IVU-BENCH: BENCHMARKING MULTIMODAL LARGE LANGUAGE MODELS FOR VIDEO UNDERSTANDING UNDER INCOMPLETE OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have demonstrated strong video understanding, yet existing evaluations largely focus on recognizing explicitly observable content. Real-world video understanding often requires inference beyond observation, including predicting missing events, recovering spatial relations from partial evidence, and reasoning under modified or hypothetical conditions. We introduce **IVU-Bench**, a benchmark for evaluating MLLMs under incomplete observations, comprising 2,032 question-answer pairs, and 13 tasks across temporal, spatial, and factual dimensions. Temporal incompleteness involves inferring missing events or temporal structure; spatial incompleteness involves recovering object locations and relations from partial observations across time and viewpoints; and factual incompleteness involves reasoning about event dynamics under modified or hypothetical conditions absent from the original video. Questions from multiple task types are grounded in shared videos, enabling controlled cross-task comparisons while reducing content-related confounds. To assess whether correct answers are supported by appropriate visual evidence, IVU-Bench provides object-level spatiotemporal annotations for target evidence and semantically related or irrelevant distractors. We further evaluate models with three complementary metrics: *Answer Accuracy*, *Evidence Discrimination*, and *Joint Answer–Evidence Score*, measuring answer correctness, target-evidence preference, and answer–evidence consistency, respectively. Experiments across representative MLLMs reveal substantial weaknesses across all three dimensions, especially in spatial reasoning from partial visual evidence, highlighting the gap between recognizing visible content and inferring unobserved information.

1 INTRODUCTION

Recent multimodal large language models (MLLMs) have made substantial progress recognizing and reasoning about events explicitly visible in videos, supporting tasks such as action recognition, event understanding, and video question answering (Li et al., 2024b; Fu et al., 2025; Mangalam et al., 2023). Real-world video understanding, however, often requires more than recognizing what is shown. In an egocentric recording, an answer-relevant action may occur before or after the observed segment, an object may move while out of view, or the question may specify an intervention that never occurs in the video. The model must therefore reconstruct an answer-relevant event or state from partial visual evidence and the conditions stated in the question. Existing studies address parts of this challenge through future-oriented egocentric understanding (Grauman et al., 2022; Jia et al., 2022), sequential spatial reasoning (Yang et al., 2025; Bang & Song, 2026), causal or counterfactual VideoQA (Li et al., 2022; Zhou et al., 2025b; Chen et al., 2026b), and goal-conditioned egocentric planning (Chen et al., 2026a; Qiu et al., 2026). Together, these lines of work point to a common challenge: models must infer events, relations, or states that are not directly exposed in the video.

We study this setting as **incomplete-observation video understanding** (IVU), where some answer-relevant information is not directly observable but can be inferred from available visual evidence and the question conditions. This can arise from a temporal gap, from object relations that are distributed across views or time, or from a factual intervention that changes the condition or outcome described by the original video. We accordingly organize IVU into three dimensions: *temporal in-*

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.

completeness, which involves missing processes, unobserved future events, or perturbed order; *spatial incompleteness*, which requires recovering object locations or relations from partial observations across time and viewpoints; and *factual incompleteness*, which requires inferring the consequences of modified or hypothetical facts that are absent from the video. Although the source of missing information differs across these dimensions, they share a common requirement: the model must infer a latent or intervened event state rather than retrieve a directly visible fact.

