VIDEO-IFBENCH: EVALUATING INSTRUCTION FOLLOWING OF MULTIMODAL LLMS IN VIDEO UNDERSTANDING SCENARIOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have shown strong performance in video understanding. However, their ability to follow instructions in this domain remains under-explored. Real-world video understanding requires models not only to interpret video content correctly, but also to satisfy diverse user-specified constraints. Existing benchmarks focus primarily on task accuracy rather than instruction adherence, leaving this capability insufficiently evaluated. To address this gap, we introduce Video-IFBench, a comprehensive benchmark for evaluating instruction following in video understanding, where models must satisfy diverse user-specified constraints, including those grounded in visual and audio content. We develop an instruction taxonomy with four templates, including single-task, multi-task, selection, and nested instructions, covering 32 task types and 39 manually designed constraint categories spanning both semantic and format requirements. To reduce annotation cost, we build a semi-automatic data construction pipeline that combines MLLMs, programmatic processing, and human verification, resulting in 1.5K samples. We conduct a large-scale evaluation of more than 20 recent MLLMs and show that video instruction following remains challenging for current models, especially for instructions with many constraints, semantic constraints, or complex conditional structures that require selecting the correct branch or path based on video content. We hope our work will facilitate future research on instruction following in video understanding scenarios.

1 INTRODUCTION

Recent Multimodal Large Language Models (MLLMs) (Team et al., 2023; Zhu et al., 2025; Liu et al., 2025; Bai et al., 2025; Team, 2026; Xu et al., 2025b) have advanced rapidly and achieved strong performance on various video understanding tasks. At the same time, the evaluation of video understanding is rapidly expanding. Existing benchmarks examine diverse capabilities, including temporal understanding (Liu et al., 2024), spatial intelligence (Yang et al., 2025), general video analysis (Fu et al., 2025; 2026; Li et al., 2024; Zhao et al., 2025), long-video comprehension (Zhou et al., 2025; Wu et al., 2024), and knowledge-intensive reasoning (Hu et al., 2025). However, evaluating video understanding solely through response correctness is increasingly insufficient for real-world applications.

An important yet underexplored dimension is instruction following. In practical video understanding scenarios, user requests are often accompanied by explicit requirements related to semantic (Li et al., 2026; Wang et al., 2026; Shen et al., 2026) and format constraints (Zhou et al., 2023; Ding et al., 2025; Pyatkin et al., 2025) on model responses. They may also vary in structural complexity, from requests that seek a single piece of information from the video to multi-aspect requests and conditional instructions: models must identify the correct condition branch or path based on the actual video content, execute the corresponding instructions, and satisfy the associated constraints. For example, when analyzing a cooking tutorial, a user may ask an MLLM to describe only the steps after the ingredients are mixed, and list them in chronological order with timestamps. Therefore, the ability to follow instructions is a critical dimension for assessing whether such models can be reliably deployed in real-world scenarios.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite its practical importance, this capability is not yet adequately evaluated. Existing video benchmarks focus primarily on response correctness (Li et al., 2024; Fu et al., 2025; Liu et al., 2024; Zhou et al., 2025; Wu et al., 2024; Fu et al., 2026). They are not designed to test whether MLLMs faithfully execute complex user instructions. Meanwhile, prior work on instruction-following evaluation focuses mainly on text-only settings (Jing et al., 2023; Jiang et al., 2024; Zhou et al., 2023; Qin et al., 2024; Wen et al., 2024; Pyatkin et al., 2025), or on relatively narrow multimodal settings such as image-based instruction following (Ding et al., 2025; Qian et al., 2024; He et al., 2026), and controllable video captioning (Li et al., 2025b). As a result, instruction following in more diverse forms, especially under richer visual/audio-grounded constraints, remains insufficiently explored.

To address this gap, we introduce **Video-IFBench**, a comprehensive benchmark for evaluating instruction-following capabilities in video understanding. Video-IFBench contains over 700 videos from public data sources, covering diverse domains with individual durations ranging from 10 seconds to 10 minutes and a total duration of approximately 49 hours. We design an instruction taxonomy with four templates, including single-task, multi-task, selection, and nested instructions, covering 32 manually curated video understanding task types and 39 semantic or format response constraints. To reduce annotation cost, we develop a semi-automatic data construction pipeline that combines MLLM-powered extraction of global and fine-grained local video information, multi-stage complex instruction and checklist generation driven by both MLLMs and programmatic rules, and rigorous human verification, resulting in 1.5K high-quality samples.

Beyond data construction, we design an evaluation protocol that combines LLM-as-Judge with programmatic verification, and conduct a large-scale empirical study on more than 20 recent MLLMs. Our results show that even the strongest video MLLMs in our evaluation still leave substantial room for improvement in instruction following, with the best model achieving only 54.5% overall score. We further find that models struggle particularly with instructions that involve many constraints, semantic constraints, and complex conditional structures that require matching the correct branch or path based on video content. These findings suggest that strong video understanding accuracy does not automatically translate into faithful adherence to user intent. We therefore argue that instruction following should be treated as a first-class evaluation dimension for video MLLMs.

2 Related Work

2.1 Instruction-Following Evaluation on LLMs.

Instruction-following evaluation has been widely studied in text-only settings. Benchmarks such as FollowEval (Jing et al., 2023), FollowBench (Jiang et al., 2024), IFEval (Zhou et al., 2023), and InFoBench (Qin et al., 2024) evaluate whether models satisfy explicit and structured instruction requirements. Later work further studies more complex and realistic settings, including compositional constraints (Wen et al., 2024), real-user multi-constraint instructions (Lior et al., 2025), unseen constraints (Pyatkin et al., 2025), sequential or multi-turn instructions (Chen et al., 2024; He et al., 2024; Li et al., 2025a), and agentic tasks (Qi et al., 2025; Chen et al., 2026).

2.2 Multimodal Instruction Following.

Recent benchmarks extend instruction-following evaluation to multimodal inputs. MIA-Bench (Qian et al., 2024), MM-IFEngine (Ding et al., 2025), VC-IFEval (He et al., 2026), MMMT-IF (Epstein et al., 2024), and MCIF (Papi et al., 2025) evaluate instruction following across image, speech, dialogue, and crosslingual settings. In the video domain, IF-VidCap (Li et al., 2025b) studies constrained video captioning. In contrast, our work evaluates instruction following across broader video understanding tasks and more complex instruction structures.

2.3 Video Understanding Benchmarks.

Existing video understanding benchmarks evaluate temporal understanding, long-video comprehension, knowledge-intensive reasoning, robustness, and streaming interaction, including MVBench (Li et al., 2024), Video-MME (Fu et al., 2025), TempCompass (Liu et al., 2024), Video-MMMU (Hu et al., 2025), MLVU (Zhou et al., 2025), MMVU (Zhao et al., 2025), LongVideoBench (Wu et al., 2024), PerceptionComp (Li et al., 2026), Video-MME-v2 (Fu et al., 2026), PhoStream (Lu et al.,

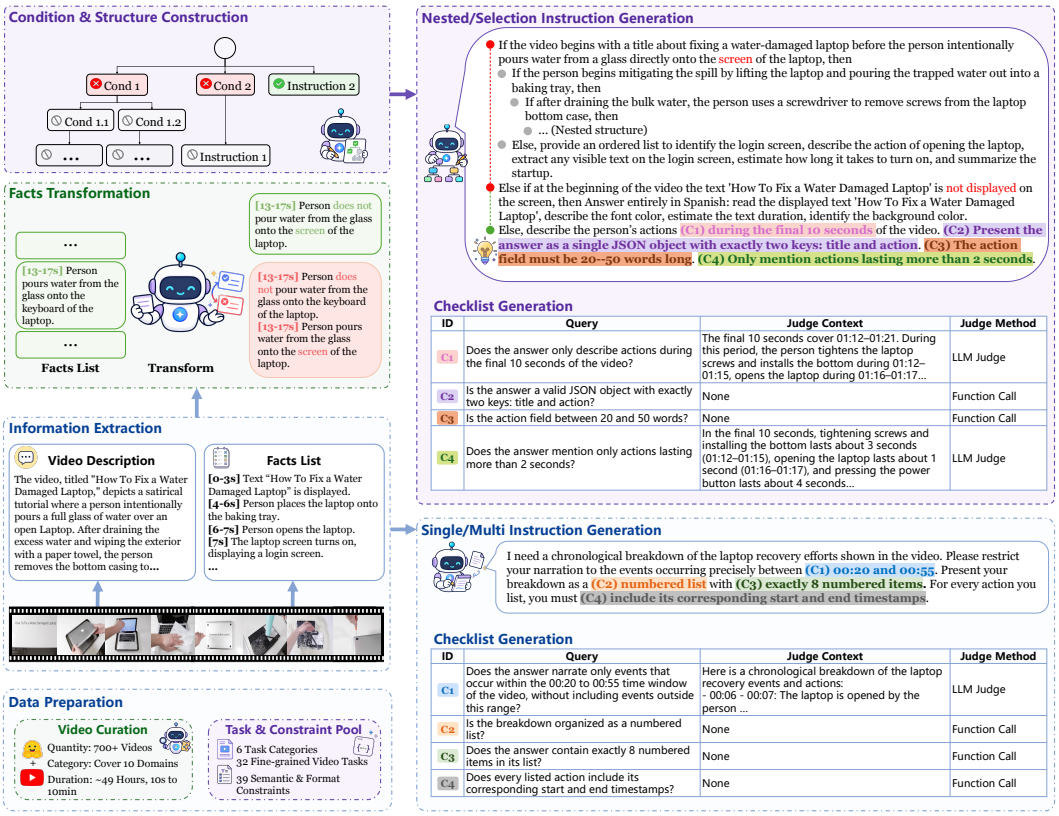


Figure 1: The data construction pipeline of Video-IFBench. Each curated video is annotated with a global description and fine-grained timestamped facts, which are used to generate Single/Multi instructions and condition-based Selection/Nested instructions. Each instruction is paired with a checklist for LLM-based or programmatic verification. More examples are provided in Appendix D.

2026), and OVO-Bench (Niu et al., 2025). These benchmarks primarily measure response correctness, but provide limited evaluation of whether models can faithfully follow complex user instructions with multiple constraints and conditional structures.

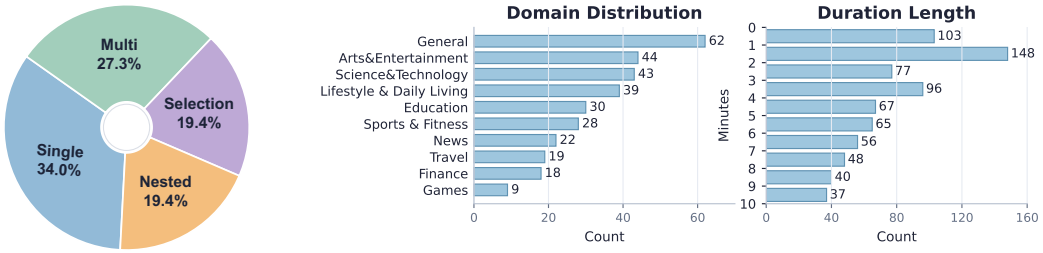
3 VIDEO-IFBENCH

3.1 BENCHMARK DESIGN

To comprehensively evaluate the video instruction-following ability of MLLMs, we design an instruction taxonomy with four instruction templates: **Single**, **Multi**, **Selection**, and **Nested**. These target two key dimensions of instruction complexity: task composition and conditional structure.

Single and Multi. Single and Multi instructions vary task-level composition. A Single instruction contains one task that asks for a specific piece of information from the video, together with a set of response constraints. We use this template to study how performance changes as the number of constraints increases, with constraint counts ranging from 1 to 15. In contrast, a Multi instruction contains two or more tasks and requires the model to retrieve or reason over multiple pieces of video information in a single request. Its constraints may apply either globally to the entire response or locally to specific tasks. To focus on the effect of task count, we keep the number of global and task-specific constraints relatively small, mostly no more than three.

Selection and Nested. Selection and Nested instructions evaluate whether models can follow video-grounded conditions. In Selection, the model chooses the correct conditional branch from



(a) Instruction types.

(b) Video domains and durations.

Figure 2: Dataset-level statistics of Video-IFBench.

several candidate branches and executes the corresponding instruction. Nested further organizes such conditions into a tree structure, requiring the model to follow the correct path before answering.

