FIFA: Unified Faithfulness Evaluation Framework for Text-to-Video and Video-to-Text Generation

Liqiang Jing[1], **Viet Lai**[2], **Seunghyun Yoon**[2], **Trung Bui** [2], **Xinya Du**[1]
[1]University of Texas at Dallas, [2]Adobe Research
jingliqiang6@gmail.com, xinya.du@utdallas.edu

Abstract

Video Multimodal Large Language Models (VideoMLLMs) have achieved remarkable progress in both Video-to-Text and Text-to-Video tasks. However, they often suffer from hallucinations, generating content that contradicts the visual input. Existing evaluation methods are limited to one task (*e.g.,* V2T) and also fail to assess hallucinations in open-ended, free-form responses. To address this gap, we propose FIFA, a unified **F**a**I**th**F**ulness ev**A**luation framework that extracts comprehensive descriptive facts, models their semantic dependencies via a Spatio-Temporal Semantic Dependency Graph, and verifies them using VideoQA models. We further introduce Post-Correction, a tool-based correction framework that revises hallucinated content. Extensive experiments demonstrate that FIFA aligns more closely with human judgment than existing evaluation methods, and that Post-Correction effectively improves factual consistency in both text and video generation. Our code is available at `https://github.com/du-nlp-lab/FIFA`.

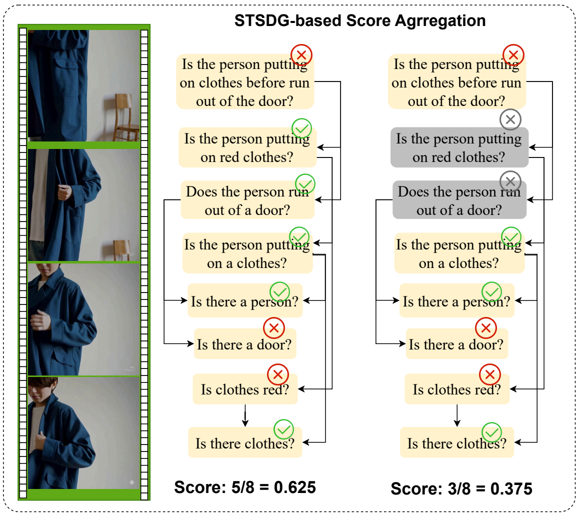

Figure 1: Illustration of Spatio-Temporal Semantic Dependency Graph-based Score Aggregation.

1 Introduction

Video Multimodal Large Language Models (VideoMLLMs) (Maaz et al., 2024; Zhang et al., 2023) have demonstrated impressive performance across a wide range of video tasks, such as Video-to-Text (V2T) (Yan et al., 2021) and Text-to-Video (T2V) (Brooks et al., 2024). Although VideoMLLMs have demonstrated remarkable performance, they are often susceptible to hallucinations, *i.e.,* the fabricated or inaccurate content generation (Wang et al., 2024; Jing and Du, 2025; Chang et al., 2024; Jing et al., 2025b,a). Such hallucinations pose serious risks, potentially leading to misinformation and safety concerns, and ultimately undermining the reliability of these models in real-world applications. Despite the criticality of this issue, limited research has focused specifically on hallucination in VideoMLLMs (Li et al., 2024a). Existing studies mainly leveraged existing Video Question Answering (VideoQA) datasets or constructed specialized datasets for hallucination evaluation in VideoMLLMs (Wang et al., 2024; Li et al., 2024a).

Although some work has been performed on hallucination evaluation in VideoMLLMs, existing efforts are relatively isolated and face notable limitations. **First**, most approaches are restricted to simplified evaluation settings, such as binary-labeled VideoQA (Wang et al., 2024). As a result, they fail to address hallucinations in complex, free-form, and long-form responses to open-ended questions, scenarios that more accurately reflect real-world usage. **Second**, current research predominantly targets the V2T models (Li et al., 2024a; Wang et al., 2024), while overlooking T2V generation. Despite their significance to the development of general artificial intelligence, hallucinations in video generation tasks remain largely unexplored.

To develop a unified evaluation framework for

Findings of the Association for Computational Linguistics: ACL 2026, pages 11418–11456
July 2-7, 2026

both T2V and V2T tasks involving free-form questions, we resort to decomposition-based evaluation methods, which first break down a response into smaller atomic information units (*i.e.,* atomic facts) and then verify each unit individually. However, designing such a framework for VideoMLLMs is non-trivial due to the following three challenges:

- **Full Semantic Coverage**: On the one hand, the existing work focuses on static scenes (Jing et al., 2024b; Hu et al., 2023), overlooking the hallucination in video dynamic scenes, such as temporal hallucination. On the other hand, they typically rely on atomic units, which may fail to capture the full meaning, potentially overlooking hallucinations during video-related tasks. For example, for the V2T task, consider a video where *"There are two people, one is wearing red clothes and the other is wearing a blue hat.",* The predicted video description is *"There are two people; one is wearing red clothes and a blue hat."* When decomposed into atomic information units such as *"two people"*, *"red clothes"*, *"blue hat"*, *"one person wears red clothes"*, and *"one person wears a blue hat"*, each individual unit might appear faithful compared to video content. However, the predicted video description contains a hallucination: it incorrectly attributes both the red clothes and the blue hat to the same person. This inter-fact contradiction is missed in the evaluation process. There are also similar situations in the T2V task.

- **Dependency Between Information Units**: In the video-related task, the correctness of facts derived from text tends to depend on the others. For instance, the statement *"The dog is white"* presumes that *"There is a dog"* is true. If the model hallucinates the existence of a dog, then any attributes ascribed to it, such as colorare also hallucinated by implication. Without explicitly modeling these dependencies, the evaluation may produce inconsistencies. For example, if *"There is a dog"* is (correctly) identified as hallucinated, yet *"The dog is white"* is (incorrectly) judged as faithful, the evaluation fails to capture the inherent dependency between the two facts.

- **Complexity of Responses**: Unlike closed-domain tasks such as binary VideoQA (Li et al., 2024a), answering open-ended video-related questions often requires not only describing visual content but also providing analytical reasoning that incorporates external commonsense knowledge. These subjective or abstract elements go beyond direct observation and can confound factuality judgments if not properly separated from descriptive content. Failing to distinguish between analytical and descriptive content inevitably distracts the factual measurement.

To tackle the above challenges, we propose FIFA, a unified faithfulness metric for T2V and V2T with a Spatio-Temporal Semantic Dependency Graph. FIFA first extracts a comprehensive set of facts from the generated text or text instruction, including both atomic facts (including temporal hallucinations) and event-level facts (a kind of composite fact including all associated atomic facts/information of core objects in an event) that better captures the full semantics of the text. We instruct LLMs to extract only descriptive facts to avoid evaluation bias caused by subjective or analytical content. Subsequently, we construct a Directed Acyclic Graph (DAG), the Spatio-Temporal Semantic Dependency Graph (STSDG), by linking fact pairs that exhibit semantic dependency relationships. Next, we transform the extracted facts into questions and utilize state-of-the-art VideoQA models to answer them based on the given video content. Finally, we aggregate the verification results of all questions using the constructed STSDG to derive the overall faithfulness score. These dependencies ensure the consistency that if the answer to a prerequisite question is negative, all downstream questions that depend on it are skipped during evaluation, thus preventing invalid fact verification and ensuring reliable scoring.

To evaluate FIFA, we conduct human annotation to assess hallucinations in both T2V and V2T tasks. We then compute the correlations between human judgments and various baseline methods. FIFA yields the highest correlation with human evaluations compared to existing metrics across T2V and V2T tasks. To further validate the key components of FIFA, we construct several dedicated evaluation sets targeting different stages of the pipeline, including Fact Extraction, Fact-to-Question Generation, VideoQA, and Dependency Generation. In addition, we introduce a unified correction framework, Post-Correction, which could utilize our Post-Correction intermediate evaluation results to mitigate hallucinations in both generated video and text outputs. Extensive experiments confirm the effectiveness of our full pipeline in enhancing the factuality and reliability of generated content.

Our contributions are summarized as: 1) To the best of our knowledge, we are the first to propose a unified evaluation metric that jointly addresses both Video-to-Text and Text-to-Video generation. 2) We construct a STSDG to explicitly model dependencies between a comprehensive set of facts, thereby enhancing the robustness and reliability of the evaluation process. 3) We are the first to develop a unified correction framework, Post-Correction, which can identify hallucinated content and revise it to improve the factual consistency of both generated text and video. 4) We conduct comprehensive experiments, and the results demonstrate the effectiveness of both our proposed FIFA metric and the hallucination mitigation strategy. 5) We created a human-annotated dataset (the test set used in experiments (Tables 1 and 4)) that could facilitate future research on video-based multimodal faithfulness evaluation.

2 Related Work

Video-to-Text Generation. Video-ChatGPT (Maaz et al., 2024) applies spatial-temporal pooling to extract relevant video features, while Video-LLaMA (Zhang et al., 2023) introduces a Video Q-Former to summarize frame-level information. Vista-LLaMA (Ma et al., 2024) enhances the alignment between visual and language modalities by maintaining equal attention distances and further proposes a temporal Q-Former for temporal reasoning. LLaMA-VID (Li et al., 2024b), on the other hand, adopts a dual-token design, assigning each frame both a context and a content token, which aids in modeling long-range temporal dependencies. Despite their promising results on several benchmarks, these models still exhibit hallucinations (Wang et al., 2024).

Text-to-Video Generation. Early studies such as TGANs-C (Pan et al., 2017) and VQ-VAE (van den Oord et al., 2017) generate short videos with some temporal coherence. Diffusion-based model, e.g., VDM (Ho et al., 2022), MagicVideo (Zhou et al., 2022), PixelDance (Zeng et al., 2024), and VideoCrafter2 (Chen et al., 2024a), leverage latent diffusion and temporal attention to generate high-fidelity videos with improved temporal consistency. In parallel, autoregressive transformers (Vaswani et al., 2023), such as NUWA (Wu et al., 2022a), Phenaki (Villegas et al., 2023), and VideoGPT (Yan et al., 2021), model video sequences as discrete latent tokens, allowing better handling of temporal structure and long-context reasoning. While these methods have greatly improved video generation quality, they often produce hallucinated content, including objects, attributes, or actions that do not faithfully reflect the input prompt. This faithfulness issue presents a serious challenge for practical applications where semantic consistency and factual grounding are essential (Wu et al., 2024; Zheng et al., 2025).

MLLM Hallucination. Despite of the advances of large language models (LLMs) (Jing et al., 2024a, 2023a; Zhang et al., 2025c) and MLLMs (Wang et al., 2025; Zhang et al., 2025b; Jing et al., 2023b; Zhang et al., 2026, 2025d), hallucination is a persistent issue in LLMs (Huang et al., 2023) and MLLM (Zhang et al., 2023). Early studies primarily focus on hallucinations in image-related tasks (Jing et al., 2024b; Hu et al., 2023; Cho et al., 2024; Liu et al., 2024a; Luo et al., 2025; Yan et al., 2024; Zhang et al., 2024b). For example, Woodpecker (Yin et al., 2024) refines generated responses using additional visual evidence. Similarly, Volcano (Lee et al., 2024) employs a self-refinement pipeline comprising critique, revision, and decision phases to enhance the factual accuracy of model outputs. Recently, the research community has investigated hallucination evaluation for video-related tasks (Zheng et al., 2025; Ullah and Mohanta, 2022; Zhang et al., 2024a; Rawte et al., 2024; Li et al., 2024a; Wang et al., 2024). Different from them, we propose a unified reference-free faithfulness evaluation framework with a spatio-temporal semantic dependency graph for both V2T and T2V. We also propose a Post-Correction method to mitigate the hallucination in the generated video and text.

3 Unified Fine-grained Faithfulness Evaluation Framework

This section presents a unified fine-grained faithfulness evaluation metric with STSDG (See Figure 2). This metric evaluates the fine-grained hallucination in T2V and V2T models. Specifically, our FIFA consists of three components: STSDG-based Question Generation, Fact Verification, and STSDG-based Score Aggregation (See Figure 2).

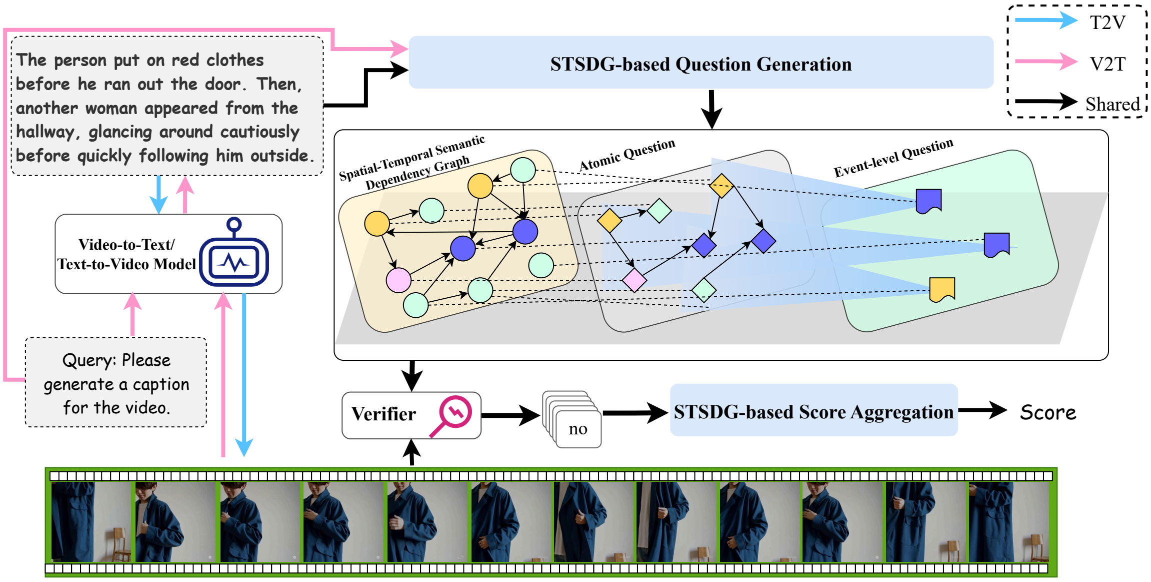

Figure 2: Illustration of our proposed FIFA metric. The blue arrows represent the information flow for T2V, the pink arrows represent the flow for V2T, and the black arrows are shared information pathways of both tasks.

3.1 Unified Faithfulness Evaluation Problem Formulation

Tasks. Firstly, we formulated the text-to-video task and the video-to-text task. 1) **Video-to-Text.** Given an input video V_t and a corresponding query Q, the video-to-text task aims to generate a response T_t from a large-scale video-language model $\mathcal{M}_t$ as follows: $\mathcal{M}_t(V_t, T_p) \rightarrow T_t$. 2) **Text-to-Video.** Given an input text T_v, the text-to-video task aims to generate a video V_v from a large-scale video generation model $\mathcal{M}_v$ as follows: $\mathcal{M}_v(T_v) \rightarrow V_v$.

Unified Evaluation Metric. Our goal is to develop a novel unified faithfulness metric, which necessitates the check of each video-text pair $a = (V, T)$, wherein V denotes either the visual input provided to a large video-language model, or the visual output synthesized by a large video generation model. Formally, the faithfulness score is defined as follows, $f = \mathcal{F}(V, T, Q)$, where f is a scalar ranging from 0.0 to 1.0–higher values indicate greater faithfulness and fewer hallucinations in the model output. $\mathcal{F}(\cdot)$ is the faithfulness estimation, which takes video, text, and input query (V, T, Q) or video and text (V, T) as inputs. $Q = \phi$ for the text-to-video task. Importantly, we make the proposed evaluation approach reference-free, meaning it does not rely on ground-truth annotations or human-written answers, making it broadly applicable across diverse video-based tasks.

3.2 STSDG-based Question Generation

We introduce how we generate various questions and the STSDG, as shown in Figure 4.

3.2.1 Extensive Semantic Fact Extraction

To enable fine-grained faithfulness evaluation, we introduce an extensive semantic fact extraction module that segments the response into atomic factual units. Inspired by prior works (Min et al., 2023), we define an atomic fact as the smallest indivisible unit of meaning. Furthermore, in the context of T2V and V2T, we categorize atomic facts as entities, attributes, relations, or scenes. This granularity ensures that each piece of information can be individually assessed for accuracy without interference from unrelated content. Specifically: **Entity** facts express the presence or absence of specific objects, including a whole entity or part of an entity(*e.g.,* door, man, and tree). **Attribute** facts refer to object characteristics, including type, material, count, color, shape, texture, and size (*e.g.,* wooden door and red chair). **Relation** facts describe interactions or spatial-temporal relationships between entities, including spatial relation, action, and temporal relation (*e.g.,* the man picks up the book). **Scene** facts reflect global properties of the scene, such as lighting condition (e.g., bright lighting), overall composition, or atmosphere (*e.g.,* the atmosphere looks happy).

To evaluate the hallucination precisely, another principle is full semantic coverage: all contents of possible hallucination for the prompt, and only the contents of the prompt, should be represented by the generated questions. However, only these kinds of hallucinations are sometimes not enough to demonstrate real faithfulness for text-video pairs. Just as we mentioned in the example in the introduction. Therefore, we additionally introduce another type of fact: the event-level fact.

Event-level facts are composite facts capturing high-level semantics that cannot be expressed by a single atomic fact alone. An event-level fact involves multiple core objects (typically an action or relation). Then, all associated semantic information about these core objects, such as their attributes, states, locations, or other relations, is integrated into a single holistic fact. This abstraction allows for disambiguation and full interpretation of complex visual events, which would otherwise be underspecified using only atomic facts. Start with an atomic fact (e.g., *a person runs out of a door*), and enrich it by aggregating other atomic facts associated with each object in that atomic fact (e.g., *"The person looks sad"* and *"The door is green"*) into one comprehensive fact (e.g., *a sad person runs out of a green door*). These facts are designed to cover the full meaning of a text, especially when multiple entities, relations, or temporal logic are involved, hence covering semantics when atomic-level representations fall short.

