

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

Anonymous ACL submission

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 **Fa**ith**Fu**llness **ev**Aluation 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.

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 generation of fabricated or inaccurate content (Wang et al., 2024). 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 multiple works have detected hallucinations 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. Consequently, hallucination in video generation tasks remains largely unexplored despite their importance for general artificial intelligence.

To develop a unified evaluation framework for both T2V and V2T tasks involving free-form questions, motivated by the existing work (Min et al., 2023; Nenkova and Passonneau, 2004), 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., 2024; 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*

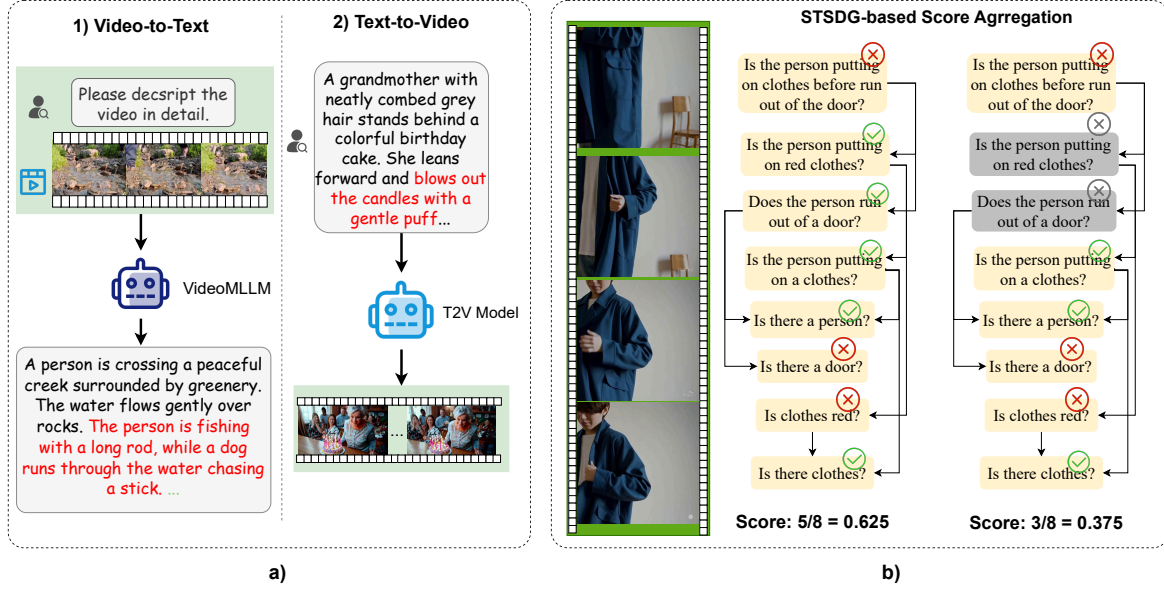


Figure 1: a) Illustration of the V2T and T2V tasks. The content in red font denotes hallucinated content. b) Illustration of Spatio-Temporal Semantic Dependency Graph-based Score Aggregation.

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 color are 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 rea-

soning 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 neg-

ative, 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 tasks. 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 that could facilitate future research on video-based multimodal hallucination and 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 objects, attributes, or actions that do not faithfully reflect the input prompt. This hallucination 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. Hallucination is a persistent issue in large language models (LLMs) (Huang et al., 2023) and MLLM (Zhang et al., 2023). Early studies primarily focus on hallucinations in image-related tasks (Jing et al., 2024; Hu et al., 2023; Cho et al., 2024; Liu et al., 2024a). 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., 2024; 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. 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 Fig. 2).

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. $Q = \phi$ (empty) 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 3.

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 tempo-

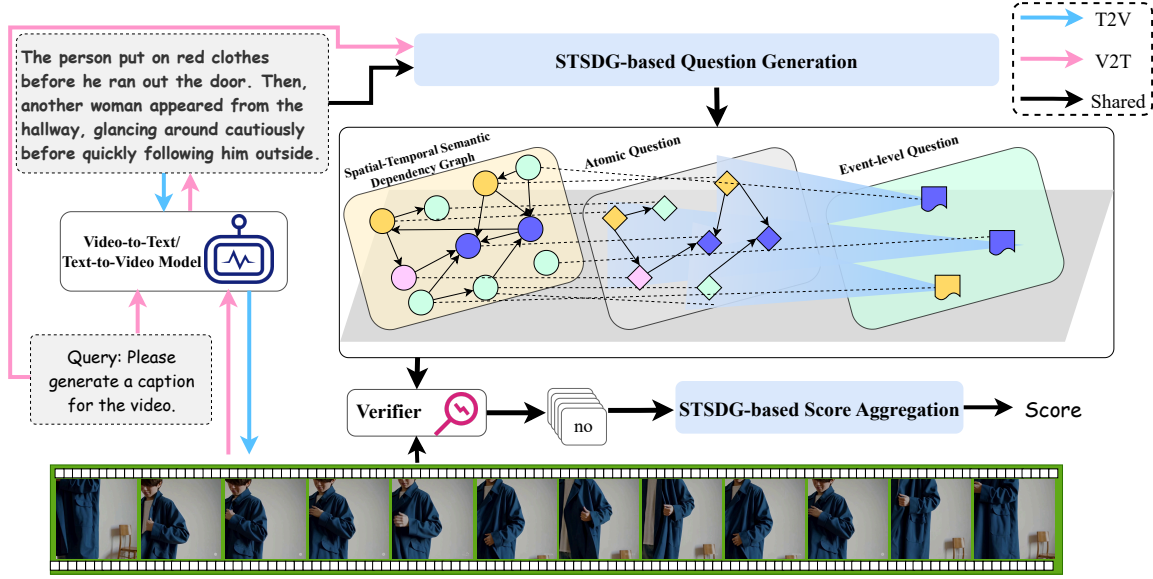


Figure 2: Illustration of our proposed FIFA metric. The blue arrows represent the information flow for T2V, the pink arrows represent the flow for V2T, and the black arrows are shared information pathways of both tasks.

ral 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:

$$G = \begin{cases} \text{LLM}(P_{t2v}, T), & \text{task} = t2v \\ \text{LLM}(P_{v2t}, T, Q), & \text{task} = v2t \end{cases} \quad (1)$$

where $G = \{g^1, \dots, g^n\}$ 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 E for detail 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 $\{q_1, \dots, q_n\} = \text{LLM}(P_q, G, T, Q)$, where $Q = \phi$ for the text-to-video task. P_q is the prompt and is shown in Appendix E. 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 3) 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 $Q = \{q_1, \dots, q_n\}$. 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 E. Specifically, we obtain semantic dependency edges between questions as an adjacency matrix $\mathbf{E} \in \mathbb{R}^{n \times n}$,

$$E_{ij} = \begin{cases} 1, & \text{if } \mathcal{S}(q_i, q_j), \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where $i, j \in [1, n]$, and $\mathcal{S}(t_i, t_j)$ 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 $\mathbf{E}_{ij} \neq \mathbf{E}_{ji}$ does not necessarily hold true.

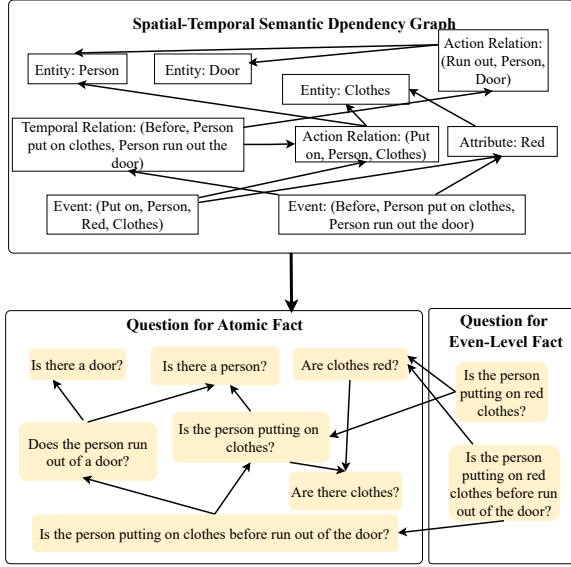


Figure 3: Illustration of STSDG-based Question Generation and STSDG-based Score Aggregation. The generation process is for text “The person put on red clothes before he ran out the door.”.

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,

$$A = \{a_1, \dots, a_n\} = \text{VideoQA}(V, q_1, \dots, q_n), \quad (3)$$

where $\text{VideoQA}(\cdot)$ is a VideoQA model. $\{q_1, \dots, q_n\}$ 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; Jing et al., 2024).

3.4 STSDG-based Score Aggregation

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

$$\hat{s}_i = \mathbb{I}(a_i = \text{“yes”}) \prod_{j \text{ s.t. } E_{ij}=1} s_j, \quad (4)$$

where $\mathbb{I}(\cdot)$ is the indicator function, and the value of $\mathbb{I}(a_i = \text{“yes”})$ 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: $\hat{f} = \sum_{i=1}^n \hat{s}_i / n$.

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 have 60 evaluation samples, resulting in a total of 120 annotated samples of hallucination across T2V and V2T. More details are in Appendix B

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.

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 details about the annotation process are given in the Appendix J. Table 1 delineates the correlation between various evaluation metrics and human judgment regarding the faithfulness of T2V and V2T. The result shows that our evaluation framework consistently achieves a significant improvement across T2V and V2T. We add more ablation studies in Appendix A and detailed benchmark results in Appendix G.

4.2 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

Table 1: Correlation between each evaluation metric and human judgment on V2T and T2V faithfulness evaluation, measured by Pearson’s r , Spearman’s ρ , and Kendall’s τ . The p-value of the significance test between our result and the baseline result is less than 0.01.

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 FIFA	4.58 58.20	-1.20 53.20	-1.01 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

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

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 $C = \{c_1, \dots, c_{n_c}\}$. Follow the definition of the last section, the generated questions are denoted as $U = \{q_1, \dots, q_n\}$. 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 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

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

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).

4.3 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., 2025), 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 tasks.

5 Post-Correction

Method. Our initial experiments show various hallucinations in the T2V and V2T models. Therefore, we devise a post-correction method to alle-

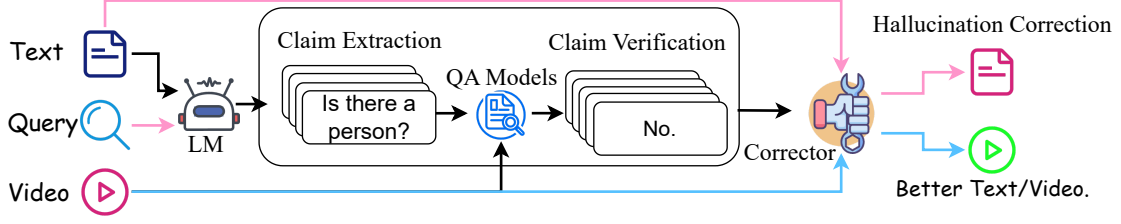


Figure 4: The proposed Post-Correction method consists of three key stages: Claim Extraction, Claim Verification, and Hallucination Correction. The Corrector takes claim-answer pairs with video for T2V, and with text for V2T.

viate 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 correction. An overview of our framework is shown in Figure 4. 1) **Claim Extraction**. Since the text usually consists of multiple claims, such as objects, attributions, and relations, we follow Eq. 1 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. 3. 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 100 videos from the MSR-VTT dataset to perform the captioning task. For the T2V task,

Task	Model	COAHA ↓		FIFA ↑	
		w/o	w/	w/o	w/
V2T	Video-LLaVA	52.45	47.23	63.43	66.08
	Video-LLaMA	53.34	45.86	60.46	65.54
	Video-LLaMA2	37.65	25.93	64.49	69.82
	Video-LLaMA3	63.25	51.27	65.28	70.41
T2V	CogVideoX	-	-	54.53	60.70

Table 4: 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.

due to the slow generation speed of video generation models and video editing models, we adopt 30 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 4 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 H.

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).

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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.
- 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.
- 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.
- Liqliang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. 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.
- 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, Duoju 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, Xincheng 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, Jiayi 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. 2023. Visual instruction tuning. *CoRR*, abs/2304.08485.
- Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tiejong 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.
- 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.
- 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.
- Ani Nenkova and Rebecca J. Passonneau. 2004. [Evaluating content selection in summarization: The pyramid method](#). In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*, pages 145–152. The 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.