3.2 Dataset Construction

As shown in Fig. 1, our pipeline includes video curation, task and constraint pool construction, information extraction, instruction generation, checklist generation, and human verification. We collect and filter public videos, build task and constraint pools, and extract a global description and fine-grained timestamped facts for each video. These annotations are then used to generate Single/Multi instructions and condition-based Selection/Nested instructions. We further generate checklists for automatic evaluation and manually verify the constructed samples.

Video Curation. We collect videos from diverse public sources, including existing video understanding benchmarks (WorldSense (Hong et al., 2025), Video-MME (Fu et al., 2025), MMBench-Video (Fang et al., 2024), FineVideo (Farré et al., 2024), and LongVALE (Geng et al., 2025)). We also use a set of manually designed keywords to retrieve additional videos from public platforms, covering ten domains. To ensure visual quality, we use MUSIQ (Ke et al., 2021) to filter out blurred videos. The domain and duration distributions are shown in Fig. 2b.

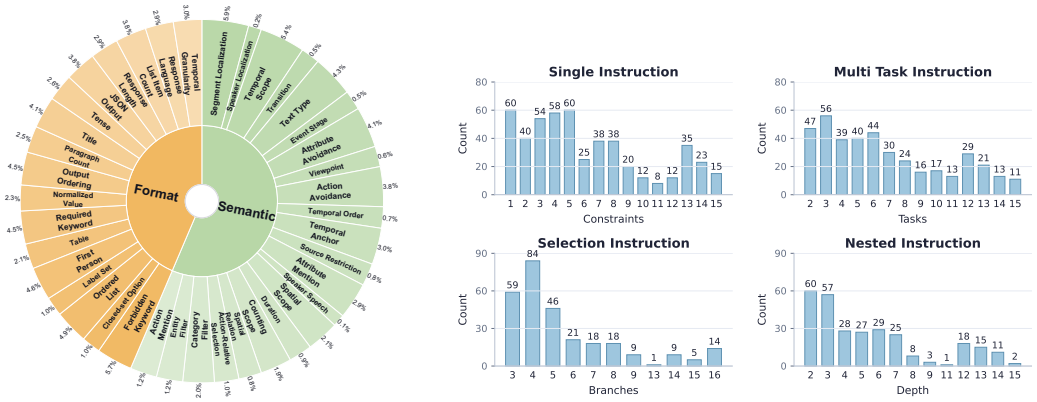
Task and Constraint Pool Construction. We analyze samples from existing video understanding benchmarks (Fu et al., 2025; Hong et al., 2025; Geng et al., 2025; Fang et al., 2024; Hu et al., 2025) and group representative tasks into six high-level categories (Perception & Recognition, Temporal Understanding, Spatial Understanding, Relationship & Interaction, Logical Reasoning, Captioning). For each category, we manually define multiple concrete task types, each with a task name, a brief description, and example questions. This design improves task diversity over directly prompting MLLMs to generate questions. The full task taxonomy is reported in Appendix B.

We construct the constraint pool by reusing and extending format constraints from prior instruction-following benchmarks (Zhou et al., 2023; Ding et al., 2025; Pyatkin et al., 2025), and designing semantic constraints that capture video-specific requirements (*e.g.*, temporal scope, relations). After multiple rounds of manual cleaning to remove ambiguous, redundant, or hard-to-verify constraints, we obtain 39 constraint types, including 22 semantic and 17 format constraints. The full constraint taxonomy is provided in Appendix C.

Information Extraction. For each video, we use Gemini-3-Pro to extract information at two granularities: a global description of the overall content and a fine-grained facts list. Each fact is defined as an atomic, timestamped visual or audio statement grounded in video content.

For longer videos, extracting the full facts list at once often produces coarse and incomplete annotations. We therefore split each long video into one-minute segments and process them sequentially. At each step, we prompt the MLLM to extract local facts and update a shared entity memory. This helps keep references to people and objects consistent across segments.

Single and Multi Instruction Generation. We sample one or more tasks according to the target instruction type and prompt the MLLM to generate instructions from the video description, facts list, and sampled constraints.



(a) Constraint distribution. (b) Instruction structure statistics.

Figure 3: Instruction and constraint statistics of Video-IFBench.

To avoid repetitive constraint patterns, we sample constraints from predefined count ranges and limit how often each constraint type can appear, which helps balance both constraint types and counts in the generated data. Examples of Single and Multi instructions are shown in Figs. 8–11 in Appendix D.

Selection and Nested Instruction Generation. Directly prompting an MLLM to generate Selection and Nested instructions often fails to produce sufficiently complex branching or nested structures. We therefore use a multi-stage pipeline that constructs the conditional structure before generating branch or leaf instructions.

As shown in the top-left of Fig. 1, we first transform facts into true and false variants to construct conditions. Then, the MLLM generates three transformations: an *edit* transformation that minimally changes the fact into a false variant (*e.g.*, "keyboard" $\rightarrow$ "screen"), a *negation* transformation that negates the original fact (*e.g.*, "pours water" $\rightarrow$ "does not pour water"), and a combined transformation that edits and negates the fact simultaneously to produce another true variant.

We then generate the instruction structure and sample one leaf as the unique true target. For each non-target leaf path, we randomly mark one intermediate node as false if the path does not already contain a false node. This ensures that each Selection or Nested instruction has exactly one valid branch or root-to-leaf path. The MLLM then writes the condition text for each intermediate node and generates a constrained instruction for each leaf node. Examples of Selection and Nested instructions are shown in Figs. 12 and 13, respectively.

Checklist Generation. During instruction generation, we record all constraints sampled for each instruction and pair each constraint with a checklist item. Given the instruction and the constraint definition, we prompt the MLLM to generate a binary checklist query, where "yes" indicates that the constraint is satisfied. For semantic constraints, we also generate a video-grounded judge context as evidence for evaluation.

Human Verification. We manually verify and clean all generated samples. Annotators check video descriptions and facts, instruction constraints, conditional branches or paths, and checklist queries and judge contexts. Problematic samples are corrected or discarded. The final distributions of instruction types, constraints, and instruction structures are shown in Fig. 2a, Fig. 3a, and Fig. 3b.

3.3 Evaluation Paradigm

We evaluate model responses with a hybrid checklist-based protocol that combines LLM-as-Judge and programmatic verification. Since a sample may contain multiple candidate instructions under conditional structures, we define the *active instruction* as the instruction that should be executed. For Single and Multi samples, the given instruction is directly active. For Selection and Nested samples, only the instruction under the correct branch or path is active.

We then evaluate the responses at two levels. Task-level coverage checks whether the response addresses all required tasks in the active instruction; for conditional instructions, this stage is especially useful because it measures whether the model follows the correct branch or path. Constraint-level adherence is evaluated using the checklist items associated with the active instruction. Each checklist item is judged either by an LLM, using the checklist query and judge context, or by a deterministic verifier for constraints that can be verified programmatically, *e.g.*, length, format, and keywords.

For each sample i, let $t_i \in \{0, 1\}$ denote task-level coverage, n_i denote the number of checklist items, and $c_{ij} \in \{0, 1\}$ denote whether the j-th item is satisfied. We report two task-gated metrics:

$$\text{TCSR} = \frac{1}{N}\sum_{i=1}^{N} t_i \cdot \frac{1}{n_i}\sum_{j=1}^{n_i} c_{ij}, \qquad \text{TISR} = \frac{1}{N}\sum_{i=1}^{N} t_i \cdot \prod_{j=1}^{n_i} c_{ij}. \tag{1}$$

TCSR measures average constraint satisfaction after task-level gating, while TISR requires the response to address the active tasks and satisfy all checklist items simultaneously.

4 Experiments

4.1 Settings

We conduct an evaluation on frontier proprietary and open-source models, including Gemini-3-Pro&Flash (Google DeepMind, 2025), GPT-5.4 (Singh et al., 2025), Doubao (ByteDance Seed Team, 2026), the Gemma 4 series (Google, 2026), the InternVL 3.5 series (Wang et al., 2025), the Qwen-Omni series (Xu et al., 2025a;b), the Qwen3-VL series (Bai et al., 2025), and the Qwen3.5 series (Qwen Team, 2026). Following prior work (Fu et al., 2026; Hong et al., 2025; Wu et al., 2024), we adopt inference settings suited best to each model for evaluation. Specifically, for Gemini and Doubao, we provide videos with audio at 1 FPS; for GPT-5.4, we sample 50 frames due to API constraints; for the Qwen series and InternVL 3.5, we sample video frames at 1 FPS. For each open-source model, we follow the recommended inference setup to ensure a representative evaluation, and run all open-source model experiments on H20 GPUs. We feed the original video with audio to omni-modal models, and provide vision-only models with video frames and timestamped subtitles transcribed by Gemini-3-Pro. In the main experiments, we use Qwen3.5-397B-A17B-Instruct (Qwen Team, 2026) as the judge model. To examine the effect of judge model choice, we also report results with a smaller Qwen3.5-35B-A3B judge in the Appendix A.1. The results show that both settings lead to consistent trends. Subsequent analyses are based on the Qwen3.5-397B-A17B-Instruct judge.

4.2 Main Results

Overall results. Table 1 reports the main results on Video-IFBench. Gemini-3-Pro achieves the strongest overall performance, with 76.5 TCSR and 54.5 TISR. Among open-source models, Qwen3.5-397B-A17B-Think performs best, reaching 69.6 TCSR and 46.1 TISR and leaving a gap of 6.9 TCSR points and 8.4 TISR points to Gemini-3-Pro, respectively. Across all models, TISR is consistently lower than TCSR, indicating that models often satisfy some constraints but fail to meet all requirements in an instruction.

The results also reveal clear differences across instruction types. Many models achieve higher scores on Multi instructions than on Single instructions, indicating that the main bottleneck is not necessarily handling multiple tasks in one request, but rather following numerous and complex constraints. The challenge becomes more pronounced for Selection and Nested instructions, where models must first select the correct branch or path based on video content before executing the corresponding instruction. Nested instructions are particularly difficult: even Gemini-3-Pro obtains only 53.7 TCSR and 46.0 TISR, while the best open-source model reaches 33.0 TCSR and 28.2 TISR. And within the same model family, larger models generally exhibit better performance, as shown in Fig. 5a.

4.3 Further Analysis and Experiments

4.3.1 Constraint-level Performance Analysis

We analyze constraint-level performance using Single and Selection instructions. Specifically, we compute the average CSR of all models for each constraint and report the 15 most challenging

Table 1: Main results on Video-IFBench. We report TCSR and TISR for each instruction type and the overall benchmark. Best results within each model group are in bold.

