FiTBench:Benchmark for Scene Graph Anticipation with Fine-grained text cues

Supplementary Material

Anonymous Author(s)
Submission Id: 88

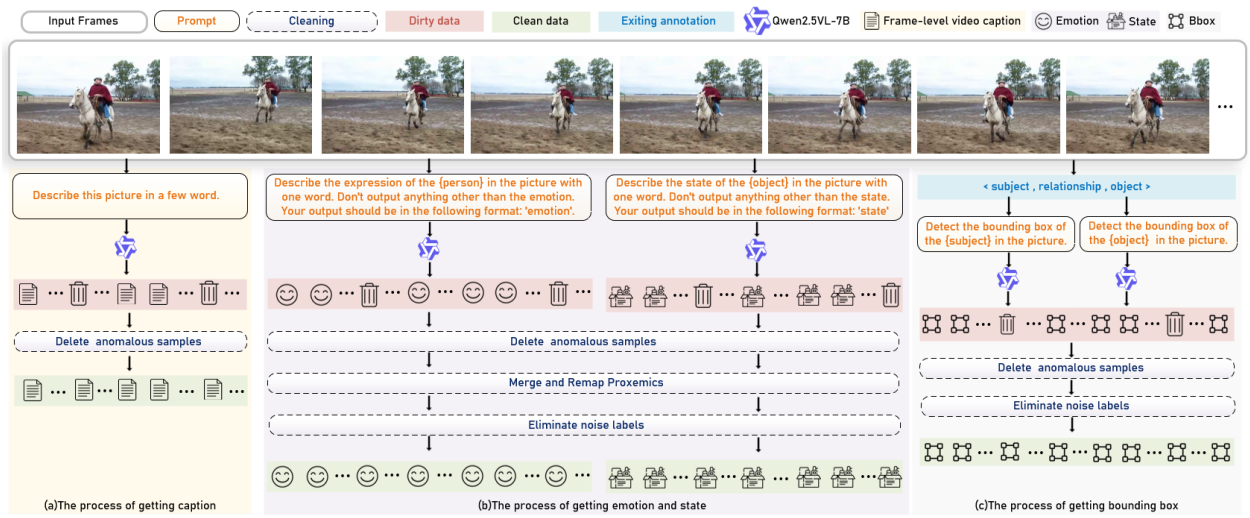

Figure 1: (a) The process of annotating frame-level caption. Relying on Qwen2.5VL-7B's powerful image description capabilities, high-quality video frame subtitles can be obtained after a round of cleaning. (b) We use a prompt-based visual language reasoning method to obtain emotion and state annotations for each frame of images, using two fixed prompts limited to word output. After cleaning up anomalous samples and noise labels, merging and remapping proxemics, we obtain high-quality emotion annotations and state annotations. (c) To ensure bounding box annotations align with scene graph prediction targets, we extract subject-object pairs from existing <subject-relationship-object> triples. These pairs, along with video frames, are input into Qwen2.5VL-7B for accurate localization, and invalid and noisy boxes are manually removed.

1 PIPELINE

In order to capture the dynamic semantic cues in each frame, we built an automatic annotation process based on the large-scale visual language model Qwen2.5VL-7B. The process is shown in Fig 1.

(a) is used to obtain frame-level video caption. Qwen2.5VL-7B has powerful image description capabilities, and we input frame-by-frame video into qwen. Most of the results obtained are compliant video subtitles, and very few placeholders (e.g., '*addCriterion*') are output. We manually removed the abnormal samples. We obtained video captions for 298,384 video frames on a single Nvidia-3090, which took about 90h and resulted in valid video captions for 262,454 video frames.

(b) is used to obtain emotion annotation and state annotation. We take each frame of image input and restrict its output to word semantic labeling with these two prompt: '*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'.*' We manually removed output placeholders and noise labels (e.g., incomplete words such as '*str*'), merged near-synonyms (e.g., '*concentrateded*' to '*concentrated*'), and adjusted the word properties of a small number of words (e.g., '*sleeping*' to '*sleepy*'). For FiT-AG dataset, we obtained emotion annotations for 2983384 video frames on a single Nvidia-3090, which took about 22h, and eventually obtained valid expression annotations for 287585 video frames of subject. It takes about 45h to obtain state annotations, and finally obtain valid state annotations for 264141 video frames. For FiT-PVSG dataset, we obtained 46,236 valid emotion annotations for 69966 exo-centric video frames, and 117,445 valid states for 149,484 video frames.

