.6	**8.6**	24.0
	STTran++	11.1	11.4	11.5	34.4	**44.1**	50.7	**1.2**	**1.3**	**1.3**	**2.5**	7.4	21.5
	STTran++ *w*/TAVS	**11.4**	**11.7**	**11.8**	**34.5**	44.0	**51.5**	**1.2**	**1.3**	**1.3**	**2.5**	**7.6**	**24.2**
	DSGDetr++	14.7	15.0	15.1	36.1	43.2	**51.4**	1.3	1.3	1.4	2.6	**6.8**	22.0
	DSGDetr++ *w*/TAVS	**38.5**	**38.8**	**38.9**	**39.8**	**43.9**	51.3	**1.8**	**1.9**	**1.9**	**2.7**	6.1	**23.8**
	SceneSayerODE	38.6	38.8	38.9	37.5	44.2	50.3	1.8	**1.9**	**1.9**	2.5	9.2	21.7
	SceneSayerODE *w*/TAVS	**38.9**	**39.1**	**39.1**	**40.1**	**46.0**	**52.2**	**1.9**	**1.9**	**1.9**	**2.7**	**9.9**	**24.7**
	SceneSayerSDE	35.0	35.2	35.3	33.8	39.8	47.2	1.6	1.6	1.7	3.0	6.0	22.8
	SceneSayerSDE *w*/TAVS	**38.9**	**39.1**	**39.1**	**39.8**	**44.0**	**50.5**	**1.9**	**1.9**	**1.9**	**3.3**	**6.8**	**23.0**

Table 7: Compare the baseline method and the baseline method with the TAVS module (baseline *w*/TAVS) on the FiT-AG and FiT-PVSG datasets at $\mathcal{F} = 0.9$.

tions, they share the common approach of modeling observed video frames to anticipate object interactions in future scenes by understanding scene evolution. Our proposed TAVS enhances this by introducing finer-grained textual cues and modeling diverse contextual drivers, thereby achieving more accurate predictions. The SceneSayerODE approach is different in that it models object relationship transitions as a nonlinear deterministic process. The rich textual cues may cause information redundancy for it, so TAVS performs less well on SceneSayerODE compared to other baseline methods. For example, at $\mathcal{F} = 0.9$, the average improvement rate is only 0.31%. Performance improve-

Dataset	Method	Recall						Mean Recall					
		With Constraint			No Constraint			With Constraint			No Constraint		
		10	20	50	10	20	50	10	20	50	10	20	50
FiT-AG	SDE	58.5	60.4	60.4	73.6	87.4	93.5	26.4	27.7	27.7	37.6	56.1	87.1
	SDE *w* / E	59.3	61.5	61.8	74.1	87.8	93.5	27.2	29.3	29.3	39.5	58.0	87.6
	SDE *w* / S	59.3	62.3	62.3	73.8	88.1	93.3	27.8	29.5	29.5	39.8	57.8	87.6
	SDE *w* / C	59.7	61.8	61.8	74.5	88.0	93.4	27.5	30.8	31.0	40.7	57.5	88.1
	SDE *w* / C.es	60.3	62.1	62.5	74.4	88.1	93.4	28.2	30.5	30.8	41.5	59.8	88.5
	SDE *w* / E+S+C (TAVS)	**60.5**	**62.8**	**62.8**	**74.7**	**88.4**	**93.6**	**28.7**	**31.8**	**31.8**	**42.2**	**60.4**	**89.0**
FiT-PVSG	SDE	35.0	35.2	35.3	33.8	39.8	47.2	1.6	1.6	1.7	3.0	6.0	22.8
	SDE *w* / E	36.8	37.0	37.0	35.2	41.5	47.9	1.5	1.7	1.7	3.1	6.0	22.7
	SDE *w* / S	36.2	36.6	36.8	35.5	41.2	47.2	1.6	1.6	1.8	3.1	6.2	22.8
	SDE *w* / C	37.1	37.5	37.5	37.8	42.8	48.8	1.7	1.7	1.7	3.2	6.2	22.8
	SDE *w* / C.es	37.5	37.7	37.7	37.2	42.2	48.0	1.7	1.7	1.8	3.2	6.5	22.9
	SDE *w* / E+S+C (TAVS)	**38.9**	**39.1**	**39.1**	**39.8**	**44.0**	**50.5**	**1.9**	**1.9**	**1.9**	**3.3**	**6.8**	**23.0**

Table 8: Ablation study of different types of text cues on the FiT-AG and FiT-PVSG dataset. The SceneSayerSDE (SDE) is used as the baseline model. E, S, and C denote emotion, state, and caption, respectively. C.es represents adding E, S, and C simultaneously as a joint caption text.

ment gaps reflect that inherent information fusion methods dictate models' use of new modal information.