Model	Single		Multi		Selection		Nested		Overall	
	TCSR	TISR	TCSR	TISR	TCSR	TISR	TCSR	TISR	TCSR	TISR
Proprietary Models										
Gemini-3-Pro (Google DeepMind, 2025)	**79.6**	**52.3**	**88.6**	**58.8**	**68.5**	**59.2**	**53.7**	**46**	**76.5**	**54.5**
Gemini-3-Flash (Google DeepMind, 2025)	76.7	49.2	87.6	56.4	63.9	54.8	39.9	30.7	72.2	49.5
Doubao-Seed-2.0-Pro-260215 (ByteDance Seed Team, 2026)	76.7	44.6	87.1	51.1	46.5	38.2	17.8	14.4	65.7	40.8
GPT-5.4 (Singh et al., 2025)	72.8	34.7	82.4	43.2	21.1	16.9	12.1	7.6	57.4	30.0
Open-source Models (Instruct)										
Qwen2.5-Omni-3B (Xu et al., 2025a)	35.2	11.6	34.0	8.6	7.7	5.4	6.5	4.5	25.8	8.6
Qwen2.5-Omni-7B (Xu et al., 2025a)	46.3	16.6	48.6	14.9	15.6	11.5	8.9	5.5	36.0	13.5
Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b)	57.9	21.4	62.9	22.9	17.2	14.3	6.1	4.1	44.3	18.0
Gemma-4-E4B-it (Google, 2026)	56.8	22.1	67.5	29.9	11.9	8.2	6.0	3.8	45.1	19.5
Gemma-4-26B-A4B-it (Google, 2026)	71.6	35.7	79.2	41.1	22.7	19.4	7.1	4.5	55.6	29.7
Gemma-4-31B-it (Google, 2026)	**75.8**	**43.0**	82.1	**44.4**	**29.9**	**25.4**	13.8	**10.0**	**60.4**	**35.4**
InternVL3.5-8B-Instruct (Wang et al., 2025)	53.4	19.1	63.6	21.8	16.3	10.1	7.6	3.5	43.1	16.0
InternVL3.5-14B-Instruct (Wang et al., 2025)	56.2	19.1	66.3	22.6	18.6	12.5	6.2	3.5	45.1	16.6
InternVL3.5-30B-A3B-Instruct (Wang et al., 2025)	57.0	19.2	65.9	24.8	21.2	13.8	**16.2**	9.6	47.4	18.5
InternVL3.5-38B-Instruct (Wang et al., 2025)	61.0	24.5	71.8	27.1	18.5	12.7	14.6	8.4	49.8	20.8
InternVL3.5-241B-A28B-Instruct (Wang et al., 2025)	64.5	24.6	75.3	32.2	16.0	12.3	13.6	10.4	51.7	22.7
Qwen3-VL-4B-Instruct (Bai et al., 2025)	55.6	19.9	65.6	20.9	15.1	11.4	8.7	5.1	44.3	16.4
Qwen3-VL-8B-Instruct (Bai et al., 2025)	60.1	23.5	68.9	24.7	17.8	13.0	13.5	8.5	48.0	19.6
Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025)	56.8	19.9	70.4	25.4	17.0	12.3	13.6	9.5	47.2	18.6
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	67.5	31.9	78.0	35.1	19.5	14.5	8.6	6.0	53.1	25.8
Qwen3.5-4B-Instruct (Qwen Team, 2026)	54.6	22.0	65.5	23.3	15.2	8.8	7.9	4.0	43.9	17.3
Qwen3.5-9B-Instruct (Qwen Team, 2026)	57.8	22.6	70.0	27.3	16.2	12.4	8.1	5.0	46.6	19.6
Qwen3.5-27B-Instruct (Qwen Team, 2026)	66.7	29.1	77.8	33.8	23.0	18.0	11.7	6.9	54.0	25.2
Qwen3.5-35B-A3B-Instruct (Qwen Team, 2026)	56.7	25.7	75.2	33.4	17.5	12.4	7.5	5.0	48.0	22.6
Qwen3.5-122B-A10B-Instruct (Qwen Team, 2026)	62.7	29.1	79.8	36.8	20.3	14.5	9.5	5.5	52.5	25.3
Qwen3.5-397B-A17B-Instruct (Qwen Team, 2026)	70.1	36.5	**83.2**	40.8	24.3	18.9	12.2	7.9	57.3	30.4
Open-source Models (Thinking)										
Qwen3-Omni-30B-A3B-Think (Xu et al., 2025b)	66.6	30.0	75.1	36.8	24.8	18.5	10.1	5.5	53.1	26.2
Qwen3-VL-30B-A3B-Think (Bai et al., 2025)	60.7	23.9	68.2	31.3	22.2	17.1	8.5	5.5	47.9	22.0
Qwen3-VL-235B-A22B-Think (Bai et al., 2025)	74.7	37.7	84.4	45.0	31.4	24.9	14.1	10.3	60.9	33.5
InternVL3.5-8B-Think (Wang et al., 2025)	53.7	18.0	63.8	21.6	18.4	13.7	9.6	6.5	43.8	16.5
InternVL3.5-14B-Think (Wang et al., 2025)	55.9	18.7	67.1	24.7	19.4	14.7	12.3	7.5	46.4	18.1
InternVL3.5-30B-A3B-Think (Wang et al., 2025)	54.4	18.0	67.8	26.1	25.6	17.1	11.9	7.6	47.0	18.7
InternVL3.5-38B-Think (Wang et al., 2025)	58.8	21.4	71.2	24.8	18.3	11.6	16.4	10.5	49.1	19.1
InternVL3.5-241B-A28B-Think (Wang et al., 2025)	64.9	22.7	76.4	33.8	20.5	15.8	11.8	6.5	52.5	22.3
Qwen3.5-9B-Think (Qwen Team, 2026)	67.1	35.9	77.6	39.8	39.1	32.0	16.9	11.9	56.0	32.2
Qwen3.5-27B-Think (Qwen Team, 2026)	74.0	40.1	82.2	46.3	42.1	34.0	24.7	20.0	62.5	37.5
Qwen3.5-35B-A3B-Think (Qwen Team, 2026)	72.4	37.9	85.2	49.1	38.2	31.5	23.3	19.8	62.8	37.2
Qwen3.5-122B-A10B-Think (Qwen Team, 2026)	77.0	45.6	84.2	50.1	46.5	39.4	23.9	19.7	65.2	41.7
Qwen3.5-397B-A17B-Think (Qwen Team, 2026)	**79.4**	**48.5**	**86.0**	**52.8**	**52.1**	**44.9**	**33.0**	**28.2**	**69.6**	**46.1**

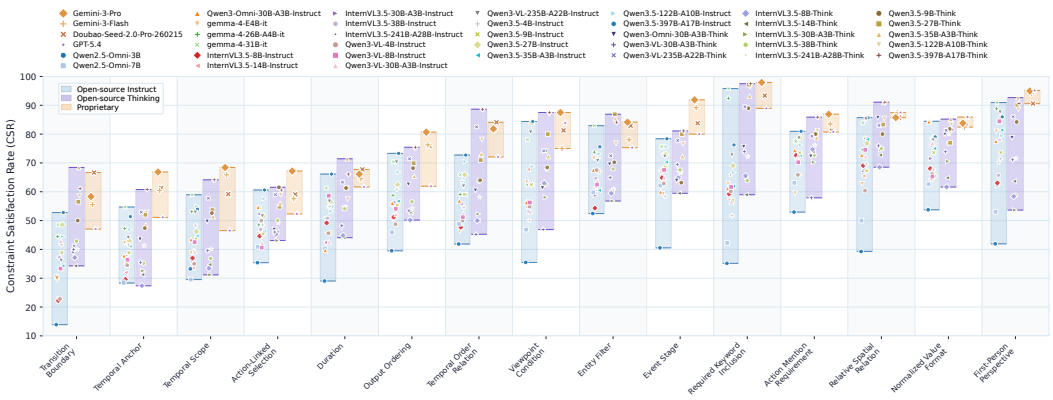


Figure 4: Performance distribution on 15 most challenging constraints. Colored bars show the performance range of different model groups, and points denote individual models.

constraints in Fig. 4. The results show that constraints related to temporal grounding and condition-dependent content selection, such as Transition Boundary, Temporal Anchor, Temporal Scope, and Action-Relative Selection, consistently achieve low scores across most models. This indicates that current models struggle to align their responses with fine-grained video evidence, especially when an instruction requires locating events or selecting content based on specific visual or audio conditions.

Fig. 4 also shows a clear separation among model groups. Proprietary models generally achieve higher scores on most constraints, while open-source models exhibit larger variation and lower

Table 2: Scaling analysis of task and constraint complexity. The left table reports task pass rate across task-count ranges, and the right table reports TISR across constraint-count ranges.

(a) Task scaling on Single and Multi instructions.

Model	1–4	5–8	9+
Gemini-3-Pro	95.6	94.9	96.3
Gemini-3-Flash	93.9	95.7	94.3
Qwen3.5-397B-A17B-Think	95.5	92.6	95.8
Gemma-4-31B-it	90.5	92.5	91.8
Qwen3-Omni-30B-A3B-Instruct	84.5	81.0	76.9
Gemma-4-E4B-it	81.0	79.1	84.1

(b) Constraint scaling on Single instructions.

Model	1–4	5–8	9+
Gemini-3-Pro	62.7	51.8	46.9
Gemini-3-Flash	59.3	47.5	46.9
Qwen3.5-397B-A17B-Think	58.2	43.9	42.5
Gemma-4-31B-it	53.7	39.6	29.2
Qwen3-Omni-30B-A3B-Instruct	40.1	11.5	3.5
Gemma-4-E4B-it	38.4	12.9	7.1

performance. Among open-source models, thinking variants improve the upper bounds on most constraints. However, this improvement is not consistent across model families, a further analysis is provided in Appendix A.2.

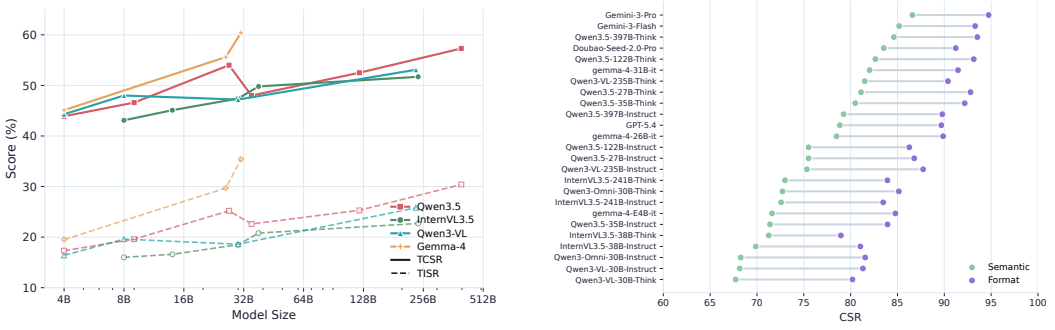
We further compare performance on semantic and format constraints in Fig. 5b. Existing models perform substantially worse on semantic constraints than on format constraints, indicating that semantic constraints are more challenging because they require precise video understanding rather than merely formatting responses.

4.3.2 Task and Constraint Scaling Analysis

We analyze how performance changes with task and constraint complexity under controlled settings. For task scaling, we use Single and Multi instructions to avoid the confounding effect of conditional branching. For constraint scaling, we use Single instructions to minimize the influence of other factors. We report representative models covering proprietary models, open-source thinking and instruct models, omni-modal models, and different model scales.

Task scaling. Table 2a reports task pass rates as the number of tasks increases. Capable models remain relatively stable across task counts, suggesting that task count itself is not the main bottleneck. This observation is consistent with the findings in Section 4.2, which suggest that the primary challenge lies in adhering to constraints rather than handling multiple tasks.

Constraint scaling. Table 2b shows that TISR drops substantially as the number of constraints increases. Gemini-3-Pro decreases from 62.7% with 1–4 constraints to 46.9% with 9+ constraints, while Qwen3.5-397B-A17B-Think drops from 58.2% to 42.5%. The decline is much sharper for weaker open-source models, with Qwen3-Omni-30B-A3B-Instruct falling from 40.1% to 3.5% and Gemma-4-E4B-it from 38.4% to 7.1%. These results show that current models struggle when many constraints are imposed within a Single instruction.



(a) Scaling across model families.

(b) Semantic vs. format constraint performance.

Figure 5: Model scaling and semantic-format constraint performance.

Table 3: Task pass rate on Selection instructions using the 3-condition-plus-else subset.

(a) Branch count.

Model	2	3	4
Gemini-3-Flash	66.7	58.0	56.5
Qwen3.5-397B-A17B-Think	68.1	59.4	59.4
Qwen3.5-397B-A17B-Instruct	58.0	50.7	39.1
Qwen3.5-9B-Instruct	37.7	24.6	27.5
Gemma-4-31B-it	50.7	46.4	36.2
Gemma-4-E4B-it	43.5	27.5	19.1
Qwen3-Omni-30B-A3B-Instruct	43.9	30.4	17.4

(b) Correct-branch position.

Model	1	2	3
Gemini-3-Flash	66.7	66.7	54.9
Qwen3.5-397B-A17B-Think	76.5	58.8	52.9
Qwen3.5-397B-A17B-Instruct	70.6	35.3	29.4
Qwen3.5-9B-Instruct	45.1	21.6	15.7
Gemma-4-31B-it	33.3	43.1	39.2
Gemma-4-E4B-it	33.3	8.3	5.9
Qwen3-Omni-30B-A3B-Instruct	23.5	12.0	18.0

4.3.3 Depth Sensitivity in Nested Instructions

We analyze depth sensitivity on Nested instructions using the same representative models. As shown in Fig. 6, task pass rate generally decreases as the maximum tree depth increases. The drop is clear for strong models: Gemini-3-Pro decreases from 64.6% at depth 2–4 to 55.6% at depth 8–14, while Qwen3.5-397B-A17B-Think drops from 40.5% to 22.2%. Weaker open-source models remain low across all depth ranges, suggesting that they already struggle with shallow nested structures. These results show that deeper instruction trees make it harder for models to recognize the correct conditional path.