(c) For obtaining character bounding boxes and object bounding boxes, 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

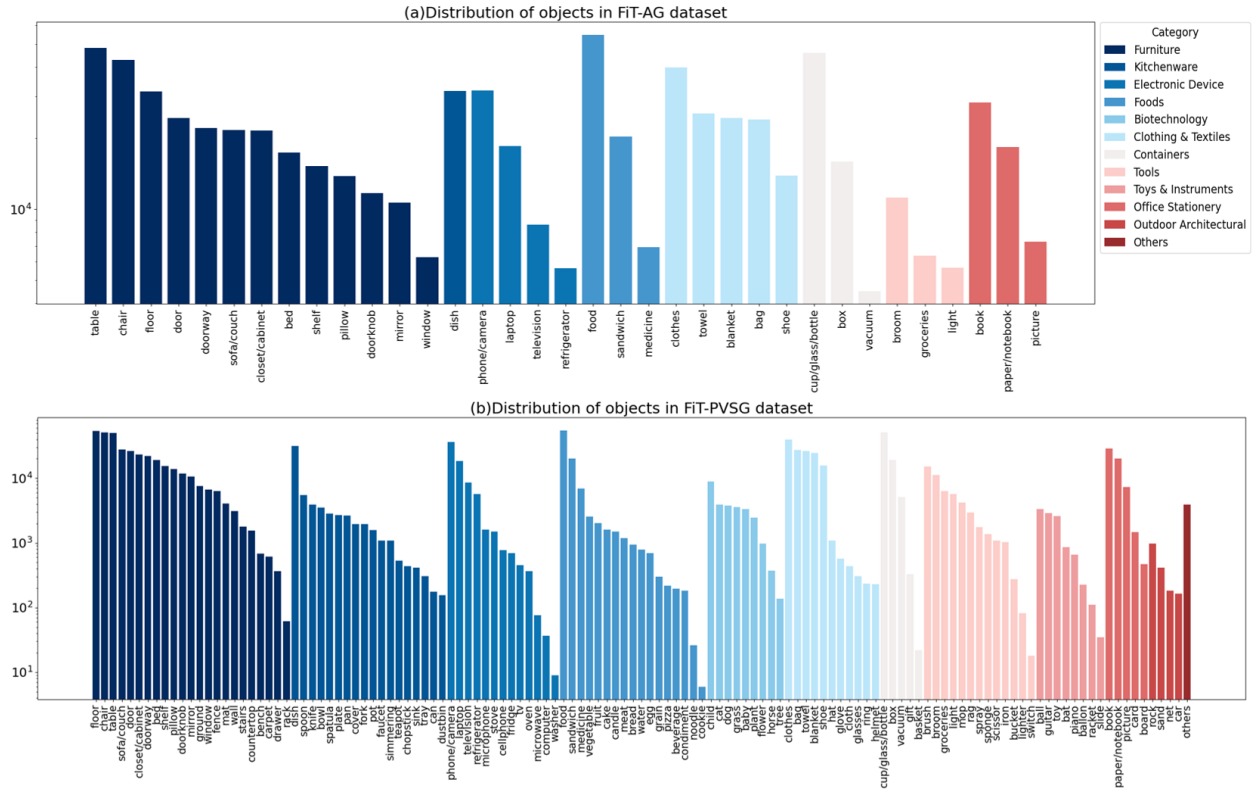

Figure 2: Distribution of objects in (a) FiT-AG dataset and (b) FiT-PVSG dataset. We divide objects into 12 types based on their attributes and usage scenarios. Each type is shown in a different color. The FiT-AG dataset includes 8 object types, while the FiT-PVSG dataset includes 12.

predicted goals of the scene graph. The subject-object pairs of interest are entered into Qwen2.5VL-7B along with the video frames, and Qwen2.5VL-7B can effectively localize the bounding boxes of the objects we are interested in. As for the few invalid placeholders and noise bounding boxes are removed manually. Finally 139655 subject bounding boxes and 216855 object bounding boxes are obtained for the FiT-pvsg dataset.