We leverage an LLM to extract all facts from the descriptive text (Min et al., 2023). Meanwhile, we explicitly instruct the LLMs to exclude any facts that involve non-descriptive content during generation. We construct a few-shot prompt by annotating a set of K_1 demonstration examples, and use them to guide the LLM in decomposing descriptive sentences into fine-grained facts. Formally, given the text T for the T2V task/the text and query (T, Q) for the V2T task, we obtain fact groups via $G = \text{LLM}(P_{t2v}, T)/\text{LLM}(P_{v2t}, T, Q)$, where $G = \{g^1, \cdots, g^n\}$ denotes the set of n generated facts. P_{t2v} and P_{v2t} are in-context instruction of fact extraction for the T2V task and V2T task, respectively (See Appendix H for prompt template).

3.2.2 STSDG Construction

To verify the faithfulness of all facts, we further convert them into a yes-or-no question in natural language format with LLM as $\{q_1, \cdots, q_n\} = LLM(P_q, G, T, Q)$, where $Q = \phi$ for the text-to-video task. P_q is the prompt and is shown in Appendix H. q_i is the generated question for the i-th fact. As we mentioned before, there are semantic relationships between different facts/questions, which could improve the reliability of our metric. Therefore, in this component, we construct an STSDG (see Figure 4) to model dependent relationships between questions.

Briefly sketched, the STSDG is a set of Text-Video alignment validation questions structured in a directed scene graph, produced from the text as the ground truth. In particular, we deem the generated question as nodes in the graph, denoted as $\mathcal{Q} = \{q_1, \cdots, q_n\}$. Next, we generate the edges for the nodes. Specifically, similar to the last step, we also implemented this stage by an LLM given task-specific in-context examples: we prompt an LLM with a preamble (with input and output sampled from manual annotations with fixed seeds) to elicit annotations of the same format for new inputs. The details on the preamble engineering is in Appendix H. Specifically, we obtain semantic dependency edges between questions as an adjacency matrix $\mathbf{E} \in \mathbb{R}^{n \times n}$,

$$E_{ij} = \begin{cases} 1, & \text{if } \mathcal{S}(q_i, q_j), \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $i, j \in [1, n]$, and $\mathcal{S}(q_i, q_j)$ is True when the semantics of the question q_i is depend on the question q_j. Notably, $\mathbf{E}$ is the adjacency matrix of a directed acyclic graph, which means $\mathbf{E}_{ij} == \mathbf{E}_{ji}$ does not necessarily hold true.

3.3 Fact Verification

Based on the video content, we verify the faithfulness of each fact by answering the generated question with a VideoQA model as follows,

$$A = \{a_1, \cdots, a_n\} = \text{VideoQA}(V, q_1, \cdots, q_n), \tag{2}$$

where $\text{VideoQA}(\cdot)$ is a video question answering model. $\{q_1, \cdots, q_n\}$ is the question set and q_i is the i-th question corresponding i-th fact. A is the corresponding answer for the question set. The reason why we use VideoQA Models to verify the consistency between fact and video, even if the VideoQA may also introduce hallucination: Our method converts the AI labeling task into a discriminative task that usually generates a short response ("yes" or "no"), and this kind of task tends to generate low hallucination (Min et al., 2023).

3.4 STSDG-based Score Aggregation

Finally, we calculate the faithfulness score FIFA for all the derived facts. In particular, we first convert answers $A = \{a_1, \cdots, a_n\}$ into scores $S = \{s_1, \cdots, s_n\}$. Thereafter, we utilize the semantic dependency relation to derive the refined scores to improve the reliability of the fact verification:

$$\hat{s}_i = \mathbb{I}(a_i = yes) \prod_{j \text{ s.t. } E_{ij}=1} s_j, \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and the value of $\mathbb{I}(a_i = yes)$ is 1 when a_i is "yes". $i, j \in [1, n]$ and $i \neq j$. Then the final faithfulness score $\hat{f}$ is the average of all refined scores: $\hat{f} = 1/n \sum_{i=1}^{n} \hat{s}_i$.

4 Meta Evaluation for FIFA

4.1 Evaluation Setup

We evaluate four widely-used models: two T2V models: CogVideoX (Yang et al., 2024) and HunyuanVideo (Kong et al., 2024), and two V2T models: Video-LLaVA (Lin et al., 2024) and Video-LLaMA (Zhang et al., 2023). For each task we curated 60 evaluation examples. Every example was independently labeled by three annotators, resulting in 360 hallucination-labeled instances in total across the T2V and V2T settings. We employ 3 workers for annotation via Amazon Mechanical Turk [1]. Every worker is a native English speaker. They are paid 15-20 USD per hour. Every worker went through a qualification test of 2 hours and was tested to be highly qualified. More details are in Appendix B

4.2 Comparison with Existing Metrics

To evaluate the superiority of our proposed metric FIFA, we compare it with several T2V and V2T evaluation metrics. For V2T metrics, we compare FIFA with 1) reference-based: BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERT-Score (Zhang* et al., 2020), and COAHA (Ullah and Mohanta, 2022); and 2) reference-free: CLIP-Score (Hessel et al., 2021). For T2V metrics, it is harder to collect ground-truth compared with the V2T task. Hence, we only select reference-free metrics for comparison. We select CLIP-Score, XCLIP-Score (Ni et al., 2022), BLIP-BLEU (Liu et al., 2024b), mPLUG-BLEU (Liu et al., 2024b) and FAST-VQA (Wu et al., 2022b) as baselines.

Table 1: Correlation between evaluation metrics and human judgment on V2T and T2V faithfulness evaluation, measured by Pearson's r, Spearmans ρ, and Kendalls τ.

Task	Type	Metrics	Pearson's r	Kendall's τ	Spearman's ρ
V2T	Reference-based	BLEU-4	41.12	35.92	45.39
		ROUGE-L	29.55	22.83	29.31
		METEOR	45.74	35.95	46.10
		BERT-Score	43.77	36.85	50.11
		COAHA	-38.15	-11.41	-13.70
	Reference-free	CLIP-Score	4.58	-1.20	-1.01
		FIFA	**58.20**	**53.20**	**62.96**
T2V	Reference-free	CLIP-Score	30.22	3.42	5.31
		XCLIP-Score	24.96	20.63	29.39
		BLIP-BLEU	57.67	43.61	60.90
		mPLUG-BLEU	-26.39	-30.07	-22.70
		FAST-VQA	7.65	4.79	5.68
		FIFA	**67.92**	**64.25**	**77.50**

To quantify the human evaluation of faithfulness, we employ the 1-5 Likert Scale (Likert, 1932) to score the faithfulness of the text-video pair on a tangible scale, ranging from 1 (worst) to 5 (best). The result in Table 1 shows that our evaluation framework consistently achieves a significant improvement across T2V and V2T. We add more fine-grained analysis, the annotation process, and case study in Appendix Q and M, and J.

4.3 STSDG-based Generation

In this section, we evaluate every key stage in Spatial-Temporal Semantic Dependency Graph Construction. We use the human evaluation to verify the reliability in each intermediate stage.

Are the generated questions reliable? The first stage of our evaluation framework is to extract all facts and then transform them into a question format. Therefore, it is very important to get high-quality questions. To evaluate the quality of the generated questions, we define the metrics precision and recall. For each text, we employ annotators to write the corresponding facts, denoted as $C = \{c_1, \cdots, c_{n_c}\}$. Follow the definition of the last section, the generated questions are denoted as $U = \{q_1, \cdots, q_n\}$. Based on the generated questions and annotated facts, we define $\sum m_{t,q}/|Q|$ as precision and $\sum m_{t,q}/|T|$ as recall. $|Q|$ and $|T|$ are the total number of questions and facts, respectively. $m_{t,q} = 1$ if t matches q, otherwise, it is 0. We show the experimental results in Table 2. Overall, the generated questions are close to perfect in matching the source semantic fact. Furthermore, we compute the consistency between 3 annotators and found Fleiss' Kappa is 0.84, which indicates

[1] https://www.mturk.com/.

Table 2: Human evaluation results of generated questions, converting facts into questions, and validity of generated dependency for T2V and V2T tasks.

Task	Question Generation		Fact Conversion	Dependency
	Precision	Recall	Accuracy	Valid Ratio
T2V	98.71	99.22	99.06	99.06
V2T	95.11	95.22	99.03	99.03
All	96.31	96.55	99.04	99.04

an almost perfect agreement between annotators.

Can the tuple be transferred into independent questions correctly? To evaluate the performance of the conversion of extracted facts into corresponding questions, we further conduct an analysis using accuracy as the evaluation metric. The results are presented in Table 2. Overall, the accuracy of converting facts into questions are close to perfect (99.88% for T2V and 99.26% for V2T). Furthermore, we compute the consistency between 3 annotators and found the Fleiss' Kappa is 0.91, which indicates an almost perfect agreement between annotators.

Are the generated dependencies between questions valid? To enhance the reliability of fact verification, our method (FIFA) introduces directed dependency edges between questions. Specifically, if question q_i depends on question q_j, then q_i is considered a valid VideoQA query only if the answer to the dependent question q_j is positive (e.g.,*"is the dog white?"* is only valid if the answer to *"is there a dog?"* is positive). To evaluate the effectiveness of the LLM in generating such dependencies, we ask human annotators to make binary judgments for each questiondependent-question pair. We show the human evaluation results in Table 2. Overall, the valid ratio of dependency generation are close to perfect (99.06% for T2V and 99.03% for V2T).

4.4 Performance on Fact Verification

As the verifier in our evaluation framework, the performance of VideoQA models plays a critical role. To assess their effectiveness, we evaluate several state-of-the-art VideoQA models, including InternVL-2.5-8b (Chen et al., 2024b), Video-LLaMA3-7b (Zhang et al., 2025a), Video-LLaVA-7b (Lin et al., 2024), Qwen2.5-VL-7b/32b/72b(Bai et al., 2025). Specifically, we collect 555 questions from the T2V evaluation set and 714 questions from the V2T evaluation set, each paired with its corresponding video. Every question is independently annotated by three annotators, and the final label is determined using majority voting. The performance of all evaluated VideoQA models is reported in Table 3. Overall, Qwen2.5-VL-72b achieves the best performance on the T2V and V2T.

Table 3: Human evaluation for fact verification.

Model	T2V Accuracy	V2T Accuracy	Average
InternVL-2.5-8b	73.86	68.21	71.56
Video-LLaVA	75.47	76.75	76.19
Video-LLaMA3	79.46	79.55	79.51
Qwen2.5-VL-7b	73.69	73.25	73.44
Qwen2.5-VL-32b	77.11	75.73	76.33
Qwen2.5-VL-72b	80.00	80.11	80.06

Table 4: Experiment results of ablation study.

Method	Pearson's r	Kendall's τ	Spearman's ρ
FIFA	**63.06**	**58.73**	**70.23**
w/o-Dependency	58.53	52.77	66.56
w/-Qwen2.5-VL-7b	56.25	45.66	58.06
w/-Qwen2.5-VL-32b	46.46	44.89	54.69
w/-Qwen3-32b	60.34	53.90	68.88

4.5 Ablation Study

Table 4 summarizes the performance of FIFA with its derivations. The details of the ablation method can be found in Appendix A. From the results, we observe that FIFA surpasses w/o-Dependency, demonstrating the importance of introducing semantic dependency relationships between facts/questions. w/-Qwen2.5-VL-7b and w/-Qwen2.5-VL-32b perform worse than our FIFA, which demonstrates the correctness of our choice of the current VideoQA model. By comparing the VideoQA accuracy in Table 3 and Table 4, we observe that models with higher VideoQA accuracy tend to achieve better correlation performance.

5 Post-Correction

Method. Our initial experiments show various hallucinations in the T2V and V2T models. Therefore, we devise a post-correction method to alleviate these issues. In particular, our goal is to identify and rectify hallucinations in texts in T2V and V2T tasks. A central challenge lies in detecting hallucinated content and identifying factual information that can serve as the basis for correction. To address this, we utilize the intermediate evaluation result of our FIFA and divide the entire process into three subtasks: key claim extraction, claim verification, and hallucination correc-

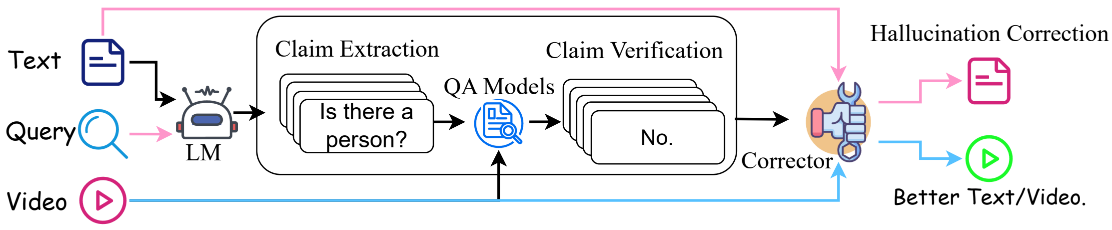

Figure 3: The proposed Post-Correction method consists of three key stages: Claim Extraction, Claim Verification, and Hallucination Correction. The Corrector takes claim-answer pairs with video for T2V, and with text for V2T.

tion. An overview of our framework is shown in Figure 3. 1) **Claim Extraction.** Since the text usually consists of multiple claims, such as objects, attributions, and relations, we follow section 3 to extract facts from the text. 2) **Claim Verification.** Then, we ask a series of questions around them to make the hallucination diagnosis following operations in Eq. 2. For all questions, we apply a VideoQA model to answer the questions conditioned on the video. The first two stages are the intermediate process in our FIFA. 3) **Hallucination Correction.** For the **V2T** task, an LLM corrects hallucinated content in the generated textual responses. Specifically, we aggregate the QA pairs into a structured prompt and instruct the LLM to generate a refined version of the response with hallucinations corrected. For the **T2V** task, a video editing model is employed to revise hallucinated visual content in generated videos. In particular, we first use an LLM to generate editing instructions based on the input prompt and corresponding QA pairs. For example, given the input prompt *"a green door"*, and QA pairs: *"Is there a door? Yes"* and *"Is the door green? No"*, the generated instruction might be *"change the door to green."* The original generated video, along with this editing instruction, is then passed to a video editing model to produce a refined video.

Experiments. We construct evaluation sets for both T2V and V2T tasks. For the V2T task, we sample 500 videos from the MSR-VTT dataset to perform the captioning task. For the T2V task, due to the slow generation speed of video generation models and video editing models, we adopt 100 prompts from the meta-evaluation benchmark for our experiments. We use Qwen2.5-VL-72b as the VideoQA model and TokenFlow as the video editing model in our Post-Correction method. Table 5 shows the performance of all the baselines without and with our correction method. For the T2V task, we found that our FIFA can improve the performance of all baselines across COAHA and FIFA metrics. For the V2T task, our method can also improve the FIFA and reduce hallucinations in the generated video, which demonstrates the effectiveness of our Post-Correction method. In addition, we show more benchmark results in Appendix O.

Task	Model	COAHA ↓		FIFA↑	
		w/o	w/	w/o	w/
V2T	Video-LLaVA	54.79	**49.57**	64.03	**66.68**
	Video-LLaMA	55.68	**48.20**	61.06	**66.14**
	Video-LLaMA2	39.99	**28.27**	65.09	**70.42**
	Video-LLaMA3	65.59	**53.61**	65.88	**71.01**
T2V	CogVideoX	-	-	55.32	**61.12**

Table 5: Results on the V2T and T2V tasks. w/ and w/o denote whether the generated content is or is not corrected by our Post-Correction method.

6 Conclusion

In this work, we propose FIFA, a unified and reference-free faithfulness evaluation framework for both V2T and T2V tasks. FIFA introduces a comprehensive fact extraction strategy and constructs an STSDG to model inter-fact relationships. These facts are then converted into questions and verified using powerful VideoQA models, with dependencies guiding the final score aggregation. Our method achieves the highest correlation with human judgments compared to existing baselines. In addition, we propose a unified correction pipeline, Post-Correction, to mitigate hallucinations in both generated videos and texts.

Limitations

FIFA focuses primarily on factual precision, ensuring that each piece of information in a text is supported by the visual input. Factual recall is more challenging and an open question (Min et al., 2023).

Acknowledgements

We thank the anonymous reviewers for valuable and insightful feedback. This research is supported in part by the National Science Foundation CAREER Grant IIS-2340435 and Open Philanthropy. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of the U.S. Government.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Summarization@ACL*, pages 65–72. ACL.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. 2024. Video generation models as world simulators. *OpenAI Blog*, 1:8.

Yue Chang, Liqiang Jing, Xiaopeng Zhang, and Yue Zhang. 2024. A unified hallucination mitigation framework for large vision-language models. *Trans. Mach. Learn. Res.*, 2024.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024a. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 7310–7320. IEEE.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Jaemin Cho, Yushi Hu, Jason M. Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2024. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Kaisi Guan, Zhengfeng Lai, Yuchong Sun, Peng Zhang, Wei Liu, Kieran Liu, Meng Cao, and Ruihua Song. 2025. Etva: Evaluation of text-to-video alignment via fine-grained question generation and answering. *arXiv preprint arXiv:2503.16867*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, pages 7514–7528. ACL.

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022. Video diffusion models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023. TIFA: accurate and interpretable text-to-image faithfulness evaluation with question answering. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20349–20360. IEEE.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.

Liqiang Jing, Guiming Hardy Chen, Ehsan Aghazadeh, Xin Eric Wang, and Xinya Du. 2025a. A comprehensive analysis for visual object hallucination in large vision-language models. *arXiv preprint arXiv:2505.01958*.

Liqiang Jing and Xinya Du. 2025. FGAIF: aligning large vision-language models with fine-grained AI feedback. *Trans. Mach. Learn. Res.*, 2025.

Liqiang Jing, Zhehui Huang, Xiaoyang Wang, Wenlin Yao, Wenhao Yu, Kaixin Ma, Hongming Zhang, Xinya Du, and Dong Yu. 2024a. Dsbench: How far are data science agents from becoming data science experts? *arXiv preprint arXiv:2409.07703*.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024b. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 5042–5063. Association for Computational Linguistics.

Liqiang Jing, Xuemeng Song, Xuming Lin, Zhongzhou Zhao, Wei Zhou, and Liqiang Nie. 2023a. Stylized data-to-text generation: A case study in the e-commerce domain. *ACM Transactions on Information Systems*, 42(1):1–24.

Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia, and Liqiang Nie. 2023b. Multi-source semantic graph-based multimodal sarcasm explanation generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11349–11361.

Liqiang Jing, Yue Zhang, and Xinya Du. 2025b. Tutorial proposal: Hallucinations in large language models and large vision-language models. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 2138–2139.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. 2024. Hunyuanvideo: A systematic framework for large video generative models. *CoRR*, abs/2412.03603.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2024. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 391–404. Association for Computational Linguistics.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. 2024a. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. *CoRR*, abs/2412.03735.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024b. Llama-vid: An image is worth 2 tokens in large language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, volume 15104 of *Lecture Notes in Computer Science*, pages 323–340. Springer.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5971–5984. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. ACL.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *CoRR*, abs/2304.08485.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024b. Evalcrafter: Benchmarking and evaluating large video generation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22139–22149. IEEE.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. 2023b. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36:62352–62387.

Meng Luo, Shengqiong Wu, Liqiang Jing, Tianjie Ju, Li Zheng, Jinxiang Lai, Tianlong Wu, Xinya Du, Jian Li, Siyuan Yan, et al. 2025. Dr. v: A hierarchical perception-temporal-cognition framework to diagnose video hallucination by fine-grained spatial-temporal grounding. *arXiv preprint arXiv:2509.11866*.

Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. 2024. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13151–13160. IEEE.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. 2025. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12585–12602. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. Association for Computational Linguistics.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, volume 13664 of *Lecture Notes in Computer Science*, pages 1–18. Springer.

Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. 2017. To create what you tell: Generating videos from captions. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1789–1798. ACM.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

Vipula Rawte, Sarthak Jain, Aarush Sinha, Garv Kaushik, Aman Bansal, Prathiksha Rumale Vishwanath, Samyak Rajesh Jain, Aishwarya Naresh Reganti, Vinija Jain, Aman Chadha, Amit P. Sheth, and Amitava Das. 2024. Vibe: A text-to-video benchmark for evaluating hallucination in large multimodal models. *CoRR*, abs/2411.10867.

Nasib Ullah and Partha Pratim Mohanta. 2022. Thinking hallucination for video captioning. In *Computer Vision - ACCV 2022 - 16th Asian Conference on Computer Vision, Macao, China, December 4-8, 2022, Proceedings, Part IV*, volume 13844 of *Lecture Notes in Computer Science*, pages 623–640. Springer.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6306–6315.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2023. Phenaki: Variable length video generation from open domain textual descriptions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Xinyu Wang, Yue Zhang, and Liqiang Jing. 2025. Can large vision-language models understand multimodal sarcasm? In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5340–5345.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *CoRR*, abs/2406.16338.

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. 2022a. Nüwa: Visual synthesis pre-training for neural visual world creation. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XVI*, volume 13676 of *Lecture Notes in Computer Science*, pages 720–736. Springer.

Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2022b. FAST-VQA: efficient end-to-end video quality assessment with fragment sampling. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, volume 13666 of *Lecture Notes in Computer Science*, pages 538–554. Springer.

Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, Weisi Lin, Wynne Hsu, Ying Shan, and Mike Zheng Shou. 2024. Towards A better metric for text-to-video generation. *CoRR*, abs/2401.07781.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society.

Bowen Yan, Zhengsong Zhang, Liqiang Jing, Eftekhar Hossain, and Xinya Du. 2024. Fiha: Autonomous hallucination evaluation in vision-language models with davidson scene graphs. *arXiv preprint arXiv:2409.13612*.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. 2021. Videogpt: Video generation using VQ-VAE and transformers. *CoRR*, abs/2104.10157.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi

Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *CoRR*, abs/2408.06072.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: hallucination correction for multimodal large language models. *Sci. China Inf. Sci.*, 67(12).

Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. 2024. Make pixels dance: High-dynamic video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 8850–8860. IEEE.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. *CoRR*, abs/2501.13106.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 543–553. Association for Computational Linguistics.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. 2024a. Eventhallusion: Diagnosing event hallucinations in video llms. *CoRR*, abs/2409.16597.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yue Zhang, Liqiang Jing, and Vibhav Gogate. 2025b. Defeasible visual entailment: Benchmark, evaluator, and reward-driven optimization. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 25976–25984. AAAI Press.

Yue Zhang, Liqiang Jing, Jia Li, Yapeng Tian, Xinya Du, Yunhui Guo, and Vibhav Gogate. 2026. A skill-augmented agentic framework and benchmark for multi-video understanding. *arXiv preprint arXiv:2603.14733*.

Yue Zhang, Seiji Maekawa, and Nikita Bhutani. 2025c. Same content, different representations: A controlled study for table qa. *arXiv preprint arXiv:2509.22983*.

Yue Zhang, Jilei Sun, Yunhui Guo, and Vibhav Gogate. 2025d. Can video large multimodal models think like doubters-or double-down: A study on defeasible video entailment. *arXiv preprint arXiv:2506.22385*.

Yue Zhang, Jingxuan Zuo, Ke Su, and Liqiang Jing. 2024b. Fine-grained and explainable factuality evaluation for multimodal summarization. *arXiv preprint arXiv:2402.11414*.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. 2025. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *CoRR*, abs/2503.21755.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. 2022. Magicvideo: Efficient video generation with latent diffusion models. *CoRR*, abs/2211.11018.

A More Ablation Study

To explore the roles of different components in our proposed evaluation framework, we compared FIFA with the following derivations. 1) w/o-Dependency. To explore the effect of the generated semantic dependency relation, we removed the STSDG in our evaluation framework. Specifically, we remove Equation 3 from our FIFA. 2) w/-Qwen2.5-VL-7b and 3) w/-Qwen2.5-VL-32b. To verify the importance of our selected VideoQA model, we replace it with Qwen2.5-VL-7b and Qwen2.5-VL-32b, respectively. 4) w/-Qwen3: To test open-source LLMs' performance in our FIFA, we replace gpt-4o with Qwen3-32b.

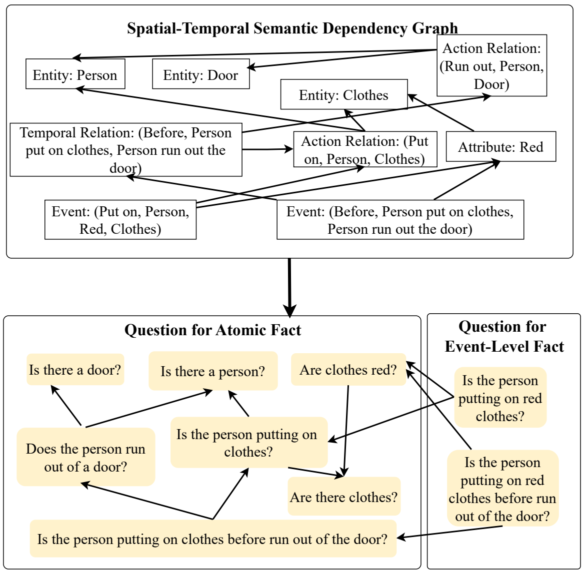

Figure 4: Illustration of STSDG-based Question Generation and STSDG-based Score Aggregation. The generation process is for the text "The person put on red clothes before he ran out the door.".

B Experimental Setups for Meta-Evaluation

V2T Data. We sampled videos from the validation set of the widely-used video captioning dataset MSR-VTT (Xu et al., 2016) for human evaluation. To enrich the diversity of question types in our dataset, we designed different types of queries for evaluation. Specifically, we selected 10 videos for the captioning task, using the query *"Please generate a brief for the video"* with the ground-truth captions from MSR-VTT serving as the reference answers.

For the remaining two tasks, i.e., detailed description and complex question answering, we sampled 10 different videos for each task and used GPT-4o to generate corresponding prompt-answer pairs, following LLaVA (Liu et al., 2023a).

T2V Data. We selected 20 captions from the validation set of MSR-VTT as inputs for the T2V task. However, these captions are typically short and contain limited semantic elements, such as objects, attributes, and temporal relationships. To address this limitation, we further sampled an additional set of 10 captions and employed GPT-4o to generate richer and more informative prompts, aiming to better evaluate the models ability to handle complex and detailed textual inputs.

C Comparison with Existing Evaluation Metrics

To evaluate the superiority of our proposed metric FIFA, we compare it with several T2V and V2T evaluation metrics. For V2T metrics, we compare FIFAwith 1) reference-based: BLEU-{1/2/3/4} (Papineni et al., 2002), ROUGE-{1/2/L} (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERT-Score (Zhang* et al., 2020), and COAHA (Ullah and Mohanta, 2022); and 2) reference-free: CLIP-Score (Hessel et al., 2021). For T2V metrics, it is harder to collect ground-truth compared with the V2T task. Hence, we only select reference-free metrics for comparison. Specifically, we select CLIP-Score, XCLIP-Score (Ni et al., 2022), BLIP-BLEU (Liu et al., 2024b), mPLUG-BLEU (Liu et al., 2024b) and FAST-VQA (Wu et al., 2022b) as baselines.

For all evaluation tasks, we employed three annotators to independently annotate each sample to ensure the reliability and consistency of the annotations. All GPT-4o outputs used in our experiments were generated with the model version `gpt-4o-2024-08-06`. We use Qwen2.5-VL-72b (Bai et al., 2025) as our videoQA model and use GPT-4o as the LLM in our evaluation framework. Details of our annotation interface are provided in Appendix Q.

D More Comparison Results

In this section, we present a comprehensive comparison of models on both the T2V and V2T tasks. Because of constraints on GPU hours, the fees charged by some closed-source systems, and the need for manual interaction with certain models, we evaluate on a small subset rather than the large-scale benchmark: 100 randomly sam-

T2V Model	Entity	Attribute	Spatial	Temporal	Action	Event
CogVideox	84.07	77.42	72.22	60.00	71.70	63.46
HunyuanVideo	86.49	76.67	66.67	20.00	54.72	52.94
V2T Model	Entity	Attribute	Spatial	Temporal	Action	Event
Video-LLaMA	88.08	73.53	71.43	68.00	73.81	58.97
Video-LLaVA	90.35	90.48	75.86	56.61	67.65	57.14

Table 6: Comparison of T2V models and V2T models.

Task	**Model**	**FIFA**
V2T	Video-LLaVA	63.43
	Video-LLaMA	60.46
	Video-LLaMA2	64.49
	Video-LLaMA3	65.28
	Qwen2.5-VL-3B	62.87
	Qwen2.5-VL-7B	71.54
T2V	CogVideoX	54.53
	HunyuanVideo	63.61
	Wan-1.3B	57.94
	Wan-14B	60.91
	Step-Video	68.10
	Sora	57.49

Table 7: Comprehensive model comparison on the V2T and T2V tasks.

pled captions from the MSR-VTT dataset for V2T, and 30 prompts from the same dataset for T2V. For V2T models, we selected Video-LLaVA, Video-LLaMA, Video-LLaMA2, Video-LLaMA3, Qwen2.5-VL-3B, and Qwen2.5-VL-7B. For T2V models, we selected Sora (Brooks et al., 2024), CogVideoX (Yang et al., 2024), Wan (Wan et al., 2025), and Step-Video (Ma et al., 2025). Table 7 show the various performance comparison across T2V and V2T.

E Cycle Detection and Acyclicity Guarantee

To ensure that the semantic dependency graph generated by the LLM is a valid Directed Acyclic Graph (DAG), we adopt an explicit cycle-checking step based on topological sorting. A directed graph is acyclic if and only if it admits a valid topological ordering of its nodes. We implement Kahns algorithm for this purpose, as described in Algorithm 1. The algorithm iteratively removes nodes with zero in-degree and appends them to a list. If, by the end of this process, all nodes have been processed (i.e., the resulting list contains all nodes), then the graph is acyclic. Otherwise, a cycle must exist.

This verification step is applied to every semantic graph produced by the LLM before proceeding to scoring. If a cycle is detected, we remove the edge with the lowest confidence within the detected cycle, and repeat the process until the graph becomes acyclic. This ensures that all graphs used in evaluation are certified DAGs both structurally and algorithmically. Empirically, all our generated graphs pass this check.

F More Annotation Results

To further improve the reliability of our evaluation, we further expanded the human evaluation to 200 samples per task (T2V and V2T), covering 100 captioning samples, 50 detailed captioning samples, 50 complex QA samples, and 200 video generation samples. This results in a total of 1200 annotation samples, with an annotation cost of approximately $1000. We believe this expanded evaluation provides a more comprehensive and reliable validation of our findings. The updated results are shown in Table 8, where FIFA continues to achieve the best performance, further demonstrating the effectiveness of our evaluation metric.

G Why unify T2V and V2T?

Although T2V and V2T are directional opposites, they share a common need: ensuring that generated content accurately aligns with the corresponding modality. Both are vulnerable to similar hallucination patternssuch as misattributing entities, temporal errors, or false compositionsand both can be evaluated using the same graph-based fact decomposition and verification process. By using a

Algorithm 1: Verify Acyclicity via Topological Sorting

Input: Directed graph $G = (V, E)$
Output: `true` if G is acyclic; `false` otherwise

1. Initialize in-degree of each node: in_deg$[v] \leftarrow 0$ for all $v \in V$
2. For each edge $(u \rightarrow v) \in E$: in_deg$[v] \leftarrow$ in_deg$[v] + 1$
3. Initialize a queue Q with all nodes v such that in_deg$[v] = 0$
4. Initialize an empty list $L \leftarrow [\,]$ to store the topological order
5. **While** Q is not empty:
    - Remove a node v from Q
    - Append v to L
    - For each neighbor u of v such that $(v \rightarrow u) \in E$:
        - in_deg$[u] \leftarrow$ in_deg$[u] - 1$
        - If in_deg$[u] = 0$, add u to Q
6. **If** $|L| = |V|$: **return** `true` // Graph is acyclic
7. **Else: return** `false` // Cycle detected

single framework to assess both tasks, we promote cross-modal consistency, reuse core components, and provide a unified metric that reflects semantic grounding quality in both directions.

H Prompts

Fact Extraction Prompt

Prompt: Task: given input prompts, describe each scene with skill-specific tuples. Do not generate the same tuples again. Do not generate tuples that are not explicitly described in the prompts.
output format: id | tuple
${In-context Examples}$

Question Generation Prompt

Prompt: Task: given input prompts and skill-specific tuples, re-write tuple each in natural language question.
output format: id | question
${In-context Examples}$

Dependency Generation Prompt

Prompt: Task: given input prompts and tuples, describe the parent tuples of each tuple.
output format: id | dependencies.
${In-context Examples}$

We show the concert in-context examples in Section P.

I Experimental Implementation

We run all experiments on a server with 4 $\times$ A100 GPUs.

In the T2V setting, we first use the QA verification results to generate structured editing instructions via an LLM. For example, if the input prompt is “a green door” and the QA pairs indicate “Is there a door? Yes” and “Is the door green? No,” the LLM produces the instruction “change the door to green.” The original generated video, together with this editing instruction, is then passed to a video editing model to locally revise the visual content. This ensures that the correction is grounded in the factual critique rather than generic prompt engineering. In the revised version, we will expand this section with more details and examples to make the methodology clearer.

Table 8: Human correlation comparison on more annotation evaluation data.

Task	Type	Metrics	Pearson's r	Kendall's τ	Spearman's ρ
V2T	**Reference-based**	BLEU-4	40.60	35.97	44.79
		ROUGE-L	30.65	21.83	29.46
		METEOR	45.73	36.98	44.47
		BERT-Score	45.19	35.92	48.76
		COAHA	-38.68	-12.49	-11.74
	Reference-free	CLIP-Score	6.57	-2.63	0.12
		FIFA	59.82	52.82	63.54
V2T	**Reference-free**	CLIP-Score	28.74	4.04	7.37
		XCLIP-Score	24.00	20.88	28.79
		BLIP-BLEU	56.25	41.62	62.62
		mPLUG-BLEU	-27.92	-27.56	-25.17
		FAST-VQA	7.35	4.17	5.67
		FIFA	69.68	61.38	79.33

J Case Study

Figure 5 provide examples of invalid VideoQA queries described in our evlauation framework. In the top image, VideoQA model answered "No" to the question Is there a person?, but the VideoQA model answered "Yes" to the question "Is the person holding a pet?". In addition, in Figure 6 we present another case highlighting a failure of our dependencygraph algorithm. In this case, because the model's answer to the root question is incorrect, we cannot use it to correct the dependent questions.

K More Baselines

We further incorporated the additional baselines UMT-score (Liu et al., 2023b) and ETVA (Guan et al., 2025) into our evaluation. The updated results are shown in Table 9. As can be observed, our proposed method still achieves the best performance across all metrics, which further strengthens the validity of our conclusions.

L Efficiency

In practice, our implementation processes 100 T2V tasks in about 10 minutes on a single A100 GPUmuch faster than large-scale Text2Video production systems (for example, WAN-14B models take about 8 hours for 100 T2V tasks). This suggests our pipeline is highly efficient relative to video-generation workloads, and leaves sufficient headroom for real-world integration.

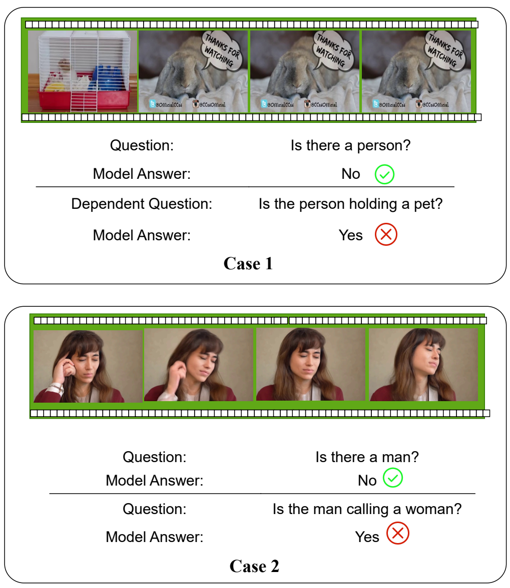

Figure 5: Examples of invalid VQA queries where a dependent question gets a positive answer when its root question gets a negative answer.

Table 9: Correlation results of different metrics.