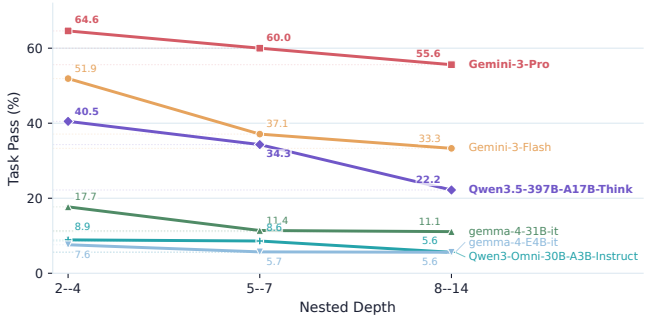


Figure 6: Depth sensitivity on Nested instructions.

4.3.4 Branch Number and Position in Selection Instructions

We conduct controlled ablation studies on the 3-condition-plus-else subset to study how branch count and correct-branch position affect performance on Selection instructions. For branch count analysis, we keep the correct branch unchanged and progressively add one or two false branches together with the else branch, resulting in 2, 3, and 4-branch variants. For position analysis, we keep the same set of condition branches but move the correct branch to different positions.

As shown in Table 3, increasing the number of candidate branches generally lowers task pass rate. For example, Gemini-3-Flash drops from 66.7% with two branches to 56.5% with four branches, and Qwen3.5-397B-A17B-Instruct drops from 58.0% to 39.1%. This suggests that models become less reliable as they must compare more video-grounded conditions.

We find that the correct-branch position has a clear impact. Most models show a preference for the first branch, with performance decreasing when the correct branch appears later. For example, Qwen3.5-397B-A17B-Think drops from 76.5% at position 1 to 52.9% at position 3, while its Instruct counterpart drops from 70.6% to 29.4%. These results indicate that current models do not evaluate conditional branches equally and may rely on positional biases.

5 Conclusion

We introduce Video-IFBench, a benchmark for evaluating instruction following in video understanding, with a hybrid protocol combining LLM-as-Judge and programmatic verification. Experiments show that current MLLMs struggle with constraint-rich and conditional video instructions. We hope Video-IFBench will facilitate future research on video MLLMs that can faithfully follow complex, constraint-rich instructions in practical applications.

AI USE STATEMENT

In this work, generative AI models were used for benchmark data construction and automated evaluation, as described in the paper. All generated benchmark samples were manually verified and cleaned, and the automatic evaluator was assessed through a human agreement study (Appendix E). We also used generative AI tools to polish the manuscript and obtain layout suggestions for figures. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We commit to fully open-sourcing the dataset, all evaluation code, and all configurations used in this work to facilitate reproducibility.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

ByteDance Seed Team. Seed2.0 model card: Towards intelligence frontier for real-world complexity. `https://seed.bytedance.com/seed2`, February 2026. Model Card.

Kaiyuan Chen, Qimin Wu, Taiyu Hou, Tianhao Tang, Xueyu Hu, Yuchen Hou, Bikun Li, Chengming Qian, Guoyin Wang, Haolin Chen, et al. Agentif-oneday: A task-level instruction-following benchmark for general ai agents in daily scenarios. *arXiv preprint arXiv:2601.20613*, 2026.

Xinyi Chen, Baohao Liao, Jirui Qi, Panagiotis Eustratiadis, Christof Monz, Arianna Bisazza, and Maarten de Rijke. The sifo benchmark: Investigating the sequential instruction following ability of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1691–1706, 2024.

Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1099–1109, 2025.

Elliot L Epstein, Kaisheng Yao, Jing Li, Xinyi Bai, and Hamid Palangi. Mmmt-if: A challenging multimodal multi-turn instruction following benchmark. *arXiv preprint arXiv:2409.18216*, 2024.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. `https://huggingface.co/datasets/HuggingFaceFV/finevideo`, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18959–18969, 2025.

Google. Gemma 4 model card, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4.

Google DeepMind. Introducing gemini 3: Our most intelligent model that helps you bring any idea to life, 2025. URL https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Google Blog.

Weilei He, Feng Ju, Zhiyuan Fan, Rui Min, Minhao Cheng, and Yi R Fung. Empowering reliable visual-centric instruction following in mllms. *arXiv preprint arXiv:2601.03198*, 2026.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. Followbench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4667–4688, 2024.

Yimin Jing, Renren Jin, Jiahao Hu, Huishi Qiu, Xiaohua Wang, Peng Wang, and Deyi Xiong. Followeval: A multi-dimensional benchmark for assessing the instruction-following capability of large language models. *arXiv preprint arXiv:2311.09829*, 2023.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Jinnan Li, Jinzhe Li, Yue Wang, Yi Chang, and Yuan Wu. Structflowbench: A structured flow benchmark for multi-turn instruction following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9322–9341, 2025a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Shaoxuan Li, Zhixuan Zhao, Hanze Deng, Zirun Ma, Shulin Tian, Zuyan Liu, Yushi Hu, Haoning Wu, Yuhao Dong, Benlin Liu, et al. Perceptioncomp: A video benchmark for complex perception-centric reasoning. *arXiv preprint arXiv:2603.26653*, 2026.

Shihao Li, Yuanxing Zhang, Jiangtao Wu, Zhide Lei, Yiwen He, Runzhe Wen, Chenxi Liao, Chengkang Jiang, An Ping, Shuo Gao, et al. If-vidcap: Can video caption models follow instructions? *arXiv preprint arXiv:2510.18726*, 2025b.

Gili Lior, Asaf Yehudai, Ariel Gera, and Liat Ein-Dor. Wildifeval: Instruction following in the wild. *arXiv preprint arXiv:2503.06573*, 2025.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv preprint arXiv:2502.04328*, 2025.

Xudong Lu, Huankang Guan, Yang Bo, Jinpeng Chen, Xintong Guo, Shuhan Li, Fang Liu, Peiwen Sun, Xueying Li, Wei Zhang, et al. Phostream: Benchmarking real-world streaming for omnimodal assistants in mobile scenarios. *arXiv preprint arXiv:2601.22575*, 2026.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18902–18913, 2025.

Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. Mcif: Multimodal crosslingual instruction-following benchmark from scientific talks. *arXiv preprint arXiv:2507.19634*, 2025.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. *arXiv preprint arXiv:2507.02833*, 2025.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Amy Xin, Youfeng Liu, Bin Xu, Lei Hou, and Juanzi Li. Agentif: Benchmarking instruction following of large language models in agentic scenarios. *arXiv preprint arXiv:2505.16944*, 2025.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*, 2024.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. Infobench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13025–13048, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`. Alibaba Cloud, Technical Report.

Haozhan Shen, Shilin Yan, Hongwei Xue, Shuaiqi Lu, Xiaojun Tang, Guannan Zhang, Tiancheng Zhao, and Jianwei Yin. Mm-condchain: A programmatically verified benchmark for visually grounded deep compositional reasoning. *arXiv preprint arXiv:2603.12266*, 2026.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Shenzhi Wang, Shixuan Liu, Jing Zhou, Chang Gao, Xiong-Hui Chen, Binghai Wang, An Yang, Shiji Song, Bowen Yu, Gao Huang, et al. Hopchain: Multi-hop data synthesis for generalizable vision-language reasoning. *arXiv preprint arXiv:2603.17024*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxin Xu, et al. Benchmarking complex instruction-following with multiple constraints composition. *Advances in Neural Information Processing Systems*, 37:137610–137645, 2024.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a. doi: 10.48550/arXiv.2503.20215. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8475–8489, 2025.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Additional Experimental Results

A.1 Main results using Qwen3.5-35B-A3B as Judge Model.

To examine the effect of judge model choice, we re-evaluate all model responses using Qwen3.5-35B-A3B as the judge. Table 4 reports the resulting scores. The values in parentheses in the Overall columns denote the difference from the main results in Table 1. Using a smaller judge generally increases the absolute scores, but the relative trends remain largely consistent with the main results, supporting the robustness of our evaluation protocol.

Table 4: Main results evaluated with Qwen3.5-35B-A3B as the judge. Best results within each model group are in bold. Parenthesized values in the Overall columns denote changes from the main results.

Model	Single		Multi		Selection		Nested		Overall	
	TCSR	TISR	TCSR	TISR	TCSR	TISR	TCSR	TISR	TCSR	TISR
Proprietary Models										
Gemini-3-Pro (Google DeepMind, 2025)	**86.6**	**64.1**	87.7	53.3	**73.1**	**63.0**	**57.8**	**51.5**	**80.1 (+3.6)**	**58.7 (+4.2)**
Gemini-3-Flash (Google DeepMind, 2025)	83.0	55.8	**87.8**	**53.4**	67.1	57.3	43.8	35.3	75.7 (+3.5)	52.2 (+2.7)
Doubao-Seed-2.0-Pro-260215 (ByteDance Seed Team, 2026)	84.1	54.3	85.9	48.7	47.7	41.2	20.8	17.8	68.6 (+2.9)	44.7 (+3.9)
GPT-5.4 (Singh et al., 2025)	78.0	42.5	80.1	40.3	24.7	20.4	13.5	9.9	59.6 (+2.2)	33.0 (+3.0)
Open-source Models (Instruct)										
Qwen2.5-Omni-3B (Xu et al., 2025a)	40.9	13.6	32.0	8.3	10.0	6.7	7.8	4.8	28.0 (+2.2)	9.5 (+0.9)
Qwen2.5-Omni-7B (Xu et al., 2025a)	53.1	19.8	48.5	13.9	16.7	10.6	12.0	7.7	39.1 (+3.1)	14.6 (+1.1)
Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b)	67.6	27.8	64.8	25.9	20.7	15.9	7.9	5.5	49.0 (+4.7)	21.6 (+3.6)
Gemma-4-E4B-it (Google, 2026)	65.0	28.8	65.8	27.7	14.2	9.6	8.0	5.5	48.2 (+3.1)	21.8 (+2.3)
Gemma-4-26B-A4B-it (Google, 2026)	77.6	44.7	78.0	39.4	25.6	21.1	10.6	7.4	58.4 (+2.8)	33.3 (+3.6)
Gemma-4-31B-it (Google, 2026)	**81.7**	**50.1**	80.9	**42.1**	**34.8**	**31.1**	17.3	**13.4**	**63.5 (+3.1)**	**38.8 (+3.4)**
InternVL3.5-8B-Instruct (Wang et al., 2025)	63.9	24.1	63.1	23.0	19.8	13.6	9.9	5.4	47.7 (+4.6)	19.1 (+3.1)
InternVL3.5-14B-Instruct (Wang et al., 2025)	65.1	24.6	68.2	25.5	19.7	12.5	10.3	7.5	49.8 (+4.7)	20.2 (+3.6)
InternVL3.5-30B-A3B-Instruct (Wang et al., 2025)	66.2	24.7	65.7	23.1	27.6	20.1	18.1	11.7	52.0 (+4.6)	21.5 (+3.0)
InternVL3.5-38B-Instruct (Wang et al., 2025)	68.9	28.2	71.9	24.9	22.3	15.9	**18.6**	12.4	54.0 (+4.2)	22.6 (+1.8)
InternVL3.5-241B-A28B-Instruct (Wang et al., 2025)	71.7	30.6	74.3	30.2	19.6	15.0	15.6	10.9	54.8 (+3.1)	24.8 (+2.1)
Qwen3-VL-4B-Instruct (Bai et al., 2025)	64.2	23.1	62.5	21.9	18.2	13.9	11.1	5.9	47.4 (+3.1)	18.5 (+2.1)
Qwen3-VL-8B-Instruct (Bai et al., 2025)	67.7	27.4	69.1	26.5	22.1	16.8	15.5	10.4	51.8 (+3.8)	22.6 (+3.0)
Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025)	67.6	25.1	69.5	25.9	18.0	11.9	14.3	10.0	51.1 (+3.9)	20.6 (+2.0)
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	75.0	36.3	79.8	37.0	20.6	16.7	9.6	6.4	57.0 (+3.9)	28.5 (+2.7)
Qwen3.5-4B-Instruct (Qwen Team, 2026)	63.8	25.4	64.4	23.6	18.9	13.3	10.0	5.5	47.9 (+4.0)	19.7 (+2.4)
Qwen3.5-9B-Instruct (Qwen Team, 2026)	65.5	26.5	67.4	28.3	18.4	15.0	10.1	5.5	49.2 (+2.6)	21.8 (+2.2)
Qwen3.5-27B-Instruct (Qwen Team, 2026)	75.7	35.9	76.8	35.8	26.2	20.2	13.9	8.5	57.9 (+3.9)	28.9 (+3.7)
Qwen3.5-35B-A3B-Instruct (Qwen Team, 2026)	67.9	29.9	72.7	31.1	20.1	14.5	10.9	7.0	52.1 (+4.1)	24.0 (+1.4)
Qwen3.5-122B-A10B-Instruct (Qwen Team, 2026)	70.9	33.1	76.9	35.4	23.3	16.2	13.6	8.5	55.7 (+3.2)	27.1 (+1.8)
Qwen3.5-397B-A17B-Instruct (Qwen Team, 2026)	79.9	43.9	**80.9**	38.3	28.3	21.6	14.3	9.0	61.3 (+4.0)	33.1 (+2.7)
Open-source Models (Thinking)										
Qwen3-Omni-30B-A3B-Think (Xu et al., 2025b)	75.0	36.0	76.2	35.4	30.3	23.0	11.7	7.0	57.8 (+4.7)	29.1 (+2.9)
Qwen3-VL-30B-A3B-Think (Bai et al., 2025)	65.4	29.3	69.2	30.2	26.0	21.2	10.8	6.7	51.1 (+3.2)	24.6 (+2.6)
Qwen3-VL-235B-A22B-Think (Bai et al., 2025)	81.4	46.6	81.7	43.2	33.9	27.6	16.8	13.3	63.3 (+2.4)	37.2 (+3.7)
InternVL3.5-8B-Think (Wang et al., 2025)	62.1	21.1	61.4	24.7	21.6	14.5	12.8	9.0	47.1 (+3.3)	19.1 (+2.6)
InternVL3.5-14B-Think (Wang et al., 2025)	64.9	24.6	66.4	22.3	24.5	16.0	17.5	10.1	51.0 (+4.6)	20.2 (+2.1)
InternVL3.5-30B-A3B-Think (Wang et al., 2025)	62.4	21.0	64.4	24.3	27.1	19.3	15.7	10.6	49.6 (+2.6)	20.1 (+1.4)
InternVL3.5-38B-Think (Wang et al., 2025)	67.1	24.1	68.4	23.7	22.6	15.5	19.4	11.6	52.5 (+3.4)	20.6 (+1.5)
InternVL3.5-241B-A28B-Think (Wang et al., 2025)	72.9	30.9	75.2	33.0	23.0	17.5	13.8	8.5	55.9 (+3.4)	25.8 (+3.5)
Qwen3.5-9B-Think (Qwen Team, 2026)	77.8	40.6	78.1	37.6	43.1	37.8	17.8	13.5	61.3 (+5.3)	34.9 (+2.7)
Qwen3.5-27B-Think (Qwen Team, 2026)	83.3	47.8	84.4	46.2	44.6	37.2	30.3	26.5	67.9 (+5.4)	42.0 (+4.5)
Qwen3.5-35B-A3B-Think (Qwen Team, 2026)	81.1	45.8	82.9	48.1	39.5	30.9	26.4	24.3	66.4 (+3.6)	40.7 (+3.5)
Qwen3.5-122B-A10B-Think (Qwen Team, 2026)	84.1	55.9	85.0	**50.5**	49.3	43.8	26.6	23.6	68.9 (+3.7)	47.0 (+5.3)
Qwen3.5-397B-A17B-Think (Qwen Team, 2026)	**86.2**	**59.6**	**85.8**	48.3	**55.7**	**47.8**	**35.5**	**31.2**	**73.0 (+3.4)**	**49.8 (+3.7)**