2 DATASETS STATISTIC

2.1 Object and Relationship Statistic

For FiT-AG dataset, we provide annotations for 234, 253 frames with a total of 476, 229 bounding boxes of 35 object classes (excluding '*person*'), and 1, 715, 568 instances of 25 relationship classes. For FiT-PVSG dataset, we provide annotations for 149, 484 frames with a total of 216, 855 bounding boxes of 128 object classes (excluding '*person*'), and 117, 964 instances of 72 relationship classes. Fig 2 and Fig 3 visualizes the distribution of object and relationship categories in FiT-AG and FiT-PVSG, respectively. As shown in these two figure, compared with the FiT-AG dataset, the FiT-VidOR dataset has more object and relationship categories, increasing the data diversity. Some objects (e.g., '*floor*' and '*food*') and relations (e.g., '*holding*' and '*in front of*') occur frequently, while others (e.g., '*cookie*' and '*squeezing*') occur only a few times. However, even with such a distribution, for FiT-AG dataset, almost all objects have at least 10K instances and every relationship as at least 1K instances. For FiT-PVSG dataset, almost all objects have at least 1.5K instances and every relationship as at least 1.5K instances.

2.2 Emotion Annotation Statistic

We annotated 53 emotion categories in the FiT-AG dataset and 54 in the FiT-PVSG dataset. Emotions often reflect intentions or behaviors. For example, '*focus*' might suggest that a person is using or focusing on an object, while '*surprise*' might suggest sudden attention. Emotions, physical states, and how objects are used can also affect relationships. Positive emotions may show liking or use of an object, while negative emotions may suggest dislike or avoidance. '*Tired*' may change how a character uses an object, and '*pain*' might require specific ones (e.g., '*medicine*'). We also handled unclear cases carefully to avoid errors while keeping the dataset reliable. Based on these considerations, we grouped the emotion labels into seven categories:

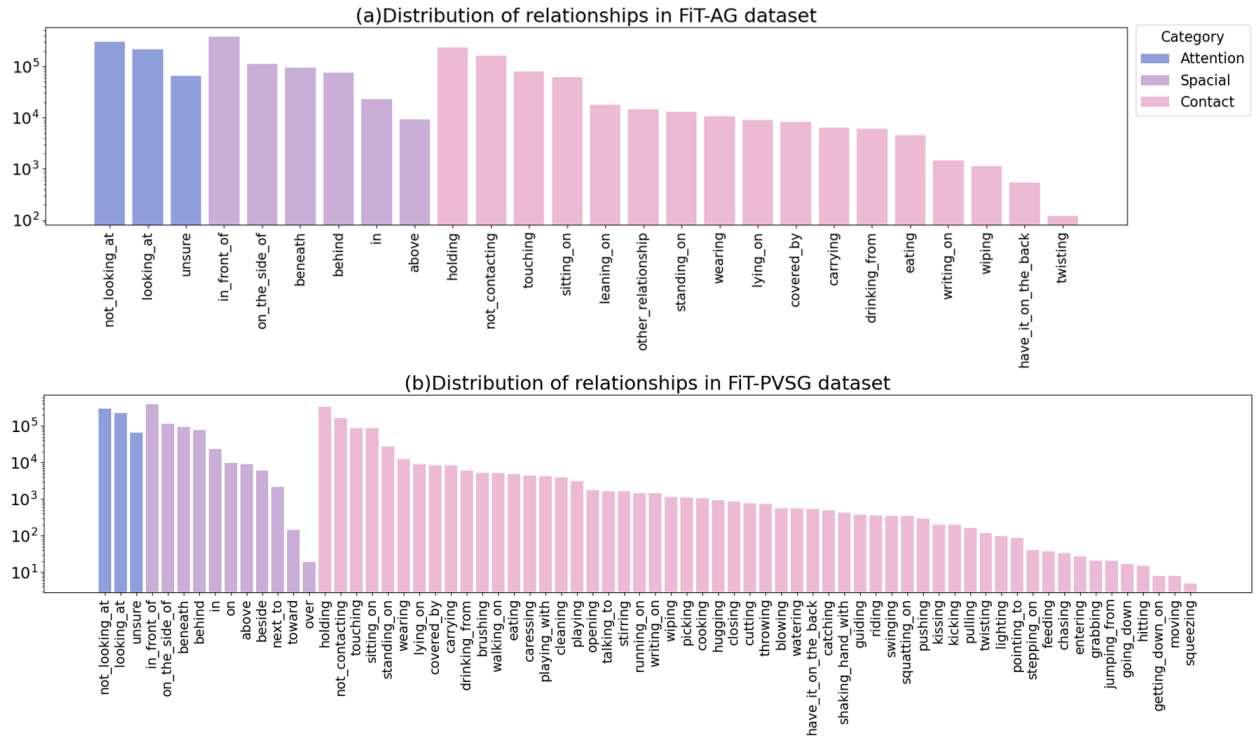

Figure 3: Distribution of relationships in (a) FiT-AG dataset and (b) FiT-PVSG dataset. The FiT-pvsg dataset contains a wider range of relationships than the FiT-AG dataset.

- Neutral Emotions: An objective state without significant emotional tendencies, reflected in a baseline mental orientation that is calm and without significant emotional fluctuations.
- Positive Emotions: Reflects positive emotional experiences, including psychologically motivating emotional expressions such as joy and fulfillment.
- Concentration & Thinking: Describes a state of cognitive resource allocation in which attention is highly focused on a specific goal or task.
- Surprise & Confusion: Labels transient emotional responses triggered by sudden stimulation, covering a spectrum of intensity from curiosity to alertness.
- Negative Emotions: Characterizes a collection of negative emotions, including sadness, anger, and other emotional patterns that have a tendency to be psychologically inhibited.
- Physiological Emotions: Perceived states of direct feedback from the body such as fatigue, pain, and other manifestations of the biological functions of the associated organism.
- Miscellaneous: Labeling uncertainty in mood determination due to incomplete information or ambiguity in expression.

Fig 4(a) visualizes the log-distribution of emotion categories in FiT-AG dataset. Fig 4(b) visualizes the log-distribution of emotion categories in FiT-PVSG dataset. People may show common emotion such as '*happy*' or '*neutral*' more frequently, while expressions that require more in-depth analysis, such as '*engaged*', may be less frequently captured or annotated.

2.3 State Annotation Statistic

We labeled the FiT-AG dataset and the FiT-PVSG dataset with 115 and 125 categories of state labels, respectively. We believe that physical attributes determine interactions: basic attributes directly affect how a character uses an object. Changes in state trigger behavioral decisions: shifts in integrity or cleanliness may lead to repair, cleaning, or replacement behaviors. The food state or container state contains domain-specific knowledge that helps the model to reason accurately in a specific scenario (e.g., '*kitchen*'). Spatial location also affects the logic of character actions. We also reserve uncertainty for fuzzy states and define unambiguous states to avoid the model from making wrong assumptions. Therefore, we distinguish status labels by the following 9 categories:

- Physical state: describes the current condition, position, or form of an object, encompassing attributes.
- Functional state: is used to describe the real-time operational state of a system, device, or object, covering whether it is in a specific functional mode.
- Cleanliness & Maintenance: used to assess the state of hygiene and maintenance of a place or object.