838	Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024. Videohalluciner: Evaluating intrinsic and extrinsic hallucinations in large video-language models . <i>CoRR</i> , abs/2406.16338.	895
839		896
840		897
841		898
842	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 <i>Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XVI</i> , volume 13676 of <i>Lecture Notes in Computer Science</i> , pages 720–736. Springer.	899
843		900
844		901
845		902
846		903
847		904
848		905
849		906
850	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 <i>Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI</i> , volume 13666 of <i>Lecture Notes in Computer Science</i> , pages 538–554. Springer.	907
851		908
852		909
853		910
854		
855		911
856		912
857		913
858		914
859	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 . <i>CoRR</i> , abs/2401.07781.	915
860		916
861		917
862		918
863		919
864		
865	Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language . In <i>2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016</i> , pages 5288–5296. IEEE Computer Society.	920
866		921
867		922
868		923
869		
870		
871	Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. 2021. Videogpt: Video generation using VQ-VAE and transformers . <i>CoRR</i> , abs/2104.10157.	
872		
873		
874	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, Wei Han Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer . <i>CoRR</i> , abs/2408.06072.	
875		
876		
877		
878		
879		
880		
881	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 . <i>Sci. China Inf. Sci.</i> , 67(12).	
882		
883		
884		
885		
886	Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. 2024. Make pixels dance: High-dynamic video generation . In <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024</i> , pages 8850–8860. IEEE.	
887		
888		
889		
890		
891		
892	Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi	
893		
894		
	Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding . <i>CoRR</i> , abs/2501.13106.	
	Hang Zhang, Xin Li, and Lidong Bing. 2023. Videollama: An instruction-tuned audio-visual language model for video understanding . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023</i> , pages 543–553. Association for Computational Linguistics.	
	Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. 2024. Eventhallusion: Diagnosing event hallucinations in video llms . <i>CoRR</i> , abs/2409.16597.	
	Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert . In <i>International Conference on Learning Representations</i> .	
	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 . <i>CoRR</i> , 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 . <i>CoRR</i> , 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 4 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. Table 5 summarizes the performance of FIFA with its derivations. From the results, we observe that: 1) Our FIFA surpasses w/o-Dependency, demonstrating the importance of introducing semantic dependency relationships between facts/questions. 2) 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. 3) By comparing the VideoQA accuracy in Table 3 and Table 5, we observe that models with higher VideoQA accuracy tend to achieve better correlation performance. This suggests that improving the accuracy of the VideoQA verifier is crucial for enhancing the overall correlation between model outputs and human judgments.

Table 5: 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

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., 2023).

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 FIFA with 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 J.

D Human Evaluation

We employ 3 workers for annotation via Amazon Mechanical Turk¹. 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.

¹<https://www.mturk.com/>.

T2V Model	Entity	Attribute	Spatial	Temporal	Action	Event
CogVidex	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 V2T models.

E 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 I.

F Experimental Details

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

G Fine-grained Benchmark 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 F1

Model	Accuracy \uparrow	
	w/o	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 7: Results on the VideoHalluciner benchmark, a V2T hallucination evaluation task.

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.

H Hallucination Mitigation on More Benchmarks

In addition to the caption task, we also conduct experiment on VideoHalluciner (Wang et al., 2024), which is a binary-QA question answering benchmark. We show the results in Table 7. 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.

I In-context Examples for Prompts

I.1 Fact Extraction for Video-to-Text

1055	query: Please generate a caption for	7 action - (two of the children,	1125
1056	↪ the video.	↪ frisbee, hold)	1126
1057	input: A male skateboarder is trying to		1127
1058	↪ pull off a trick on the ramp.	query: Please generate a caption for	1128
1059	output: 1 entity - whole	↪ the video.	1129
1060	↪ (skateboarder)	input: the word 'START' written in	1130
1061	2 entity - whole (ramp)	↪ chalk on a sidewalk	1131
1062	3 attribute - type (skateboarder,	output: 1 entity - whole (word)	1132
1063	↪ male)	2 entity - whole (sidewalk)	1133
1064	4 action - (male skateboarder, pull	3 other - text rendering (word,	1134
1065	↪ off a trick)	↪ "START")	1135
1066	5 relation - spatial (male	4 attribute - texture (word, chalk)	1136
1067	↪ skateboarder, ramp, on)	5 relation - spatial (word 'START',	1137
1068	6 event - ambiguity (skateboarder,	↪ sidewalk, on)	1138
1069	↪ male, pull off a trick)		1139
1070	7 event - ambiguity (male	query: Please generate a caption for	1140
1071	↪ skateboarder, ramp, on)	↪ the video.	1141
1072	8 event - ambiguity (skateboarder,	input: A pear, orange, and two bananas	1142
1073	↪ pull off a trick, ramp, on)	↪ in a wooden bowl.	1143
1074		output: 1 entity - whole (pear)	1144
1075	query: Please generate a caption for	2 entity - whole (orange)	1145
1076	↪ the video.	3 entity - whole (bananas)	1146
1077	input: A car playing soccer, digital	4 other - count (bananas, ==2)	1147
1078	↪ art.	5 entity - whole (bowl)	1148
1079	output: 1 entity - whole (car)	6 attribute - material (bowl, wood)	1149
1080	2 global - (digital art)	7 relation - spatial (pear, bowl, in)	1150
1081	3 action - (car, soccer, play)	8 relation - spatial (orange, bowl,	1151
1082		↪ in)	1152
1083	query: Please generate a caption for	9 relation - spatial (bananas, bowl,	1153
1084	↪ the video.	↪ in)	1154
1085	input: A set of 2x2 emoji icons with	10 relation - spatial (bananas, bowl,	1155
1086	↪ happy, angry, surprised and	↪ in)	1156
1087	↪ sobbing faces. The emoji icons	11 event - ambiguity (pear, orange,	1157
1088	↪ look like pigs. All of the pigs	↪ bananas, ==2, bowl, in)	1158
1089	↪ are wearing crowns.		1159
1090	output: 1 entity - whole (emoji icons)	query: Please generate a caption for	1160
1091	2 other - count (emoji icons, ==4)	↪ the video.	1161
1092	3 attribute - state (emoji icons, 2x2	input: Closeup picture of the front of	1162
1093	↪ grid)	↪ a clean motorcycle.	1163
1094	4 attribute - type (emoji icons, pig)	output: 1 entity - whole (motorcycle)	1164
1095	5 attribute - state (emoji_1, happy)	2 global - (closeup)	1165
1096	6 attribute - state (emoji_2, angry)	3 global - (picture)	1166
1097	7 attribute - state (emoji_3,	4 attribute - state (motorcycle,	1167
1098	↪ surprised)	↪ clean)	1168
1099	8 attribute - state (emoji_4, sobbing	5 entity - part (front of the clean	1169
1100	↪ face)	↪ motorcycle)	1170
1101	9 entity - part (pig's crown)		1171
1102		query: Please generate a caption for	1172
1103	query: Please generate a caption for	↪ the video.	1173
1104	↪ the video.	input: a sad man with green hair	1174
1105	input: a photo of bear and dining	output: 1 entity - whole (man)	1175
1106	↪ table; dining table is below bear	2 entity - part (man's hair)	1176
1107	output: 1 global - (photo)	3 attribute - state (man, sad)	1177
1108	2 entity - whole (bear)	4 attribute - color (man's hair,	1178
1109	3 entity - whole (dining table)	↪ green)	1179
1110	4 relation - spatial (dining table,	5 event - ambiguity (man, sad, man's	1180
1111	↪ bear, below)	↪ hair, green)	1181
1112			1182
1113	query: Please generate a caption for	query: Please generate a caption for	1183
1114	↪ the video.	↪ the video.	1184
1115	input: A group of children sitting in	input: A commercial airplane with	1185
1116	↪ the grass with two of them	↪ propellers flying through the air.	1186
1117	↪ holding a Frisbee .	output: 1 entity - whole (airplane)	1187
1118	output: 1 entity - whole (children)	2 entity - part (airplane's	1188
1119	2 entity - whole (grass)	↪ propellers)	1189
1120	3 entity - whole (frisbee)	3 action - (airplane, air, fly	1190
1121	4 attribute - state (children, sit)	↪ through)	1191
1122	5 relation - spatial (a group of	4 event - ambiguity (airplane, with	1192
1123	↪ children, grass, sitting in)	↪ propellers, air, fly through)	1193
1124	6 entity - part (two of the children)		1194

1195	query: Please generate a caption for	input: A realistic photo of a	1264
1196	→ the video.	→ Pomeranian dressed up like a	1265
1197	input: A little boy grips a soccer ball	→ 1980s professional wrestler with	1266
1198	→ in his arms surrounded by other	→ neon green and neon orange face	1267
1199	→ youth soccer players.	→ paint and bright green wrestling	1268
1200	output: 1 entity - whole (boy)	→ tights with bright orange boots.	1269
1201	2 entity - whole (ball)	output: 1 global - (photo)	1270
1202	3 entity - whole (soccer players)	2 entity - whole (Pomeranian)	1271
1203	4 entity - part (boy's arms)	3 global - (realistic)	1272
1204	5 entity - scale (boy, little)	4 entity - part (Pomeranian's costume)	1273
1205	6 attribute - type (ball, soccer)	5 attribute - type (Pomeranian's	1274
1206	7 attribute - state (soccer players,	→ costume, 1980s professional	1275
1207	→ youth)	→ wrestler)	1276
1208	8 relation - spatial (little boy,	6 entity - part (Pomeranian's	1277
1209	→ ball, grip in his arms)	→ costume's wrestling tights)	1278
1210	9 relation - spatial (little boy	7 entity - part (Pomeranian's	1279
1211	→ gripping the ball in his arms,	→ costume's wrestling tights' boots)	1280
1212	→ soccer players, surrounded by)	8 entity - part (Pomeranian's	1281
1213	10 event - ambiguity (boy's arm,	→ facepaint)	1282
1214	→ little, ball, soccer, grip in his	9 attribute - color (Pomeranian's	1283
1215	→ arms)	→ facepaint, neon green)	1284
1216	11 event - ambiguity (boy, little,	10 attribute - color (Pomeranian's	1285
1217	→ soccer players, youth, surrounded	→ facepaint, neon orange)	1286
1218	→ by)	11 attribute - color (Pomeranian's	1287
1219		→ costume's wrestling tights,	1288
1220	query: Please generate a caption for	→ bright green)	1289
1221	→ the video.	12 attribute - color (Pomeranian's	1290
1222	input: A traffic light and a signpost	→ costume's wrestling tights'	1291
1223	→ at a crossroads intersection near	→ boots, bright orange)	1292
1224	→ a waterway.		1293
1225	output: 1 entity - whole (traffic	query: Please generate a caption for	1294
1226	→ light)	→ the video.	1295
1227	2 entity - whole (signpost)	input: a four-piece band on a stage in	1296
1228	3 entity - whole (crossroads	→ front of a small crowd	1297
1229	→ intersection)	output: 1 entity - whole (band)	1298
1230	4 entity - whole (waterway)	2 entity - whole (stage)	1299
1231	5 relation - spatial (traffic light,	3 entity - whole (crowd)	1300
1232	→ crossroads intersection, at)	4 other - count (band members, ==4)	1301
1233	6 relation - spatial (signpost,	5 attribute - shape (crowd, small)	1302
1234	→ crossroads intersection, at)	6 relation - spatial (four-piece	1303
1235	7 relation - spatial (traffic light,	→ band, stage, on)	1304
1236	→ waterway, near)	7 relation - spatial (four-piece	1305
1237	8 relation - spatial (signpost,	→ band, crowd, in front of)	1306
1238	→ waterway, near)	8 relation - spatial (stage, crowd,	1307
1239	9 relation - spatial (crossroads	→ in front of)	1308
1240	→ intersection, waterway, near)	9 event - ambiguity (band, ==4	1309
1241	10 event - ambiguity (traffic light,	→ picece, stage, on)	1310
1242	→ signpost, crossroads	10 event - ambiguity (band, ==4	1311
1243	→ intersection, at)	→ picece, crowd, small, in front of)	1312
1244	11 event - ambiguity (traffic light,	11 event - ambiguity (stage, crowd,	1313
1245	→ crossroads intersection, at,	→ small, in front off)	1314
1246	→ waterway, near)		1315
1247	12 event - spatial (signpost,	query: Please generate a caption for	1316
1248	→ crossroads intersection, at,	→ the video.	1317
1249	→ waterway, near)	input: two laptops, a mouse cord, and a	1318
1250		→ monitor	1319
1251	query: Please generate a caption for	output: 1 entity - whole (laptops)	1320
1252	→ the video.	2 other - count (laptops, ==2)	1321
1253	input: a photo of dining table and	3 entity - whole (mouse coord)	1322
1254	→ traffic light; traffic light is	4 entity - whole (monitor)	1323
1255	→ below dining table		1324
1256	output: 1 global - (photo)	query: Please generate a caption for	1325
1257	2 entity - whole (dining table)	→ the video.	1326
1258	3 entity - whole (traffic light)	input: A red motorcycle parked by paint	1327
1259	4 relation - spatial (traffic light,	→ chipped doors.	1328
1260	→ dining table, below)	output: 1 entity - whole (motorcycle)	1329
1261		2 entity - whole (doors)	1330
1262	query: Please generate a caption for	3 attribute - color (motorcycle, red)	1331
1263	→ the video.	4 attribute - state (door, paint	1332
		→ chipped)	1333