A.2 Effect of Thinking Mode on Video-IFBench.

As shown in Fig. 7, thinking mode brings clear gains for Qwen3 and Qwen3.5 models, but yields little improvement or even performance drops for InternVL3.5. We hypothesize that the model-dependent effect may be related to how instruction-following rewards are designed during reinforcement learning. Moreover, the gains are generally larger on semantic constraints than on format constraints, suggesting that thinking-style inference is more helpful when constraint following requires visual or audio semantic understanding.

B Task Taxonomy

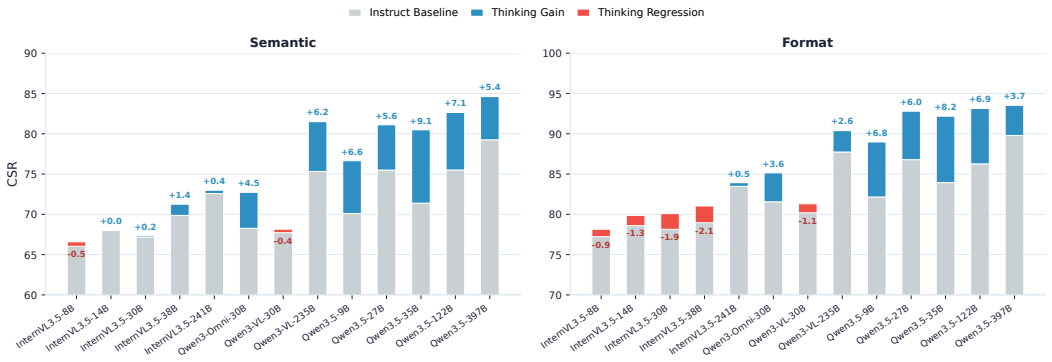


Figure 7: Effect of thinking mode on semantic and format constraints.

Table 5: Video understanding task taxonomy in Video-IFBench.

Category	Task Type	Description	Example Question
Perception & Recognition	Entity Recognition	To describe specific objects, individuals, animals, or elements present within the video frames.	What type of vehicle is parked in front of the house at 0:15?
Perception & Recognition	Action Recognition	To classify the specific movements, gestures, or activities being performed by entities in the video.	What is the man in the blue shirt doing when the black car stops?
Perception & Recognition	Attribute Recognition	To describe specific properties or characteristics of an entity, such as its color, size, shape, or material.	What color is the backpack the child is carrying?
Perception & Recognition	Scene Description	To recognize and describe the global environment, setting, or background where the video takes place.	Count the number of scenes appeared in the video and describe each of them in detail.
Perception & Recognition	Text Recognition (OCR)	To detect, extract, and read visible text within the video, such as street signs, subtitles, documents, or labels.	What does the neon sign above the shop door say?
Perception & Recognition	Dialogue Recognition	To recognize and describe spoken content, speaker turns, or dialogue exchanged in the video.	What are the exact words the news anchor says right after the loud crash is heard?
Perception & Recognition	Counting	To determine how many times a target speech pattern occurs or how many target objects appear throughout the video.	How many times does the host say 'thank you' in the video?
Perception & Recognition	Emotion Recognition	To infer the emotional state, mood, or facial expressions of the people or characters in the video.	Does the woman look happy or frustrated when she opens the cardboard box?
Perception & Recognition	Event Recognition	To categorize the overarching macroscopic event or social activity happening in the video clip.	Describe, in chronological order, the events experienced by the man wearing red in the video.
Temporal Understanding	Event Grounding	To pinpoint the exact start and end timestamps of a specific event or action, including all instances of its occurrence.	At what exact timestamps does the dog bark during the video?
Temporal Understanding	Occurrence Localization	To locate the precise time interval when an entity, action, or state appears for the first, last, or N-th time.	When does the blue car appear for the third time in the footage?
Temporal Understanding	Temporal Ordering	To determine the chronological sequence of two or more distinct actions or events.	Did the character put on his hat before or after he picked up the umbrella?
Temporal Understanding	Temporal Adjacent Retrieval	To retrieve the event or action that occurs immediately preceding or following a specified anchor event.	What happens immediately after the glass shatters on the floor?
Temporal Understanding	Local Temporal Understanding	To sequentially narrate the list of actions or events that take place within a strictly defined time window.	List all the actions performed by the chef between 1:00 and 1:30 in chronological order.
Temporal Understanding	Temporal Sorting	To logically reorder a randomized or shuffled list of video events into their correct chronological sequence.	Given the events 'eating dinner', 'cooking food', and 'washing dishes', order them as they actually occurred in the video.
Temporal Understanding	Duration Estimation	To calculate or estimate the length of time a specific action, event, or state persists.	How many seconds does the traffic light remain red?
Temporal Understanding	Audio Change	To identify and describe how spoken audio evolves over time, such as changes in active speaker, speaking state, turn-taking, or stretches of silence.	Describe how the active speaker changes from the beginning of the video to the end.

Category	Task Type	Description	Example Question
Temporal Understanding	Audio-Visual Temporal Alignment	To determine and describe the temporal correspondence between spoken audio cues and visual events, identifying which utterances or speaker turns align with which visible actions, moments, or scene transitions.	Describe which visible actions coincide with the moments when the host starts speaking.
Spatial Understanding	Relative Direction Judgment	To determine the directional position of an entity relative to the camera's viewpoint or another reference entity.	Describe the clothing of all people to the left of the entrance.
Spatial Understanding	Relative Distance Comparison	To compare and determine which entity is closer to or further from a given reference point.	Order the oak tree, park bench, motorcycle, and statue from farthest to nearest relative to the black trash can.
Spatial Understanding	Spatial Relationship Judgment	To classify the spatial relations between entities, such as inside, on, under, beside, or attached to.	Describe the cat's spatial relationship to the table, chair, rug, and fireplace while it is sleeping in the video.
Spatial Understanding	Perspective-based Spatial Judgment	To reinterpret spatial relationships from a hypothetical or shifted orientation (e.g., 'standing at X facing Y').	If you are standing at the doorway facing the window, describe all the objects on your left-hand side and their relative positions.
Relationship & Interaction	Entity Interaction Recognition	To describe dynamic interactions between multiple entities, such as holding, passing, attacking, talking, or hugging.	How do the two players interact immediately after the goal is scored?
Relationship & Interaction	Action-Entity Binding	To accurately associate actions with their specific agents and targets to determine who did what to whom.	Describe the attributes of the player who passes the basketball to the player wearing a yellow jersey.
Relationship & Interaction	Speech Source Recognition	To visually localize the source of spoken audio, such as identifying which person in a crowd is speaking.	Which person on screen is speaking when the announcement is heard?
Logical Reasoning	Causal Reasoning	To deduce the cause-and-effect relationship between different events, actions, or states in the video.	Why did the fire alarm go off in the hallway?
Logical Reasoning	Counterfactual Reasoning	To imagine and reason about alternative outcomes if a specific past event had not occurred or happened differently.	If the man hadn't dropped his keys on the sidewalk, would he have missed the bus?
Logical Reasoning	Multi-hop Reasoning	To answer complex queries by synthesizing evidence across multiple timestamps, events, or logical steps.	Based on the ingredients the chef gathered at the start and the cooking method used in the middle, what specific dish will be plated at the end?
Logical Reasoning	Knowledge Transfer	To extract a rule, concept, or physical principle shown in the video and apply it to a new, unseen hypothetical scenario.	Based on the physics principle demonstrated with the pendulum in the video, what would happen if the string were twice as long?
Captioning	Segmented Captioning	To generate individual, coherent textual descriptions for pre-defined or uniformly segmented time intervals of the video.	Provide a descriptive caption for the events occurring between 0:00-0:30, 0:30-1:00, and 1:00-1:30.
Captioning	Dense Captioning	To generate highly detailed, fine-grained descriptions covering all visible actions, entities, and micro-events throughout the video.	Provide a dense, second-by-second narrative of every action happening in this street view camera footage.
Captioning	Video Summarization	To generate a concise, high-level overview that captures the core theme, main plot, or key highlights of the entire video.	Write a two-sentence summary of the main story and outcome shown in this short film.

C CONSTRAINT TAXONOMY

Table 6: Constraint taxonomy in Video-IFBench.

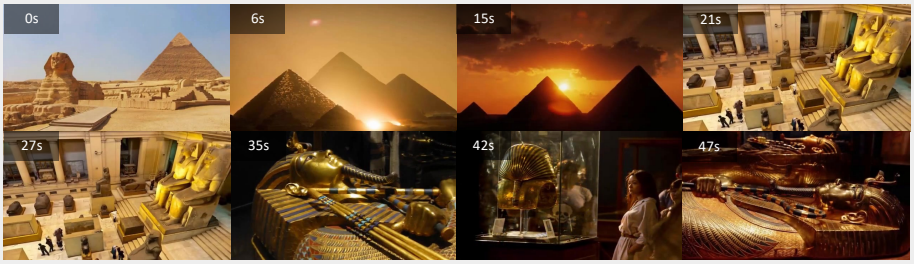
Group	Constraint Type	Description	Example
Semantic	Temporal Scope	Restrict the task to evidence, entities, events, or judgments within a specified absolute or segment-defined time window of the video, such as a timestamp range, beginning/middle/end portion, or last N seconds.	Use only the visual events that occur in the last 10 seconds of the video in your answer.
Semantic	Temporal Anchor	Constrain a target relative to one explicit reference event, action, state, or timestamp, such as before, after, during, or at the moment of that anchor.	Mention only the video content before the door opens.
Semantic	Temporal Order Relation	Require the response to determine or preserve a temporal ordering relation among two or more already-specified events or targets, such as whether one happens before, after, between, or immediately adjacent to another.	Indicate the temporal order of all mentioned visual events in your answer.