Metric	Pearson's r	Kendall's τ	Spearman's ρ
CLIP-Score	30.22	3.42	5.31
XCLIP-Score	24.96	20.63	29.39
BLIP-BLEU	57.67	43.61	60.90
mPLUG-BLEU	-26.39	-30.07	-22.70
FAST-VQA	7.65	4.79	5.68
UMT-Score	41.52	18.34	26.90
ETVA	59.28	50.81	63.06
FIFA (Ours)	**67.92**	**64.25**	**77.50**

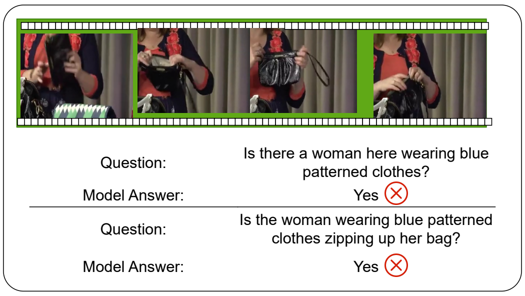

Figure 6: Examples of queries where both the dependent question and its root question get negative answers.

Model	Accuracy ↑ w/o	Accuracy ↑ w/
Video-LLaVA	55.75	**56.68**
Video-LLaMA	53.75	**62.25**
Video-LLaMA2	56.25	**61.75**
Video-LLaMA3	65.42	**67.02**

Table 10: Results on the VideoHallucer benchmark, a V2T hallucination evaluation task.

M Fine-grained Analysis for V2T and T2V

Table 6 presents a comparison of T2V and V2T models across different fact categories in our human evaluation. We observe that entity and attribute categories achieve relatively high FIFA scores across all models, indicating that hallucinations related to objects and their properties are less frequent in video-related tasks. In contrast, action and relation categories (particularly spatial and temporal relations) tend to have lower scores, suggesting these are the main sources of hallucination. Notably, the temporal category shows the lowest accuracy in T2V settings, highlighting the importance of modeling temporal hallucinations explicitly. Additionally, the low scores in the event-level facts underscore the necessity of incorporating composite, high-level semantic facts to better capture and evaluate complex visual events.

N Evaluation Scope

We would like to clarify that the scope of our current evaluation framework is indeed factual faithfulness, i.e., whether the generated content is semantically consistent with the source modality at the level of entities, attributes, relations, scenes, and higher-order events. Other aspects of T2V generation, including transitions, camera movements, and visual style, remain important perceptual and aesthetic dimensions that we leave for future exploration. In fact, we see our metric as complementary to such assessments: FIFA addresses factual correctness, while perceptual/video-quality metrics better capture style, smoothness, or cinematography.

O Hallucination Mitigation on More Benchmarks

In addition to the caption task, we also conduct experiment on VideoHallucer (Wang et al., 2024), which is a binary-QA question answering benchmark. We show the results in Table 10. From this table, we found that our method could mitigate hallucination for all baselines. In addition, we observe a positive correlation between the performance of our method and the baseline models. In general, the stronger the baseline, the greater the improvement achieved by our approach.

P In-context Examples for Prompts

P.1 Fact Extraction for Video-to-Text

```
query: Please generate a caption for
    ↪ the video.
input: A male skateboarder is trying to
    ↪ pull off a trick on the ramp.
```

```
output: 1 | entity - whole
    ↪ (skateboarder)
2 | entity - whole (ramp)
3 | attribute - type (skateboarder,
    ↪ male)
4 | action - (male skateboarder, pull
    ↪ off a trick)
5 | relation - spatial (male
    ↪ skateboarder, ramp, on)
6 | event - ambigity (skateboarder,
    ↪ male, pull off a trick)
7 | event - ambiguity (male
    ↪ skateboarder, ramp, on)
8 | event - ambiguity (skateboarder,
    ↪ pull off a trick, ramp, on)

query: Please generate a caption for
    ↪ the video.
input: A car playing soccer, digital
    ↪ art.
output: 1 | entity - whole (car)
2 | global - (digital art)
3 | action - (car, soccer, play)

query: Please generate a caption for
    ↪ the video.
input: A set of 2x2 emoji icons with
    ↪ happy, angry, surprised and
    ↪ sobbing faces. The emoji icons
    ↪ look like pigs. All of the pigs
    ↪ are wearing crowns.
output: 1 | entity - whole (emoji icons)
2 | other - count (emoji icons, ==4)
3 | attribute - state (emoji icons, 2x2
    ↪ grid)
4 | attribute - type (emoji icons, pig)
5 | attribute - state (emoji_1, happy)
6 | attribute - state (emoji_2, angry)
7 | attribute - state (emoji_3,
    ↪ surprised)
8 | attribute - state (emoji_4, sobbing
    ↪ face)
9 | entity - part (pig's crown)

query: Please generate a caption for
    ↪ the video.
input: a photo of bear and dining
    ↪ table; dining table is below bear
output: 1 | global - (photo)
2 | entity - whole (bear)
3 | entity - whole (dining table)
4 | relation - spatial (dining table,
    ↪ bear, below)

query: Please generate a caption for
    ↪ the video.
input: A group of children sitting in
    ↪ the grass with two of them
    ↪ holding a Frisbee .
output: 1 | entity - whole (children)
2 | entity - whole (grass)
3 | entity - whole (frisbee)
4 | attribute - state (children, sit)
5 | relation - spatial (a group of
    ↪ children, grass, sitting in)
6 | entity - part (two of the children)
7 | action - (two of the children,
    ↪ frisbee, hold)

query: Please generate a caption for
    ↪ the video.
input: the word 'START' written in
    ↪ chalk on a sidewalk
output: 1 | entity - whole (word)
2 | entity - whole (sidewalk)
3 | other - text rendering (word,
    ↪ "START")
4 | attribute - texture (word, chalk)
5 | relation - spatial (word 'START',
    ↪ sidewalk, on)

query: Please generate a caption for
    ↪ the video.
input: A pear, orange, and two bananas
    ↪ in a wooden bowl.
output: 1 | entity - whole (pear)
2 | entity - whole (orange)
3 | entity - whole (bananas)
4 | other - count (bananas, ==2)
5 | entity - whole (bowl)
6 | attribute - material (bowl, wood)
7 | relation - spatial (pear, bowl, in)
8 | relation - spatial (orange, bowl,
    ↪ in)
9 | relation - spatial (bananas, bowl,
    ↪ in)
10 | relation - spatial (bananas, bowl,
    ↪ in)
11 | event - ambiguity (pear, orange,
    ↪ bananas, ==2, bowl, in)

query: Please generate a caption for
    ↪ the video.
input: Closeup picture of the front of
    ↪ a clean motorcycle.
output: 1 | entity - whole (motorcycle)
2 | global - (closeup)
3 | global - (picture)
4 | attribute - state (motorcycle,
    ↪ clean)
5 | entity - part (front of the clean
    ↪ motorcycle)

query: Please generate a caption for
    ↪ the video.
input: a sad man with green hair
output: 1 | entity - whole (man)
2 | entity - part (man's hair)
3 | attribute - state (man, sad)
4 | attribute - color (man's hair,
    ↪ green)
5 | event - ambiguity (man, sad, man's
    ↪ hair, green)

query: Please generate a caption for
    ↪ the video.
input: A commercial airplane with
    ↪ propellers flying through the air.
output: 1 | entity - whole (airplane)
2 | entity - part (airplane's
    ↪ propellers)
3 | action - (airplane, air, fly
    ↪ through)
4 | event - ambiguity (airplane, with
    ↪ propellers, air, fly through)

query: Please generate a caption for
    ↪ the video.
```

```
input: A little boy grips a soccer ball
    ↪ in his arms surrounded by other
    ↪ youth soccer players.
output: 1 | entity - whole (boy)
2 | entity - whole (ball)
3 | entity - whole (soccer players)
4 | entity - part (boy's arms)
5 | entity - scale (boy, little)
6 | attribute - type (ball, soccer)
7 | attribute - state (soccer players,
    ↪ youth)
8 | relation - spatial (little boy,
    ↪ ball, grip in his arms)
9 | relation - spatial (little boy
    ↪ gripping the ball in his arms,
    ↪ soccer players, surrounded by)
10 | event - ambiguity (boy's arm,
    ↪ little, ball, soccer, grip in his
    ↪ arms)
11 | event - ambiguity (boy, little,
    ↪ soccer players, youth, surrounded
    ↪ by)

query: Please generate a caption for
    ↪ the video.
input: A traffic light and a signpost
    ↪ at a crossroads intersection near
    ↪ a waterway.
output: 1 | entity - whole (traffic
    ↪ light)
2 | entity - whole (signpost)
3 | entity - whole (crossroads
    ↪ intersection)
4 | entity - whole (waterway)
5 | relation - spatial (traffic light,
    ↪ crossroads intersection, at)
6 | relation - spatial (signpost,
    ↪ crossroads intersection, at)
7 | relation - spatial (traffic light,
    ↪ waterway, near)
8 | relation - spatial (signpost,
    ↪ waterway, near)
9 | relation - spatial (crossroads
    ↪ intersection, waterway, near)
10 | event - ambiguity (traffic light,
    ↪ signpost, crossroads
    ↪ intersection, at)
11 | event - ambiguity (traffic light,
    ↪ crossroads intersection, at,
    ↪ waterway, near)
12 | event - spatial (signpost,
    ↪ crossroads intersection, at,
    ↪ waterway, near)

query: Please generate a caption for
    ↪ the video.
input: a photo of dining table and
    ↪ traffic light; traffic light is
    ↪ below dining table
output: 1 | global - (photo)
2 | entity - whole (dining table)
3 | entity - whole (traffic light)
4 | relation - spatial (traffic light,
    ↪ dining table, below)

query: Please generate a caption for
    ↪ the video.
input: A realistic photo of a
    ↪ Pomeranian dressed up like a
    ↪ 1980s professional wrestler with
    ↪ neon green and neon orange face
    ↪ paint and bright green wrestling
    ↪ tights with bright orange boots.
output: 1 | global - (photo)
2 | entity - whole (Pomeranian)
3 | global - (realistic)
4 | entity - part (Pomeranian's costume)
5 | attribute - type (Pomeranian's
    ↪ costume, 1980s professional
    ↪ wrestler)
6 | entity - part (Pomeranian's
    ↪ costume's wrestling tights)
7 | entity - part (Pomeranian's
    ↪ costume's wrestling tights' boots)
8 | entity - part (Pomeranian's
    ↪ facepaint)
9 | attribute - color (Pomeranian's
    ↪ facepaint, neon green)
10 | attribute - color (Pomeranian's
    ↪ facepaint, neon orange)
11 | attribute - color (Pomeranian's
    ↪ costume's wrestling tights,
    ↪ bright green)
12 | attribute - color (Pomeranian's
    ↪ costume's wrestling tights'
    ↪ boots, bright orange)

query: Please generate a caption for
    ↪ the video.
input: a four-piece band on a stage in
    ↪ front of a small crowd
output: 1 | entity - whole (band)
2 | entity - whole (stage)
3 | entity - whole (crowd)
4 | other - count (band members, ==4)
5 | attribute - shape (crowd, small)
6 | relation - spatial (four-piece
    ↪ band, stage, on)
7 | relation - spatial (four-piece
    ↪ band, crowd, in front of)
8 | relation - spatial (stage, crowd,
    ↪ in front of)
9 | event - ambiguity (band, ==4
    ↪ picece, stage, on)
10 | event - ambiguity (band, ==4
    ↪ picece, crowd, small, in front of)
11 | event - ambiguity (stage, crowd,
    ↪ small, in front off)

query: Please generate a caption for
    ↪ the video.
input: two laptops, a mouse cord, and a
    ↪ monitor
output: 1 | entity - whole (laptops)
2 | other - count (laptops, ==2)
3 | entity - whole (mouse coord)
4 | entity - whole (monitor)

query: Please generate a caption for
    ↪ the video.
input: A red motorcycle parked by paint
    ↪ chipped doors.
output: 1 | entity - whole (motorcycle)
2 | entity - whole (doors)
3 | attribute - color (motorcycle, red)
4 | attribute - state (door, paint
    ↪ chipped)
5 | relation - spatial (red motorcycle,
    ↪ paint chipped door, next to)
```

```
6 | attribute - state (motorcycle, parked)
7 | event- ambiguity (motorcycle, red, door, paint chipped, next to)
8 | event- ambiguity (motorcycle, red, parked)

query: Please generate a caption for the video.
input: A cube made of denim. A cube with the texture of denim.
output: 1 | entity - whole (cube)
2 | attribute - material (cube, denim)
3 | attribute - texture (cube, denim)

query: Please generate a caption for the video.
input: an espresso machine that makes coffee from human souls
output: 1 | entity - whole (espresso machine)
2 | entity - whole (coffee)
3 | entity - whole (human souls)
4 | action - (espresso machine, coffee, make)
5 | attribute - material (coffee, human souls)
6 | event - ambiguity (espresso machine, coffee, make, human souls)

query: Please generate a caption for the video.
input: Three people standing next to an elephant along a river.
output: 1 | entity - whole (people)
2 | other - count (people, ==3)
3 | entity - whole (elephant)
4 | entity - whole (river)
5 | attribute - state (people, stand)
6 | relation - spatial (three people, elephant, next to)
7 | relation - spatial (people, river, next to)
8 | relation - spatial (elephant, river, next to)
9 | event - ambiguity (people, ==3, stand)
10 | event - ambiguity (people, ==3, elephant, next to)
11 | event - ambiguity (people, ==3, river, next to)
12 | event - ambiguity (people, stand, elephant, next to)
13 | event - ambiguity (people, stand, river, next to)
14 | event - ambiguity (people, elephant, next to, river, next to)

query: Please generate a caption for the video.
input: Aerial view of downtown Manhattan, but with Millennium Wheel next to the Statue of Liberty. The Great Pyramid is on a sandy island near the buildings.
output: 1 | entity - (downtown Manhattan)
2 | entity - (Millennium Wheel)
3 | entity - (the Statue of the Liberty)
4 | entity - (the Great Pyramid)
5 | entity - (island)
6 | entity - (buildings)
7 | global - (aerial view)
8 | attribute - texture (island, sandy)
9 | relation - spatial (Millennium Wheel, the Statue of Liberty, next to)
10 | relation - spatial (the Great Pyramid, island, on)
11 | relation - spatial (the Great Pyramid, buildings, near)
12 | event - ambiguity (the Great Pyramid, island, on, buildings, near)

query: Please generate a caption for the video.
input: A laptop with external keyboard, mouse, phone and photo on a desk.
output: 1 | entity - whole (laptop)
2 | entity - whole (keyboard)
3 | entity - whole (mouse)
4 | entity - whole (phone)
5 | entity - whole (photo)
6 | entity - whole (desk)
7 | attribute - type (keyboard, external)
8 | relation - spatial (laptop, desk, on)
9 | relation - spatial (keyboard, desk, on)
10 | relation - spatial (mouse, desk, on)
11 | relation - spatial (phone, desk, on)
12 | relation - spatial (photo, desk, on)
13 | event - ambiguity (laptop, external keyboard, mouse, phone, photo, desk, on)

query: Please generate a caption for the video.
input: A white slope covers the background, while the foreground features a grassy slope with several rams grazing and one measly and underdeveloped evergreen in the foreground.
output: 1 | entity - whole (slopes)
2 | other - count (slopes, ==2)
3 | entity - whole (rams)
4 | entity - whole (evergreen)
5 | attribute - color (slope_1, white)
6 | attribute - texture (slope_2, grassy)
7 | attribute - state (evergreen, measly and underdeveloped)
8 | relation - spatial (slope_1, background, in)
9 | relation - spatial (slope_2, foreground, in)
10 | relation - spatial (several, rams, grassy slope_2, on)
11 | attribute - state (several rams, graze)
12 | event - ambiguity (slope_1, white, background, in)
```

```
13 | event - ambiguity (slope_2,
    ↪ grassy, foreground, in)
14 | event - ambiguity (several, rams,
    ↪ slope_2, grassy, on)

query: Please generate a caption for
    ↪ the video.
input: A man walks into a room and sits
    ↪ on a chair. A dog follows him.
output: 1 | entity - whole (man)
2 | entity - whole (room)
3 | entity - whole (chair)
4 | entity - whole (dog)
5 | action - (man, walk, room)
6 | action - (man, sit on, chair)
7 | action - (dog, follow, man)
8 | relation - temporal (man, sit,
    ↪ before, walk)
9 | relation - temporal (dog, follows,
    ↪ after, man, sit)
10 | event - temporal (man, walks into
    ↪ a room and sits on a chair, dog
    ↪ follows him)

query: Please generate a caption for
    ↪ the video.
input: A car is parked by the roadside.
    ↪ Later, it starts moving and
    ↪ drives away.
output: 1 | entity - whole (car)
2 | entity - whole (roadside)
3 | relation - spatial (car, roadside,
    ↪ park)
4 | action - (car, move)
5 | action - (car, drives away)
6 | relation - temporal (car, starts,
    ↪ after, parked)
7 | relation - temporal (car, drive
    ↪ away, after, parked)
8 | event - temporal (car, move,
    ↪ roadside, park, after)
9 | event - temporal (car, drive away,
    ↪ roadside, park, after)
10 | event - temporal (car, starts,
    ↪ parked, move, drive away)

query: What's unusual in this video?
input: A man is running across a street
    ↪ while carrying a large bag. This
    ↪ is unusual because people
    ↪ typically do not carry large bags
    ↪ while running across streets.
output: 1 | entity - whole (man)
2 | entity - whole (street)
3 | entity - whole (bag)
4 | relation - spatial (man, run,
    ↪ street)
5 | entity - scale (large bag)
6 | relation - spatial (man, carry,
    ↪ large bag)
7 | relation - temporal (man, carry,
    ↪ while, running)
8 | event - ambiguity (man, large bag,
    ↪ carry)
9 | event - temporal (man, run, street,
    ↪ while, carry, large bag)
```