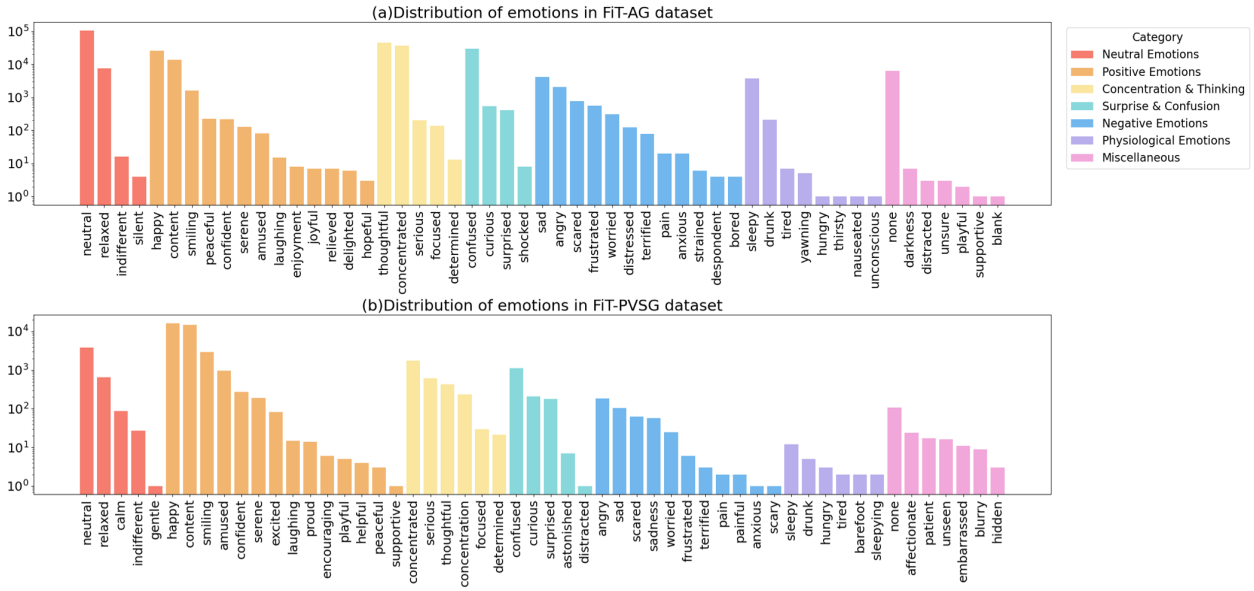

Figure 4: Distribution of emotions in (a) FiT-AG dataset and (b) FiT-PVSG dataset. Different types of emotions are marked with different colors. The different types of emotions are distributed evenly, with positive emotions being the most common and neutral emotions less frequent. However, emotions within the same category are not evenly distributed. For example, in both datasets, happy is the most frequent positive emotion. This may be because most of the video data comes from everyday life scenes.

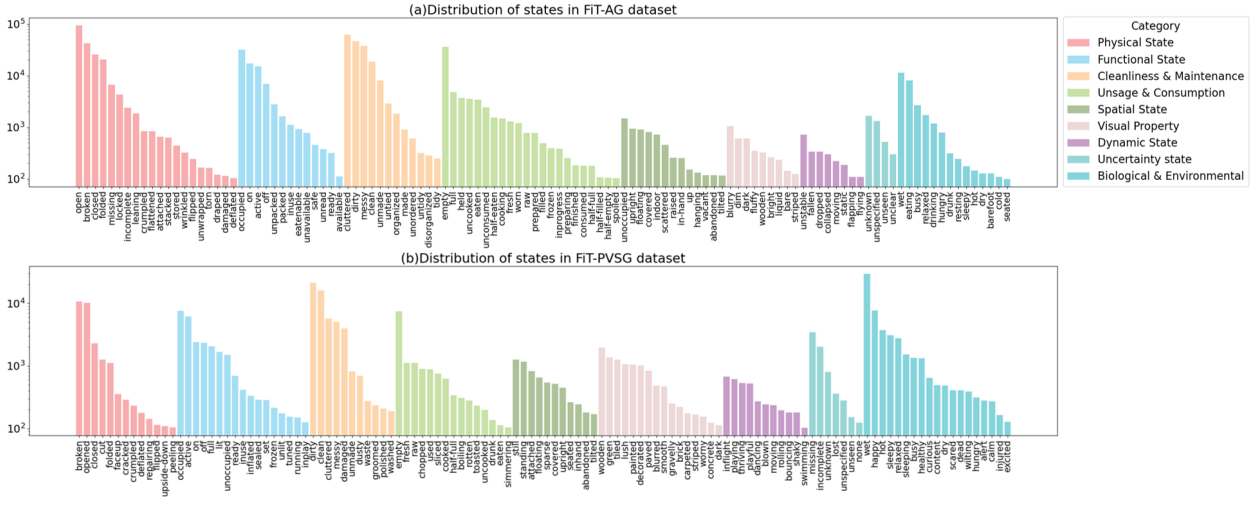

Figure 5: Distribution of states in (a) FiT-AG dataset and (b) FiT-PVSG dataset. Different types of states are marked with different colors. The types of states are evenly distributed overall. In the FiT-AG dataset, states related to usage and consumption are more common, while in the FiT-FiT dataset, states related to biological and environmental aspects appear most frequently. This may be because FiT-AG mainly contains indoor scenes, whereas FiT-PVSG includes more complex scenes.