Group	Constraint Type	Description	Example
Semantic	Event Stage	Restrict the target to an intrinsic phase of a single continuous event or process, such as onset, peak, hold, completion, or recovery, rather than an arbitrary external time window.	Do not mention any video content after the man leaves the room in your answer.
Semantic	Duration Constraint	Require the target event, state, relation, or occurrence to satisfy a duration condition such as minimum length, maximum length, longest, shortest, brief, or sustained.	Mention only actions lasting more than 2 seconds.
Semantic	Temporal Granularity	Constrain the temporal unit used in the answer to a specified granularity, such as seconds only, minutes only, or shot indices only.	Use shot numbers for every mentioned visual segment in your answer; do not use other timestamps.
Semantic	Spatial Region Restriction	Restrict valid candidates to a designated region.	Do not mention any outdoor content in your answer.
Semantic	Relative Spatial Relation	Select or interpret the target based on its spatial relation to another object, person, landmark, or scene element.	For all stationary objects mentioned in your answer, describe their spatial relationships relative to the leftmost streetlight at the 3-second mark of the video.
Semantic	Viewpoint Condition	Require the judgment to be made only under a specified camera view, shot type, or visual perspective when interpretation depends on viewpoint.	From a bird's-eye view, describe the positions of all objects mentioned in the answer.
Semantic	Category Filter	Restrict valid targets to a specified semantic category or predefined set of categories.	In the answer, only food-related entities may be mentioned; all other entities must be replaced with uppercase English letters starting from C.
Semantic	Entity Filter	Require the valid entity or speaker targets mentioned in the answer to satisfy a specified attribute, appearance property, acoustic trait, or state qualifier.	Mention only people wearing red jackets in the answer.
Semantic	Attribute Avoidance	Restrict the output from mentioning specific attributes, physical traits, acoustic traits, or sensitive characteristics of the entities.	Do not mention any person's gender or age in the answer.
Semantic	Attribute Mention Requirement	Require the answer to explicitly mention one or more specified attributes, appearance properties, acoustic traits, or state descriptors for each relevant entity or speaker mentioned in the response, rather than only selecting targets by those attributes.	For every person mentioned in the answer, include the color of their pants.
Semantic	Action-Relative Selection	Restrict valid targets to those participating in a specified action, interaction, role, or event involvement condition.	Only mention players who were involved in passing in the answer.
Semantic	Action Mention Requirement	Require the answer to explicitly mention one or more specified actions, interactions, or motion types performed by a designated entity, entity group, body part, or target set, optionally covering all valid instances rather than merely identifying participants.	The answer must include all actions performed by the woman in the video.
Semantic	Action Avoidance	Restrict the answer from mentioning specified actions, interactions, or motion types associated with a designated entity, entity group, body part, instrument, or target set, even if those actions are visible in the video.	Do not mention any actions performed by the woman's hands in the answer.
Semantic	Counting Scope	Constrain how many valid items, entities, events, or occurrences should be returned or counted, such as exactly k, at most k, all, or one.	Only two people may be explicitly mentioned in the answer; all others must be referred to using uppercase English letters.
Semantic	Transition Boundary	Focus on the boundary where presence, state, relation, or scene status changes, such as appears, disappears, turns on, becomes inside, or shifts environment.	The answer must include every moment when the man in blue enters or exits the frame.
Format	Output Ordering	Require outputs to follow a specified ordering rule such as chronological order, reading order, numeric order, order of appearance, or speaker-turn order.	All objects mentioned in the answer must be ordered from farthest to nearest relative to the man.
Format	Normalized Value Format	Constrain scalar, label, date, string, or top-k outputs to a specified normalized representation without imposing a full table, list, or JSON structure.	Return dates in [YYYY-MM-DD] format.
Format	Structured JSON Output	Require the answer to be a valid JSON object or JSON array, optionally with required top-level keys or a fixed schema skeleton.	Return the answer as a JSON object with keys reason and evidence_segments.
Format	Table Output	Require the answer to be presented as a table with an explicit row-column structure, such as Markdown table or CSV-style table.	Summarize each stage in a two-column table.
Format	Ordered List Output	Require the answer to be organized as an ordered list or bullet list rather than a free-form paragraph.	List the key events as an ordered list.
Format	Closed-Set Option Output	Require the answer to be exactly one option from a provided closed set, optionally with no extra explanation.	Answer with exactly one of: yes, no, unknown.
Format	Response Language	Require the entire answer to be produced in a specified language or bilingual pair.	answer entirely in Spanish

Group	Constraint Type	Description	Example
Format	Response Length Constraint	Constrain the overall answer length by an exact value, a lower bound, an upper bound, or a closed range, typically measured in words, or in project-defined Chinese length units for Chinese responses.	Answer in 80 to 120 words.
Format	Paragraph Count Constraint	Require the answer to contain a specified number of paragraphs or paragraph-like blocks.	Write the explanation in exactly three paragraphs.
Format	List Item Count Constraint	Require a list-style answer to contain exactly or at most a specified number of bullet items or numbered items.	Give exactly five bullet points.
Format	Title Requirement	Require the answer to include a title with a specified style, scope, or placement.	add a short title before the summary
Format	Required Keyword Inclusion	Require the answer to include one or more specified words or phrases.	The summary must include the word touchdown exactly three times.
Format	Forbidden Keyword Exclusion	Require the answer to avoid a specified banned word set, phrase set, or lexical family.	Do not mention she, he, it, they, or we.
Format	First-Person Perspective	Require the answer to be phrased from a first-person point of view rather than an external narrator perspective.	Answer from the first-person perspective of the woman wearing red in the scene.
Format	Text Type Restriction	Restrict text, quoted speech, or symbol targets by semantic type, lexical pattern, script, style, or character class.	Any on-screen text quoted in the answer must be written entirely in uppercase.
Format	Entity Label Set Restriction	Restrict the entity, speaker, or category labels used within the answer to an explicitly given closed-set label inventory or a specified branch of a taxonomy, without requiring the entire answer to be a single closed-set option.	All entities mentioned in the answer may only be referred to using the following names: [Anna, John].
Semantic	Source Restriction	Restrict the sources cited in the answer, such as particular speakers, visible text, on-screen graphics, documents, or other designated information sources.	The answer must cite only text shown on the scoreboard and must not include any spoken dialogue.
Semantic	Speaker-Speech Binding	Require every referenced speech item in the answer to be explicitly bound to a specific speaker.	Every speech item must be bound to a specific speaker.
Semantic	Speaker Localization Constraint	Require each referenced speaker in the answer to be localized using a coarse on-screen position or relative visual location.	For each speech item, indicate whether the speaker is on the left, center, or right side of the frame.
Semantic	Segment Localization	Require each referenced speech item or visual segment in the answer to be localized to a specific start timestamp and end timestamp in the video.	Every speech item must include its corresponding start timestamp and end timestamp.
Format	Temporal/Tense Constraint	Restrict the output to be generated using specific grammatical tenses or within a specific timeline context.	Use the past tense in the first sentence of the answer, and the past perfect tense in the rest.

D EXAMPLE VISUALIZATION

Example 1



Instruction

Focusing only on the first 45 seconds of the video, describe the different scenes and settings shown. You must completely exclude any outdoor content and avoid using the words 'sun', 'pyramid', and 'sphinx'. Do not mention any colors in your descriptions. Start your response with the title 'Indoor Scenes'. Present your answer as an ordered list with exactly three numbered items. For each scene you describe, include its corresponding start timestamp and end timestamp, formatted strictly in seconds only (for example, 21s - 29s).

Gemma-4-31B-it ✗

Indoor Scenes
1. 22s - 28s: A museum gallery featuring large statues and various artifacts on display.
2. 30s - 36s: A close-up view of a golden burial mask.
3. 38s - 44s: A person observing a gold object inside a glass display case.

Gemini-3-Pro ✓

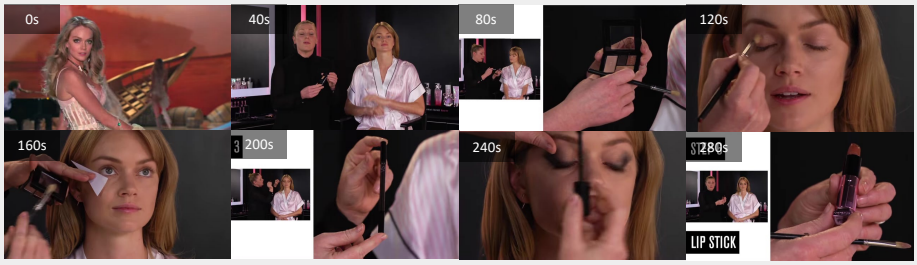
Indoor Scenes
1. 21s - 29s: A high-angle view of a spacious museum hall filled with various ancient artifacts. Visitors stroll across the floor, observing multiple historical items, including two massive, seated statues positioned on a raised platform.
2. 29s - 37s: A close-up shot focusing on the upper portion of a highly ornate sarcophagus. The intricately detailed figure is depicted holding ceremonial objects crossed over its chest, with distinct patterns visible on its headdress and braided beard.
3. 37s - 45s: A museum exhibit showcasing an ancient pharaonic burial mask protected within a glass display case. Several visitors are gathered around the exhibit, including a woman in the foreground who is carefully observing the artifact.

Qwen3-Omni-30B-A3B-Instruct ✗

Indoor Scenes
1. 00s - 04s: The scene opens with a wide shot of a large, spacious museum hall. The architecture features tall, classical columns and high ceilings. The floor is made of light-colored stone tiles. Numerous large, ancient stone statues and sarcophagi are displayed on raised platforms and in glass cases. Several visitors are seen walking around, observing the exhibits.
2. 05s - 10s: The scene transitions to a close-up of a magnificent golden sarcophagus. The sarcophagus is intricately decorated with detailed patterns and features a lifelike, gilded face of a pharaoh on the lid. The craftsmanship is highly detailed, showcasing the artistry of the period.
3. 11s - 15s: The scene shifts to a different part of the museum, showing a large, ornate golden mask displayed in a glass case. The mask is highly detailed and appears to be made of gold. A woman is standing in front of the case, looking at the mask with interest. Other visitors can be seen in the background, also observing the exhibits.

Figure 8: A Single instruction example with 6 constraints. Gemma-4-31B-it and Qwen3-Omni-30B-A3B-Instruct fail to avoid mentioning words that denote colors.

Example 2



Instruction

Please sequentially narrate the events taking place in the video strictly between the 01:16 and 02:59 marks. Format your narration as an ordered list containing exactly 5 numbered items, and order the events chronologically from earliest to latest. You must write the narration from the first-person perspective of the makeup artist, Polly Osmond (e.g., 'I was applying...'). Additionally, write the descriptions of the events entirely in the past continuous tense. Ensure that you include the corresponding start and end timestamps in brackets for every event you list. Include the exact title 'Eye Shadow Application' right before your ordered list. Finally, do not mention any actions performed by Lindsay in your narration, and absolutely avoid using the words 'put', 'place', or 'add' anywhere in your response.

Gemma-4-31B-it ✗

Eye Shadow Application
1. I was applying a light base color to the eyelid and brow bone [01:16 - 01:34].
2. I was using a champagne shade in the inner corner of the eye [01:42 - 02:02].
3. I was highlighting the inner corner and the area underneath the eye [02:14 - 02:22].
4. I was utilizing a small piece of paper to create a wing while working a darker color into the crease [02:23 - 03:06].
5. I was blending the shadow into the lid and the lower wing [03:06 - 03:18].