1334	5 relation - spatial (red motorcycle,	2 entity - (Millennium Wheel)	1404
1335	↪ paint chipped door, next to)	3 entity - (the Statue of the Liberty)	1405
1336	6 attribute - state (motorcycle,	4 entity - (the Great Pyramid)	1406
1337	↪ parked)	5 entity - (island)	1407
1338	7 event- ambiguity (motorcycle, red,	6 entity - (buildings)	1408
1339	↪ door, paint chipped, next to)	7 global - (aerial view)	1409
1340	8 event- ambiguity (motorcycle, red,	8 attribute - texture (island, sandy)	1410
1341	↪ parked)	9 relation - spatial (Millennium	1411
1342		↪ Wheel, the Statue of Liberty,	1412
1343	query: Please generate a caption for	↪ next to)	1413
1344	↪ the video.	10 relation - spatial (the Great	1414
1345	input: A cube made of denim. A cube	↪ Pyramid, island, on)	1415
1346	↪ with the texture of denim.	11 relation - spatial (the Great	1416
1347	output: 1 entity - whole (cube)	↪ Pyramid, buildings, near)	1417
1348	2 attribute - material (cube, denim)	12 event - ambiguity (the Great	1418
1349	3 attribute - texture (cube, denim)	↪ Pyramid, island, on, buildings,	1419
1350		↪ near)	1420
1351	query: Please generate a caption for		1421
1352	↪ the video.	query: Please generate a caption for	1422
1353	input: an espresso machine that makes	↪ the video.	1423
1354	↪ coffee from human souls	input: A laptop with external keyboard,	1424
1355	output: 1 entity - whole (espresso	↪ mouse, phone and photo on a desk.	1425
1356	↪ machine)	output: 1 entity - whole (laptop)	1426
1357	2 entity - whole (coffee)	2 entity - whole (keyboard)	1427
1358	3 entity - whole (human souls)	3 entity - whole (mouse)	1428
1359	4 action - (espresso machine, coffee,	4 entity - whole (phone)	1429
1360	↪ make)	5 entity - whole (photo)	1430
1361	5 attribute - material (coffee, human	6 entity - whole (desk)	1431
1362	↪ souls)	7 attribute - type (keyboard,	1432
1363	6 event - ambiguity (espresso	↪ external)	1433
1364	↪ machine, coffee, make, human	8 relation - spatial (laptop, desk,	1434
1365	↪ souls)	↪ on)	1435
1366		9 relation - spatial (keyboard, desk,	1436
1367	query: Please generate a caption for	↪ on)	1437
1368	↪ the video.	10 relation - spatial (mouse, desk,	1438
1369	input: Three people standing next to an	↪ on)	1439
1370	↪ elephant along a river.	11 relation - spatial (phone, desk,	1440
1371	output: 1 entity - whole (people)	↪ on)	1441
1372	2 other - count (people, ==3)	12 relation - spatial (photo, desk,	1442
1373	3 entity - whole (elephant)	↪ on)	1443
1374	4 entity - whole (river)	13 event - ambiguity (laptop,	1444
1375	5 attribute - state (people, stand)	↪ external keyboard, mouse, phone,	1445
1376	6 relation - spatial (three people,	↪ photo, desk, on)	1446
1377	↪ elephant, next to)		1447
1378	7 relation - spatial (people, river,	query: Please generate a caption for	1448
1379	↪ next to)	↪ the video.	1449
1380	8 relation - spatial (elephant,	input: A white slope covers the	1450
1381	↪ river, next to)	↪ background, while the foreground	1451
1382	9 event - ambiguity (people, ==3,	↪ features a grassy slope with	1452
1383	↪ stand)	↪ several rams grazing and one	1453
1384	10 event - ambiguity (people, ==3,	↪ measly and underdeveloped	1454
1385	↪ elephant, next to)	↪ evergreen in the foreground.	1455
1386	11 event - ambiguity (people, ==3,	output: 1 entity - whole (slopes)	1456
1387	↪ river, next to)	2 other - count (slopes, ==2)	1457
1388	12 event - ambiguity (people, stand,	3 entity - whole (rams)	1458
1389	↪ elephant, next to)	4 entity - whole (evergreen)	1459
1390	13 event - ambiguity (people, stand,	5 attribute - color (slope_1, white)	1460
1391	↪ river, next to)	6 attribute - texture (slope_2,	1461
1392	14 event - ambiguity (people,	↪ grassy)	1462
1393	↪ elephant, next to, river, next to)	7 attribute - state (evergreen,	1463
1394		↪ measly and underdeveloped)	1464
1395	query: Please generate a caption for	8 relation - spatial (slope_1,	1465
1396	↪ the video.	↪ background, in)	1466
1397	input: Aerial view of downtown	9 relation - spatial (slope_2,	1467
1398	↪ Manhattan, but with Millennium	↪ foreground, in)	1468
1399	↪ Wheel next to the Statue of	10 relation - spatial (several, rams,	1469
1400	↪ Liberty. The Great Pyramid is on	↪ grassy slope_2, on)	1470
1401	↪ a sandy island near the buildings.	11 attribute - state (several rams,	1471
1402	output: 1 entity - (downtown	↪ graze)	1472
1403	↪ Manhattan)		

1473 12 | event - ambiguity (slope_1, white,
1474 ↪ background, in)
1475 13 | event - ambiguity (slope_2,
1476 ↪ grassy, foreground, in)
1477 14 | event - ambiguity (several, rams,
1478 ↪ slope_2, grassy, on)
1479
1480 query: Please generate a caption for
1481 ↪ the video.
1482 input: A man walks into a room and sits
1483 ↪ on a chair. A dog follows him.
1484 output: 1 | entity - whole (man)
1485 2 | entity - whole (room)
1486 3 | entity - whole (chair)
1487 4 | entity - whole (dog)
1488 5 | action - (man, walk, room)
1489 6 | action - (man, sit on, chair)
1490 7 | action - (dog, follow, man)
1491 8 | relation - temporal (man, sit,
1492 ↪ before, walk)
1493 9 | relation - temporal (dog, follows,
1494 ↪ after, man, sit)
1495 10 | event - temporal (man, walks into
1496 ↪ a room and sits on a chair, dog
1497 ↪ follows him)
1498
1499 query: Please generate a caption for
1500 ↪ the video.
1501 input: A car is parked by the roadside.
1502 ↪ Later, it starts moving and
1503 ↪ drives away.
1504 output: 1 | entity - whole (car)
1505 2 | entity - whole (roadside)
1506 3 | relation - spatial (car, roadside,
1507 ↪ park)
1508 4 | action - (car, move)
1509 5 | action - (car, drives away)
1510 6 | relation - temporal (car, starts,
1511 ↪ after, parked)
1512 7 | relation - temporal (car, drive
1513 ↪ away, after, parked)
1514 8 | event - temporal (car, move,
1515 ↪ roadside, park, after)
1516 9 | event - temporal (car, drive away,
1517 ↪ roadside, park, after)
1518 10 | event - temporal (car, starts,
1519 ↪ parked, move, drive away)
1520
1521 query: What's unusual in this video?
1522 input: A man is running across a street
1523 ↪ while carrying a large bag. This
1524 ↪ is unusual because people
1525 ↪ typically do not carry large bags
1526 ↪ while running across streets.
1527 output: 1 | entity - whole (man)
1528 2 | entity - whole (street)
1529 3 | entity - whole (bag)
1530 4 | relation - spatial (man, run,
1531 ↪ street)
1532 5 | entity - scale (large bag)
1533 6 | relation - spatial (man, carry,
1534 ↪ large bag)
1535 7 | relation - temporal (man, carry,
1536 ↪ while, running)
1537 8 | event - ambiguity (man, large bag,
1538 ↪ carry)
1539 9 | event - temporal (man, run, street,
1540 ↪ while, carry, large bag)

1.2 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,

1610	↪ "START")	↪ arms)	1680
1611	4 attribute - texture (word, chalk)	11 event - ambiguity (boy, little,	1681
1612	5 relation - spatial (word 'START',	↪ soccer players, youth, surrounded	1682
1613	↪ sidewalk, on)	↪ by)	1683
1614			1684
1615	input: A pear, orange, and two bananas	input: A traffic light and a signpost	1685
1616	↪ in a wooden bowl.	↪ at a crossroads intersection near	1686
1617	output: 1 entity - whole (pear)	↪ a waterway.	1687
1618	2 entity - whole (orange)	output: 1 entity - whole (traffic	1688
1619	3 entity - whole (bananas)	↪ light)	1689
1620	4 other - count (bananas, ==2)	2 entity - whole (signpost)	1690
1621	5 entity - whole (bowl)	3 entity - whole (crossroads	1691
1622	6 attribute - material (bowl, wood)	↪ intersection)	1692
1623	7 relation - spatial (pear, bowl, in)	4 entity - whole (waterway)	1693
1624	8 relation - spatial (orange, bowl,	5 relation - spatial (traffic light,	1694
1625	↪ in)	↪ crossroads intersection, at)	1695
1626	9 relation - spatial (bananas, bowl,	6 relation - spatial (signpost,	1696
1627	↪ in)	↪ crossroads intersection, at)	1697
1628	10 relation - spatial (bananas, bowl,	7 relation - spatial (traffic light,	1698
1629	↪ in)	↪ waterway, near)	1699
1630	11 event - ambiguity (pear, orange,	8 relation - spatial (signpost,	1700
1631	↪ bananas, ==2, bowl, in)	↪ waterway, near)	1701
1632		9 relation - spatial (crossroads	1702
1633	input: Closeup picture of the front of	↪ intersection, waterway, near)	1703
1634	↪ a clean motorcycle.	10 event - ambiguity (traffic light,	1704
1635	output: 1 entity - whole (motorcycle)	↪ signpost, crossroads	1705
1636	2 global - (closeup)	↪ intersection, at)	1706
1637	3 global - (picture)	11 event - ambiguity (traffic light,	1707
1638	4 attribute - state (motorcycle,	↪ crossroads intersection, at,	1708
1639	↪ clean)	↪ waterway, near)	1709
1640	5 entity - part (front of the clean	12 event - spatial (signpost,	1710
1641	↪ motorcycle)	↪ crossroads intersection, at,	1711
1642		↪ waterway, near)	1712
1643	input: a sad man with green hair	input: a photo of dining table and	1713
1644	output: 1 entity - whole (man)	↪ traffic light; traffic light is	1714
1645	2 entity - part (man's hair)	↪ below dining table	1715
1646	3 attribute - state (man, sad)	output: 1 global - (photo)	1716
1647	4 attribute - color (man's hair,	2 entity - whole (dining table)	1717
1648	↪ green)	3 entity - whole (traffic light)	1718
1649	5 event - ambiguity (man, sad, man's	4 relation - spatial (traffic light,	1719
1650	↪ hair, green)	↪ dining table, below)	1720
1651			1721
1652	input: A commercial airplane with	input: A realistic photo of a	1722
1653	↪ propellers flying through the air.	↪ Pomeranian dressed up like a	1723
1654	output: 1 entity - whole (airplane)	↪ 1980s professional wrestler with	1724
1655	2 entity - part (airplane's	↪ neon green and neon orange face	1725
1656	↪ propellers)	↪ paint and bright green wrestling	1726
1657	3 action - (airplane, air, fly	↪ tights with bright orange boots.	1727
1658	↪ through)	output: 1 global - (photo)	1728
1659	4 event - ambiguity (airplane, with	2 entity - whole (Pomeranian)	1729
1660	↪ propellers, air, fly through)	3 global - (realistic)	1730
1661		4 entity - part (Pomeranian's costume)	1731
1662	input: A little boy grips a soccer ball	5 attribute - type (Pomeranian's	1732
1663	↪ in his arms surrounded by other	↪ costume, 1980s professional	1733
1664	↪ youth soccer players.	↪ wrestler)	1734
1665	output: 1 entity - whole (boy)	6 entity - part (Pomeranian's	1735
1666	2 entity - whole (ball)	↪ costume's wrestling tights)	1736
1667	3 entity - whole (soccer players)	7 entity - part (Pomeranian's	1737
1668	4 entity - part (boy's arms)	↪ costume's wrestling tights' boots)	1738
1669	5 entity - scale (boy, little)	8 entity - part (Pomeranian's	1739
1670	6 attribute - type (ball, soccer)	↪ facepaint)	1740
1671	7 attribute - state (soccer players,	9 attribute - color (Pomeranian's	1741
1672	↪ youth)	↪ facepaint, neon green)	1742
1673	8 relation - spatial (little boy,	10 attribute - color (Pomeranian's	1743
1674	↪ ball, grip in his arms)	↪ facepaint, neon orange)	1744
1675	9 relation - spatial (little boy	11 attribute - color (Pomeranian's	1745
1676	↪ gripping the ball in his arms,	↪ costume's wrestling tights,	1746
1677	↪ soccer players, surrounded by)	↪ bright green)	1747
1678	10 event - ambiguity (boy's arm,	12 attribute - color (Pomeranian's	1748
1679	↪ little, ball, soccer, grip in his		1749