P.2 Fact Extraction for Text-to-Video

```
input: A male skateboarder is trying to
    ↪ pull off a trick on the ramp.
output: 1 | entity - whole
    ↪ (skateboarder)
2 | entity - whole (ramp)
3 | attribute - type (skateboarder,
    ↪ male)
4 | action - (male skateboarder, pull
    ↪ off a trick)
5 | relation - spatial (male
    ↪ skateboarder, ramp, on)
6 | event - ambiguity (skateboarder,
    ↪ male, pull off a trick)
7 | event - ambiguity (male
    ↪ skateboarder, ramp, on)
8 | event - ambiguity (skateboarder,
    ↪ pull off a trick, ramp, on)

input: A car playing soccer, digital
    ↪ art.
output: 1 | entity - whole (car)
2 | global - (digital art)
3 | action - (car, soccer, play)

input: A set of 2x2 emoji icons with
    ↪ happy, angry, surprised and
    ↪ sobbing faces. The emoji icons
    ↪ look like pigs. All of the pigs
    ↪ are wearing crowns.
output: 1 | entity - whole (emoji icons)
2 | other - count (emoji icons, ==4)
3 | attribute - state (emoji icons, 2x2
    ↪ grid)
4 | attribute - type (emoji icons, pig)
5 | attribute - state (emoji_1, happy)
6 | attribute - state (emoji_2, angry)
7 | attribute - state (emoji_3,
    ↪ surprised)
8 | attribute - state (emoji_4, sobbing
    ↪ face)
9 | entity - part (pig's crown)

input: a photo of bear and dining
    ↪ table; dining table is below bear
output: 1 | global - (photo)
2 | entity - whole (bear)
3 | entity - whole (dining table)
4 | relation - spatial (dining table,
    ↪ bear, below)

input: A group of children sitting in
    ↪ the grass with two of them
    ↪ holding a Frisbee .
output: 1 | entity - whole (children)
2 | entity - whole (grass)
3 | entity - whole (frisbee)
4 | attribute - state (children, sit)
5 | relation - spatial (a group of
    ↪ children, grass, sitting in)
6 | entity - part (two of the children)
7 | action - (two of the children,
    ↪ frisbee, hold)

input: the word 'START' written in
    ↪ chalk on a sidewalk
output: 1 | entity - whole (word)
2 | entity - whole (sidewalk)
3 | other - text rendering (word,
    ↪ "START")
4 | attribute - texture (word, chalk)
```

```
5 | relation - spatial (word 'START', sidewalk, on)

input: A pear, orange, and two bananas in a wooden bowl.
output: 1 | entity - whole (pear)
2 | entity - whole (orange)
3 | entity - whole (bananas)
4 | other - count (bananas, ==2)
5 | entity - whole (bowl)
6 | attribute - material (bowl, wood)
7 | relation - spatial (pear, bowl, in)
8 | relation - spatial (orange, bowl, in)
9 | relation - spatial (bananas, bowl, in)
10 | relation - spatial (bananas, bowl, in)
11 | event - ambiguity (pear, orange, bananas, ==2, bowl, in)

input: Closeup picture of the front of a clean motorcycle.
output: 1 | entity - whole (motorcycle)
2 | global - (closeup)
3 | global - (picture)
4 | attribute - state (motorcycle, clean)
5 | entity - part (front of the clean motorcycle)

input: a sad man with green hair
output: 1 | entity - whole (man)
2 | entity - part (man's hair)
3 | attribute - state (man, sad)
4 | attribute - color (man's hair, green)
5 | event - ambiguity (man, sad, man's hair, green)

input: A commercial airplane with propellers flying through the air.
output: 1 | entity - whole (airplane)
2 | entity - part (airplane's propellers)
3 | action - (airplane, air, fly through)
4 | event - ambiguity (airplane, with propellers, air, fly through)

input: A little boy grips a soccer ball in his arms surrounded by other youth soccer players.
output: 1 | entity - whole (boy)
2 | entity - whole (ball)
3 | entity - whole (soccer players)
4 | entity - part (boy's arms)
5 | entity - scale (boy, little)
6 | attribute - type (ball, soccer)
7 | attribute - state (soccer players, youth)
8 | relation - spatial (little boy, ball, grip in his arms)
9 | relation - spatial (little boy gripping the ball in his arms, soccer players, surrounded by)
10 | event - ambiguity (boy's arm, little, ball, soccer, grip in his arms)
11 | event - ambiguity (boy, little, soccer players, youth, surrounded by)

input: A traffic light and a signpost at a crossroads intersection near a waterway.
output: 1 | entity - whole (traffic light)
2 | entity - whole (signpost)
3 | entity - whole (crossroads intersection)
4 | entity - whole (waterway)
5 | relation - spatial (traffic light, crossroads intersection, at)
6 | relation - spatial (signpost, crossroads intersection, at)
7 | relation - spatial (traffic light, waterway, near)
8 | relation - spatial (signpost, waterway, near)
9 | relation - spatial (crossroads intersection, waterway, near)
10 | event - ambiguity (traffic light, signpost, crossroads intersection, at)
11 | event - ambiguity (traffic light, crossroads intersection, at, waterway, near)
12 | event - spatial (signpost, crossroads intersection, at, waterway, near)

input: a photo of dining table and traffic light; traffic light is below dining table
output: 1 | global - (photo)
2 | entity - whole (dining table)
3 | entity - whole (traffic light)
4 | relation - spatial (traffic light, dining table, below)

input: A realistic photo of a Pomeranian dressed up like a 1980s professional wrestler with neon green and neon orange face paint and bright green wrestling tights with bright orange boots.
output: 1 | global - (photo)
2 | entity - whole (Pomeranian)
3 | global - (realistic)
4 | entity - part (Pomeranian's costume)
5 | attribute - type (Pomeranian's costume, 1980s professional wrestler)
6 | entity - part (Pomeranian's costume's wrestling tights)
7 | entity - part (Pomeranian's costume's wrestling tights' boots)
8 | entity - part (Pomeranian's facepaint)
9 | attribute - color (Pomeranian's facepaint, neon green)
10 | attribute - color (Pomeranian's facepaint, neon orange)
11 | attribute - color (Pomeranian's costume's wrestling tights, bright green)
12 | attribute - color (Pomeranian's costume's wrestling tights' boots, bright orange)
```

```
input: a four-piece band on a stage in
    ↪ front of a small crowd
output: 1 | entity - whole (band)
2 | entity - whole (stage)
3 | entity - whole (crowd)
4 | other - count (band members, ==4)
5 | attribute - shape (crowd, small)
6 | relation - spatial (four-piece
    ↪ band, stage, on)
7 | relation - spatial (four-piece
    ↪ band, crowd, in front of)
8 | relation - spatial (stage, crowd,
    ↪ in front of)
9 | event - ambiguity (band, ==4
    ↪ picece, stage, on)
10 | event - ambiguity (band, ==4
    ↪ picece, crowd, small, in front of)
11 | event - ambiguity (stage, crowd,
    ↪ small, in front off)

input: two laptops, a mouse cord, and a
    ↪ monitor
output: 1 | entity - whole (laptops)
2 | other - count (laptops, ==2)
3 | entity - whole (mouse coord)
4 | entity - whole (monitor)

input: A red motorcycle parked by paint
    ↪ chipped doors.
output: 1 | entity - whole (motorcycle)
2 | entity - whole (doors)
3 | attribute - color (motorcycle, red)
4 | attribute - state (door, paint
    ↪ chipped)
5 | relation - spatial (red motorcycle,
    ↪ paint chipped door, next to)
6 | attribute - state (motorcycle,
    ↪ parked)
7 | event- ambiguity (motorcycle, red,
    ↪ door, paint chipped, next to)
8 | event- ambiguity (motorcycle, red,
    ↪ parked)

input: A cube made of denim. A cube
    ↪ with the texture of denim.
output: 1 | entity - whole (cube)
2 | attribute - material (cube, denim)
3 | attribute - texture (cube, denim)

input: an espresso machine that makes
    ↪ coffee from human souls
output: 1 | entity - whole (espresso
    ↪ machine)
2 | entity - whole (coffee)
3 | entity - whole (human souls)
4 | action - (espresso machine, coffee,
    ↪ make)
5 | attribute - material (coffee, human
    ↪ souls)
6 | event - ambiguity (espresso
    ↪ machine, coffee, make, human
    ↪ souls)

input: Three people standing next to an
    ↪ elephant along a river.
output: 1 | entity - whole (people)
2 | other - count (people, ==3)
3 | entity - whole (elephant)
4 | entity - whole (river)
5 | attribute - state (people, stand)
6 | relation - spatial (three people,
    ↪ elephant, next to)
7 | relation - spatial (people, river,
    ↪ next to)
8 | relation - spatial (elephant,
    ↪ river, next to)
9 | event - ambiguity (people, ==3,
    ↪ stand)
10 | event - ambiguity (people, ==3,
    ↪ elephant, next to)
11 | event - ambiguity (people, ==3,
    ↪ river, next to)
12 | event - ambiguity (people, stand,
    ↪ elephant, next to)
13 | event - ambiguity (people, stand,
    ↪ river, next to)
14 | event - ambiguity (people,
    ↪ elephant, next to, river, next to)

input: Aerial view of downtown
    ↪ Manhattan, but with Millennium
    ↪ Wheel next to the Statue of
    ↪ Liberty. The Great Pyramid is on
    ↪ a sandy island near the buildings.
output: 1 | entity - (downtown
    ↪ Manhattan)
2 | entity - (Millennium Wheel)
3 | entity - (the Statue of the Liberty)
4 | entity - (the Great Pyramid)
5 | entity - (island)
6 | entity - (buildings)
7 | global - (aerial view)
8 | attribute - texture (island, sandy)
9 | relation - spatial (Millennium
    ↪ Wheel, the Statue of Liberty,
    ↪ next to)
10 | relation - spatial (the Great
    ↪ Pyramid, island, on)
11 | relation - spatial (the Great
    ↪ Pyramid, buildings, near)
12 | event - ambiguity (the Great
    ↪ Pyramid, island, on, buildings,
    ↪ near)

input: A laptop with external keyboard,
    ↪ mouse, phone and photo on a desk.
output: 1 | entity - whole (laptop)
2 | entity - whole (keyboard)
3 | entity - whole (mouse)
4 | entity - whole (phone)
5 | entity - whole (photo)
6 | entity - whole (desk)
7 | attribute - type (keyboard,
    ↪ external)
8 | relation - spatial (laptop, desk,
    ↪ on)
9 | relation - spatial (keyboard, desk,
    ↪ on)
10 | relation - spatial (mouse, desk,
    ↪ on)
11 | relation - spatial (phone, desk,
    ↪ on)
12 | relation - spatial (photo, desk,
    ↪ on)
13 | event - ambiguity (laptop,
    ↪ external keyboard, mouse, phone,
    ↪ photo, desk, on)

input: A white slope covers the
```

```
    ↪ background, while the foreground
    ↪ features a grassy slope with
    ↪ several rams grazing and one
    ↪ measly and underdeveloped
    ↪ evergreen in the foreground.
output: 1 | entity - whole (slopes)
2 | other - count (slopes, ==2)
3 | entity - whole (rams)
4 | entity - whole (evergreen)
5 | attribute - color (slope_1, white)
6 | attribute - texture (slope_2,
    ↪ grassy)
7 | attribute - state (evergreen,
    ↪ measly and underdeveloped)
8 | relation - spatial (slope_1,
    ↪ background, in)
9 | relation - spatial (slope_2,
    ↪ foreground, in)
10 | relation - spatial (several, rams,
    ↪ grassy slope_2, on)
11 | attribute - state (several rams,
    ↪ graze)
12 | event - ambiguity (slope_1, white,
    ↪ background, in)
13 | event - ambiguity (slope_2,
    ↪ grassy, foreground, in)
14 | event - ambiguity (several, rams,
    ↪ slope_2, grassy, on)

input: A man walks into a room and sits
    ↪ on a chair. A dog follows him.
output: 1 | entity - whole (man)
2 | entity - whole (room)
3 | entity - whole (chair)
4 | entity - whole (dog)
5 | action - (man, walk, room)
6 | action - (man, sit on, chair)
7 | action - (dog, follow, man)
8 | relation - temporal (man, sit,
    ↪ before, walk)
9 | relation - temporal (dog, follows,
    ↪ after, man, sit)
10 | event - temporal (man, walks into
    ↪ a room and sits on a chair, dog
    ↪ follows him)

input: A car is parked by the roadside.
    ↪ Later, it starts moving and
    ↪ drives away.
output: 1 | entity - whole (car)
2 | entity - whole (roadside)
3 | relation - spatial (car, roadside,
    ↪ park)
4 | action - (car, move)
5 | action - (car, drives away)
6 | relation - temporal (car, starts,
    ↪ after, parked)
7 | relation - temporal (car, drive
    ↪ away, after, parked)
8 | event - temporal (car, move,
    ↪ roadside, park, after)
9 | event - temporal (car, drive away,
    ↪ roadside, park, after)
10 | event - temporal (car, starts,
    ↪ parked, move, drive away)

input: A man is running across a street
    ↪ while carrying a large bag. This
    ↪ is unusual because people
    ↪ typically do not carry large bags
    ↪ while running across streets.
output: 1 | entity - whole (man)
2 | entity - whole (street)
3 | entity - whole (bag)
4 | relation - spatial (man, run,
    ↪ street)
5 | entity - scale (large bag)
6 | relation - spatial (man, carry,
    ↪ large bag)
7 | relation - temporal (man, carry,
    ↪ while, running)
8 | event - ambiguity (man, large bag,
    ↪ carry)
9 | event - temporal (man, run, street,
    ↪ while, carry, large bag)
```