Existing video benchmarks cover temporal reasoning, long-video understanding, causal reasoning, and spatial relation reasoning, but typically evaluate these capabilities in isolation (Xiao et al., 2021; Li et al., 2024b; Mangalam et al., 2023; Li et al., 2022; Yang et al., 2025). Broad benchmarks emphasize recognition, retrieval, temporal comprehension, or long-video aggregation (Fu et al., 2025; Wu et al., 2024; Zhou et al., 2025a), whereas prediction, planning, spatial reconstruction, and counterfactual reasoning are often treated as separate problems. Differences in task format, available context, and evaluation targets, together with confounds from language knowledge, static image semantics, and temporal cues (Liu et al., 2024; Zhou et al., 2025c; Ataallah et al., 2025), make results difficult to compare and obscure the intended reasoning ability. These gaps motivate a unified and controlled benchmark spanning multiple forms of incomplete observation. However, unified coverage alone is insufficient if evaluation stops at answer correctness. A model may exploit language priors or incidental visual cues without identifying observations supporting its answer (Xiao et al., 2024; Liu et al., 2024); conversely, it may recognize relevant objects but select a semantically similar distractor as evidence. Although recent grounded VideoQA benchmarks examine evidence integration and temporal or spatial verification (Nagrani et al., 2025; Cheng et al., 2026; Meng et al., 2026), answer correctness and evidence support are often assessed separately. IVU therefore evaluates both whether the answer is correct and whether selected observations support it.

We introduce **IVU-Bench**, a diagnostic benchmark for **I**ncomplete-observation **V**ideo **U**nderstanding. IVU-Bench is designed around two principles: covering multiple forms of incomplete information and diagnosing whether answers are supported by the appropriate visual evidence. It contains 127 egocentric videos from EPIC-KITCHENS-100, HD-EPIC, and Ego4D (Damen et al., 2022; Perrett et al., 2025; Grauman et al., 2022), with 2,032 question-answer pairs organized into 13 tasks. The temporal tasks cover prediction, abduction, goal-conditioned planning, and ordering; the spatial tasks assess object locations and pairwise relations before and after movement; and the factual tasks require reasoning about modified or hypothetical facts and alternative outcomes. Importantly, each video is paired with questions spanning all three task types, allowing different reasoning abilities to be evaluated under shared visual contexts. This shared-video design enables controlled cross-task comparisons while reducing confounding effects from differences in video content.

For its evidence-based task suite, IVU-Bench provides object-level spatiotemporal annotations of *target evidence*, *semantically related distractors*, and *semantically irrelevant distractors*. These annotations provide the supervision needed to evaluate evidence selection and to separate whether a model answers correctly from whether it selects the visual evidence that justifies the answer. Besides, we introduce three complementary metrics: *Answer Accuracy*, which measures whether the predicted answer is correct; *Evidence Discrimination*, which measures whether the model favors target evidence over semantically related and irrelevant distractors; and the *Joint Answer–Evidence Score*, which rewards a correct answer when it is accompanied by appropriate target evidence. Together, these metrics distinguish answer correctness, evidence selection, and answer-evidence consistency. We evaluate multiple representative MLLMs, together with analyses of cross-task correlations, model scaling, frame scaling, and evidence quality. Experiments show that representative MLLMs remain weak across all three dimensions, with spatial reasoning from partial visual evidence being particularly challenging. These findings reveal a persistent gap between recognizing visible content and inferring what is not directly observed. Our main contributions are:

- We formulate incomplete-observation video understanding as a unified evaluation problem and operationalize it through temporal, spatial, and factual dimensions.
- We introduce IVU-Bench with 2,032 question–answer pairs, 13 tasks, and object-level annotations of target evidence and distractors, enabling joint answer-evidence evaluation.
- We evaluate representative MLLMs and identify substantial limitations, particularly in spatial reasoning from partial evidence.

2 Related Work

Video Understanding Benchmarks. Video benchmarks evaluate increasingly diverse capabilities of multimodal large language models (MLLMs) (Patraucean et al., 2023; Li et al., 2024c). SEED-Bench assesses multimodal comprehension (Li et al., 2024a), while MVBench emphasizes temporal and dynamic understanding (Li et al., 2024b). Video-MME (Fu et al., 2025) and Video-MME-v2 (Fu et al., 2026) broaden evaluation across video domains and capabilities, and EgoSchema targets long-term temporal comprehension in egocentric videos (Mangalam et al., 2023). These benchmarks emphasize broad capability coverage or extended temporal context. IVU-Bench instead centers evaluation on reasoning when answer-relevant information is only partially observable.