1750	↪ costume's wrestling tights'	3 entity - whole (elephant)	1820
1751	↪ boots, bright orange)	4 entity - whole (river)	1821
1752		5 attribute - state (people, stand)	1822
1753	input: a four-piece band on a stage in	6 relation - spatial (three people,	1823
1754	↪ front of a small crowd	↪ elephant, next to)	1824
1755	output: 1 entity - whole (band)	7 relation - spatial (people, river,	1825
1756	2 entity - whole (stage)	↪ next to)	1826
1757	3 entity - whole (crowd)	8 relation - spatial (elephant,	1827
1758	4 other - count (band members, ==4)	↪ river, next to)	1828
1759	5 attribute - shape (crowd, small)	9 event - ambiguity (people, ==3,	1829
1760	6 relation - spatial (four-piece	↪ stand)	1830
1761	↪ band, stage, on)	10 event - ambiguity (people, ==3,	1831
1762	7 relation - spatial (four-piece	↪ elephant, next to)	1832
1763	↪ band, crowd, in front of)	11 event - ambiguity (people, ==3,	1833
1764	8 relation - spatial (stage, crowd,	↪ river, next to)	1834
1765	↪ in front of)	12 event - ambiguity (people, stand,	1835
1766	9 event - ambiguity (band, ==4	↪ elephant, next to)	1836
1767	↪ picece, stage, on)	13 event - ambiguity (people, stand,	1837
1768	10 event - ambiguity (band, ==4	↪ river, next to)	1838
1769	↪ picece, crowd, small, in front of)	14 event - ambiguity (people,	1839
1770	11 event - ambiguity (stage, crowd,	↪ elephant, next to, river, next to)	1840
1771	↪ small, in front off)		1841
1772		input: Aerial view of downtown	1842
1773	input: two laptops, a mouse cord, and a	↪ Manhattan, but with Millennium	1843
1774	↪ monitor	↪ Wheel next to the Statue of	1844
1775	output: 1 entity - whole (laptops)	↪ Liberty. The Great Pyramid is on	1845
1776	2 other - count (laptops, ==2)	↪ a sandy island near the buildings.	1846
1777	3 entity - whole (mouse coord)	output: 1 entity - (downtown	1847
1778	4 entity - whole (monitor)	↪ Manhattan)	1848
1779		2 entity - (Millennium Wheel)	1849
1780	input: A red motorcycle parked by paint	3 entity - (the Statue of the Liberty)	1850
1781	↪ chipped doors.	4 entity - (the Great Pyramid)	1851
1782	output: 1 entity - whole (motorcycle)	5 entity - (island)	1852
1783	2 entity - whole (doors)	6 entity - (buildings)	1853
1784	3 attribute - color (motorcycle, red)	7 global - (aerial view)	1854
1785	4 attribute - state (door, paint	8 attribute - texture (island, sandy)	1855
1786	↪ chipped)	9 relation - spatial (Millennium	1856
1787	5 relation - spatial (red motorcycle,	↪ Wheel, the Statue of Liberty,	1857
1788	↪ paint chipped door, next to)	↪ next to)	1858
1789	6 attribute - state (motorcycle,	10 relation - spatial (the Great	1859
1790	↪ parked)	↪ Pyramid, island, on)	1860
1791	7 event- ambiguity (motorcycle, red,	11 relation - spatial (the Great	1861
1792	↪ door, paint chipped, next to)	↪ Pyramid, buildings, near)	1862
1793	8 event- ambiguity (motorcycle, red,	12 event - ambiguity (the Great	1863
1794	↪ parked)	↪ Pyramid, island, on, buildings,	1864
1795		↪ near)	1865
1796	input: A cube made of denim. A cube		1866
1797	↪ with the texture of denim.	input: A laptop with external keyboard,	1867
1798	output: 1 entity - whole (cube)	↪ mouse, phone and photo on a desk.	1868
1799	2 attribute - material (cube, denim)	output: 1 entity - whole (laptop)	1869
1800	3 attribute - texture (cube, denim)	2 entity - whole (keyboard)	1870
1801		3 entity - whole (mouse)	1871
1802	input: an espresso machine that makes	4 entity - whole (phone)	1872
1803	↪ coffee from human souls	5 entity - whole (photo)	1873
1804	output: 1 entity - whole (espresso	6 entity - whole (desk)	1874
1805	↪ machine)	7 attribute - type (keyboard,	1875
1806	2 entity - whole (coffee)	↪ external)	1876
1807	3 entity - whole (human souls)	8 relation - spatial (laptop, desk,	1877
1808	4 action - (espresso machine, coffee,	↪ on)	1878
1809	↪ make)	9 relation - spatial (keyboard, desk,	1879
1810	5 attribute - material (coffee, human	↪ on)	1880
1811	↪ souls)	10 relation - spatial (mouse, desk,	1881
1812	6 event - ambiguity (espresso	↪ on)	1882
1813	↪ machine, coffee, make, human	11 relation - spatial (phone, desk,	1883
1814	↪ souls)	↪ on)	1884
1815		12 relation - spatial (photo, desk,	1885
1816	input: Three people standing next to an	↪ on)	1886
1817	↪ elephant along a river.	13 event - ambiguity (laptop,	1887
1818	output: 1 entity - whole (people)	↪ external keyboard, mouse, phone,	1888
1819	2 other - count (people, ==3)	↪ photo, desk, on)	1889

1890
1891 input: A white slope covers the
1892 ↪ background, while the foreground
1893 ↪ features a grassy slope with
1894 ↪ several rams grazing and one
1895 ↪ measly and underdeveloped
1896 ↪ evergreen in the foreground.
1897 output: 1 | entity - whole (slopes)
1898 2 | other - count (slopes, ==2)
1899 3 | entity - whole (rams)
1900 4 | entity - whole (evergreen)
1901 5 | attribute - color (slope_1, white)
1902 6 | attribute - texture (slope_2,
1903 ↪ grassy)
1904 7 | attribute - state (evergreen,
1905 ↪ measly and underdeveloped)
1906 8 | relation - spatial (slope_1,
1907 ↪ background, in)
1908 9 | relation - spatial (slope_2,
1909 ↪ foreground, in)
1910 10 | relation - spatial (several, rams,
1911 ↪ grassy slope_2, on)
1912 11 | attribute - state (several rams,
1913 ↪ graze)
1914 12 | event - ambiguity (slope_1, white,
1915 ↪ background, in)
1916 13 | event - ambiguity (slope_2,
1917 ↪ grassy, foreground, in)
1918 14 | event - ambiguity (several, rams,
1919 ↪ slope_2, grassy, on)
1920
1921 input: A man walks into a room and sits
1922 ↪ on a chair. A dog follows him.
1923 output: 1 | entity - whole (man)
1924 2 | entity - whole (room)
1925 3 | entity - whole (chair)
1926 4 | entity - whole (dog)
1927 5 | action - (man, walk, room)
1928 6 | action - (man, sit on, chair)
1929 7 | action - (dog, follow, man)
1930 8 | relation - temporal (man, sit,
1931 ↪ before, walk)
1932 9 | relation - temporal (dog, follows,
1933 ↪ after, man, sit)
1934 10 | event - temporal (man, walks into
1935 ↪ a room and sits on a chair, dog
1936 ↪ follows him)
1937
1938 input: A car is parked by the roadside.
1939 ↪ Later, it starts moving and
1940 ↪ drives away.
1941 output: 1 | entity - whole (car)
1942 2 | entity - whole (roadside)
1943 3 | relation - spatial (car, roadside,
1944 ↪ park)
1945 4 | action - (car, move)
1946 5 | action - (car, drives away)
1947 6 | relation - temporal (car, starts,
1948 ↪ after, parked)
1949 7 | relation - temporal (car, drive
1950 ↪ away, after, parked)
1951 8 | event - temporal (car, move,
1952 ↪ roadside, park, after)
1953 9 | event - temporal (car, drive away,
1954 ↪ roadside, park, after)
1955 10 | event - temporal (car, starts,
1956 ↪ parked, move, drive away)
1957
1958 input: A man is running across a street
1959 ↪ 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)

I.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 - ambiguity (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)

2027	6 attribute - state (emoji_2, angry)	2 entity - whole (orange)	2097
2028	7 attribute - state (emoji_3,	3 entity - whole (bananas)	2098
2029	↪ surprised)	4 other - count (bananas, ==2)	2099
2030	8 attribute - state (emoji_4, sobbing	5 entity - whole (bowl)	2100
2031	↪ face)	6 attribute - material (bowl, wood)	2101
2032	9 entity - part (pig's crown)	7 relation - spatial (pear, bowl, in)	2102
2033	output: 1 nan	8 relation - spatial (orange, bowl,	2103
2034	2 Is there a total of four emoji	↪ in)	2104
2035	↪ icons?	9 relation - spatial (bananas, bowl,	2105
2036	3 Were the emojis in a 2x2 grid?	↪ in)	2106
2037	4 Did emojis look like pigs?	10 relation - spatial (bananas, bowl,	2107
2038	5 Did one emoji look happy?	↪ in)	2108
2039	6 Did one emoji look angry?	11 event - ambiguity (pear, orange,	2109
2040	7 Did one emoji look surprised?	↪ bananas, ==2, bowl, in)	2110
2041	8 Did the emoji have a sobbing face?	output: 1 Is there a pear?	2111
2042	9 Are all the emoji wearing crowns?	2 Is there an orange?	2112
2043		3 Are there bananas?	2113
2044	input: a photo of bear and dining	4 Are there two bananas?	2114
2045	↪ table; dining table is below bear	5 Is there a bowl?	2115
2046	1 global - (photo)	6 Is the bowl made of wood?	2116
2047	2 entity - whole (bear)	7 Is the pear in the wooden bowl?	2117
2048	3 entity - whole (dining table)	8 Is the orange in the wooden bowl?	2118
2049	4 relation - spatial (dining table,	9 Are bananas in the wooden bowl?	2119
2050	↪ bear, below)	10 Are bananas in the wooden bowl?	2120
2051	output: 1 Is this a photo?	11 Are the pear, the orange and two	2121
2052	2 Is there a bear?	↪ bananas bananas in the same	2122
2053	3 Is there a dining table?	↪ wooden bowl?	2123
2054	4 Is the dining table below the bear?		2124
2055		input: Closeup picture of the front of	2125
2056	input: A group of children sitting in	↪ a clean motorcycle.	2126
2057	↪ the grass with two of them	1 entity - whole (motorcycle)	2127
2058	↪ holding a Frisbee .	2 global - (closeup)	2128
2059	1 entity - whole (children)	3 global - (picture)	2129
2060	2 entity - whole (grass)	4 attribute - state (motorcycle,	2130
2061	3 entity - whole (frisbee)	↪ clean)	2131
2062	4 attribute - state (children, sit)	5 entity - part (front of the clean	2132
2063	5 relation - spatial (a group of	↪ motorcycle)	2133
2064	↪ children, grass, sitting in)	output: 1 Is there a motorcycle?	2134
2065	6 entity - part (two of the children)	2 Is this a closeup image?	2135
2066	7 action - (two of the children,	3 Is this a picture?	2136
2067	↪ frisbee, hold)	4 Is the motorcycle clean?	2137
2068	output: 1 Are there a group of	5 Is the closeup picture in the front	2138
2069	↪ children?	↪ of the clean motorcycle?	2139
2070	2 Is there grass?		2140
2071	3 Is there a frisbee?	input: a sad man with green hair	2141
2072	4 Are the children sitting?	1 entity - whole (man)	2142
2073	5 Are a group of children sitting in	2 entity - part (man's hair)	2143
2074	↪ the grass?	3 attribute - state (man, sad)	2144
2075	6 Are there two of the children?	4 attribute - color (man's hair,	2145
2076	7 Are two of the children holding a	↪ green)	2146
2077	↪ frisbee?	5 event - ambiguity (man, sad, man's	2147
2078		↪ hair, green)	2148
2079	input: the word 'START' written in	output: 1 Is there a man?	2149
2080	↪ chalk on a sidewalk	2 Is there hair?	2150
2081	1 entity - whole (word)	3 Was the man sad?	2151
2082	2 entity - whole (sidewalk)	4 Is the hair green?	2152
2083	3 other - text rendering (word,	5 Is the sad man with hair green?	2153
2084	↪ "START")		2154
2085	4 attribute - texture (word, chalk)	input: A commercial airplane with	2155
2086	5 relation - spatial (word 'START',	↪ propellers flying through the air.	2156
2087	↪ sidewalk, on)	1 entity - whole (airplane)	2157
2088	output: 1 Is there a word?	2 entity - part (airplane's	2158
2089	2 Is there a sidewalk?	↪ propellers)	2159
2090	3 Does the word say "START"?	3 action - (airplane, air, fly	2160
2091	4 Is the word written in chalk?	↪ through)	2161
2092	5 Is the word 'START' on the sidewalk?	4 event - ambiguity (airplane, with	2162
2093		↪ propellers, air, fly through)	2163
2094	input: A pear, orange, and two bananas	output: 1 Is there an airplane?	2164
2095	↪ in a wooden bowl.	2 Does the airplane have propellers?	2165
2096	1 entity - whole (pear)	3 Is the airplane flying through the	2166