P.3 Question Generation

```
input: A male skateboarder is trying to
    ↪ pull off a trick on the ramp.
1 | entity - whole (skateboarder)
2 | entity - whole (ramp)
3 | attribute - type (skateboarder,
    ↪ male)
4 | action - (male skateboarder, pull
    ↪ off a trick)
5 | relation - spatial (male
    ↪ skateboarder, ramp, on)
6 | event - ambigity (skateboarder,
    ↪ male, pull off a trick)
7 | event - ambiguity (male
    ↪ skateboarder, ramp, on)
8 | event - ambiguity (skateboarder,
    ↪ pull off a trick, ramp, on)
output: 1 | Is there a skateboarder?
2 | Is there a ramp?
3 | Is the skateboarder male?
4 | Is the skateboarder pulling off a
    ↪ trick?
5 | Is the skateboarder on the ramp?
6 | Is the male skateboarder on the
    ↪ ramp?
7 | Is the male skateboarder on the
    ↪ ramp?
8 | Is the skateboarder pulling off a
    ↪ trick on the ramp?

input: A car playing soccer, digital
    ↪ art.
1 | entity - whole (car)
2 | global - (digital art)
3 | action - (car, soccer, play)
output: 1 | Is there a car?
2 | Is this digital art?
3 | Is the car playing soccer?

input: A set of 2x2 emoji icons with
    ↪ happy, angry, surprised and
    ↪ sobbing faces. The emoji icons
    ↪ look like pigs. All of the pigs
    ↪ are wearing crowns.
1 | entity - whole (emoji icons)
2 | other - count (emoji icons, ==4)
3 | attribute - state (emoji icons, 2x2
    ↪ grid)
4 | attribute - type (emoji icons, pig)
5 | attribute - state (emoji_1, happy)
6 | attribute - state (emoji_2, angry)
```

```
7 | attribute - state (emoji_3,
    ↪ surprised)
8 | attribute - state (emoji_4, sobbing
    ↪ face)
9 | entity - part (pig's crown)
output: 1 | nan
2 | Is there a total of four emoji
    ↪ icons?
3 | Were the emojis in a 2x2 grid?
4 | Did emojis look like pigs?
5 | Did one emoji look happy?
6 | Did one emoji look angry?
7 | Did one emoji look surprised?
8 | Did the emoji have a sobbing face?
9 | Are all the emoji wearing crowns?

input: a photo of bear and dining
    ↪ table; dining table is below bear
1 | global - (photo)
2 | entity - whole (bear)
3 | entity - whole (dining table)
4 | relation - spatial (dining table,
    ↪ bear, below)
output: 1 | Is this a photo?
2 | Is there a bear?
3 | Is there a dining table?
4 | Is the dining table below the bear?

input: A group of children sitting in
    ↪ the grass with two of them
    ↪ holding a Frisbee .
1 | entity - whole (children)
2 | entity - whole (grass)
3 | entity - whole (frisbee)
4 | attribute - state (children, sit)
5 | relation - spatial (a group of
    ↪ children, grass, sitting in)
6 | entity - part (two of the children)
7 | action - (two of the children,
    ↪ frisbee, hold)
output: 1 | Are there a group of
    ↪ children?
2 | Is there grass?
3 | Is there a frisbee?
4 | Are the children sitting?
5 | Are a group of children sitting in
    ↪ the grass?
6 | Are there two of the children?
7 | Are two of the children holding a
    ↪ frisbee?

input: the word 'START' written in
    ↪ chalk on a sidewalk
1 | entity - whole (word)
2 | entity - whole (sidewalk)
3 | other - text rendering (word,
    ↪ "START")
4 | attribute - texture (word, chalk)
5 | relation - spatial (word 'START',
    ↪ sidewalk, on)
output: 1 | Is there a word?
2 | Is there a sidewalk?
3 | Does the word say "START"?
4 | Is the word written in chalk?
5 | Is the word 'START' on the sidewalk?

input: A pear, orange, and two bananas
    ↪ in a wooden bowl.
1 | entity - whole (pear)
2 | entity - whole (orange)
3 | entity - whole (bananas)
4 | other - count (bananas, ==2)
5 | entity - whole (bowl)
6 | attribute - material (bowl, wood)
7 | relation - spatial (pear, bowl, in)
8 | relation - spatial (orange, bowl,
    ↪ in)
9 | relation - spatial (bananas, bowl,
    ↪ in)
10 | relation - spatial (bananas, bowl,
    ↪ in)
11 | event - ambiguity (pear, orange,
    ↪ bananas, ==2, bowl, in)
output: 1 | Is there a pear?
2 | Is there an orange?
3 | Are there bananas?
4 | Are there two bananas?
5 | Is there a bowl?
6 | Is the bowl made of wood?
7 | Is the pear in the wooden bowl?
8 | Is the orange in the wooden bowl?
9 | Are bananas in the wooden bowl?
10 | Are bananas in the wooden bowl?
11 | Are the pear, the orange and two
    ↪ bananas bananas in the same
    ↪ wooden bowl?

input: Closeup picture of the front of
    ↪ a clean motorcycle.
1 | entity - whole (motorcycle)
2 | global - (closeup)
3 | global - (picture)
4 | attribute - state (motorcycle,
    ↪ clean)
5 | entity - part (front of the clean
    ↪ motorcycle)
output: 1 | Is there a motorcycle?
2 | Is this a closeup image?
3 | Is this a picture?
4 | Is the motorcycle clean?
5 | Is the closeup picture in the front
    ↪ of the clean motorcycle?

input: a sad man with green hair
1 | entity - whole (man)
2 | entity - part (man's hair)
3 | attribute - state (man, sad)
4 | attribute - color (man's hair,
    ↪ green)
5 | event - ambiguity (man, sad, man's
    ↪ hair, green)
output: 1 | Is there a man?
2 | Is there hair?
3 | Was the man sad?
4 | Is the hair green?
5 | Is the sad man with hair green?

input: A commercial airplane with
    ↪ propellers flying through the air.
1 | entity - whole (airplane)
2 | entity - part (airplane's
    ↪ propellers)
3 | action - (airplane, air, fly
    ↪ through)
4 | event - ambiguity (airplane, with
    ↪ propellers, air, fly through)
output: 1 | Is there an airplane?
2 | Does the airplane have propellers?
3 | Is the airplane flying through the
    ↪ air?
```

```
4 | Is the airplane with propellers
    ↪ flying through the air?

input: A little boy grips a soccer ball
    ↪ in his arms surrounded by other
    ↪ youth soccer players.
1 | entity - whole (boy)
2 | entity - whole (ball)
3 | entity - whole (soccer players)
4 | entity - part (boy's arms)
5 | entity - scale (boy, little)
6 | attribute - type (ball, soccer)
7 | attribute - state (soccer players,
    ↪ youth)
8 | relation - spatial (little boy,
    ↪ ball, grip in his arms)
9 | relation - spatial (little boy
    ↪ gripping the ball in his arms,
    ↪ soccer players, surrounded by)
10 | event - ambiguity (boy's arm,
    ↪ little, ball, soccer, grip in his
    ↪ arms)
11 | event - ambiguity (boy, little,
    ↪ soccer players, youth, surrounded
    ↪ by)
output: 1 | Is there a boy?
2 | Is there a ball?
3 | Are there other soccer players?
4 | Does the boy have arms?
5 | Is the boy little?
6 | Is the ball a soccer ball?
7 | Are the other soccer players young?
8 | Is the boy gripping the ball in his
    ↪ arms?
9 | Is the little boy surrounded by the
    ↪ other soccer players?
10 | Is the little boy gripping the
    ↪ soccer ball in his arms?
11 | Is the little boy surrounded by
    ↪ the other youth soccer players?

input: A traffic light and a signpost
    ↪ at a crossroads intersection near
    ↪ a waterway.
1 | entity - whole (traffic light)
2 | entity - whole (signpost)
3 | entity - whole (crossroads
    ↪ intersection)
4 | entity - whole (waterway)
5 | relation - spatial (traffic light,
    ↪ crossroads intersection, at)
6 | relation - spatial (signpost,
    ↪ crossroads intersection, at)
7 | relation - spatial (traffic light,
    ↪ waterway, near)
8 | relation - spatial (signpost,
    ↪ waterway, near)
9 | relation - spatial (crossroads
    ↪ intersection, waterway, near)
10 | event - ambiguity (traffic light,
    ↪ signpost, crossroads
    ↪ intersection, at)
11 | event - ambiguity (traffic light,
    ↪ crossroads intersection, at,
    ↪ waterway, near)
12 | event - spatial (signpost,
    ↪ crossroads intersection, at,
    ↪ waterway, near)
output: 1 | Is there a light?
2 | Is there a signpost?
3 | Is there an intersection?
4 | Is there a waterway?
5 | Is the light a traffic light?
6 | Is the intersection a crossroads
    ↪ intersection?
7 | Is the traffic light at the
    ↪ crossroads intersection?
8 | Is the signpost at the crossroads
    ↪ intersection?
9 | Is the intersection near the
    ↪ waterway?
10 | Are the traffic light and signpost
    ↪ at a crossrodas intersection?
11 | Is the traffic light at a
    ↪ crossrodas intersection near
    ↪ waterway?
12 | Is the signpost at a crossrodas
    ↪ intersection near waterway?

input: a photo of dining table and
    ↪ traffic light; traffic light is
    ↪ below dining table
1 | global - (photo)
2 | entity - whole (dining table)
3 | entity - whole (traffic light)
4 | relation - spatial (traffic light,
    ↪ dining table, below)
output: 1 | Is this a photo?
2 | Is there a dining table?
3 | Is there a traffic light?
4 | Is the traffice light below the
    ↪ dining table?

input: A realistic photo of a
    ↪ Pomeranian dressed up like a
    ↪ 1980s professional wrestler with
    ↪ neon green and neon orange face
    ↪ paint and bright green wrestling
    ↪ tights with bright orange boots.
1 | global - (photo)
2 | entity - whole (Pomeranian)
3 | global - (realistic)
4 | entity - part (Pomeranian's costume)
5 | attribute - type (Pomeranian's
    ↪ costume, 1980s professional
    ↪ wrestler)
6 | entity - part (Pomeranian's
    ↪ costume's wrestling tights)
7 | entity - part (Pomeranian's
    ↪ costume's wrestling tights' boots)
8 | entity - part (Pomeranian's
    ↪ facepaint)
9 | attribute - color (Pomeranian's
    ↪ facepaint, neon green)
10 | attribute - color (Pomeranian's
    ↪ facepaint, neon orange)
11 | attribute - color (Pomeranian's
    ↪ costume's wrestling tights,
    ↪ bright green)
12 | attribute - color (Pomeranian's
    ↪ costume's wrestling tights'
    ↪ boots, bright orange)
output: 1 | Is this a photo?
2 | Is there a Pomeranian?
3 | Is the photo realistic?
4 | Is the Pomeranian dressed up?
5 | Is the costume of a 1980s
    ↪ professional wrestler?
6 | Are wrestling tights included in
    ↪ the costume?
```

```
7 | Did the costume come with boots?
8 | Does the Pomeranian has a facepaint?
9 | Is the facepaint neon green?
10 | Is the facepaint neon orange?
11 | Are the wrestling tights bright
    ↪ green?
12 | Are the boots bright orange?

input: a four-piece band on a stage in
    ↪ front of a small crowd
1 | entity - whole (band)
2 | entity - whole (stage)
3 | entity - whole (crowd)
4 | other - count (band members, ==4)
5 | attribute - shape (crowd, small)
6 | relation - spatial (four-piece
    ↪ band, stage, on)
7 | relation - spatial (four-piece
    ↪ band, crowd, in front of)
8 | relation - spatial (stage, crowd,
    ↪ in front of)
9 | event - ambiguity (band, ==4
    ↪ picece, stage, on)
10 | event - ambiguity (band, ==4
    ↪ picece, crowd, small, in front of)
11 | event - ambiguity (stage, crowd,
    ↪ small, in front off)
output: 1 | Is there a band?
2 | Is there a stage?
3 | Is there a crowd?
4 | Is the band a fourpiece band?
5 | Is the crowd small?
6 | Is the band on the stage?
7 | Is the band in front of the crowd?
8 | Is the stage in front of the crowd?
9 | Are the four-piece band on the
    ↪ stage?
10 | Is the four-piece band in front of
    ↪ the small crowd?
11 | Is the stage in front of the small
    ↪ crowd?

input: two laptops, a mouse cord, and a
    ↪ monitor
1 | entity - whole (laptops)
2 | other - count (laptops, ==2)
3 | entity - whole (mouse coord)
4 | entity - whole (monitor)
output: 1 | Are there laptops?
2 | Are there two laptops?
3 | Is there a cord?
4 | Is there a monitor?

input: A red motorcycle parked by paint
    ↪ chipped doors.
1 | entity - whole (motorcycle)
2 | entity - whole (doors)
3 | attribute - color (motorcycle, red)
4 | attribute - state (door, paint
    ↪ chipped)
5 | relation - spatial (red motorcycle,
    ↪ paint chipped door, next to)
6 | attribute - state (motorcycle,
    ↪ parked)
7 | event- ambiguity (motorcycle, red,
    ↪ door, paint chipped, next to)
8 | event- ambiguity (motorcycle, red,
    ↪ parked)
output: 1 | Is there a motorcycle?
2 | Are there any doors?
3 | Are the doors painted?
4 | Is the paint chipped?
5 | Is the motorcycle next to doors?
6 | Is the motorcycle parked?
7 | Is the red motorcycle next to paint
    ↪ chipped doors?
8 | Is the red motorcycle parked?

input: A cube made of denim. A cube
    ↪ with the texture of denim.
1 | entity - whole (cube)
2 | attribute - material (cube, denim)
3 | attribute - texture (cube, denim)
output: 1 | Is there a cube?
2 | Is the cube made of denim?
3 | Does the cube have texture of denim?

input: an espresso machine that makes
    ↪ coffee from human souls
1 | entity - whole (espresso machine)
2 | entity - whole (coffee)
3 | entity - whole (human souls)
4 | action - (espresso machine, coffee,
    ↪ make)
5 | attribute - material (coffee, human
    ↪ souls)
6 | event - ambiguity (espresso
    ↪ machine, coffee, make, human
    ↪ souls)
output: 1 | Do we have an espresso
    ↪ machine?
2 | Do we have coffee?
3 | Do human beings have souls?
4 | Is the espresso machine making
    ↪ coffee?
5 | Is the expersso made of human souls?
6 | Is the expersso machine making
    ↪ coffe with human souls?

input: Three people standing next to an
    ↪ elephant along a river.
1 | entity - whole (people)
2 | other - count (people, ==3)
3 | entity - whole (elephant)
4 | entity - whole (river)
5 | attribute - state (people, stand)
6 | relation - spatial (three people,
    ↪ elephant, next to)
7 | relation - spatial (people, river,
    ↪ next to)
8 | relation - spatial (elephant,
    ↪ river, next to)
9 | event - ambiguity (people, ==3,
    ↪ stand)
10 | event - ambiguity (people, ==3,
    ↪ elephant, next to)
11 | event - ambiguity (people, ==3,
    ↪ river, next to)
12 | event - ambiguity (people, stand,
    ↪ elephant, next to)
13 | event - ambiguity (people, stand,
    ↪ river, next to)
14 | event - ambiguity (people,
    ↪ elephant, next to, river, next to)
output: 1 | Are there people?
2 | Are there three people?
3 | Is there an elephant?
4 | Is there a river?
5 | Are people standing?
6 | Are people next to the elephant?
```

```
7 | Are people next to the river?
8 | Is the elephant next to the river?
9 | Are the three people standing?
10 | Are the three people next to the
    ↪ elephant?
11 | Are the three people next to the
    ↪ river?
12 | Are people stadning next to an
    ↪ elephant?
13 | Are people standing next to the
    ↪ river?
14 | Are people next to the river and
    ↪ an elephant?

input: Aerial view of downtown
    ↪ Manhattan, but with Millennium
    ↪ Wheel next to the Statue of
    ↪ Liberty. The Great Pyramid is on
    ↪ a sandy island near the buildings.
1 | entity - (downtown Manhattan)
2 | entity - (Millennium Wheel)
3 | entity - (the Statue of the Liberty)
4 | entity - (the Great Pyramid)
5 | entity - (island)
6 | entity - (buildings)
7 | global - (aerial view)
8 | attribute - texture (island, sandy)
9 | relation - spatial (Millennium
    ↪ Wheel, the Statue of Liberty,
    ↪ next to)
10 | relation - spatial (the Great
    ↪ Pyramid, island, on)
11 | relation - spatial (the Great
    ↪ Pyramid, buildings, near)
12 | event - ambiguity (the Great
    ↪ Pyramid, island, on, buildings,
    ↪ near)
output: 1 | Is downtown Manhattan there?
2 | Is Millennium Wheel there?
3 | Is the Statue of Liberty there?
4 | Is the Great Pyramid there?
5 | Is there an island?
6 | Are there buildings?
7 | Is this an aerial view?
8 | Is there the island sandy?
9 | Is the Millennium Wheel next to the
    ↪ Statue of Liberty?
10 | Is the Great Pyramid on the sandy
    ↪ island?
11 | Is the Great Pyramid near the
    ↪ buildings?
12 | Is the Great Pyramid on a sady
    ↪ island near the buildings?

input: A laptop with external keyboard,
    ↪ mouse, phone and photo on a desk.
1 | entity - whole (laptop)
2 | entity - whole (keyboard)
3 | entity - whole (mouse)
4 | entity - whole (phone)
5 | entity - whole (photo)
6 | entity - whole (desk)
7 | attribute - type (keyboard,
    ↪ external)
8 | relation - spatial (laptop, desk,
    ↪ on)
9 | relation - spatial (keyboard, desk,
    ↪ on)
10 | relation - spatial (mouse, desk,
    ↪ on)
11 | relation - spatial (phone, desk,
    ↪ on)
12 | relation - spatial (photo, desk,
    ↪ on)
13 | event - ambiguity (laptop,
    ↪ external keyboard, mouse, phone,
    ↪ photo, desk, on)
output: 1 | Is there a laptop?
2 | Is there a keyboard?
3 | Is there a mouse?
4 | Is there a phone?
5 | Is there a photo?
6 | Is there a desk?
7 | Is the keyboard external?
8 | Is the laptop on the desk?
9 | Is the keyboard on the desk?
10 | Is the mouse on the desk?
11 | Is the phone on the desk?
12 | Is the photo on the desk?
13 | Is all laptop, external keyboard,
    ↪ mouse, phone, photo on the same
    ↪ desk?

input: A white slope covers the
    ↪ background, while the foreground
    ↪ features a grassy slope with
    ↪ several rams grazing and one
    ↪ measly and underdeveloped
    ↪ evergreen in the foreground.
1 | entity - whole (slopes)
2 | other - count (slopes, ==2)
3 | entity - whole (rams)
4 | entity - whole (evergreen)
5 | attribute - color (slope_1, white)
6 | attribute - texture (slope_2,
    ↪ grassy)
7 | attribute - state (evergreen,
    ↪ measly and underdeveloped)
8 | relation - spatial (slope_1,
    ↪ background, in)
9 | relation - spatial (slope_2,
    ↪ foreground, in)
10 | relation - spatial (several, rams,
    ↪ grassy slope_2, on)
11 | attribute - state (several rams,
    ↪ graze)
12 | event - ambiguity (slope_1, white,
    ↪ background, in)
13 | event - ambiguity (slope_2,
    ↪ grassy, foreground, in)
14 | event - ambiguity (several, rams,
    ↪ slope_2, grassy, on)
output: 1 | Are there slopes?
2 | Are there two slopes?
3 | Are there rams?
4 | Is there evergreen?
5 | Is one slope white?
6 | Is one slope grassy?
7 | Is the evergreen measly and
    ↪ underdeveloped?
8 | Is the slope in the background?
9 | Is the slope in the foreground?
10 | Are the several rams on the slope?
11 | Are the several rams grazing on
    ↪ grass?
12 | Is the white slope in the
    ↪ background?
13 | Is the grassy slope in the
    ↪ foreground?
14 | Are the several rams on the grassy
```

```
    ↪ slope?

input: A man walks into a room and sits
    ↪ on a chair. A dog follows him.
1 | entity - whole (man)
2 | entity - whole (room)
3 | entity - whole (chair)
4 | entity - whole (dog)
5 | action - (man, walk, room)
6 | action - (man, sit on, chair)
7 | action - (dog, follow, man)
8 | relation - temporal (man, sit,
    ↪ before, walk)
9 | relation - temporal (dog, follows,
    ↪ after, man, sit)
10 | event - temporal (man, walks into
    ↪ a room and sits on a chair, dog
    ↪ follows him)
output: 1 | Is there a man?
2 | Is there a room?
3 | Is there a chair?
4 | Is there a dog?
5 | Does the man walk into the room?
6 | Does the man sit on the chair?
7 | Does the dog follow the man?
8 | Does the man sit before walking?
9 | Does the dog follow after the man
    ↪ sat?
10 | Does the dog follow after the man
    ↪ who walked into a room and sits
    ↪ on a chair?

input: A car is parked by the roadside.
    ↪ Later, it starts moving and
    ↪ drives away.
1 | entity - whole (car)
2 | entity - whole (roadside)
3 | relation - spatial (car, roadside,
    ↪ park)
4 | action - (car, move)
5 | action - (car, drives away)
6 | relation - temporal (car, starts,
    ↪ after, parked)
7 | relation - temporal (car, drive
    ↪ away, after, parked)
8 | event - temporal (car, move,
    ↪ roadside, park, after)
9 | event - temporal (car, drive away,
    ↪ roadside, park, after)
10 | event - temporal (car, starts,
    ↪ parked, move, drive away)
output: 1 | Is there a car?
2 | Is there a roadside?
3 | Does the car park near the roadside?
4 | Des the car move?
5 | Does the car drive away?
6 | Does the car move after being
    ↪ parked?
7 | Does the car drive away after being
    ↪ parked?
8 | Does the car move after being
    ↪ parked near roadside?
9 | Does the car drive away after being
    ↪ parked near roadside?
10 | Is that a same car which parked by
    ↪ the roadsid and then starts
    ↪ moving and drives away?

input: A man is running across a street
    ↪ while carrying a large bag. This
    ↪ is unusual because people
    ↪ typically do not carry large bags
    ↪ while running across streets.
1 | entity - whole (man)
2 | entity - whole (street)
3 | entity - whole (bag)
4 | relation - spatial (man, run,
    ↪ street)
5 | entity - scale (large bag)
6 | relation - spatial (man, carry,
    ↪ large bag)
7 | relation - temporal (man, carry,
    ↪ while, running)
8 | event - ambiguity (man, large bag,
    ↪ carry)
9 | event - temporal (man, run, street,
    ↪ while, carry, large bag)
output: 1 | Is there a man?
2 | Is there a street?
3 | Is there a bag?
4 | Is the man running across the
    ↪ street?
5 | Is a bag large?
6 | Is the man carrying a bag?
7 | Is the man carrying a bag while
    ↪ running?
8 | Is the man carrying a large bag?
9 | Is the man carrying a big bag while
    ↪ running across a street?
```