Performance on FiT-PVSG. The DSGDetr+, STTran++, DSGDetr++, SceneSayerSDE, and SceneSayerODE methods demonstrate significant improvements at any $\mathcal{F}$ value after adding the TAVS module. Such as DSGDetr+ *w*/TAVS achieves an average improvement of 5.06% at $\mathcal{F} = 0.5$, while SceneSayerODE *w*/TAVS achieves an average improvement of 1.56% at $\mathcal{F} = 0.3$. STTran+ *w*/TAVS and DSGDetr+ *w*/TAVS also improved, though less substantially than other baseline methods. For example, STTran+ *w*/TAVS achieved an average improvement of 0.04% at $\mathcal{F} = 0.5$. This is due to STTran+ and DSGDetr+ employing prediction Transformers built upon foundational Transformer architectures, which generate future relation representations by processing temporal and contextual relationships within observations. When processing datasets like PVSG with more categories and complex scenarios, more detailed contextual clues may be required.

2) Different $\mathcal{F}$ Values Experimental Results Analysis

Performance on FiT-AG. We observe that when $\mathcal{F} = 0.3$, our proposed TAVS module achieves an average improvement of 3.09% over six baseline methods. As $\mathcal{F}$ increases ($\mathcal{F} = 0.5$, $\mathcal{F} = 0.7$, $\mathcal{F} = 0.9$), the average improvement also grows (to 4.26%, 4.84%, and 5.13%, respectively). This is due to the higher initial clip rate of the input video providing more fine-grained text cues from known context, enabling TAVS to enhance the model's narrative reasoning capabilities by leveraging a greater number of text cues.

Performance on FiT-PVSG. Unlike the FiT-AG dataset, the FiT-PVSG dataset contains more diverse scenes and unique ego-centric view samples, making predictions significantly more challenging. Particularly under the with constraint strategy, the dramatic increase in relationship categories within the FiT-PVSG dataset substantially elevates prediction difficulty. When $\mathcal{F}$ is set to 0.3, 0.5, 0.7, and 0.9, respectively, our proposed TAVS module achieves average improvements of 3.63%, 3.03%, 3.04%, and 2.76% over six baseline methods. The video durations in the FiT-PVSG dataset are generally longer. Consequently, even when the initial input video has the lowest clip rate, TAVS can still leverage various contextual drivers within the evolving video scenes to enable more accurate predictions.

$\mathcal{F} = 0.3$	Recall						Mean Recall					
	With Constraint			No Constraint			With Constraint			No Constraint		
	10	20	50	10	20	50	10	20	50	10	20	50
SDE	23.3	23.4	23.7	28.7	32.3	49.9	2.4	2.7	2.7	3.9	**7.6**	21.5
SDE *w* / TAVS	**24.5**	**24.7**	**25.1**	**29.0**	**37.3**	**56.7**	**2.6**	**3.3**	**3.3**	**4.4**	7.5	**22.1**

Table 9: Transfer experiments between exo-centric to ego-centric videos on FiTBench. Both models are only trained on the FiT-AG dataset (which only contains exo-centric video) and evaluated on the ego-centric subset of the FiT-PVSG dataset. Results demonstrate that incorporating fine-grained text cues benefits knowledge transfer between ego-centric and exo-centric videos.