2167	↪ air?	2 Is there a signpost?	2237
2168	4 Is the airplane with propellers	3 Is there an intersection?	2238
2169	↪ flying through the air?	4 Is there a waterway?	2239
2170		5 Is the light a traffic light?	2240
2171	input: A little boy grips a soccer ball	6 Is the intersection a crossroads	2241
2172	↪ in his arms surrounded by other	↪ intersection?	2242
2173	↪ youth soccer players.	7 Is the traffic light at the	2243
2174	1 entity - whole (boy)	↪ crossroads intersection?	2244
2175	2 entity - whole (ball)	8 Is the signpost at the crossroads	2245
2176	3 entity - whole (soccer players)	↪ intersection?	2246
2177	4 entity - part (boy's arms)	9 Is the intersection near the	2247
2178	5 entity - scale (boy, little)	↪ waterway?	2248
2179	6 attribute - type (ball, soccer)	10 Are the traffic light and signpost	2249
2180	7 attribute - state (soccer players,	↪ at a crossroads intersection?	2250
2181	↪ youth)	11 Is the traffic light at a	2251
2182	8 relation - spatial (little boy,	↪ crossroads intersection near	2252
2183	↪ ball, grip in his arms)	↪ waterway?	2253
2184	9 relation - spatial (little boy	12 Is the signpost at a crossroads	2254
2185	↪ gripping the ball in his arms,	↪ intersection near waterway?	2255
2186	↪ soccer players, surrounded by)		2256
2187	10 event - ambiguity (boy's arm,	input: a photo of dining table and	2257
2188	↪ little, ball, soccer, grip in his	↪ traffic light; traffic light is	2258
2189	↪ arms)	↪ below dining table	2259
2190	11 event - ambiguity (boy, little,	1 global - (photo)	2260
2191	↪ soccer players, youth, surrounded	2 entity - whole (dining table)	2261
2192	↪ by)	3 entity - whole (traffic light)	2262
2193	output: 1 Is there a boy?	4 relation - spatial (traffic light,	2263
2194	2 Is there a ball?	↪ dining table, below)	2264
2195	3 Are there other soccer players?	output: 1 Is this a photo?	2265
2196	4 Does the boy have arms?	2 Is there a dining table?	2266
2197	5 Is the boy little?	3 Is there a traffic light?	2267
2198	6 Is the ball a soccer ball?	4 Is the traffic light below the	2268
2199	7 Are the other soccer players young?	↪ dining table?	2269
2200	8 Is the boy gripping the ball in his		2270
2201	↪ arms?	input: A realistic photo of a	2271
2202	9 Is the little boy surrounded by the	↪ Pomeranian dressed up like a	2272
2203	↪ other soccer players?	↪ 1980s professional wrestler with	2273
2204	10 Is the little boy gripping the	↪ neon green and neon orange face	2274
2205	↪ soccer ball in his arms?	↪ paint and bright green wrestling	2275
2206	11 Is the little boy surrounded by	↪ tights with bright orange boots.	2276
2207	↪ the other youth soccer players?	1 global - (photo)	2277
2208		2 entity - whole (Pomeranian)	2278
2209	input: A traffic light and a signpost	3 global - (realistic)	2279
2210	↪ at a crossroads intersection near	4 entity - part (Pomeranian's costume)	2280
2211	↪ a waterway.	5 attribute - type (Pomeranian's	2281
2212	1 entity - whole (traffic light)	↪ costume, 1980s professional	2282
2213	2 entity - whole (signpost)	↪ wrestler)	2283
2214	3 entity - whole (crossroads	6 entity - part (Pomeranian's	2284
2215	↪ intersection)	↪ costume's wrestling tights)	2285
2216	4 entity - whole (waterway)	7 entity - part (Pomeranian's	2286
2217	5 relation - spatial (traffic light,	↪ costume's wrestling tights' boots)	2287
2218	↪ crossroads intersection, at)	8 entity - part (Pomeranian's	2288
2219	6 relation - spatial (signpost,	↪ facepaint)	2289
2220	↪ crossroads intersection, at)	9 attribute - color (Pomeranian's	2290
2221	7 relation - spatial (traffic light,	↪ facepaint, neon green)	2291
2222	↪ waterway, near)	10 attribute - color (Pomeranian's	2292
2223	8 relation - spatial (signpost,	↪ facepaint, neon orange)	2293
2224	↪ waterway, near)	11 attribute - color (Pomeranian's	2294
2225	9 relation - spatial (crossroads	↪ costume's wrestling tights,	2295
2226	↪ intersection, waterway, near)	↪ bright green)	2296
2227	10 event - ambiguity (traffic light,	12 attribute - color (Pomeranian's	2297
2228	↪ signpost, crossroads	↪ costume's wrestling tights'	2298
2229	↪ intersection, at)	↪ boots, bright orange)	2299
2230	11 event - ambiguity (traffic light,	output: 1 Is this a photo?	2300
2231	↪ crossroads intersection, at,	2 Is there a Pomeranian?	2301
2232	↪ waterway, near)	3 Is the photo realistic?	2302
2233	12 event - spatial (signpost,	4 Is the Pomeranian dressed up?	2303
2234	↪ crossroads intersection, at,	5 Is the costume of a 1980s	2304
2235	↪ waterway, near)	↪ professional wrestler?	2305
2236	output: 1 Is there a light?	6 Are wrestling tights included in	2306

2307	↪ the costume?	2 Are there any doors?	2377
2308	7 Did the costume come with boots?	3 Are the doors painted?	2378
2309	8 Does the Pomeranian has a facepaint?	4 Is the paint chipped?	2379
2310	9 Is the facepaint neon green?	5 Is the motorcycle next to doors?	2380
2311	10 Is the facepaint neon orange?	6 Is the motorcycle parked?	2381
2312	11 Are the wrestling tights bright	7 Is the red motorcycle next to paint	2382
2313	↪ green?	↪ chipped doors?	2383
2314	12 Are the boots bright orange?	8 Is the red motorcycle parked?	2384
2315			2385
2316	input: a four-piece band on a stage in	input: A cube made of denim. A cube	2386
2317	↪ front of a small crowd	↪ with the texture of denim.	2387
2318	1 entity - whole (band)	1 entity - whole (cube)	2388
2319	2 entity - whole (stage)	2 attribute - material (cube, denim)	2389
2320	3 entity - whole (crowd)	3 attribute - texture (cube, denim)	2390
2321	4 other - count (band members, ==4)	output: 1 Is there a cube?	2391
2322	5 attribute - shape (crowd, small)	2 Is the cube made of denim?	2392
2323	6 relation - spatial (four-piece	3 Does the cube have texture of denim?	2393
2324	↪ band, stage, on)		2394
2325	7 relation - spatial (four-piece	input: an espresso machine that makes	2395
2326	↪ band, crowd, in front of)	↪ coffee from human souls	2396
2327	8 relation - spatial (stage, crowd,	1 entity - whole (espresso machine)	2397
2328	↪ in front of)	2 entity - whole (coffee)	2398
2329	9 event - ambiguity (band, ==4	3 entity - whole (human souls)	2399
2330	↪ picece, stage, on)	4 action - (espresso machine, coffee,	2400
2331	10 event - ambiguity (band, ==4	↪ make)	2401
2332	↪ picece, crowd, small, in front of)	5 attribute - material (coffee, human	2402
2333	11 event - ambiguity (stage, crowd,	↪ souls)	2403
2334	↪ small, in front off)	6 event - ambiguity (espresso	2404
2335	output: 1 Is there a band?	↪ machine, coffee, make, human	2405
2336	2 Is there a stage?	↪ souls)	2406
2337	3 Is there a crowd?	output: 1 Do we have an espresso	2407
2338	4 Is the band a fourpiece band?	↪ machine?	2408
2339	5 Is the crowd small?	2 Do we have coffee?	2409
2340	6 Is the band on the stage?	3 Do human beings have souls?	2410
2341	7 Is the band in front of the crowd?	4 Is the espresso machine making	2411
2342	8 Is the stage in front of the crowd?	↪ coffee?	2412
2343	9 Are the four-piece band on the	5 Is the experssso made of human souls?	2413
2344	↪ stage?	6 Is the experssso machine making	2414
2345	10 Is the four-piece band in front of	↪ coffe with human souls?	2415
2346	↪ the small crowd?		2416
2347	11 Is the stage in front of the small	input: Three people standing next to an	2417
2348	↪ crowd?	↪ elephant along a river.	2418
2349		1 entity - whole (people)	2419
2350	input: two laptops, a mouse cord, and a	2 other - count (people, ==3)	2420
2351	↪ monitor	3 entity - whole (elephant)	2421
2352	1 entity - whole (laptops)	4 entity - whole (river)	2422
2353	2 other - count (laptops, ==2)	5 attribute - state (people, stand)	2423
2354	3 entity - whole (mouse coord)	6 relation - spatial (three people,	2424
2355	4 entity - whole (monitor)	↪ elephant, next to)	2425
2356	output: 1 Are there laptops?	7 relation - spatial (people, river,	2426
2357	2 Are there two laptops?	↪ next to)	2427
2358	3 Is there a cord?	8 relation - spatial (elephant,	2428
2359	4 Is there a monitor?	↪ river, next to)	2429
2360		9 event - ambiguity (people, ==3,	2430
2361	input: A red motorcycle parked by paint	↪ stand)	2431
2362	↪ chipped doors.	10 event - ambiguity (people, ==3,	2432
2363	1 entity - whole (motorcycle)	↪ elephant, next to)	2433
2364	2 entity - whole (doors)	11 event - ambiguity (people, ==3,	2434
2365	3 attribute - color (motorcycle, red)	↪ river, next to)	2435
2366	4 attribute - state (door, paint	12 event - ambiguity (people, stand,	2436
2367	↪ chipped)	↪ elephant, next to)	2437
2368	5 relation - spatial (red motorcycle,	13 event - ambiguity (people, stand,	2438
2369	↪ paint chipped door, next to)	↪ river, next to)	2439
2370	6 attribute - state (motorcycle,	14 event - ambiguity (people,	2440
2371	↪ parked)	↪ elephant, next to, river, next to)	2441
2372	7 event- ambiguity (motorcycle, red,	output: 1 Are there people?	2442
2373	↪ door, paint chipped, next to)	2 Are there three people?	2443
2374	8 event- ambiguity (motorcycle, red,	3 Is there an elephant?	2444
2375	↪ parked)	4 Is there a river?	2445
2376	output: 1 Is there a motorcycle?	5 Are people standing?	2446

2447	6 Are people next to the elephant?	↪ on)	2517
2448	7 Are people next to the river?	11 relation - spatial (phone, desk,	2518
2449	8 Is the elephant next to the river?	↪ on)	2519
2450	9 Are the three people standing?	12 relation - spatial (photo, desk,	2520
2451	10 Are the three people next to the	↪ on)	2521
2452	↪ elephant?	13 event - ambiguity (laptop,	2522
2453	11 Are the three people next to the	↪ external keyboard, mouse, phone,	2523
2454	↪ river?	↪ photo, desk, on)	2524
2455	12 Are people standing next to an	output: 1 Is there a laptop?	2525
2456	↪ elephant?	2 Is there a keyboard?	2526
2457	13 Are people standing next to the	3 Is there a mouse?	2527
2458	↪ river?	4 Is there a phone?	2528
2459	14 Are people next to the river and	5 Is there a photo?	2529
2460	↪ an elephant?	6 Is there a desk?	2530
2461		7 Is the keyboard external?	2531
2462	input: Aerial view of downtown	8 Is the laptop on the desk?	2532
2463	↪ Manhattan, but with Millennium	9 Is the keyboard on the desk?	2533
2464	↪ Wheel next to the Statue of	10 Is the mouse on the desk?	2534
2465	↪ Liberty. The Great Pyramid is on	11 Is the phone on the desk?	2535
2466	↪ a sandy island near the buildings.	12 Is the photo on the desk?	2536
2467	1 entity - (downtown Manhattan)	13 Is all laptop, external keyboard,	2537
2468	2 entity - (Millennium Wheel)	↪ mouse, phone, photo on the same	2538
2469	3 entity - (the Statue of the Liberty)	↪ desk?	2539
2470	4 entity - (the Great Pyramid)		2540
2471	5 entity - (island)	input: A white slope covers the	2541
2472	6 entity - (buildings)	↪ background, while the foreground	2542
2473	7 global - (aerial view)	↪ features a grassy slope with	2543
2474	8 attribute - texture (island, sandy)	↪ several rams grazing and one	2544
2475	9 relation - spatial (Millennium	↪ measly and underdeveloped	2545
2476	↪ Wheel, the Statue of Liberty,	↪ evergreen in the foreground.	2546
2477	↪ next to)	1 entity - whole (slopes)	2547
2478	10 relation - spatial (the Great	2 other - count (slopes, ==2)	2548
2479	↪ Pyramid, island, on)	3 entity - whole (rams)	2549
2480	11 relation - spatial (the Great	4 entity - whole (evergreen)	2550
2481	↪ Pyramid, buildings, near)	5 attribute - color (slope_1, white)	2551
2482	12 event - ambiguity (the Great	6 attribute - texture (slope_2,	2552
2483	↪ Pyramid, island, on, buildings,	↪ grassy)	2553
2484	↪ near)	7 attribute - state (evergreen,	2554
2485	output: 1 Is downtown Manhattan there?	↪ measly and underdeveloped)	2555
2486	2 Is Millennium Wheel there?	8 relation - spatial (slope_1,	2556
2487	3 Is the Statue of Liberty there?	↪ background, in)	2557
2488	4 Is the Great Pyramid there?	9 relation - spatial (slope_2,	2558
2489	5 Is there an island?	↪ foreground, in)	2559
2490	6 Are there buildings?	10 relation - spatial (several, rams,	2560
2491	7 Is this an aerial view?	↪ grassy slope_2, on)	2561
2492	8 Is there the island sandy?	11 attribute - state (several rams,	2562
2493	9 Is the Millennium Wheel next to the	↪ graze)	2563
2494	↪ Statue of Liberty?	12 event - ambiguity (slope_1, white,	2564
2495	10 Is the Great Pyramid on the sandy	↪ background, in)	2565
2496	↪ island?	13 event - ambiguity (slope_2,	2566
2497	11 Is the Great Pyramid near the	↪ grassy, foreground, in)	2567
2498	↪ buildings?	14 event - ambiguity (several, rams,	2568
2499	12 Is the Great Pyramid on a sady	↪ slope_2, grassy, on)	2569
2500	↪ island near the buildings?	output: 1 Are there slopes?	2570
2501		2 Are there two slopes?	2571
2502	input: A laptop with external keyboard,	3 Are there rams?	2572
2503	↪ mouse, phone and photo on a desk.	4 Is there evergreen?	2573
2504	1 entity - whole (laptop)	5 Is one slope white?	2574
2505	2 entity - whole (keyboard)	6 Is one slope grassy?	2575
2506	3 entity - whole (mouse)	7 Is the evergreen measly and	2576
2507	4 entity - whole (phone)	↪ underdeveloped?	2577
2508	5 entity - whole (photo)	8 Is the slope in the background?	2578
2509	6 entity - whole (desk)	9 Is the slope in the foreground?	2579
2510	7 attribute - type (keyboard,	10 Are the several rams on the slope?	2580
2511	↪ external)	11 Are the several rams grazing on	2581
2512	8 relation - spatial (laptop, desk,	↪ grass?	2582
2513	↪ on)	12 Is the white slope in the	2583
2514	9 relation - spatial (keyboard, desk,	↪ background?	2584
2515	↪ on)	13 Is the grassy slope in the	2585
2516	10 relation - spatial (mouse, desk,	↪ foreground?	2586