P.4 Dependency Generation

```
input: A male skateboarder is trying to
    ↪ pull off a trick on the ramp.
1 | entity - whole (skateboarder)
2 | entity - whole (ramp)
3 | attribute - type (skateboarder,
    ↪ male)
4 | action - (male skateboarder, pull
    ↪ off a trick)
5 | relation - spatial (male
    ↪ skateboarder, ramp, on)
6 | event - ambigity (skateboarder,
    ↪ male, pull off a trick)
7 | event - ambiguity (male
    ↪ skateboarder, ramp, on)
8 | event - ambiguity (skateboarder,
    ↪ pull off a trick, ramp, on)
output: 1 | 0
2 | 0
3 | 1
4 | 1
5 | 1,3
6 | 3,4
7 | 3,5
8 | 4,5

input: A car playing soccer, digital
    ↪ art.
1 | entity - whole (car)
2 | global - (digital art)
3 | action - (car, soccer, play)
output: 1 | 0
2 | 0
3 | 1

input: A set of 2x2 emoji icons with
    ↪ happy, angry, surprised and
    ↪ sobbing faces. The emoji icons
```

```
    ↪ look like pigs. All of the pigs
    ↪ are wearing crowns.
1 | entity - whole (emoji icons)
2 | other - count (emoji icons, ==4)
3 | attribute - state (emoji icons, 2x2
    ↪ grid)
4 | attribute - type (emoji icons, pig)
5 | attribute - state (emoji_1, happy)
6 | attribute - state (emoji_2, angry)
7 | attribute - state (emoji_3,
    ↪ surprised)
8 | attribute - state (emoji_4, sobbing
    ↪ face)
9 | entity - part (pig's crown)
output: 1 | 0
2 | 0
3 | 1
4 | 1
5 | 1
6 | 1
7 | 1
8 | 1
9 | 1,4

input: a photo of bear and dining
    ↪ table; dining table is below bear
1 | global - (photo)
2 | entity - whole (bear)
3 | entity - whole (dining table)
4 | relation - spatial (dining table,
    ↪ bear, below)
output: 1 | 0
2 | 0
3 | 0
4 | 2,3

input: A group of children sitting in
    ↪ the grass with two of them
    ↪ holding a Frisbee .
1 | entity - whole (children)
2 | entity - whole (grass)
3 | entity - whole (frisbee)
4 | attribute - state (children, sit)
5 | relation - spatial (a group of
    ↪ children, grass, sitting in)
6 | entity - part (two of the children)
7 | action - (two of the children,
    ↪ frisbee, hold)
output: 1 | 0
2 | 0
3 | 0
4 | 1
5 | 1,2
6 | 1
7 | 3,6

input: the word 'START' written in
    ↪ chalk on a sidewalk
1 | entity - whole (word)
2 | entity - whole (sidewalk)
3 | other - text rendering (word,
    ↪ "START")
4 | attribute - texture (word, chalk)
5 | relation - spatial (word 'START',
    ↪ sidewalk, on)
output: 1 | 0
2 | 0
3 | 1
4 | 1
5 | 2,3
```

```
input: A pear, orange, and two bananas
    ↪ in a wooden bowl.
1 | entity - whole (pear)
2 | entity - whole (orange)
3 | entity - whole (bananas)
4 | other - count (bananas, ==2)
5 | entity - whole (bowl)
6 | attribute - material (bowl, wood)
7 | relation - spatial (pear, bowl, in)
8 | relation - spatial (orange, bowl,
    ↪ in)
9 | relation - spatial (bananas, bowl,
    ↪ in)
10 | relation - spatial (bananas, bowl,
    ↪ in)
11 | event - ambiguity (pear, orange,
    ↪ bananas, ==2, bowl, in)
output: 1 | 0
2 | 0
3 | 0
4 | 0
5 | 0
6 | 0
7 | 1,5
8 | 2,5
9 | 3,5
10 | 4,9
11 | 7,8,10

input: Closeup picture of the front of
    ↪ a clean motorcycle.
1 | entity - whole (motorcycle)
2 | global - (closeup)
3 | global - (picture)
4 | attribute - state (motorcycle,
    ↪ clean)
5 | entity - part (front of the clean
    ↪ motorcycle)
output: 1 | 0
2 | 0
3 | 0
4 | 0
5 | 1

input: a sad man with green hair
1 | entity - whole (man)
2 | entity - part (man's hair)
3 | attribute - state (man, sad)
4 | attribute - color (man's hair,
    ↪ green)
5 | event - ambiguity (man, sad, man's
    ↪ hair, green)
output: 1 | 0
2 | 1
3 | 1
4 | 2
5 | 3,4

input: A commercial airplane with
    ↪ propellers flying through the air.
1 | entity - whole (airplane)
2 | entity - part (airplane's
    ↪ propellers)
3 | action - (airplane, air, fly
    ↪ through)
4 | event - ambiguity (airplane, with
    ↪ propellers, air, fly through)
output: 1 | 0
2 | 0
```

```
3 | 1
4 | 2,3

input: A little boy grips a soccer ball
    ↪ in his arms surrounded by other
    ↪ youth soccer players.
1 | entity - whole (boy)
2 | entity - whole (ball)
3 | entity - whole (soccer players)
4 | entity - part (boy’s arms)
5 | entity - scale (boy, little)
6 | attribute - type (ball, soccer)
7 | attribute - state (soccer players,
    ↪ youth)
8 | relation - spatial (little boy,
    ↪ ball, grip in his arms)
9 | relation - spatial (little boy
    ↪ gripping the ball in his arms,
    ↪ soccer players, surrounded by)
10 | event - ambiguity (boy’s arm,
    ↪ little, ball, soccer, grip in his
    ↪ arms)
11 | event - ambiguity (boy, little,
    ↪ soccer players, youth, surrounded
    ↪ by)
output: 1 | 0
2 | 0
3 | 0
4 | 0
5 | 1
6 | 1
7 | 3
8 | 2,4
9 | 1,3
10 | 4,5,6,8
11 | 5,7,9

input: A traffic light and a signpost
    ↪ at a crossroads intersection near
    ↪ a waterway.
1 | entity - whole (traffic light)
2 | entity - whole (signpost)
3 | entity - whole (crossroads
    ↪ intersection)
4 | entity - whole (waterway)
5 | relation - spatial (traffic light,
    ↪ crossroads intersection, at)
6 | relation - spatial (signpost,
    ↪ crossroads intersection, at)
7 | relation - spatial (traffic light,
    ↪ waterway, near)
8 | relation - spatial (signpost,
    ↪ waterway, near)
9 | relation - spatial (crossroads
    ↪ intersection, waterway, near)
10 | event - ambiguity (traffic light,
    ↪ signpost, crossroads
    ↪ intersection, at)
11 | event - ambiguity (traffic light,
    ↪ crossroads intersection, at,
    ↪ waterway, near)
12 | event - spatial (signpost,
    ↪ crossroads intersection, at,
    ↪ waterway, near)
output: 1 | 0
2 | 0
3 | 0
4 | 0
5 | 1,3
6 | 2,3
7 | 1,4
8 | 2,4
9 | 3,4
10 | 5,6
11 | 5,7
12 | 6,8

input: a photo of dining table and
    ↪ traffic light; traffic light is
    ↪ below dining table
1 | global - (photo)
2 | entity - whole (dining table)
3 | entity - whole (traffic light)
4 | relation - spatial (traffic light,
    ↪ dining table, below)
output: 1 | 0
2 | 0
3 | 0
4 | 2,3

input: A realistic photo of a
    ↪ Pomeranian dressed up like a
    ↪ 1980s professional wrestler with
    ↪ neon green and neon orange face
    ↪ paint and bright green wrestling
    ↪ tights with bright orange boots.
1 | global - (photo)
2 | entity - whole (Pomeranian)
3 | global - (realistic)
4 | entity - part (Pomeranian’s costume)
5 | attribute - type (Pomeranian’s
    ↪ costume, 1980s professional
    ↪ wrestler)
6 | entity - part (Pomeranian’s
    ↪ costume’s wrestling tights)
7 | entity - part (Pomeranian’s
    ↪ costume’s wrestling tights’ boots)
8 | entity - part (Pomeranian’s
    ↪ facepaint)
9 | attribute - color (Pomeranian’s
    ↪ facepaint, neon green)
10 | attribute - color (Pomeranian’s
    ↪ facepaint, neon orange)
11 | attribute - color (Pomeranian’s
    ↪ costume’s wrestling tights,
    ↪ bright green)
12 | attribute - color (Pomeranian’s
    ↪ costume’s wrestling tights’
    ↪ boots, bright orange)
output: 1 | 0
2 | 0
3 | 0
4 | 2
5 | 4
6 | 4
7 | 4
8 | 2
9 | 8
10 | 8
11 | 6
12 | 7

input: a four-piece band on a stage in
    ↪ front of a small crowd
1 | entity - whole (band)
2 | entity - whole (stage)
3 | entity - whole (crowd)
4 | other - count (band members, ==4)
5 | attribute - shape (crowd, small)
6 | relation - spatial (four-piece
```

```
    ↪ band, stage, on)
7 | relation - spatial (four-piece
    ↪ band, crowd, in front of)
8 | relation - spatial (stage, crowd,
    ↪ in front of)
9 | event - ambiguity (band, ==4
    ↪ picece, stage, on)
10 | event - ambiguity (band, ==4
    ↪ picece, crowd, small, in front of)
11 | event - ambiguity (stage, crowd,
    ↪ small, in front off)
output: 1 | 0
2 | 0
3 | 0
4 | 1
5 | 3
6 | 2,4
7 | 3,4
8 | 2,3
9 | 2,4
10 | 4,5,7
11 | 2,5,8

input: two laptops, a mouse cord, and a
    ↪ monitor
1 | entity - whole (laptops)
2 | other - count (laptops, ==2)
3 | entity - whole (mouse coord)
4 | entity - whole (monitor)
output: 1 | 0
2 | 0
3 | 0
4 | 0

input: A red motorcycle parked by paint
    ↪ chipped doors.
1 | entity - whole (motorcycle)
2 | entity - whole (doors)
3 | attribute - color (motorcycle, red)
4 | attribute - state (door, paint
    ↪ chipped)
5 | relation - spatial (red motorcycle,
    ↪ paint chipped door, next to)
6 | attribute - state (motorcycle,
    ↪ parked)
7 | event- ambiguity (motorcycle, red,
    ↪ door, paint chipped, next to)
8 | event- ambiguity (motorcycle, red,
    ↪ parked)
output: 1 | 0
2 | 0
3 | 0
4 | 1
5 | 2
6 | 2,3
7 | 3,4,5
8 | 3,6

input: A cube made of denim. A cube
    ↪ with the texture of denim.
1 | entity - whole (cube)
2 | attribute - material (cube, denim)
3 | attribute - texture (cube, denim)
output: 1 | 0
2 | 1
3 | 1

input: an espresso machine that makes
    ↪ coffee from human souls
1 | entity - whole (espresso machine)
2 | entity - whole (coffee)
3 | entity - whole (human souls)
4 | action - (espresso machine, coffee,
    ↪ make)
5 | attribute - material (coffee, human
    ↪ souls)
6 | event - ambiguity (espresso
    ↪ machine, coffee, make, human
    ↪ souls)
output: 1 | 0
2 | 0
3 | 0
4 | 1,2
5 | 2,3
6 | 4,5

input: Three people standing next to an
    ↪ elephant along a river.
1 | entity - whole (people)
2 | other - count (people, ==3)
3 | entity - whole (elephant)
4 | entity - whole (river)
5 | attribute - state (people, stand)
6 | relation - spatial (three people,
    ↪ elephant, next to)
7 | relation - spatial (people, river,
    ↪ next to)
8 | relation - spatial (elephant,
    ↪ river, next to)
9 | event - ambiguity (people, ==3,
    ↪ stand)
10 | event - ambiguity (people, ==3,
    ↪ elephant, next to)
11 | event - ambiguity (people, ==3,
    ↪ river, next to)
12 | event - ambiguity (people, stand,
    ↪ elephant, next to)
13 | event - ambiguity (people, stand,
    ↪ river, next to)
14 | event - ambiguity (people,
    ↪ elephant, next to, river, next to)
output: 1 | 0
2 | 1
3 | 0
4 | 0
5 | 1
6 | 1,3
7 | 1,4
8 | 2,4
9 | 2,5
10 | 2,6
11 | 2,7
12 | 5,6
13 | 5,7
14 | 6,7

input: Aerial view of downtown
    ↪ Manhattan, but with Millennium
    ↪ Wheel next to the Statue of
    ↪ Liberty. The Great Pyramid is on
    ↪ a sandy island near the buildings.
1 | entity - (downtown Manhattan)
2 | entity - (Millennium Wheel)
3 | entity - (the Statue of the Liberty)
4 | entity - (the Great Pyramid)
5 | entity - (island)
6 | entity - (buildings)
7 | global - (aerial view)
8 | attribute - texture (island, sandy)
9 | relation - spatial (Millennium
```

```
    ↪ Wheel, the Statue of Liberty,
    ↪ next to)
10 | relation - spatial (the Great
    ↪ Pyramid, island, on)
11 | relation - spatial (the Great
    ↪ Pyramid, buildings, near)
12 | event - ambiguity (the Great
    ↪ Pyramid, island, on, buildings,
    ↪ near)
output: 1 | 0
2 | 0
3 | 0
4 | 0
5 | 0
6 | 0
7 | 0
8 | 5
9 | 2,3
10 | 4,5
11 | 4,6
12 | 10,11

input: A laptop with external keyboard,
    ↪ mouse, phone and photo on a desk.
1 | entity - whole (laptop)
2 | entity - whole (keyboard)
3 | entity - whole (mouse)
4 | entity - whole (phone)
5 | entity - whole (photo)
6 | entity - whole (desk)
7 | attribute - type (keyboard,
    ↪ external)
8 | relation - spatial (laptop, desk,
    ↪ on)
9 | relation - spatial (keyboard, desk,
    ↪ on)
10 | relation - spatial (mouse, desk,
    ↪ on)
11 | relation - spatial (phone, desk,
    ↪ on)
12 | relation - spatial (photo, desk,
    ↪ on)
13 | event - ambiguity (laptop,
    ↪ external keyboard, mouse, phone,
    ↪ photo, desk, on)
output: 1 | 0
2 | 0
3 | 0
4 | 0
5 | 0
6 | 0
7 | 0
8 | 1,6
9 | 2,6
10 | 3,6
11 | 4,6
12 | 5,6
13 | 8,9,10,11,12

input: A white slope covers the
    ↪ background, while the foreground
    ↪ features a grassy slope with
    ↪ several rams grazing and one
    ↪ measly and underdeveloped
    ↪ evergreen in the foreground.
1 | entity - whole (slopes)
2 | other - count (slopes, ==2)
3 | entity - whole (rams)
4 | entity - whole (evergreen)
5 | attribute - color (slope_1, white)
```

```
6 | attribute - texture (slope_2,
    ↪ grassy)
7 | attribute - state (evergreen,
    ↪ measly and underdeveloped)
8 | relation - spatial (slope_1,
    ↪ background, in)
9 | relation - spatial (slope_2,
    ↪ foreground, in)
10 | relation - spatial (several, rams,
    ↪ grassy slope_2, on)
11 | attribute - state (several rams,
    ↪ graze)
12 | event - ambiguity (slope_1, white,
    ↪ background, in)
13 | event - ambiguity (slope_2,
    ↪ grassy, foreground, in)
14 | event - ambiguity (several, rams,
    ↪ slope_2, grassy, on)
output: 1 | 0
2 | 1
3 | 0
4 | 0
5 | 1
6 | 1
7 | 4
8 | 5
9 | 1
10 | 1
11 | 1,3
12 | 5,8
13 | 6,9
14 | 6,10

input: A man walks into a room and sits
    ↪ on a chair. A dog follows him.
1 | entity - whole (man)
2 | entity - whole (room)
3 | entity - whole (chair)
4 | entity - whole (dog)
5 | action - (man, walk, room)
6 | action - (man, sit on, chair)
7 | action - (dog, follow, man)
8 | relation - temporal (man, sit,
    ↪ before, walk)
9 | relation - temporal (dog, follows,
    ↪ after, man, sit)
10 | event - temporal (man, walks into
    ↪ a room and sits on a chair, dog
    ↪ follows him)
output: 1 | 0
2 | 0
3 | 0
4 | 0
5 | 1,2
6 | 1,3
7 | 1,4
8 | 5,7
9 | 6,7
10 | 8,9

input: A car is parked by the roadside.
    ↪ Later, it starts moving and
    ↪ drives away.
1 | entity - whole (car)
2 | entity - whole (roadside)
3 | relation - spatial (car, roadside,
    ↪ park)
4 | action - (car, move)
5 | action - (car, drives away)
6 | relation - temporal (car, starts,
```

```
    ↪ after, parked)
7 | relation - temporal (car, drive
    ↪ away, after, parked)
8 | event - temporal (car, move,
    ↪ roadside, park, after)
9 | event - temporal (car, drive away,
    ↪ roadside, park, after)
10 | event - temporal (car, starts,
    ↪ parked, move, drive away)
output: 1 | 0
2 | 0
3 | 1,2
4 | 1
5 | 1
6 | 1,4
7 | 1, 5
8 | 3,6
9 | 3,7
10 | 6,7

input: A man is running across a street
    ↪ while carrying a large bag. This
    ↪ is unusual because people
    ↪ typically do not carry large bags
    ↪ while running across streets.
1 | entity - whole (man)
2 | entity - whole (street)
3 | entity - whole (bag)
4 | relation - spatial (man, run,
    ↪ street)
5 | entity - scale (large bag)
6 | relation - spatial (man, carry,
    ↪ large bag)
7 | relation - temporal (man, carry,
    ↪ while, running)
8 | event - ambiguity (man, large bag,
    ↪ carry)
9 | event - temporal (man, run, street,
    ↪ while, carry, large bag)
output: 1 | 0
2 | 0
3 | 0
4 | 1, 2
5 | 3
6 | 1, 5
7 | 4, 7
8 | 5,6
9 | 4, 7
```

Q Annotation Details

We show UI for all human evaluation tasks in Figure 7, Figure 8, Figure 9, Figure 10, and Figure 11.