5.5. Ablation Study

To evaluate the impact of the three annotations, we conducted ablation experiments under With-Constraint and No-Constraint strategies on the FiT-AG dataset. Table 8 shows the results of differnt ways of adding text features to the baseline SceneSayerSDE model: adding 1) emotion, 2) state, and 3) caption text features separately improved performance on both the FiT-AG and FiT-PVSG datasets; 4) adding subtitle text features that integrate emotion and state information, which achieved an average improvement of 1.97% on the FiT-AG dataset and an average improvement of 1.26% on the FiT-PVSG dataset; and 5) using the TAVS module to simultaneously introduce emotion, state, and subtitle text features, which achieved an average im-

Figure 9: Scene graph anticipation results of SceneSayerSDE and SceneSayerSDE *w*/TAVS on the proposed FiTBench. We report performance for both short-term (blue box) and long-term (yellow box) future predictions. By incorporating fine-grained text cues with TAVS, the model predicts more coherent and plausible relations for both short-term and long-term futures.

provement of 2.53% on the FiT-AG dataset and an average improvement of 2.27% on the FiT-PVSG dataset. The results demonstrate that incorporating the TAVS module alongside three types of fine-grained text cues yields the greatest improvement to the model. This is due to the fact that integrating emotion and state features into captions may interfere with the context of the caption. The model may struggle to distinguish between the content of the caption itself and the emotion and state features during training, leading to poorer performance in capturing key information compared to when these features are input separately.

5.6. Transfer Setting

Ego-centric videos and exo-centric videos hold different senses and relations. To analyze the role of fine-grained text cues in knowledge transfer between exo-centric and ego-centric videos, we conducted transfer experiments on FiT-Bench. The models are trained on FiT-AG, which only contains exo-centric videos, and evaluated on the ego-centric videos in FiT-PVSG. As shown in Table 9, we found that with the assistance of fine-grained text cues, the model's performance in the transfer setting improved by an average of 0.4%. We argue that fine-grained text features exhibit better generalization properties, remaining applicable

across different viewpoints. Meanwhile, multiple feature inputs enhance the model's representation ability and stability.

5.7. Visualization Results

To investigate the effectiveness of TAVS in the proposed FITBench scene graph anticipation task, we compare the two approaches—SceneSayerSDE and SceneSayerSDE *w* /TAVS—for short-term (blue box) and long-term (yellow box) future predictions, respectively.

Performance on FiT-AG. As expected, long-term future predictions prove more challenging due to larger time intervals and higher uncertainty. However, adding the TAVS module enhances the prediction performance of the baseline models in both situations. This suggests that fine-grained text cues help the model better understand the scene evolution. For example, as shown in Fig. 9 (a), since the television is usually active, our method infers that the person is watching it, leading to the prediction <person-not contacting-television> instead of <person-holding-television> in the short-term. Similarly, when the cup becomes empty, our method predicts <person-not looking at-cup> rather than <person-looking at-cup> for the long-term.

Performance on FiT-PVSG. Multi-scenario, multi-object interactions characterize the FiT-PVSG dataset. Fine-grained textual cues not only assist models in understanding scenario evolution for short-term and long-term future predictions, but also prompt models to focus on the subject-object pairs currently interacting within the scenario. As shown in Fig. 9 (c), the SceneSayerSDE method overlooked <adult-blowing-candle>, whereas fine-grained textual cues enabled the model to capture the interaction between the adult and the candle. Furthermore, as shown in Fig. 9 (d), while the adult maintains spatial relationships with others, the active interaction within the scene does not involve the adult and others. In contrast, SceneSayerSDE *w* /TAVS effectively selects the key interacting objects in the scene.

6. Conclusion and Future Work

In this paper, we propose FiTBench, a scene graph anticipation benchmark that introduces fine-grained contextual drivers to facilitate the understanding of relation evolution and support relation anticipation in the unseen future. In particular, each video sample is enhanced with three types of fine-grained annotations revealing scene evolution: person emotions, object states, and frame-level captions. Moreover, we have designed TAVS, a model-agnostic, lightweight, proof-of-concept module that is specifically intended to clearly verify the intrinsic effectiveness of fine-grained text cues. Extensive experiments confirm that incorporating fine-grained text cues with TAVS consistently improve base models' performance, indicating that decomposing fine-grained contextual drivers is a key step towards understanding relation dynamics and video scene evolution. We anticipate that this work will lay a solid foundation for research on multimodal fusion mechanisms, temporal causal reasoning, and robust scene understanding.