2587	14 Are the several rams on the grassy	→ while carrying a large bag. This	2657
2588	→ slope?	→ is unusual because people	2658
2589		→ typically do not carry large bags	2659
2590	input: A man walks into a room and sits	→ while running across streets.	2660
2591	→ on a chair. A dog follows him.	1 entity - whole (man)	2661
2592	1 entity - whole (man)	2 entity - whole (street)	2662
2593	2 entity - whole (room)	3 entity - whole (bag)	2663
2594	3 entity - whole (chair)	4 relation - spatial (man, run,	2664
2595	4 entity - whole (dog)	→ street)	2665
2596	5 action - (man, walk, room)	5 entity - scale (large bag)	2666
2597	6 action - (man, sit on, chair)	6 relation - spatial (man, carry,	2667
2598	7 action - (dog, follow, man)	→ large bag)	2668
2599	8 relation - temporal (man, sit,	7 relation - temporal (man, carry,	2669
2600	→ before, walk)	→ while, running)	2670
2601	9 relation - temporal (dog, follows,	8 event - ambiguity (man, large bag,	2671
2602	→ after, man, sit)	→ carry)	2672
2603	10 event - temporal (man, walks into	9 event - temporal (man, run, street,	2673
2604	→ a room and sits on a chair, dog	→ while, carry, large bag)	2674
2605	→ follows him)	output: 1 Is there a man?	2675
2606	output: 1 Is there a man?	2 Is there a street?	2676
2607	2 Is there a room?	3 Is there a bag?	2677
2608	3 Is there a chair?	4 Is the man running across the	2678
2609	4 Is there a dog?	→ street?	2679
2610	5 Does the man walk into the room?	5 Is a bag large?	2680
2611	6 Does the man sit on the chair?	6 Is the man carrying a bag?	2681
2612	7 Does the dog follow the man?	7 Is the man carrying a bag while	2682
2613	8 Does the man sit before walking?	→ running?	2683
2614	9 Does the dog follow after the man	8 Is the man carrying a large bag?	2684
2615	→ sat?	9 Is the man carrying a big bag while	2685
2616	10 Does the dog follow after the man	→ running across a street?	2686
2617	→ who walked into a room and sits		
2618	→ on a chair?		
2619			
2620	input: A car is parked by the roadside.	I.4 Dependency Generation	2687
2621	→ Later, it starts moving and	input: A male skateboarder is trying to	2688
2622	→ drives away.	→ pull off a trick on the ramp.	2689
2623	1 entity - whole (car)	1 entity - whole (skateboarder)	2690
2624	2 entity - whole (roadside)	2 entity - whole (ramp)	2691
2625	3 relation - spatial (car, roadside,	3 attribute - type (skateboarder,	2692
2626	→ park)	→ male)	2693
2627	4 action - (car, move)	4 action - (male skateboarder, pull	2694
2628	5 action - (car, drives away)	→ off a trick)	2695
2629	6 relation - temporal (car, starts,	5 relation - spatial (male	2696
2630	→ after, parked)	→ skateboarder, ramp, on)	2697
2631	7 relation - temporal (car, drive	6 event - ambiguity (skateboarder,	2698
2632	→ away, after, parked)	→ male, pull off a trick)	2699
2633	8 event - temporal (car, move,	7 event - ambiguity (male	2700
2634	→ roadside, park, after)	→ skateboarder, ramp, on)	2701
2635	9 event - temporal (car, drive away,	8 event - ambiguity (skateboarder,	2702
2636	→ roadside, park, after)	→ pull off a trick, ramp, on)	2703
2637	10 event - temporal (car, starts,	output: 1 0	2704
2638	→ parked, move, drive away)	2 0	2705
2639	output: 1 Is there a car?	3 1	2706
2640	2 Is there a roadside?	4 1	2707
2641	3 Does the car park near the roadside?	5 1,3	2708
2642	4 Does the car move?	6 3,4	2709
2643	5 Does the car drive away?	7 3,5	2710
2644	6 Does the car move after being	8 4,5	2711
2645	→ parked?	input: A car playing soccer, digital	2712
2646	7 Does the car drive away after being	→ art.	2713
2647	→ parked?	1 entity - whole (car)	2714
2648	8 Does the car move after being	2 global - (digital art)	2715
2649	→ parked near roadside?	3 action - (car, soccer, play)	2716
2650	9 Does the car drive away after being	output: 1 0	2717
2651	→ parked near roadside?	2 0	2718
2652	10 Is that a same car which parked by	3 1	2719
2653	→ the roadsid and then starts		2720
2654	→ moving and drives away?		2721
2655		input: A set of 2x2 emoji icons with	2722
2656	input: A man is running across a street	→ happy, angry, surprised and	2723

2724	↪ sobbing faces. The emoji icons	5 2,3	2794
2725	↪ look like pigs. All of the pigs		2795
2726	↪ are wearing crowns.	input: A pear, orange, and two bananas	2796
2727	1 entity - whole (emoji icons)	↪ in a wooden bowl.	2797
2728	2 other - count (emoji icons, ==4)	1 entity - whole (pear)	2798
2729	3 attribute - state (emoji icons, 2x2	2 entity - whole (orange)	2799
2730	↪ grid)	3 entity - whole (bananas)	2800
2731	4 attribute - type (emoji icons, pig)	4 other - count (bananas, ==2)	2801
2732	5 attribute - state (emoji_1, happy)	5 entity - whole (bowl)	2802
2733	6 attribute - state (emoji_2, angry)	6 attribute - material (bowl, wood)	2803
2734	7 attribute - state (emoji_3,	7 relation - spatial (pear, bowl, in)	2804
2735	↪ surprised)	8 relation - spatial (orange, bowl,	2805
2736	8 attribute - state (emoji_4, sobbing	↪ in)	2806
2737	↪ face)	9 relation - spatial (bananas, bowl,	2807
2738	9 entity - part (pig's crown)	↪ in)	2808
2739	output: 1 0	10 relation - spatial (bananas, bowl,	2809
2740	2 0	↪ in)	2810
2741	3 1	11 event - ambiguity (pear, orange,	2811
2742	4 1	↪ bananas, ==2, bowl, in)	2812
2743	5 1	output: 1 0	2813
2744	6 1	2 0	2814
2745	7 1	3 0	2815
2746	8 1	4 0	2816
2747	9 1,4	5 0	2817
2748		6 0	2818
2749	input: a photo of bear and dining	7 1,5	2819
2750	↪ table; dining table is below bear	8 2,5	2820
2751	1 global - (photo)	9 3,5	2821
2752	2 entity - whole (bear)	10 4,9	2822
2753	3 entity - whole (dining table)	11 7,8,10	2823
2754	4 relation - spatial (dining table,		2824
2755	↪ bear, below)	input: Closeup picture of the front of	2825
2756	output: 1 0	↪ a clean motorcycle.	2826
2757	2 0	1 entity - whole (motorcycle)	2827
2758	3 0	2 global - (closeup)	2828
2759	4 2,3	3 global - (picture)	2829
2760		4 attribute - state (motorcycle,	2830
2761	input: A group of children sitting in	↪ clean)	2831
2762	↪ the grass with two of them	5 entity - part (front of the clean	2832
2763	↪ holding a Frisbee .	↪ motorcycle)	2833
2764	1 entity - whole (children)	output: 1 0	2834
2765	2 entity - whole (grass)	2 0	2835
2766	3 entity - whole (frisbee)	3 0	2836
2767	4 attribute - state (children, sit)	4 0	2837
2768	5 relation - spatial (a group of	5 1	2838
2769	↪ children, grass, sitting in)		2839
2770	6 entity - part (two of the children)	input: a sad man with green hair	2840
2771	7 action - (two of the children,	1 entity - whole (man)	2841
2772	↪ frisbee, hold)	2 entity - part (man's hair)	2842
2773	output: 1 0	3 attribute - state (man, sad)	2843
2774	2 0	4 attribute - color (man's hair,	2844
2775	3 0	↪ green)	2845
2776	4 1	5 event - ambiguity (man, sad, man's	2846
2777	5 1,2	↪ hair, green)	2847
2778	6 1	output: 1 0	2848
2779	7 3,6	2 1	2849
2780		3 1	2850
2781	input: the word 'START' written in	4 2	2851
2782	↪ chalk on a sidewalk	5 3,4	2852
2783	1 entity - whole (word)		2853
2784	2 entity - whole (sidewalk)	input: A commercial airplane with	2854
2785	3 other - text rendering (word,	↪ propellers flying through the air.	2855
2786	↪ "START")	1 entity - whole (airplane)	2856
2787	4 attribute - texture (word, chalk)	2 entity - part (airplane's	2857
2788	5 relation - spatial (word 'START',	↪ propellers)	2858
2789	↪ sidewalk, on)	3 action - (airplane, air, fly	2859
2790	output: 1 0	↪ through)	2860
2791	2 0	4 event - ambiguity (airplane, with	2861
2792	3 1	↪ propellers, air, fly through)	2862
2793	4 1	output: 1 0	2863