Reasoning under Incomplete Visual Observations. Existing benchmarks examine several relevant reasoning abilities. NExT-QA evaluates causal and temporal relationships (Xiao et al., 2021), EgoTaskQA studies task structure and dependencies in egocentric activities (Jia et al., 2022), and Black Swan investigates abductive and defeasible reasoning (Chinchure et al., 2025). VSI-Bench evaluates spatial reasoning from sequential observations (Yang et al., 2025), while PhysBench (Chow et al., 2025) and SOK-Bench (Wang et al., 2024) examine physical understanding and situated reasoning. Building on these complementary directions, IVU-Bench evaluates temporal prediction and reconstruction, spatial state recovery, and reasoning under modified factual conditions within a common protocol for incomplete observations.

Visual Evidence Grounding. Correct answers do not necessarily imply reliable grounding in supporting visual observations. NExT-GQA demonstrates this gap through joint evaluation of question answering and temporal evidence localization (Xiao et al., 2024). VideoZeroBench further examines spatiotemporal evidence verification (Meng et al., 2026). IVU-Bench extends evidence-based evaluation to incomplete observations with object-level spatiotemporal annotations that distinguish target evidence from semantically related and semantically irrelevant distractors. This supports evaluation of answer correctness, evidence selection and discrimination, and answer–evidence consistency, revealing whether correct answers coincide with appropriate evidence selection.

3 Benchmark Design and Construction

We introduce IVU-Bench, a diagnostic benchmark for video understanding under incomplete observations. Rather than focusing only on information explicitly presented in the available visual input, IVU-Bench evaluates whether models can infer unobserved or altered states from limited, local, or discontinuous visual evidence. Each question is associated with a task-specific temporal interval defined by start and end timestamps, allowing the evaluated visual context to range from a local clip to the entire source video when required by the task. The benchmark is organized around a capability hierarchy, three dimensions of incompleteness, object-level visual evidence, and a unified construction and evaluation protocol.

3.1 Capability Hierarchy

We characterize video understanding under incomplete observations using three progressively more demanding capabilities. **Visual Information Aggregation** requires extracting and integrating entities, attributes, and visual–textual correspondences from local observations. **Temporal Entity Tracking** requires maintaining consistent entity identities and states across time despite viewpoint changes, occlusion, or missing observations. **Reasoning under Incomplete Observations** further requires inferring unobserved or altered world states from the available evidence. Together, these levels describe a progression from observable information acquisition, through dynamic state modeling, to latent world-state inference.

3.2 Tasks and Evidence

IVU-Bench contains 13 tasks across three dimensions: temporal, spatial, and factual incompleteness. All tasks use an eight-way multiple-choice format. Each question is associated with a task-specific temporal interval $[t_s, t_e]$, which determines the visual context available for that question.

Depending on the task instance, this interval may correspond to a local portion of the source video or span the entire video.

Temporal Incompleteness. The temporal dimension contains four tasks. **Prediction** asks the model to infer the most likely future action sequence from preceding observations. **Abduction** hides an intermediate process and requires reconstructing it from observations before and after the missing interval. **Planning** additionally provides a high-level goal and asks for the actions needed to achieve it. **Ordering** shuffles action segments and requires recovering their original temporal order.

Spatial Incompleteness. The spatial dimension evaluates whether models can integrate local observations across time and viewpoints into a consistent spatial representation. It contains five task types: **Object Position Before Movement**, **Object Position After Movement**, **Object Position Before vs. After Movement**, **Pairwise Position Before Movement**, and **Pairwise Position After Movement**. The first three object-position task types are instantiated for two target objects, while the two pairwise tasks each contribute one question, yielding eight spatial questions per source video. Questions are defined under a spatial reference frame and require reasoning about object locations, state changes, and pairwise spatial relations.