References

[1] Q. Cao and H. Huang. Video visual relation detection with contextual knowledge embedding. *IEEE T Know Data En*, 35(12):13083–13095, 2023. 2

[2] S. Chen, Y. Du, P. Mettes, and C. G. Snoek. Multi-label meta weighting for long-tailed dynamic scene graph generation. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 39–47, 2023. 2

[3] Y. Cong, W. Liao, H. Ackermann, B. Rosenhahn, and M. Y. Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16372–16382, 2021. 2, 8

[4] D. Damen, H. Doughty, G. Maria Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, pages 720–736, 2018. 3

[5] C. Diller, T. Funkhouser, and A. Dai. Futurehuman3d: Forecasting complex long-term 3d human behavior from video observations. In *CVPR*, pages 19902–19914, 2024. 1, 3

[6] S. Feng, H. Mostafa, M. Nassar, S. Majumdar, and S. 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. 2, 8

[7] D. Gong, J. Lee, M. Kim, S. J. Ha, and M. Cho. Future transformer for long-term action anticipation. In *CVPR*, pages 3052–3061, 2022. 1, 3

[8] J. Ji, R. Krishna, F.-F. Li, and J. C. Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pages 10236–10247, 2020. 1, 2, 3, 4

[9] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, and L. Saulnier. Mistral 7b. arxiv preprint. *arXiv preprint arXiv:2310.06825*, 100, 2023. 3

[10] X. Jiang, C. Zheng, X. Xu, B. Liu, W. Zheng, H. Zhang, and S. He. Vrdone: One-stage video visual relation detection. *ACMMM*, 2024. 2

[11] A. Khandelwal. Flocode: Unbiased dynamic scene graph generation with temporal consistency and correlation debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2516–2526, 2024. 2

[12] K. Kim, K. Yoon, Y. In, J. Jeon, J. Moon, D. Kim, and C. Park. Weakly supervised video scene graph generation via natural language supervision. *arXiv preprint arXiv:2502.15370*, 2025. 2, 8

[13] K. Kim, K. Yoon, Y. In, J. Moon, D. Kim, and C. Park. Adaptive self-training framework for fine-grained scene graph generation. In *ICLR*, 2024. 2

[14] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 3

[15] J. Liang, Y. Wang, Z. Wang, M. Liu, R. Fu, Z. Wang, and

B. Qin. Gtr: A grafting-then-reassembling framework for dynamic scene graph generation. In *IJCAI*, pages 1177–1185, 2023. 2

[16] X. Liao, W. Wei, D. Chen, and Y. Fu. Uniq: Unified decoder with task-specific queries for efficient scene graph generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8815–8824, 2024. 2

[17] X. Lin, C. Shi, Y. Zhan, Z. Yang, Y. Wu, and D. Tao. Td2-net: toward denoising and debiasing for dynamic scene graph generation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, pages 3495–3503, 2024. 2

[18] L. Liu, S. Sun, S. Zhi, F. Shi, Z. Liu, J. Heikkilä, and Y. Liu. A causal adjustment module for debiasing scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[19] P. Lv, H. Wei, T. Gu, Y. Zhang, X. Jiang, B. Zhou, and M. Xu. Trajectory distributions: A new description of movement for trajectory prediction. *Computational visual media*, 8(2):213–224, 2022. 3

[20] E. V. Mascaró, H. Ahn, and D. Lee. Intention-conditioned long-term human egocentric action anticipation. In *WACV*, pages 6048–6057, 2023. 1, 3

[21] T.-T. Nguyen, P. Nguyen, J. Cothren, A. Yilmaz, and K. Luu. Hyperglm: Hypergraph for video scene graph generation and anticipation. *arXiv preprint arXiv:2411.18042*, 2024. 1, 2, 3

[22] T.-T. Nguyen, P. Nguyen, and K. Luu. Hig: Hierarchical interlacement graph approach to scene graph generation in video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18384–18394, 2024. 2

[23] T. T. Nguyen, X. Wu, Y. Bin, C.-D. T. Nguyen, S.-K. Ng, and A. T. Luu. Motion-aware contrastive learning for temporal panoptic scene graph generation. In *AAAI*, pages 6218–6226, 2025. 2