- Unsage & Consumption: describes the various states of items in relation to their preparation, condition, or depletion.
- Spatial State: describes the position, direction, motion state and interaction with the environment of an object in three-dimensional space.

Table 1: Compare baseline to the method with the addition of the TAVS module (baseline *w* /TAVS) on the FiT-AG dataset and FiT-PVSG dataset at $\mathcal{F}$ = 0.5

Method	FiT-AG												FiT-PVSG											
	Recall						Mean Recall						Recall						Mean Recall					
	With constrain			No constrain			With constrain			No constrain			With constrain			No constrain			With constrain			No constrain		
	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50
$\mathcal{F}$=0.5																								
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	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	**30.3**	**33.6**	**33.6**	**43.1**	**61.1**	**73.0**	**12.6**	**14.8**	**14.8**	**19.1**	**30.8**	**58.4**	**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+	**35.0**	37.1	37.1	34.4	53.2	70.8	8.0	8.7	8.8	10.5	21.4	48.9	13.6	14	14.1	30.7	**42.4**	**47.6**	1.1	1.1	1.1	2.2	**9.6**	**18.9**
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**	**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++	27.8	30.1	30.1	38.2	55.3	71.8	10.3	11.5	11.5	14.8	27.5	52.2	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	**38.8**	**42.5**	**42.6**	**49.4**	**65.1**	**73.3**	**15.7**	**18.5**	**18.5**	**22.7**	**36.0**	**61.8**	**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++	28.1	30.5	30.6	37.4	55.8	72.0	10.7	12.0	12.0	15.3	28.6	52.6	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	**38.3**	**41.3**	**41.3**	**47.4**	**63.5**	**73.0**	**14.7**	**16.8**	**16.9**	**19.9**	**34.2**	**61.0**	**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	43.1	45.8	45.8	49.6	64.1	72.8	**20**	**21.9**	**21.9**	**26.0**	**39.7**	**62.9**	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	**43.3**	**46.9**	**46.9**	**51.4**	**65.6**	**73.2**	17.9	21.3	21.5	25.2	38.3	61.7	**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	43.8	46.6	46.6	51.4	65.7	73.4	18.9	21.1	21.2	24.4	40.0	63.8	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	**43.9**	**47.5**	**47.6**	**52.1**	**66.5**	**73.5**	**20.2**	**24.1**	**24.3**	**28.0**	**43.2**	**65.2**	**35.0**	**35.4**	**35.5**	**29.9**	**38.8**	**44**	**1.6**	1.6	1.6	**4.4**	**7.0**	12.5
$\mathcal{F}$=0.7																								
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	**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	**35.1**	**38.1**	**38.1**	**51.3**	**70.9**	**82.9**	**16.7**	**16.7**	**16.7**	**22.6**	**36.9**	**69.0**	**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+	40.0	41.8	41.8	41.0	62.1	80.5	9.1	9.8	9.8	12.6	26.1	57.8	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	**40.1**	**42.9**	**42.9**	**51.9**	**70.3**	**82.7**	**13.9**	**15.7**	**15.7**	**20.3**	**34.5**	**67.1**	**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++	32.0	34.1	34.1	45.8	64.9	81.8	12.4	13.7	13.7	18.2	34.1	62.0	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	**44.9**	**48.1**	**48.1**	**58.0**	**75.2**	**83.2**	**18.5**	**21.1**	**21.1**	**27.4**	**43.2**	**72.3**	**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++	34.6	37.0	37.0	46.0	66.0	81.9	13.4	14.9	14.9	19.4	36.3	62.4	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	**45.3**	**48.3**	**48.3**	**56.9**	**73.9**	**83.2**	**18.3**	**20.8**	**20.8**	**25.5**	**41.4**	**72.6**	**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	50.9	53.3	53.3	59.6	75.0	83.0	**23.3**	25.0	25.0	**31.9**	**48.7**	**73.0**	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	**51.0**	**54.3**	**54.3**	**61.2**	**76.1**	**83.2**	21.5	**25.1**	**25.1**	29.6	46.1	71.6	**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	51.2	53.7	53.7	61.0	75.8	83.3	22.4	24.5	24.5	29.7	46.1	74.0	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	**51.5**	**54.7**	**54.8**	**61.9**	**77.0**	**83.5**	**23.5**	**27.6**	**27.6**	**34.1**	**51.6**	**76.6**	**38.8**	**39.6**	**39.7**	**33.8**	**42.0**	**48.8**	**1.8**	1.9	1.9	**2.4**	**5.4**	15.6
$\mathcal{F}$=0.9																								
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	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	**40.1**	**42.1**	**42.1**	**63.4**	**83.4**	**93.1**	**16.8**	**18.4**	**18.4**	**28.3**	**46.5**	**78.2**	**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+	44.7	45.9	45.9	50.9	74.7	91.0	10.3	10.8	10.8	16.3	22.7	71.5	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	**46.4**	**48.7**	**48.7**	**64.9**	**83.5**	**93.0**	**17.0**	**18.5**	**18.5**	**27.9**	**46.9**	**78.4**	**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++	38.5	40.1	40.1	58.3	78.9	92.1	17.7	19.1	19.1	26.5	47.1	75.6	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	**52.3**	**54.6**	**54.6**	**70.9**	**87.0**	**93.4**	**23.4**	**25.4**	**25.4**	**36.7**	**54.9**	**82.6**	**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++	41.6	43.5	43.5	58.2	79.6	92.1	17.3	18.7	18.7	27.6	48.5	76.2	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	**53.3**	**55.5**	**55.5**	**70.6**	**86.7**	**93.4**	**23.9**	**26.1**	**26.1**	**35.8**	**54.2**	**82.4**	**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	**59.7**	61.3	61.3	73.2	87.2	93.3	**27.9**	28.9	28.9	**39.5**	**57.5**	83.4	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	59.6	**61.9**	**61.9**	**74.0**	**87.7**	**93.4**	26.8	**29.4**	**29.4**	38.6	56.5	**86.6**	**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	58.5	60.4	60.4	73.6	87.4	93.5	26.4	27.7	27.7	37.6	56.1	87.1	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	**60.5**	**62.8**	**62.8**	**74.7**	**88.4**	**93.6**	**28.7**	**31.8**	**31.8**	**42.2**	**60.4**	**89.0**	**38.9**	**39.1**	**39.1**	**39.8**	**44.0**	**50.5**	**1.9**	**1.9**	**1.9**	**3.3**	**6.8**	21.0