Factual Incompleteness. The factual dimension evaluates reasoning under factual conditions or outcomes that differ from those observed in the original video. **Single-Fact Perturbation** modifies one existing fact, **Dual-Fact Perturbation** modifies two facts jointly, **Counterfactual Fact Injection** introduces a hypothetical fact absent from the observed process, and **Outcome Fact Modification** specifies an alternative outcome and requires reasoning about a process consistent with that outcome. These tasks require models to reason under modified factual conditions rather than simply reproduce the observed video trajectory.

Visual Evidence. Except for Ordering, each sample additionally includes object-level spatiotemporal evidence represented by bounding boxes in selected video frames. Candidate evidence is divided into **Target Evidence**, which supports the correct answer; **Semantically Related Distractors**, which are semantically similar to the target evidence but do not support the correct answer; and **Semantically Irrelevant Distractors**, which are unrelated to the required inference. This design allows us to evaluate not only answer correctness but also whether models can distinguish supporting observations from plausible distractors.

3.3 Benchmark Construction

Our source videos are drawn from four representative egocentric datasets: **EPIC-KITCHENS-100**, **EPIC-KITCHENS-55**, **HD-EPIC**, and **Ego4D** (Damen et al., 2022; 2018; Perrett et al., 2025; Grauman et al., 2022). We select videos containing clear action processes, trackable object-state changes, sufficient temporal context, and substantial viewpoint variation. The resulting benchmark contains **127 videos** and **2,032 question–answer pairs**, with 16 questions per video: four temporal questions, eight spatial questions instantiated from five spatial task types, and four factual questions.

Benchmark construction consists of three stages. First, annotators inspect the complete source video and instantiate task-specific templates using temporal ranges, entities, object states, task goals, and factual interventions. Each question is assigned a start timestamp t_s and an end timestamp t_e, defining the visual interval for evaluation. This interval may cover a local portion of the source video or, when required, the entire video. Temporal ranges are defined according to per-task information requirements; spatial questions use consistent object references and spatial coordinate definitions; factual interventions are expressed through the question text without modifying the source video.

Second, annotators determine the correct answer and construct seven distractors. Distractors are designed to capture meaningful error modes, including object confusion, incorrect actions or action ordering, inconsistent spatial relations, failure to account for factual interventions, and contextually plausible but incorrect alternatives. Options are kept similar in form and length whenever possible to reduce superficial linguistic cues.

Third, annotators identify visual observations supporting the correct answer and annotate the corresponding objects with bounding boxes. Related and unrelated distractors are then added to form the

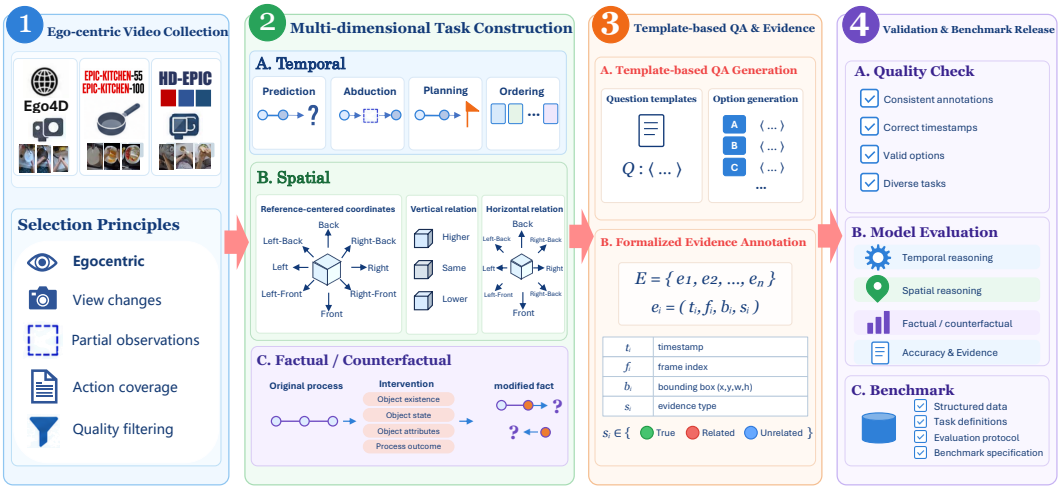


Figure 1: Construction pipeline of IVU-Bench. Source videos are transformed into structured question–answer samples through template-based question annotation, answer and option construction, and spatiotemporal evidence annotation.

evidence candidate set. Representative frames are preferred over redundant adjacent frames, while additional evidence is annotated when object states or viewpoints change.