[24] R. Peddi, S. Singh, Saurabh, P. Singla, and V. Gogate. Towards scene graph anticipation. In *European Conference on Computer Vision*, pages 159–175. Springer, 2024. 1, 2, 3, 8

[25] T. Pu, T. Chen, H. Wu, Y. Lu, and L. Lin. Spatial–temporal knowledge-embedded transformer for video scene graph generation. *IEEE Transactions on Image Processing*, 33:556–568, 2023. 2

[26] Z. Qi, S. Wang, C. Su, L. Su, Q. Huang, and Q. Tian. Self-regulated learning for egocentric video activity anticipation. *IEEE T-PAMI*, 45(6):6715–6730, 2021. 3

[27] T.-W. Qian, Y. Wang, Y.-J. Xu, Z. Zhang, L. Wu, Q. Qiu, and F. Wang. A model-agnostic hierarchical framework towards trajectory prediction. *Journal of Computer Science and Technology*, 40(2):322–339, 2025. 3

[28] X. Qian, Y. Zhuang, Y. Li, S. Xiao, S. Pu, and J. Xiao. Video relation detection with spatio-temporal graph. In *ACM MM*, pages 84–93, 2019. 2

[29] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, pages 549–565, 2016. 3

[30] X. Shang, D. Di, J. Xiao, Y. Cao, X. Yang, and T.-S. Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287. ACM, 2019. 1, 2, 3

[31] X. Shang, T. Ren, J. Guo, H. Zhang, and T.-S. Chua. Video visual relation detection. In *ACM International Conference on Multimedia*, Mountain View, CA USA, October 2017. 3

[32] X. Sun, T. Ren, Y. Zi, and G. Wu. Video visual relation detection via multi-modal feature fusion. In *ACMMM*, pages 2657–2661, 2019. 2

[33] Q. Team. Qwen2.5-vl, January 2025. 3, 4

[34] Y.-H. H. Tsai, S. Divvala, L.-P. Morency, R. Salakhutdinov, and A. Farhadi. Video relationship reasoning using gated spatio-temporal energy graph. In *CVPR*, pages 10424–10433, 2019. 2

[35] V. Voleti, A. Jolicoeur-Martineau, and C. Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *NeurIPS*, 35:23371–23385, 2022. 1, 3

[36] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *CVPR*, pages 3302–3309, 2014. 3

[37] A. Wang, B. Wu, S. Chen, Z. Chen, H. Guan, W.-N. Lee, L. E. Li, and C. Gan. Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13384–13394, 2024. 1

[38] G. Wang, Z. Li, Q. Chen, and Y. Liu. OED: towards one-stage end-to-end dynamic scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27938–27947, 2024. 2

[39] W. Wang, K. Gao, Y. Luo, T. Jiang, F. Gao, J. Shao, J. Sun, and J. Xiao. Triple correlations-guided label supplementation for unbiased video scene graph generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5153–5163, 2023. 2

[40] W. Wang, Y. Luo, Z. Chen, T. Jiang, Y. Yang, and J. Xiao. Taking a closer look at visual relation: Unbiased video scene graph generation with decoupled label learning. *IEEE Transactions on Multimedia*, 26:5718–5728, 2023. 2

[41] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang, Z. Xie, Y. Wu, K. Hu, J. Wang, Y. Sun, Y. Li, Y. Piao, K. Guan, A. Liu, X. Xie, Y. You, K. Dong, X. Yu, H. Zhang, L. Zhao, Y. Wang, and C. Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. 3

[42] J. Yang, W. Peng, X. Li, Z. Guo, L. Chen, B. Li, Z. Ma, K. Zhou, W. Zhang, C. C. Loy, et al. Panoptic video scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18675–18685, 2023. 1, 2, 3, 4

[43] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *IJCV*, 126(2-4):375–389, 2018. 3

[44] K. Yuan, M. Kattel, J. L. Lavanchy, N. Navab, V. Srivastav, and N. Padoy. Advancing surgical vqa with scene graph knowledge. *International Journal of Computer Assisted Radiology and Surgery*, 19(7):1409–1417, 2024. 1

[45] Y. Zhang, X. Li, H. Xie, W. Zhuang, S. Guo, and Z. Li. Multi-label action anticipation for real-world videos with scene understanding. *IEEE TIP*, 2024. 3