- Visual Property: describes the visual characteristics of an object or scene, covering a wide range of properties such as material, color, surface condition, environmental features, and visual effects, which are used to accurately convey its appearance.
- Dynamic State: describes an object or system in a constantly changing, active state.
- Biological & Environmental: dynamically interact to influence states ranging from physiological needs and emotional responses to environmental adaptations.
- Uncertainty state is a special state characterization used to describe the situation when there is missing or incomplete information in the system, reflecting the uncertain knowledge and incomplete mastery of the current situation or data reliability.

Fig 5(a) visualizes the log-distribution of state categories in FiT-AG dataset. Fig 5(b) visualizes the log-distribution of state categories in FiT-PVSG dataset. Some states (e.g., '*open*', '*dirty*') occur frequently, while others (e.g., '*boiled*') occur only a few times. This is because clean or open environments may be more prevalent, while broken or dead are relatively rare. These high-frequency categories are more common in the real world and are more likely to be captured during the data collection process.

3 EXPERIMENTS RESULTS

$\mathcal{F}$ denotes the overall proportion of initial video segments input to the model, which serves to assess the model's predictive ability over different time spans by controlling the length of video content observed by the model. Specifically, the value of $\mathcal{F}$ (e.g., 0.3, 0.5, 0.7, 0.9) represents the fact that

Figure 6: Qualitative Results. To the left, we show a sampled subset of the frames observed by the models. The second column provides future predicted frames, and the third column provides a ground truth scene graph corresponding to a future frame. In the subsequent columns, we compare the performance of the baseline and the SceneSayer *w* /TAVS model. We report performance for both short-term and long-term future predictions. In each graph above, correct anticipations of relationships are denoted with text in black and incorrect anticipation of the relationships are highlighted with text in red.

Table 2: Average improvement rate (%) on FiT-AG and FiT-PVSG datasets after incorporating TAVS module under different initial fraction of the video $\mathcal{F}$.