Qwen3-Omni-30B-A3B-Instruct ✗

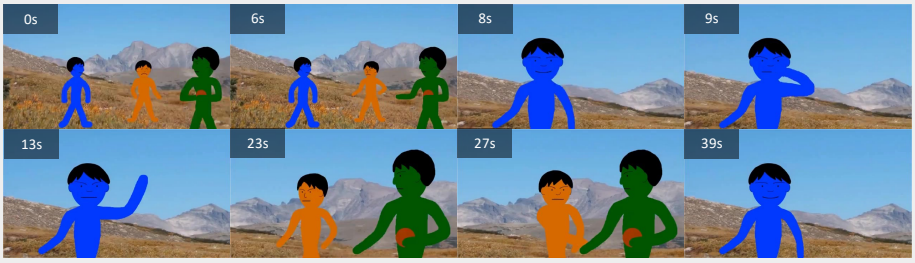
Eye Shadow Application
1. [01:16 - 01:28] I was applying the base color from the VS Eye Shadow Quad in Eye Contact to the model's eyelid and extending it up to the brow bone to create a highlight.
2. [01:29 - 01:40] I was using a champagne-colored shadow on the inner corner of the model's eye to enhance the luminous effect.
3. [01:41 - 01:53] I was applying a highlighter into the inner corner of the eye and slightly underneath to define the eye area.
4. [01:54 - 02:05] I was using a triangular piece of paper to create a sharp wing on the model's eyeliner, holding it in place with my fingers.
5. [02:06 - 02:59] I was blending the dark shadow into the crease and along the wing of the eyeliner, then using a brush to blend the bottom of the wing into the lid.

Qwen3.5-397B-A17B-Think ✓

Eye Shadow Application
1. [01:16 - 01:33] I was holding the eyeshadow quad and brushing the base color onto the lid and up towards the eyebrow.
2. [01:42 - 01:51] I was using the champagne color on the inner eye and taking it right into the corner.
3. [02:16 - 02:23] I was guiding the highlighter into the corner and slightly underneath the eye.
4. [02:23 - 02:37] I was showing how to use a tiny piece of triangular paper to create a wing without using tape.
5. [02:37 - 02:59] I was working the darker color into the eye and into the crease while holding the paper guide.

Figure 9: A Single instruction example with 9 constraints. Gemma-4-31B-it includes content outside the requested time window of 01:16 to 02:59, and Qwen3-Omni-30B-A3B-Instruct fails to avoid mentioning the word "place".

Example 3



Instruction

Add the title 'Colored Figures Analysis' at the top. I need you to analyze the social dynamics of the characters by addressing three tasks. Format your response as an ordered list with exactly three numbered items. For your overall response, limit all your evidence to events before the 0:40 mark, mention exactly three figures in total, and explicitly state the body color of every figure you mention. Restrict your mentioned entities strictly to animate figures and their body parts, meaning you must not mention inanimate objects like the circular object. Additionally, do not describe the background region, such as the mountain. Make sure the exact word 'gesture' appears at least once in your response.
1. Audio-Visual Temporal Alignment: Describe the visual actions of the orange and green figures that align temporally with the narrator referring to the blue figure as 'Charlie' (0:03 - 0:06).
2. Segmented Captioning: Provide a descriptive caption for the sequence between 0:07 and 0:21. Write this caption purely from the first-person perspective of the blue figure (using 'I' and 'my'). You must explicitly describe the blue figure touching his head, but only include actions or states of the blue figure that last more than 2 seconds.
3. Counterfactual Reasoning: Imagine if the green figure had kept his arm down at 0:26 instead of extending it; how would that have logically changed the subsequent physical interaction? For this item, restrict your focus solely to the specific figures who originally participated in the handshaking action.

Gemini-3-Pro ✗

Colored Figures Analysis
1. Between the 0:03 and 0:06 marks, the orange figure and the green figure both lift their arms and point their hands directly at the blue figure.
2. I, the blue figure, maintain a constant smile on my face while I bring my hand up to touch my head in a thinking gesture. Afterward, I raise my arm and hold up two of my fingers, keeping them extended.
3. If the green figure had kept his arm lowered at 0:26, the subsequent physical interaction would have been altered because the orange figure would not have been able to grasp his hand. Consequently, after the orange figure walks over, the orange figure and the green figure would merely stand facing each other without completing their handshaking action.

Qwen3.5-397B-A17B-Think ✓

Analysis of the Figures
1. I am Charlie, the wise arbitrator, and my distinct color is blue.
2. In my first close-up, I scratch my head, raise my arm, and make a peace sign gesture.
3. The orange figure begins walking towards the green one right after my shot ends.
4. Relative to the orange circular object held tightly by the green figure, the order from farthest is orange then green.
5. When the orange figure stops and extends his arm, the stationary green figure extends his hand back.
6. Without this specific response, the final cooperative handshake gesture between the two would have been impossible.

Figure 10: A Multi instruction example with 12 constraints. Gemini-3-Pro mentions an action that lasts less than 2 seconds in the video, violating the constraint in red.

Example 4



Instruction

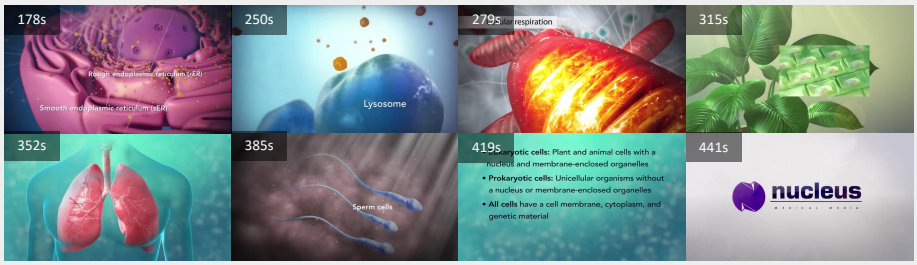
I need you to perform a detailed analysis of this live musical performance. Please complete the following eight tasks concurrently:
1. Describe the overall stage and audience environment as seen in the wide shots.
2. Identify the facial expression of the female singer during the loud audience cheers at the beginning.
3. Describe the clothing and accessories of the drummer. Important: avoid mentioning his hair or facial hair.
4. Count the exact number of separate visual instances where the drummer is actively shown playing the drum kit.
5. Locate the precise time interval when the large screen behind the band displays the female singer with angel wings.
6. State whether this angel wings projection is shown before or after the singer first stands on stage facing away from the audience.
7. Pinpoint the exact timestamps when the keyboardist is visible playing the keyboard.
8. Describe the interaction the singer has with her acoustic guitar immediately after she stops singing and turns away from the microphone near the end. In this description, do not mention her walking.
To structure your findings, format your entire response as a single valid JSON object using the exact keys "task_1" through "task_8" for the respective answers.

Qwen2.5-Omni-7B ✗

```json
{
 "task_1": "The wide shots show a large, dark stage with a crowd of people. The audience is seated in rows, and the stage is illuminated by spotlights. The atmosphere is energetic, with the audience cheering and clapping. The stage is set up with various musical instruments and equipment, and the lighting creates a dramatic effect, highlighting the performers and the crowd.",
 "task_2": "The female singer has a joyful and excited facial expression during the loud audience cheers at the beginning of the performance.",
 "task_3": "The drummer is wearing a black shirt and black pants. He is also wearing a black wristband on his left wrist. He is actively playing the drum kit throughout the performance.",
 "task_4": "The drummer is actively shown playing the drum kit in 5 separate visual instances.",
 "task_5": "The large screen behind the band displays the female singer with angel wings from 1:15 to 1:20.",
 "task_6": "The angel wings projection is shown after the singer first stands on stage facing away from the audience.",
 "task_7": "The keyboardist is visible playing the keyboard from 2:00 to 2:10.",
 "task_8": "Immediately after the singer stops singing and turns away from the microphone, she interacts with her acoustic guitar by strumming it gently while looking down at it. She then begins to walk away from the microphone, still holding the guitar."
}
```

Figure 11: A Multi instruction example with 3 constraints. Qwen2.5-Omni-7B fails to avoid mentioning her walking.

Example 5



Instruction

If the video shows a blue, blob-like lysosome absorbing orange particles before introducing red, bean-shaped mitochondria that glow with blue (yellow) energy to represent cellular respiration, then Identify the exact on-screen text labels that appear between 04:10 and 04:39. Return your answer as a valid JSON array containing exactly three text labels. Do not mention any shapes or colors in your response. Do not use the words 'organelle', 'cell', or 'energy' anywhere. Only include the exact text labels as they are written, preserving their original capitalization.
Else if an animal cell is depicted as a spherical structure containing a web-like cytoskeleton made up of thread-like microfilaments and hollow, tube-like microtubules, then As an observer looking through a microscope, please answer the following five questions in the exact order they are asked: First, at what exact timestamps in seconds does the web-like structure first appear? Second, what is the on-screen text label displayed next to it? Third, for how many seconds are the hollow tubes visible? Fourth, what text label accompanies these hollow tubes? Finally, what object is shown immediately after the hollow tubes disappear? For this final question, you must answer with exactly one of the following options: 'potted plant', 'lungs', 'bacteria'. For all extracted text, write it entirely in uppercase. Do not mention the color of any structure. Write all your descriptions in the present tense. Provide your answer as an ordered list with exactly five numbered list items. Use only seconds for all timestamps and durations. Do not use the word 'biology' anywhere. Include the word 'microscopic' exactly once in your response. Keep your total response under 150 words.
Else if the video zooms into a potted plant to reveal plant cells arranged in a scattered, random pattern (organized pattern) featuring a thick outer cell wall, a thinner inner membrane, and green, stacked disc-like chloroplasts, then Complete the following six steps in order. First, at what exact timestamps is the text 'Other unique STRUCTURES' displayed? Second, identify the organ shown inside a body outline right after the end timestamp of that text. Third, after zooming in on that organ, what hair-like structures are seen waving in a fluid? Fourth, what is the color of those hair-like structures? Fifth, following the scene with those hair-like structures, what pill-shaped organism is shown next? Sixth, what is the exact on-screen text label pointing to the long tail attached to that organism? Write your final answer in exactly six paragraphs, one for each step. Use strictly minutes and seconds (MM:SS) for all timestamps. Include the word 'subsequent' at least once.
Else if blue (pink) anatomical lungs are displayed inside a body outline just before revealing a microscopic view of pink, hair-like cilia waving in a fluid, then Identify the two specific types of swimming entities that appear after the waving structures are shown. Format your answer as a JSON object with a single key 'entites' containing a list of exactly two entities. Write your entire response in French. Do not mention any colors. Keep your response strictly under 40 words. Do not use the words 'tail', 'tails', or 'queue' anywhere in your response.
Else if single-celled and reproductive cells are shown swimming using long tails, specifically a pink, pill-shaped bacteria with flagella and three red (blue) sperm cells, then From the first-person perspective of a student taking notes, please answer the following six questions in the exact order provided: First, what topics are being compared in the summary text slides shown near the end? Second, at what exact timestamps in seconds do these summary slides appear? Third, what specific logo is shown at the very end of the footage? Fourth, for how many seconds is this logo displayed? Fifth, do the summary slides appear before or after the logo? Sixth, what exact text is written inside the logo? Format your response as an ordered list with exactly six list items. Express all timestamps and durations strictly in seconds. Write all extracted text from the footage entirely in lowercase. Use the past perfect tense for all your descriptions. Do not mention the color of the logo. Do not use the word 'video' anywhere in your response. Include the word 'conclusion' exactly once. Ensure your entire response is between 50 and 150 words long.
Else if the video concludes by presenting bulleted summary text slides that compare animal and plant cells (compare eukaryotic and prokaryotic cells) before showing the Nucleus Medical Media logo, then Trace the final events by completing these eight dependent steps: First, identify the logo shown at the very end of the footage. Second, what are the exact three words written in that logo? Third, what is shown immediately before the scene containing that logo? Fourth, what two broad categories of cells are being compared in those summary slides? Fifth, at what exact timestamps are the slides comparing those two cell types shown? Sixth, identify the specific entities shown swimming immediately before those summary slides. Seventh, how many of those swimming entities are shown? Eighth, what color are those entities? Provide your final output as a JSON object where the keys are 'step_1' through 'step_8'. Ensure your total answer is under 200 words. Do not use the word 'backward' or 'reverse'.
Else, Summarize the different types of microscopic structures and cells presented in the footage. Format your answer as exactly one paragraph. Write from the first-person perspective of a tour guide. Ensure your summary is between 30 and 80 words. Use only the future tense for all verbs. Write your entire response in uppercase letters. Do not mention any specific colors. Include the exact phrase 'AMAZING JOURNEY' exactly once. Do not use the words 'VIDEO', 'CLIP', or 'FOOTAGE'.