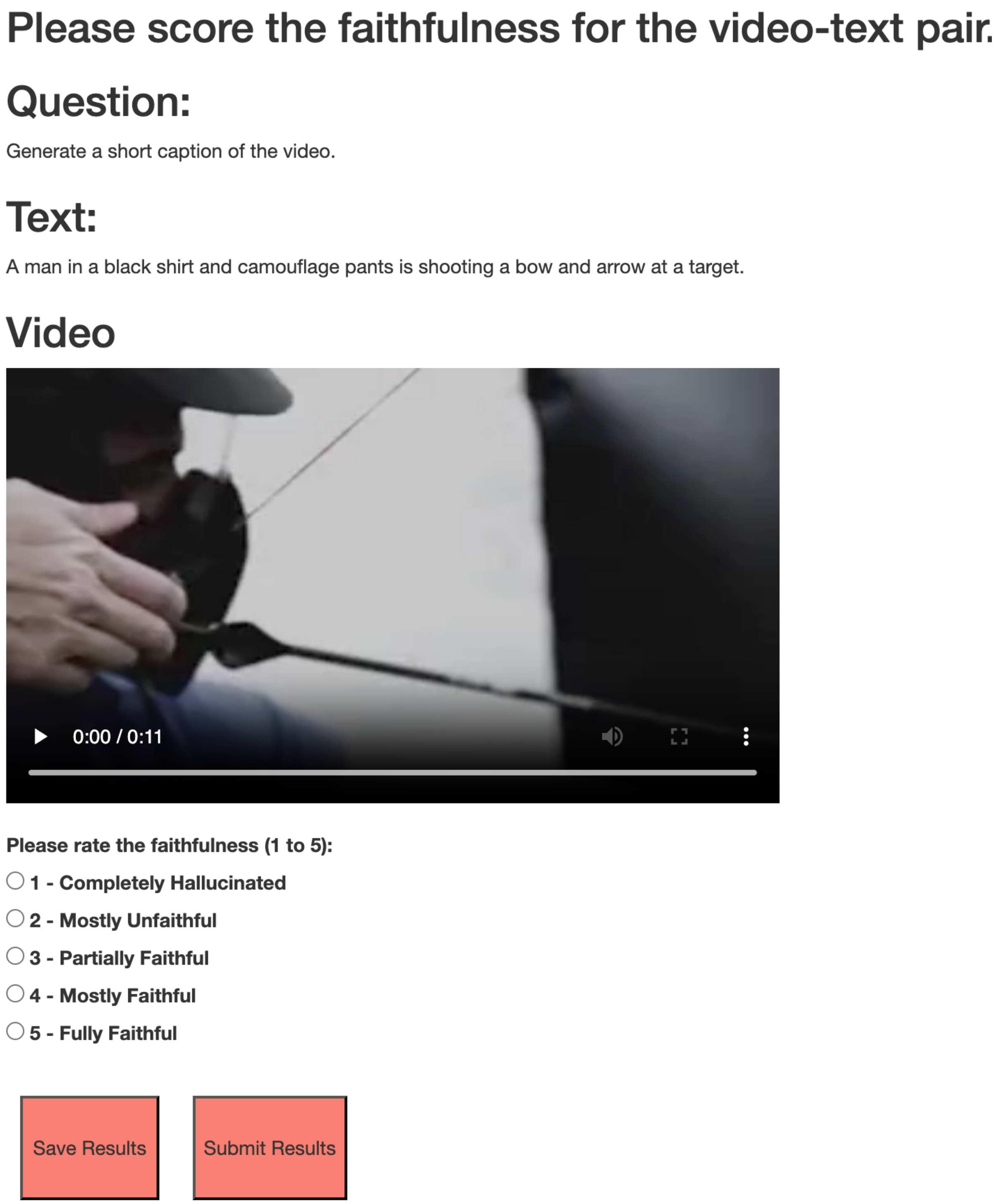

Figure 7: UI for faithfulness evaluation of human annotation.

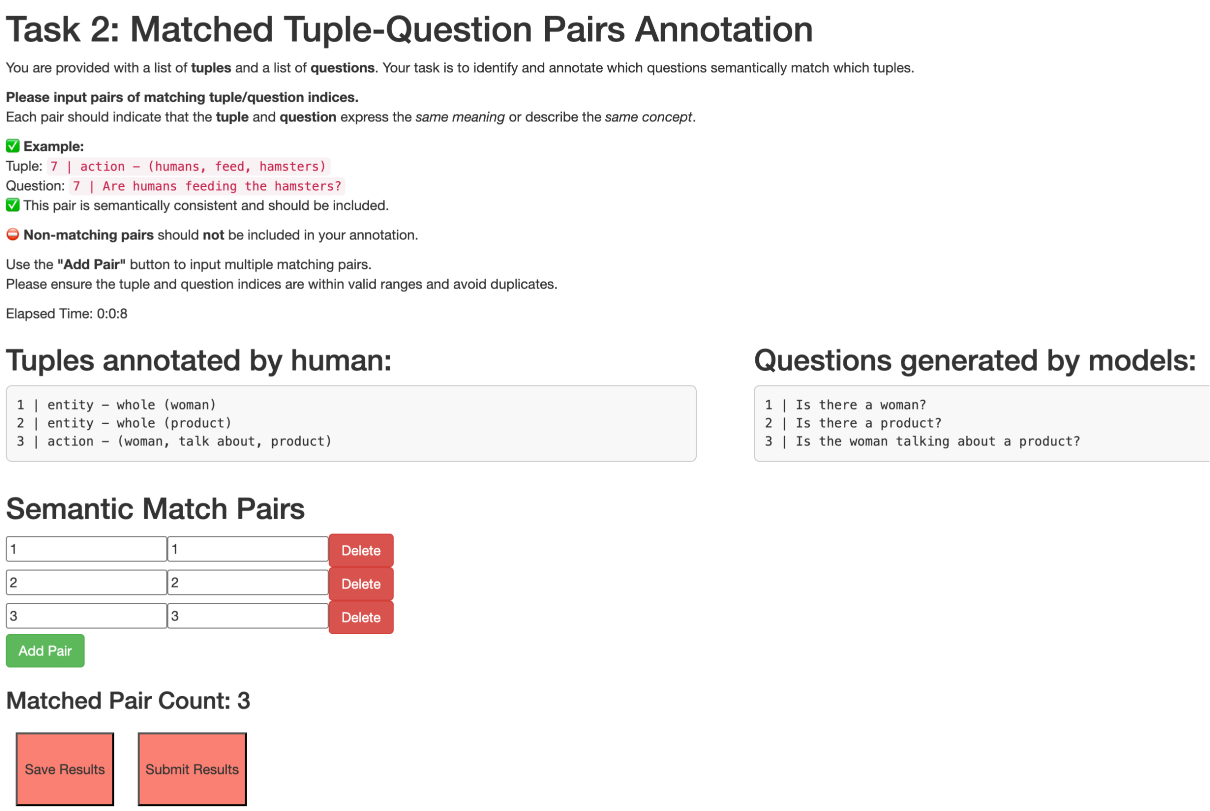

Figure 8: UI for question quality evaluation of human annotation.

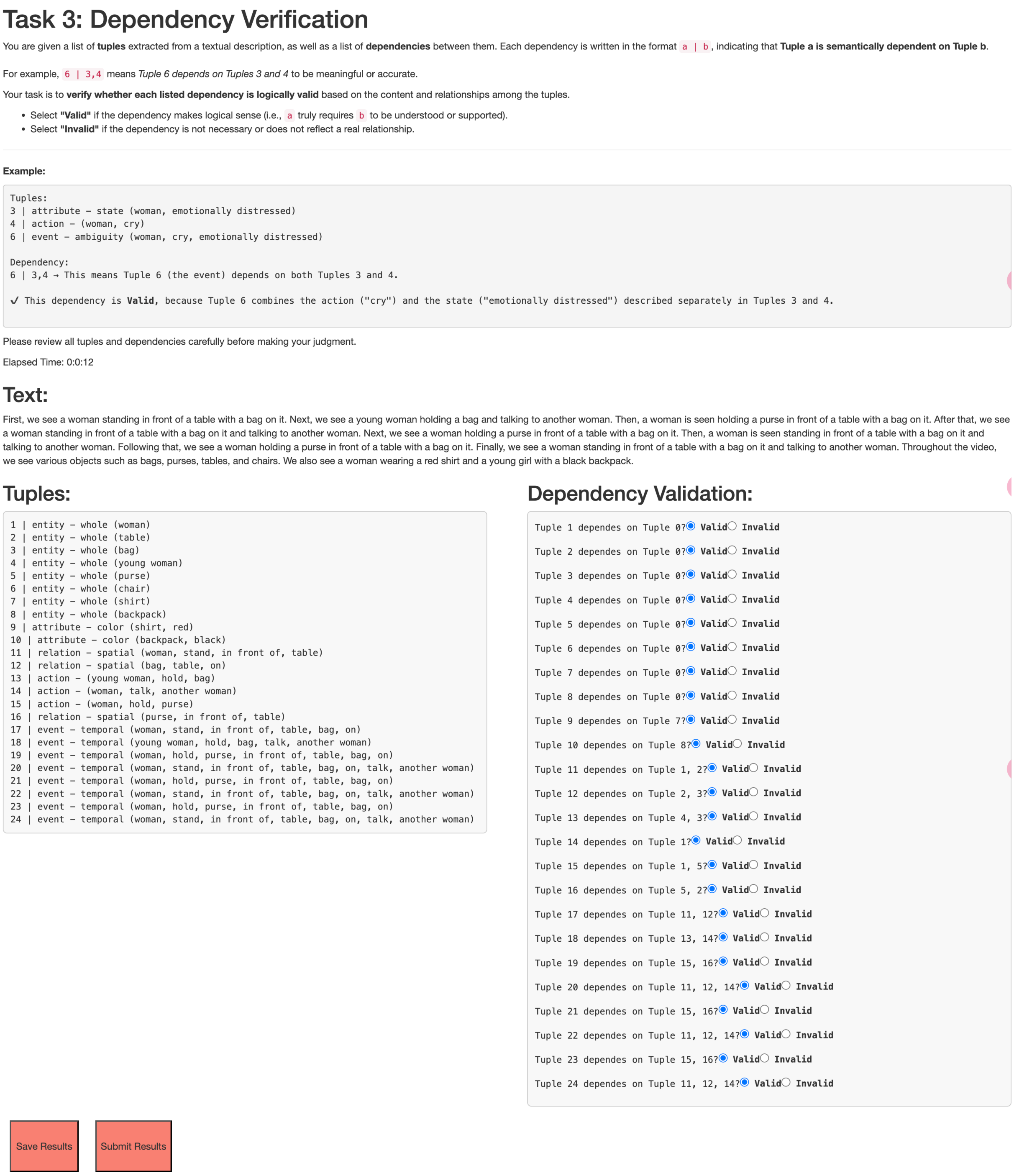

Figure 9: UI for dependency verification of human annotation.

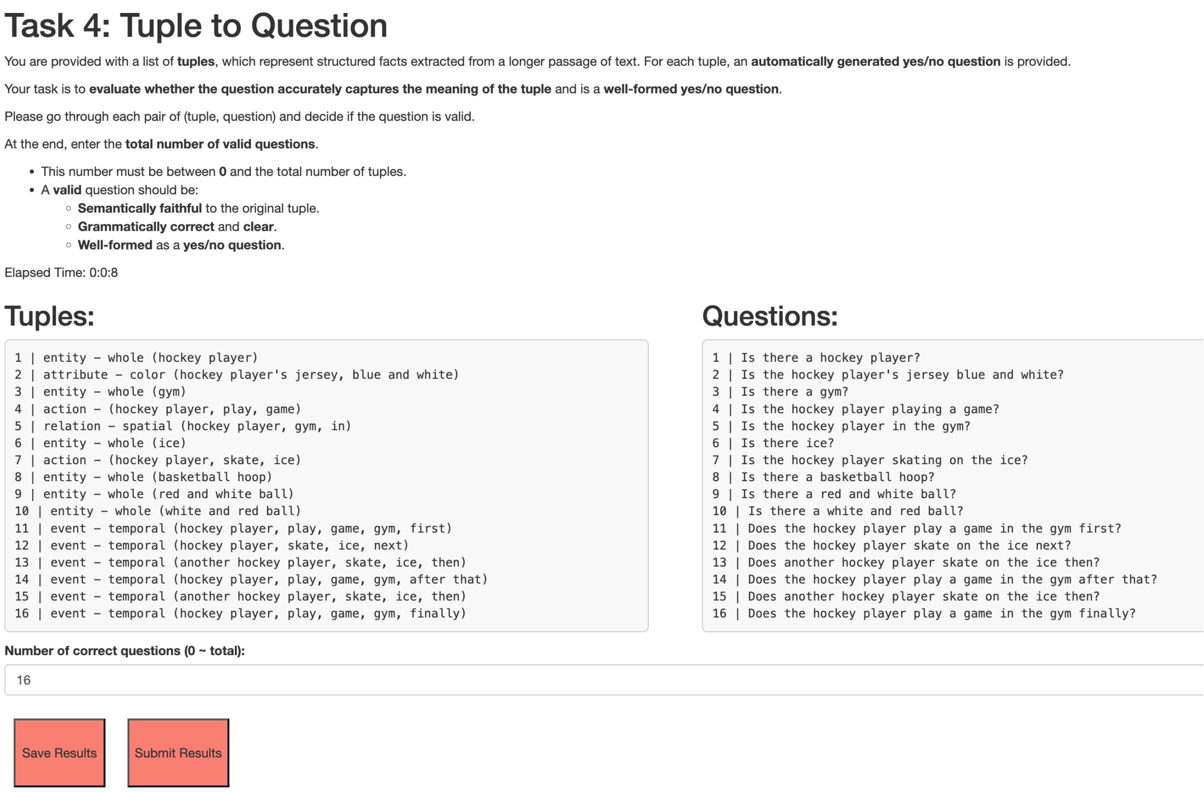

Figure 10: UI for the fact-to-question task of human evaluation.

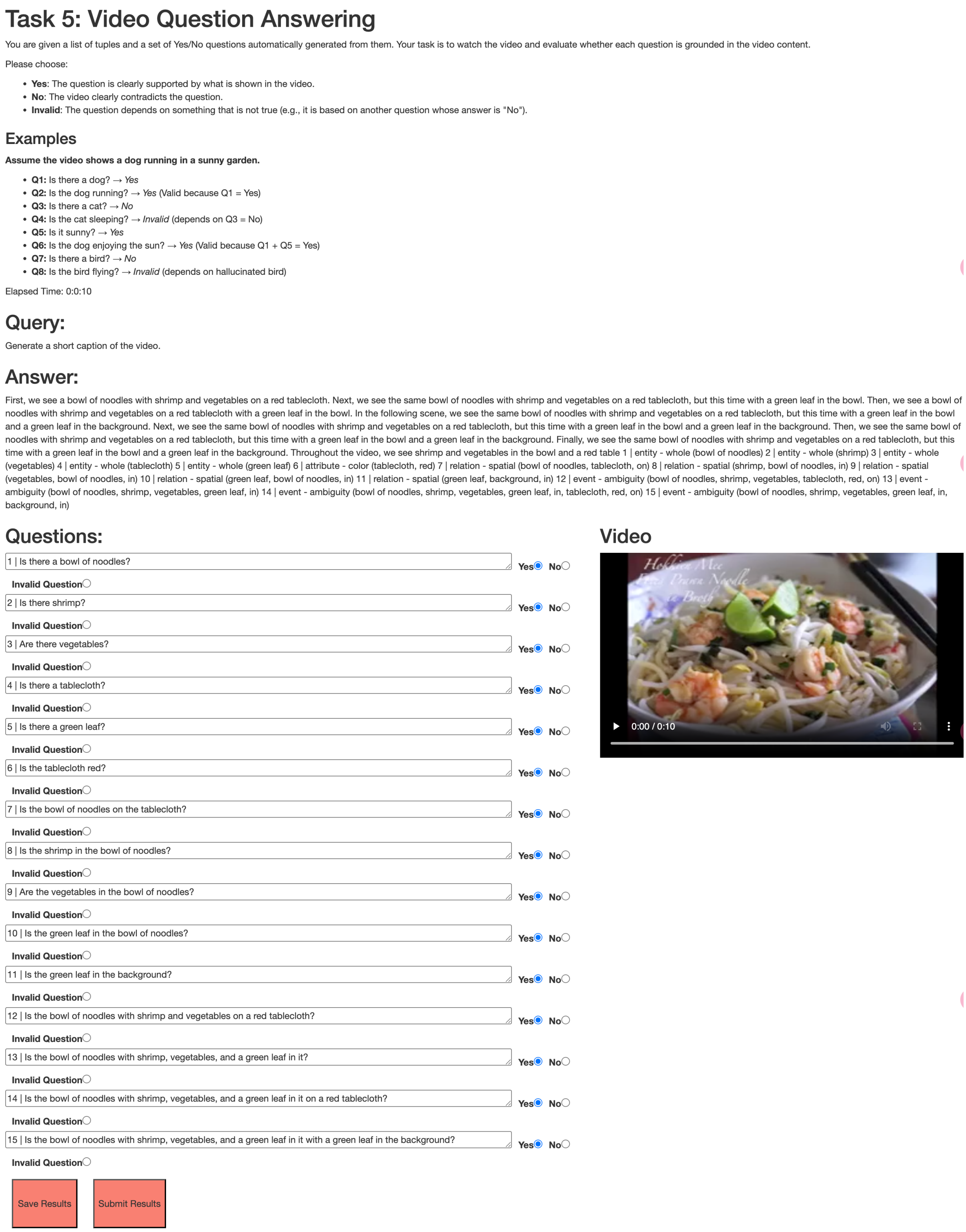

Figure 11: UI for Video Question Answering of human annotation.