2864	2 0	6 2,3	2934
2865	3 1	7 1,4	2935
2866	4 2,3	8 2,4	2936
2867		9 3,4	2937
2868	input: A little boy grips a soccer ball	10 5,6	2938
2869	↪ in his arms surrounded by other	11 5,7	2939
2870	↪ youth soccer players.	12 6,8	2940
2871	1 entity - whole (boy)		2941
2872	2 entity - whole (ball)	input: a photo of dining table and	2942
2873	3 entity - whole (soccer players)	↪ traffic light; traffic light is	2943
2874	4 entity - part (boy's arms)	↪ below dining table	2944
2875	5 entity - scale (boy, little)	1 global - (photo)	2945
2876	6 attribute - type (ball, soccer)	2 entity - whole (dining table)	2946
2877	7 attribute - state (soccer players,	3 entity - whole (traffic light)	2947
2878	↪ youth)	4 relation - spatial (traffic light,	2948
2879	8 relation - spatial (little boy,	↪ dining table, below)	2949
2880	↪ ball, grip in his arms)	output: 1 0	2950
2881	9 relation - spatial (little boy	2 0	2951
2882	↪ gripping the ball in his arms,	3 0	2952
2883	↪ soccer players, surrounded by)	4 2,3	2953
2884	10 event - ambiguity (boy's arm,		2954
2885	↪ little, ball, soccer, grip in his	input: A realistic photo of a	2955
2886	↪ arms)	↪ Pomeranian dressed up like a	2956
2887	11 event - ambiguity (boy, little,	↪ 1980s professional wrestler with	2957
2888	↪ soccer players, youth, surrounded	↪ neon green and neon orange face	2958
2889	↪ by)	↪ paint and bright green wrestling	2959
2890	output: 1 0	↪ tights with bright orange boots.	2960
2891	2 0	1 global - (photo)	2961
2892	3 0	2 entity - whole (Pomeranian)	2962
2893	4 0	3 global - (realistic)	2963
2894	5 1	4 entity - part (Pomeranian's costume)	2964
2895	6 1	5 attribute - type (Pomeranian's	2965
2896	7 3	↪ costume, 1980s professional	2966
2897	8 2,4	↪ wrestler)	2967
2898	9 1,3	6 entity - part (Pomeranian's	2968
2899	10 4,5,6,8	↪ costume's wrestling tights)	2969
2900	11 5,7,9	7 entity - part (Pomeranian's	2970
2901		↪ costume's wrestling tights' boots)	2971
2902	input: A traffic light and a signpost	8 entity - part (Pomeranian's	2972
2903	↪ at a crossroads intersection near	↪ facepaint)	2973
2904	↪ a waterway.	9 attribute - color (Pomeranian's	2974
2905	1 entity - whole (traffic light)	↪ facepaint, neon green)	2975
2906	2 entity - whole (signpost)	10 attribute - color (Pomeranian's	2976
2907	3 entity - whole (crossroads	↪ facepaint, neon orange)	2977
2908	↪ intersection)	11 attribute - color (Pomeranian's	2978
2909	4 entity - whole (waterway)	↪ costume's wrestling tights,	2979
2910	5 relation - spatial (traffic light,	↪ bright green)	2980
2911	↪ crossroads intersection, at)	12 attribute - color (Pomeranian's	2981
2912	6 relation - spatial (signpost,	↪ costume's wrestling tights'	2982
2913	↪ crossroads intersection, at)	↪ boots, bright orange)	2983
2914	7 relation - spatial (traffic light,	output: 1 0	2984
2915	↪ waterway, near)	2 0	2985
2916	8 relation - spatial (signpost,	3 0	2986
2917	↪ waterway, near)	4 2	2987
2918	9 relation - spatial (crossroads	5 4	2988
2919	↪ intersection, waterway, near)	6 4	2989
2920	10 event - ambiguity (traffic light,	7 4	2990
2921	↪ signpost, crossroads	8 2	2991
2922	↪ intersection, at)	9 8	2992
2923	11 event - ambiguity (traffic light,	10 8	2993
2924	↪ crossroads intersection, at,	11 6	2994
2925	↪ waterway, near)	12 7	2995
2926	12 event - spatial (signpost,		2996
2927	↪ crossroads intersection, at,	input: a four-piece band on a stage in	2997
2928	↪ waterway, near)	↪ front of a small crowd	2998
2929	output: 1 0	1 entity - whole (band)	2999
2930	2 0	2 entity - whole (stage)	3000
2931	3 0	3 entity - whole (crowd)	3001
2932	4 0	4 other - count (band members, ==4)	3002
2933	5 1,3	5 attribute - shape (crowd, small)	3003

3004	6 relation - spatial (four-piece	1 entity - whole (espresso machine)	3074
3005	↪ band, stage, on)	2 entity - whole (coffee)	3075
3006	7 relation - spatial (four-piece	3 entity - whole (human souls)	3076
3007	↪ band, crowd, in front of)	4 action - (espresso machine, coffee,	3077
3008	8 relation - spatial (stage, crowd,	↪ make)	3078
3009	↪ in front of)	5 attribute - material (coffee, human	3079
3010	9 event - ambiguity (band, ==4	↪ souls)	3080
3011	↪ picece, stage, on)	6 event - ambiguity (espresso	3081
3012	10 event - ambiguity (band, ==4	↪ machine, coffee, make, human	3082
3013	↪ picece, crowd, small, in front of)	↪ souls)	3083
3014	11 event - ambiguity (stage, crowd,	output: 1 0	3084
3015	↪ small, in front off)	2 0	3085
3016	output: 1 0	3 0	3086
3017	2 0	4 1,2	3087
3018	3 0	5 2,3	3088
3019	4 1	6 4,5	3089
3020	5 3		3090
3021	6 2,4	input: Three people standing next to an	3091
3022	7 3,4	↪ elephant along a river.	3092
3023	8 2,3	1 entity - whole (people)	3093
3024	9 2,4	2 other - count (people, ==3)	3094
3025	10 4,5,7	3 entity - whole (elephant)	3095
3026	11 2,5,8	4 entity - whole (river)	3096
3027		5 attribute - state (people, stand)	3097
3028	input: two laptops, a mouse cord, and a	6 relation - spatial (three people,	3098
3029	↪ monitor	↪ elephant, next to)	3099
3030	1 entity - whole (laptops)	7 relation - spatial (people, river,	3100
3031	2 other - count (laptops, ==2)	↪ next to)	3101
3032	3 entity - whole (mouse coord)	8 relation - spatial (elephant,	3102
3033	4 entity - whole (monitor)	↪ river, next to)	3103
3034	output: 1 0	9 event - ambiguity (people, ==3,	3104
3035	2 0	↪ stand)	3105
3036	3 0	10 event - ambiguity (people, ==3,	3106
3037	4 0	↪ elephant, next to)	3107
3038		11 event - ambiguity (people, ==3,	3108
3039	input: A red motorcycle parked by paint	↪ river, next to)	3109
3040	↪ chipped doors.	12 event - ambiguity (people, stand,	3110
3041	1 entity - whole (motorcycle)	↪ elephant, next to)	3111
3042	2 entity - whole (doors)	13 event - ambiguity (people, stand,	3112
3043	3 attribute - color (motorcycle, red)	↪ river, next to)	3113
3044	4 attribute - state (door, paint	14 event - ambiguity (people,	3114
3045	↪ chipped)	↪ elephant, next to, river, next to)	3115
3046	5 relation - spatial (red motorcycle,	output: 1 0	3116
3047	↪ paint chipped door, next to)	2 1	3117
3048	6 attribute - state (motorcycle,	3 0	3118
3049	↪ parked)	4 0	3119
3050	7 event- ambiguity (motorcycle, red,	5 1	3120
3051	↪ door, paint chipped, next to)	6 1,3	3121
3052	8 event- ambiguity (motorcycle, red,	7 1,4	3122
3053	↪ parked)	8 2,4	3123
3054	output: 1 0	9 2,5	3124
3055	2 0	10 2,6	3125
3056	3 0	11 2,7	3126
3057	4 1	12 5,6	3127
3058	5 2	13 5,7	3128
3059	6 2,3	14 6,7	3129
3060	7 3,4,5		3130
3061	8 3,6	input: Aerial view of downtown	3131
3062		↪ Manhattan, but with Millennium	3132
3063	input: A cube made of denim. A cube	↪ Wheel next to the Statue of	3133
3064	↪ with the texture of denim.	↪ Liberty. The Great Pyramid is on	3134
3065	1 entity - whole (cube)	↪ a sandy island near the buildings.	3135
3066	2 attribute - material (cube, denim)	1 entity - (downtown Manhattan)	3136
3067	3 attribute - texture (cube, denim)	2 entity - (Millennium Wheel)	3137
3068	output: 1 0	3 entity - (the Statue of the Liberty)	3138
3069	2 1	4 entity - (the Great Pyramid)	3139
3070	3 1	5 entity - (island)	3140
3071		6 entity - (buildings)	3141
3072	input: an espresso machine that makes	7 global - (aerial view)	3142
3073	↪ coffee from human souls	8 attribute - texture (island, sandy)	3143

3144	9 relation - spatial (Millennium	5 attribute - color (slope_1, white)	3214
3145	↪ Wheel, the Statue of Liberty,	6 attribute - texture (slope_2,	3215
3146	↪ next to)	↪ grassy)	3216
3147	10 relation - spatial (the Great	7 attribute - state (evergreen,	3217
3148	↪ Pyramid, island, on)	↪ measly and underdeveloped)	3218
3149	11 relation - spatial (the Great	8 relation - spatial (slope_1,	3219
3150	↪ Pyramid, buildings, near)	↪ background, in)	3220
3151	12 event - ambiguity (the Great	9 relation - spatial (slope_2,	3221
3152	↪ Pyramid, island, on, buildings,	↪ foreground, in)	3222
3153	↪ near)	10 relation - spatial (several, rams,	3223
3154	output: 1 0	↪ grassy slope_2, on)	3224
3155	2 0	11 attribute - state (several rams,	3225
3156	3 0	↪ graze)	3226
3157	4 0	12 event - ambiguity (slope_1, white,	3227
3158	5 0	↪ background, in)	3228
3159	6 0	13 event - ambiguity (slope_2,	3229
3160	7 0	↪ grassy, foreground, in)	3230
3161	8 5	14 event - ambiguity (several, rams,	3231
3162	9 2,3	↪ slope_2, grassy, on)	3232
3163	10 4,5	output: 1 0	3233
3164	11 4,6	2 1	3234
3165	12 10,11	3 0	3235
3166		4 0	3236
3167	input: A laptop with external keyboard,	5 1	3237
3168	↪ mouse, phone and photo on a desk.	6 1	3238
3169	1 entity - whole (laptop)	7 4	3239
3170	2 entity - whole (keyboard)	8 5	3240
3171	3 entity - whole (mouse)	9 1	3241
3172	4 entity - whole (phone)	10 1	3242
3173	5 entity - whole (photo)	11 1,3	3243
3174	6 entity - whole (desk)	12 5,8	3244
3175	7 attribute - type (keyboard,	13 6,9	3245
3176	↪ external)	14 6,10	3246
3177	8 relation - spatial (laptop, desk,		3247
3178	↪ on)	input: A man walks into a room and sits	3248
3179	9 relation - spatial (keyboard, desk,	↪ on a chair. A dog follows him.	3249
3180	↪ on)	1 entity - whole (man)	3250
3181	10 relation - spatial (mouse, desk,	2 entity - whole (room)	3251
3182	↪ on)	3 entity - whole (chair)	3252
3183	11 relation - spatial (phone, desk,	4 entity - whole (dog)	3253
3184	↪ on)	5 action - (man, walk, room)	3254
3185	12 relation - spatial (photo, desk,	6 action - (man, sit on, chair)	3255
3186	↪ on)	7 action - (dog, follow, man)	3256
3187	13 event - ambiguity (laptop,	8 relation - temporal (man, sit,	3257
3188	↪ external keyboard, mouse, phone,	↪ before, walk)	3258
3189	↪ photo, desk, on)	9 relation - temporal (dog, follows,	3259
3190	output: 1 0	↪ after, man, sit)	3260
3191	2 0	10 event - temporal (man, walks into	3261
3192	3 0	↪ a room and sits on a chair, dog	3262
3193	4 0	↪ follows him)	3263
3194	5 0	output: 1 0	3264
3195	6 0	2 0	3265
3196	7 0	3 0	3266
3197	8 1,6	4 0	3267
3198	9 2,6	5 1,2	3268
3199	10 3,6	6 1,3	3269
3200	11 4,6	7 1,4	3270
3201	12 5,6	8 5,7	3271
3202	13 8,9,10,11,12	9 6,7	3272
3203		10 8,9	3273
3204	input: A white slope covers the		3274
3205	↪ background, while the foreground	input: A car is parked by the roadside.	3275
3206	↪ features a grassy slope with	↪ Later, it starts moving and	3276
3207	↪ several rams grazing and one	↪ drives away.	3277
3208	↪ measly and underdeveloped	1 entity - whole (car)	3278
3209	↪ evergreen in the foreground.	2 entity - whole (roadside)	3279
3210	1 entity - whole (slopes)	3 relation - spatial (car, roadside,	3280
3211	2 other - count (slopes, ==2)	↪ park)	3281
3212	3 entity - whole (rams)	4 action - (car, move)	3282
3213	4 entity - whole (evergreen)	5 action - (car, drives away)	3283

```

3284 6 | relation - temporal (car, starts,
3285   ↪ after, parked)
3286 7 | relation - temporal (car, drive
3287   ↪ away, after, parked)
3288 8 | event - temporal (car, move,
3289   ↪ roadside, park, after)
3290 9 | event - temporal (car, drive away,
3291   ↪ roadside, park, after)
3292 10 | event - temporal (car, starts,
3293   ↪ parked, move, drive away)
3294 output: 1 | 0
3295 2 | 0
3296 3 | 1,2
3297 4 | 1
3298 5 | 1
3299 6 | 1,4
3300 7 | 1, 5
3301 8 | 3,6
3302 9 | 3,7
3303 10 | 6,7
3304
3305 input: A man is running across a street
3306   ↪ while carrying a large bag. This
3307   ↪ is unusual because people
3308   ↪ typically do not carry large bags
3309   ↪ while running across streets.
3310 1 | entity - whole (man)
3311 2 | entity - whole (street)
3312 3 | entity - whole (bag)
3313 4 | relation - spatial (man, run,
3314   ↪ street)
3315 5 | entity - scale (large bag)
3316 6 | relation - spatial (man, carry,
3317   ↪ large bag)
3318 7 | relation - temporal (man, carry,
3319   ↪ while, running)
3320 8 | event - ambiguity (man, large bag,
3321   ↪ carry)
3322 9 | event - temporal (man, run, street,
3323   ↪ while, carry, large bag)
3324 output: 1 | 0
3325 2 | 0
3326 3 | 0
3327 4 | 1, 2
3328 5 | 3
3329 6 | 1, 5
3330 7 | 4, 7
3331 8 | 5,6
3332 9 | 4, 7

```

3333 J Annotation Details

3334 We show UI for all human evaluation tasks in Fig-
3335 ure 5, Figure 6, Figure 7, Figure 8, and Figure 9.

Please score the faithfulness for the video-text pair.

Question:

Generate a short caption of the video.