Gemini-3-Pro ✓

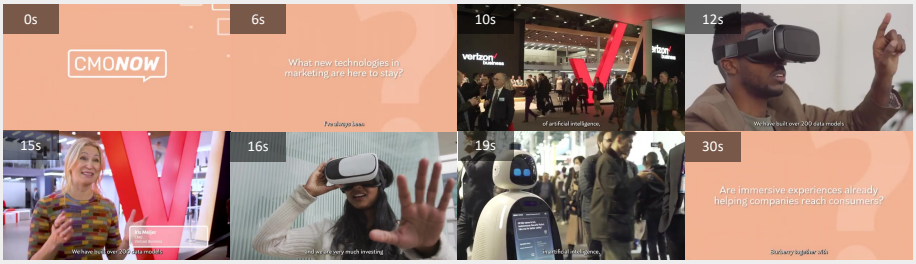
1. I observe that the web-like structure first appears at 287 seconds.
2. The on-screen text label displayed next to it is "CYTOSKELETON".
3. I note that the hollow tubes remain visible for 5 seconds.
4. The text label that accompanies these hollow tubes is "MICROTUBULES".
5. After the hollow tubes disappear from my microscopic view, the next object shown is a potted plant.

Qwen3.5-397B-A17B-Think (Wrong Branch) ✗

Cellular Exploration Log
I documented the following structures during my microscopic journey.
1. I examined a LYSOSOME, which appeared as a translucent blue, irregular sphere floating freely within the cellular fluid surrounding environment [04:10-04:20].
2. Nearby, a red, bean-shaped MITOCHONDRION displayed distinct internal folded membranes glowing with bright energy particles inside [04:20-04:45].
3. I observed a CYTOSKELETON forming a complex white and blue thread-like network throughout the interior space completely [04:49-05:03].
4. Within this network, twisted white MICROFILAMENTS provided structural support and maintained cell shape effectively [04:55-04:58].
5. Hollow blue MICROTUBULES also extended through the cytoplasm as thin, hollow tubes spanning the area [04:59-05:02].
6. A green, rectangular PLANT CELL was shown containing various internal components and distinct organelles inside clearly visible [05:15-05:20].
7. Inside, a green, oval CHLOROPLAST contained stacked disc structures for capturing light energy efficiently during the process of photosynthesis [05:20-05:29].
8. The CELL WALL surrounded the membrane as a rigid green outer layer for protection and support [05:32-05:35].
9. Finally, I noted pink, tail-like FLAGELLA propelling rod-shaped organisms through the surrounding medium very rapidly [06:15-06:23].

Figure 12: A Selection instruction example with 7 branches. Qwen3.5-397B-A17B-Think selects the wrong branch.

Example 6



Instruction

If the video begins with the CMO NOW logo on a peach background and an on-screen question about what new marketing technologies are here to stay, then
- If a speaker begins to address this introductory question by discussing artificial intelligence, then
-- If the speaker's introduction is accompanied by a wide shot of a Verizon Business exhibition booth with people gathered around, then
--- If a man in this Verizon Business booth uses a VR headset and lowers (raises) his hand, then
---- If the speaker is explicitly identified by on-screen text as 'Iris Meijer, CMO, Verizon Business', connecting her to the booth, then
----- If another individual, a woman, is shown wearing a similar VR headset and reaching out with her hand, then
------ If Iris Meijer continues by speaking about orchestrating communication while a robot navigates through a crowded hallway, then
------- If a new on-screen question asks if these immersive experiences are already helping companies reach consumers, then
-------- If Iris Meijer answers by mentioning virtual fashion shows, illustrating this point with a screenshot of a Burberry article, then
--------- If she provides a second illustration of these experiences via a screenshot of a The Met article showing virtual art galleries, then
---------- If she concludes her point about people being immersed in these experiences while a woman is seen looking at and using a smartphone, then
---------- If the topic then pivots to how companies can position themselves as early adopters of such technologies, as Iris mentions serving Fortune 500 customers, then
---------- If visuals of people walking through a large exhibition hall and close-ups of a server rack with blinking lights accompany this discussion on early adopters, then
---------- If Iris mentions enterprise intelligence providing platforms for these technologies, while a man in a busy area looks at a smartphone and a woman wears an AR headset, then
---------- If the video wraps up the insights on these platforms by displaying the CMO NOW logo again with text about exclusive insights, followed by the NBCUniversal logo, then 1. Locate the timestamp when Iris mentions enterprise intelligence platforms. 2. Bind the actions of the man with the smartphone and the woman with the AR headset to their specific agents. 3. Caption the on-screen text that appears with the CMO NOW logo at the end. 4. Deduce the causal link between the enterprise platforms dialogue and the final text. 5. Summarize the video's conclusion. Output as an ordered list, entirely in English, exactly 5 items, use past tense, and include the word 'NBCUniversal'.
---------- Else, Describe the visual events that follow the shot of the woman wearing the AR headset. Provide the response as a single JSON object, entirely in German, containing exactly one paragraph, bounded to a maximum of 50 words, and written in the past tense.
---------- Else, Analyze the shots of the server rack and exhibition hall. 1. Describe the overarching event in the hall. 2. Determine the spatial relationship between the connected cables and the server rack. 3. Judge the relative distance of the blinking lights to the camera. 4. Narrate the local temporal events in these shots. 5. Summarize the technological theme. Provide exactly 5 bullet points, entirely in Spanish, focus only on these specific shots, avoid mentioning 'platforms' in your vocabulary, and write in the past tense.
---------- Else, Transcribe exactly what Iris Meijer says regarding early adopters and Fortune 500 customers. Output must be enclosed in a JSON object, written entirely in French, be exactly one paragraph long, include the word 'Fortune', and be limited temporally to the specific time she discusses this topic.
---------- Else, 1. Locate the exact timestamp when the woman looking at the smartphone is shown. 2. Identify the dialogue Iris Meijer speaks immediately after this visual. 3. Analyze the causal reason for Iris's statement. 4. Summarize the visual scene accompanying this dialogue. 5. Explain how this connects to the prior smartphone shot. Use a structured JSON output, entirely in English, use past tense, include the keyword 'smartphone', and anchor the recognized dialogue strictly to the localized timestamp.
---------- Else, Focus on The Met article and virtual art galleries segment. 1. Identify the overarching event depicted in the galleries. 2. Describe the spatial relationship between the artworks and the gallery walls. 3. Compare the relative distance of the artworks to the viewer's perspective. 4. Describe the relative direction of the artworks. 5. Summarize the visual theme. Output as an ordered list, exactly 5 items, entirely in Spanish, use the past tense, and include the word 'virtual' twice.
---------- Else, Provide a dense caption of the screen showing the Burberry article. Format the output as a valid JSON object, entirely in German, write exactly one paragraph, keep it under 80 words, and restrict descriptions strictly to text and images visible on the article.
-------- Else, 1. Locate the moment the question about immersive experiences appears. 2. Transcribe Iris Meijer's exact spoken answer following this question. 3. Identify the visual events shown during her answer. 4. Deduce the causal relationship between her answer and the visual events. 5. Summarize her core argument. Output as an ordered list, entirely in English, use exactly 5 list items, use the past tense, and restrict the answer to the 0:28-0:40 timeframe.
------- Else, Regarding the robot moving in the hallway: 1. Determine its relative direction to the crowd. 2. Compare its relative distance to the walls versus the people. 3. Describe its spatial relationship to the floor. 4. Narrate the local temporal events happening around it. 5. Summarize its role in the scene. Provide the response as a table, entirely in Spanish, focus only on the timeframe of the robot's shot, restrict entities mentioned to physical objects, and include the word 'hallway' exactly once.
------ Else, Caption the visual events that happen immediately after the woman wearing the VR headset is shown. Limit the caption to exactly one paragraph, entirely in French, use the past tense, and ensure the text is under 60 words.
----- Else, 1. Locate the timestamp when Iris Meijer is identified on screen. 2. Identify the dialogue she speaks right after this timestamp. 3. Explain the causal reason for her statement. 4. Describe the visual events shown while she makes this statement. 5. Summarize how her statement connects to the previous VR headset scene. Format as a JSON object, entirely in English, include the keyword 'Iris', use the past tense, and ensure the recognized dialogue is anchored exactly after the localized timestamp.
---- Else, Focus on the man using the VR headset. 1. Describe the overarching event he is participating in. 2. Infer his emotional state. 3. Compare the relative distance of the people standing around him. 4. Describe his spatial relationship to the booth counter. 5. Summarize his interaction with the headset. Output as an ordered list, containing exactly 5 items, entirely in Spanish, restricted to the timeframe he is visible, and do not mention his clothing.
--- Else, Provide a dense caption of all visible actions and entities in the wide shot of the Verizon Business exhibition booth. Output the response as a table, write entirely in German, restrict your descriptions to people only, ensure the table covers only the 0:08-0:09 timeframe, and explicitly include the word 'booth' in your text.
-- Else, 1. Locate the precise timestamp when the speaker begins discussing AI. 2. Based on that timestamp, recognize the exact words she says next. 3. Visually localize the source of this speech on screen. 4. Narrate the background actions happening during this speech. 5. Conclude why this speech is relevant to the intro question. Ensure the output is an ordered list, written in English, containing exactly 5 list items, avoiding any mention of the speaker's accent, and explicitly using the located timestamp to anchor the recognized dialogue.
- Else, Analyze the introductory part of the video where the logo appears. 1. Describe the overarching event happening in the first 10 seconds. 2. Identify the emotional state of the speaker. 3. Compare the relative distance of the objects behind the speaker. 4. Describe the spatial relationship between the speaker and the background. 5. Summarize the text shown on screen. Provide the answer in French, as a JSON object, keeping the total response under 150 words. Ensure the events described in the first task are in chronological order.
Else, Describe the visual events in the first 3 seconds of the video since it does not begin with the CMO NOW logo. Write exactly one paragraph under 50 words entirely in Spanish. Format the answer as a JSON object.

Gemini-3-Pro (Wrong Path) ✗

1. Iris mentioned enterprise intelligence platforms at the 1:19 timestamp.
2. The man looked down at his smartphone in a crowded area, while the woman wore an AR headset and appeared to speak.
3. The final on-screen text displayed "Exclusive insights from global marketing and business leaders" along with the CMO NOW logo and a website URL.
4. The dialogue regarding enterprise platforms provided a concrete example of the expert knowledge that the concluding text subsequently promoted.
5. The video concluded by establishing the CMO NOW series as a premier source for professional insights, finishing with the display of the NBCUniversal logo.

Figure 13: A Nested instruction example with a maximum depth of 15. Gemini-3-Pro follows the wrong path.

E HUMAN AGREEMENT ANALYSIS

Table 7: Human agreement with the automatic evaluator on sampled examples.

Annotator	Task Pass Agreement	TCSR Agreement	TISR Agreement
Annotator 1	90.6	84.1	82.4
Annotator 2	88.7	86.4	76.5
Annotator 3	92.3	83.0	88.2
Annotator 4	89.5	87.5	82.4
Overall	90.3	85.3	82.4

To examine the reliability of our automatic evaluation protocol, we conduct a human agreement study on a randomly sampled subset of examples. Four human annotators independently review the model responses and judge whether each response satisfies the task-level requirement and each checklist constraint.

We compare human annotations with the automatic evaluator at three levels. For task pass agreement, we compare whether the human annotator and the automatic evaluator give the same task-level pass/fail decision. For TCSR agreement, we compare human and automatic judgments at the constraint level and compute the fraction of constraints on which they give the same pass/fail decision. For TISR agreement, we first aggregate each example into a binary instruction-level decision by checking whether all constraints are satisfied, and then compute agreement between human and automatic decisions. The final score is averaged over the four annotators.

As shown in Table 7, the automatic evaluator reaches 85.3% at TCSR agreement, and 82.4% at the TISR agreement. These results indicate that the automatic evaluation is sufficiently aligned with human judgments.

F LIMITATIONS

Video-IFBench is designed to evaluate instruction following in video understanding, but it has several limitations. First, although the benchmark covers diverse domains, tasks, and constraint types, it is still built from a finite set of public videos and manually designed instruction templates. Therefore, it may not cover all real-world video instruction-following scenarios, such as very long videos, live streaming interaction, multi-turn dialogue, or domain-specific professional applications.

Second, our evaluation protocol relies on a hybrid of LLM-as-Judge and programmatic verification. Programmatic checks provide deterministic assessment for many format constraints, but semantic constraints often require LLM judges. Although we use a strong judge model in the main experiments and provide an additional robustness analysis with a smaller judge, LLM-based evaluation can still introduce judgment errors.

Third, different model families support different video input interfaces. For example, some proprietary models can consume videos directly, while GPT-5.4 is evaluated with uniformly sampled frames due to API constraints, and vision-only open-source models are provided with video frames and timestamped subtitles. We follow the recommended or feasible inference setup for each model, but these differences may affect absolute scores.