	$\mathcal{F}=0.3$		$\mathcal{F}=0.5$		$\mathcal{F}=0.7$		$\mathcal{F}=0.9$	
	FiT-AG	FiT-PVSG	FiT-AG	FiT-PVSG	FiT-AG	FiT-PVSG	FiT-AG	FiT-PVSG
STTran+	23.7	3.4	20.7	-0.4	17.6	-1.4	6.7	-1.4
DSGDetr+	19.2	49.0	22.8	46.1	27.6	41.6	37.7	61.4
STTran++	34.2	3.7	37.4	5.2	34.8	5.2	25.3	2.1
DSGDetr++	26.3	45.9	27.7	45.6	25.6	44.0	23.5	51.0
SceneSayerODE	0.8	10.4	-1.0	9.3	-1.1	4.4	0.2	4.3
SceneSayerSDE	5.3	48.3	5.7	34.5	5.7	19.5	6.2	11.1
Average	18.2	26.8	18.9	23.4	18.4	14.3	16.6	21.4

the model can access only the first 30%, 50%, 70%, or 90% of the video as input and needs to predict the subsequent unobserved portions based on this. By adjusting the size of $\mathcal{F}$, the researcher is able to analyze the model's ability to capture and reason about short-term relationships (e.g., when $\mathcal{F}$ is large, the model relies on longer known segments for proximity prediction) and long-term relationships (e.g., when $\mathcal{F}$ is small, the model needs to infer potential correlations farther into the future from limited information), thus comprehensively validating the model's adaptability and generalization in different time-series scenarios. Tables 1 show the results for $\mathcal{F}$=0.5, $\mathcal{F}$=0.7, and $\mathcal{F}$=0.9, respectively. Table 2 compares the average improvement rates on the FiT-AG and FiT-PVSG datasets after incorporating the TAVS module into six baselines.

To better understand the role of dynamic semantic semantic annotations such as subject emotion, object state, and caption, we evaluate multiple scene graph anticipation baselines on our proposed FiTBench, which consists of two augmented datasets: FiT-AG and FiT-PVSG. We added the Text-Augmented Visual Semantic (TAVS) module to the six baselines and measured the relative performance changes over the two datasets. The results reveal different trends across models.

In the FiT-AG dataset, the model basically shows a boosting trend. At $\mathcal{F}$=0.9, the DSGdetr+w /TAVS method boosts substantially. It is possible that when visual features extracted by ORPU (Object Representation Processing Unit) and SCPU (Spatial Context Processing Unit) are aligned with textual descriptions at the semantic level, the textual module can improve the performance by enhancing the object localization and relational reasoning. In particular, when the model is equipped with a dual-decoding approach (e.g., STTran++w /TAVS and DSGDetr++ w /TAVS.), significant improvements of 37.36% and 45.64% are obtained. On the SceneSayerODEw /TAVS method there is also a boost (e.g., at $\mathcal{F}$=0.9, the average boost is 0.24%), but it is not as effective as the other methods.

In the FiT-PVSG dataset, most of the models performed well, with the DSGdetr+w /TAVS method improving by 51.41% at $\mathcal{F}$=0.9. The rich semantic information embedded in textual features is universally valuable for predicting future relationships, as confirmed by the significant enhancements in the DSGDetr+ and DSGDetr++. In particular, the additional temporal encoder and double decoding loss introduced by DSGDetr++ provide a powerful architectural and optimization foundation for effectively fusing and exploiting text features. The small decrease in STTran+ highlights the fact that multimodal fusion is not a simple splicing of features, and that its effectiveness is significantly influenced by the model architecture and training objectives.

4 QUALITATIVE EXPERIMENT

We show some of the results of the SceneSayerSDE method and the SceneSayerSDE w /TAVS method on long-term future prediction and short-term future prediction in Fig 6. Short-term predictions usually rely on immediate contextual information, such as the current action or the position of an object. Textual features such as emotion, state, or caption may provide additional semantic information to help the model more accurately capture the relationships in the current scene. For example, a text describing 'handing a cup' can assist the model in recognizing the intent of the action and thus inferring the next action in short-term prediction, e.g., 'picking up the cup'. Long-term prediction requires an understanding of more complex temporal relationships and underlying intentions. Text features may contain higher-level information about the whole scene or character's goals. In this case, text features provide long-term structural cues for the model to predict events further in the future. In this case, text features provide long-term structured cues for the model to predict events further into the future.