Text:

A man in a black shirt and camouflage pants is shooting a bow and arrow at a target.

Video



Please rate the faithfulness (1 to 5):

- ☐ 1 - Completely Hallucinated
- ☐ 2 - Mostly Unfaithful
- ☐ 3 - Partially Faithful
- ☐ 4 - Mostly Faithful
- ☐ 5 - Fully Faithful

Save Results

Submit Results

Figure 5: UI for faithfulness evaluation of human annotation.

Task 2: Matched Tuple-Question Pairs Annotation

You are provided with a list of **tuples** and a list of **questions**. Your task is to identify and annotate which questions semantically match which tuples.

Please input pairs of matching tuple/question indices.

Each pair should indicate that the **tuple** and **question** express the *same meaning* or describe the *same concept*.

✔ **Example:**

Tuple: 7 | action - (humans, feed, hamsters)

Question: 7 | Are humans feeding the hamsters?

✔ This pair is semantically consistent and should be included.

✖ **Non-matching pairs** should **not** be included in your annotation.

Use the "Add Pair" button to input multiple matching pairs.

Please ensure the tuple and question indices are within valid ranges and avoid duplicates.

Elapsed Time: 0:0:8

Tuples annotated by human:

1 | entity - whole (woman)

2 | entity - whole (product)

3 | action - (woman, talk about, product)

Questions generated by models:

1 | Is there a woman?

2 | Is there a product?

3 | Is the woman talking about a product?

Semantic Match Pairs

1	1	Delete
2	2	Delete
3	3	Delete

Add Pair

Matched Pair Count: 3

Save Results

Submit Results

Figure 6: UI for question quality evaluation of human annotation.

Task 3: Dependency Verification

You are given a list of **tuples** extracted from a textual description, as well as a list of **dependencies** between them. Each dependency is written in the format **a | b**, indicating that **tuple a** is **semantically dependent on tuple b**.

For example, **6 | 3,4** means *tuple 6 depends on tuples 3 and 4* to be meaningful or accurate.

Your task is to **verify whether each listed dependency is logically valid** based on the content and relationships among the tuples.

- Select **"Valid"** if the dependency makes logical sense (i.e., **a** truly requires **b** to be understood or supported).
- Select **"Invalid"** if the dependency is not necessary or does not reflect a real relationship.

Example:

Tuples:

3 | attribute – state (woman, emotionally distressed)

4 | action – (woman, cry)

6 | event – ambiguity (woman, cry, emotionally distressed)

Dependency:

6 | 3,4 → This means Tuple 6 (the event) depends on both Tuples 3 and 4.

✓ This dependency is **Valid**, because Tuple 6 combines the action ("cry") and the state ("emotionally distressed") described separately in Tuples 3 and 4.

Please review all tuples and dependencies carefully before making your judgment.

Elapsed Time: 0:0:12

Text:

First, we see a woman standing in front of a table with a bag on it. Next, we see a young woman holding a bag and talking to another woman. Then, a woman is seen holding a purse in front of a table with a bag on it. After that, we see a woman standing in front of a table with a bag on it and talking to another woman. Next, we see a woman holding a purse in front of a table with a bag on it. Then, a woman is seen standing in front of a table with a bag on it and talking to another woman. Following that, we see a woman holding a purse in front of a table with a bag on it. Finally, we see a woman standing in front of a table with a bag on it and talking to another woman. Throughout the video, we see various objects such as bags, purses, tables, and chairs. We also see a woman wearing a red shirt and a young girl with a black backpack.

Tuples:

1 | entity – whole (woman)

2 | entity – whole (table)

3 | entity – whole (bag)

4 | entity – whole (young woman)

5 | entity – whole (purse)

6 | entity – whole (chair)

7 | entity – whole (shirt)

8 | entity – whole (backpack)

9 | attribute – color (shirt, red)

10 | attribute – color (backpack, black)

11 | relation – spatial (woman, stand, in front of, table)

12 | relation – spatial (bag, table, on)

13 | action – (young woman, hold, bag)

14 | action – (woman, talk, another woman)

15 | action – (woman, hold, purse)

16 | relation – spatial (purse, in front of, table)

17 | event – temporal (woman, stand, in front of, table, bag, on)

18 | event – temporal (young woman, hold, bag, talk, another woman)

19 | event – temporal (woman, hold, purse, in front of, table, bag, on)

20 | event – temporal (woman, stand, in front of, table, bag, on, talk, another woman)

21 | event – temporal (woman, hold, purse, in front of, table, bag, on)

22 | event – temporal (woman, stand, in front of, table, bag, on, talk, another woman)

23 | event – temporal (woman, hold, purse, in front of, table, bag, on)

24 | event – temporal (woman, stand, in front of, table, bag, on, talk, another woman)

Dependency Validation:

Tuple 1 dependes on Tuple 0? ☒ Valid ☐ Invalid

Tuple 2 dependes on Tuple 0? ☒ Valid ☐ Invalid

Tuple 3 dependes on Tuple 0? ☒ Valid ☐ Invalid

Tuple 4 dependes on Tuple 0? ☒ Valid ☐ Invalid

Tuple 5 dependes on Tuple 0? ☒ Valid ☐ Invalid

Tuple 6 dependes on Tuple 0? ☒ Valid ☐ Invalid

Tuple 7 dependes on Tuple 0? ☒ Valid ☐ Invalid

Tuple 8 dependes on Tuple 0? ☒ Valid ☐ Invalid

Tuple 9 dependes on Tuple 7? ☒ Valid ☐ Invalid

Tuple 10 dependes on Tuple 8? ☒ Valid ☐ Invalid

Tuple 11 dependes on Tuple 1, 2? ☒ Valid ☐ Invalid

Tuple 12 dependes on Tuple 2, 3? ☒ Valid ☐ Invalid

Tuple 13 dependes on Tuple 4, 3? ☒ Valid ☐ Invalid

Tuple 14 dependes on Tuple 1? ☒ Valid ☐ Invalid

Tuple 15 dependes on Tuple 1, 5? ☒ Valid ☐ Invalid

Tuple 16 dependes on Tuple 5, 2? ☒ Valid ☐ Invalid

Tuple 17 dependes on Tuple 11, 12? ☒ Valid ☐ Invalid

Tuple 18 dependes on Tuple 13, 14? ☒ Valid ☐ Invalid

Tuple 19 dependes on Tuple 15, 16? ☒ Valid ☐ Invalid

Tuple 20 dependes on Tuple 11, 12, 14? ☒ Valid ☐ Invalid

Tuple 21 dependes on Tuple 15, 16? ☒ Valid ☐ Invalid

Tuple 22 dependes on Tuple 11, 12, 14? ☒ Valid ☐ Invalid

Tuple 23 dependes on Tuple 15, 16? ☒ Valid ☐ Invalid

Tuple 24 dependes on Tuple 11, 12, 14? ☒ Valid ☐ Invalid

Save Results

Submit Results

Figure 7: UI for dependency verification of human annotation.

Task 4: Tuple to Question

You are provided with a list of **tuples**, which represent structured facts extracted from a longer passage of text. For each tuple, an **automatically generated yes/no question** is provided.

Your task is to **evaluate whether the question accurately captures the meaning of the tuple** and is a **well-formed yes/no question**.

Please go through each pair of (tuple, question) and decide if the question is valid.

At the end, enter the **total number of valid questions**.

- This number must be between 0 and the total number of tuples.
- A **valid** question should be:
 - **Semantically faithful** to the original tuple.
 - **Grammatically correct** and **clear**.
 - **Well-formed** as a **yes/no question**.

Elapsed Time: 0:0:8

Tuples:

```
1 | entity - whole (hockey player)
2 | attribute - color (hockey player's jersey, blue and white)
3 | entity - whole (gym)
4 | action - (hockey player, play, game)
5 | relation - spatial (hockey player, gym, in)
6 | entity - whole (ice)
7 | action - (hockey player, skate, ice)
8 | entity - whole (basketball hoop)
9 | entity - whole (red and white ball)
10 | entity - whole (white and red ball)
11 | event - temporal (hockey player, play, game, gym, first)
12 | event - temporal (hockey player, skate, ice, next)
13 | event - temporal (another hockey player, skate, ice, then)
14 | event - temporal (hockey player, play, game, gym, after that)
15 | event - temporal (another hockey player, skate, ice, then)
16 | event - temporal (hockey player, play, game, gym, finally)
```

Questions:

```
1 | Is there a hockey player?
2 | Is the hockey player's jersey blue and white?
3 | Is there a gym?
4 | Is the hockey player playing a game?
5 | Is the hockey player in the gym?
6 | Is there ice?
7 | Is the hockey player skating on the ice?
8 | Is there a basketball hoop?
9 | Is there a red and white ball?
10 | Is there a white and red ball?
11 | Does the hockey player play a game in the gym first?
12 | Does the hockey player skate on the ice next?
13 | Does another hockey player skate on the ice then?
14 | Does the hockey player play a game in the gym after that?
15 | Does another hockey player skate on the ice then?
16 | Does the hockey player play a game in the gym finally?
```

Number of correct questions (0 ~ total):

16

Save Results

Submit Results

Figure 8: UI for the fact-to-question task of human evaluation.

Task 5: Video Question Answering

You are given a list of tuples and a set of Yes/No questions automatically generated from them. Your task is to watch the video and evaluate whether each question is grounded in the video content.

Please choose:

- **Yes:** The question is clearly supported by what is shown in the video.
- **No:** The video clearly contradicts the question.
- **Invalid:** The question depends on something that is not true (e.g., it is based on another question whose answer is "No").

Examples

Assume the video shows a dog running in a sunny garden.

- **Q1:** Is there a dog? → *Yes*
- **Q2:** Is the dog running? → *Yes* (Valid because Q1 = Yes)
- **Q3:** Is there a cat? → *No*
- **Q4:** Is the cat sleeping? → *Invalid* (depends on Q3 = No)
- **Q5:** Is it sunny? → *Yes*
- **Q6:** Is the dog enjoying the sun? → *Yes* (Valid because Q1 + Q5 = Yes)
- **Q7:** Is there a bird? → *No*
- **Q8:** Is the bird flying? → *Invalid* (depends on hallucinated bird)

Elapsed Time: 0:0:10

Query:

Generate a short caption of the video.

Answer:

First, we see a bowl of noodles with shrimp and vegetables on a red tablecloth. Next, we see the same bowl of noodles with shrimp and vegetables on a red tablecloth, but this time with a green leaf in the bowl. Then, we see a bowl of noodles with shrimp and vegetables on a red tablecloth with a green leaf in the bowl. In the following scene, we see the same bowl of noodles with shrimp and vegetables on a red tablecloth, but this time with a green leaf in the bowl and a green leaf in the background. Next, we see the same bowl of noodles with shrimp and vegetables on a red tablecloth, but this time with a green leaf in the bowl and a green leaf in the background. Then, we see the same bowl of noodles with shrimp and vegetables on a red tablecloth, but this time with a green leaf in the bowl and a green leaf in the background. Finally, we see the same bowl of noodles with shrimp and vegetables on a red tablecloth, but this time with a green leaf in the bowl and a green leaf in the background. Throughout the video, we see shrimp and vegetables in the bowl and a red table 1 | entity - whole (bowl of noodles) 2 | entity - whole (shrimp) 3 | entity - whole (vegetables) 4 | entity - whole (tablecloth) 5 | entity - whole (green leaf) 6 | attribute - color (tablecloth, red) 7 | relation - spatial (bowl of noodles, tablecloth, on) 8 | relation - spatial (shrimp, bowl of noodles, in) 9 | relation - spatial (vegetables, bowl of noodles, in) 10 | relation - spatial (green leaf, bowl of noodles, in) 11 | relation - spatial (green leaf, background, in) 12 | event - ambiguity (bowl of noodles, shrimp, vegetables, tablecloth, red, on) 13 | event - ambiguity (bowl of noodles, shrimp, vegetables, green leaf, in) 14 | event - ambiguity (bowl of noodles, shrimp, vegetables, green leaf, in, tablecloth, red, on) 15 | event - ambiguity (bowl of noodles, shrimp, vegetables, green leaf, in, background, in)

Questions:

1 Is there a bowl of noodles?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
2 Is there shrimp?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
3 Are there vegetables?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
4 Is there a tablecloth?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
5 Is there a green leaf?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
6 Is the tablecloth red?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
7 Is the bowl of noodles on the tablecloth?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
8 Is the shrimp in the bowl of noodles?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
9 Are the vegetables in the bowl of noodles?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
10 Is the green leaf in the bowl of noodles?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
11 Is the green leaf in the background?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
12 Is the bowl of noodles with shrimp and vegetables on a red tablecloth?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
13 Is the bowl of noodles with shrimp, vegetables, and a green leaf in it?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
14 Is the bowl of noodles with shrimp, vegetables, and a green leaf in it on a red tablecloth?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		
15 Is the bowl of noodles with shrimp, vegetables, and a green leaf in it with a green leaf in the background?	Yes <input checked="" type="radio"/>	No <input type="radio"/>
Invalid Question <input type="radio"/>		

Save Results

Submit Results

Video



Figure 9: UI for Video Question Answering of human annotation.