We apply two complementary quality-control procedures. **Text-only validation** identifies questions that can be solved without visual input due to language priors or option artifacts. **Inter-annotator verification** cross-checks questions, answers, and annotations against the source videos.

3.4 Evaluation Metrics

We evaluate models using answer accuracy, evidence discrimination, and joint answer–evidence performance. Evidence-related quantities are computed at the sample level and then macro-averaged over tasks and dimensions. Ordering is excluded from evidence-related evaluation because it does not contain object-level evidence annotations.

Answer Accuracy. Let Acc_k denote the answer accuracy for task k. Each dimension-level accuracy is the macro-average over its corresponding tasks. The overall answer score Avg equally averages the three dimensions:

$$\mathrm{Avg} = \frac{\mathrm{Acc}_{\mathrm{Temp}} + \mathrm{Acc}_{\mathrm{Spa}} + \mathrm{Acc}_{\mathrm{Fact}}}{3}. \tag{1}$$

Discrimination. For task k, the Evidence Discrimination score is

$$D_k = \frac{1}{N_k} \sum_{i=1}^{N_k} d_i, \tag{2}$$

where N_k is the number of samples and d_i measures how strongly the model favors target evidence over distractors:

$$d_i = \frac{d_i^R + d_i^U}{2}, \qquad d_i^R = \mathrm{esr}_i^T - \mathrm{esr}_i^R, \qquad d_i^U = \mathrm{esr}_i^T - \mathrm{esr}_i^U. \tag{3}$$

Here, d_i^R and d_i^U measure discrimination against semantically related and irrelevant distractors, respectively. For sample i, let S_i be the evidence selected by the model, and let T_i, R_i, and U_i denote the sets of Target Evidence, Semantically Related Distractors, and Semantically Irrelevant Distractors. The selection rate for evidence type $C \in \{T, R, U\}$ is

$$\mathrm{esr}_i^C = \frac{|S_i \cap C_i|}{|C_i|}, \qquad C \in \{T, R, U\}. \tag{4}$$

A higher D_k indicates more reliable selection of target evidence and stronger rejection of distractors.

Joint Answer–Evidence Score. For task k, we define

$$J_k = \frac{1}{N_k} \sum_{i=1}^{N_k} a_i \max(d_i, 0), \tag{5}$$

where $a_i = 1$ if the predicted answer is correct and 0 otherwise. A higher J_k indicates that correct answers more frequently coincide with positive evidence discrimination. This metric measures answer–evidence consistency rather than the causal faithfulness of the model's internal reasoning.

4 EXPERIMENTS

Evaluation Setup. We evaluate representative commercial and open-source MLLMs under a unified evaluation protocol. Each question is evaluated using the visual interval defined by its annotated start and end timestamps $[t_s, t_e]$. Depending on the task instance, this interval may cover only part of the source video or span the entire video. Unless otherwise specified, 128 video frames are sampled from this task-specific interval using the same sampling and decoding settings across models to minimize performance differences caused by evaluation configurations. For samples with evidence annotations, models additionally return the identifiers of the evidence candidates they select, without being constrained to select a fixed number of candidates.

Results. We report model performance from two perspectives. Table 1 reports answer accuracy for all 13 tasks, together with macro-averaged results for temporal, spatial, and factual incompleteness and the equal-weight average across the three dimensions. Table 2 reports task-level evidence-related performance using the Evidence Discrimination score (D) and Joint Answer–Evidence score (J). Ordering evaluation is excluded because it does not contain evidence annotations.

Table 1: Main benchmark results (%). We report task accuracy, dimension-level macro averages, and their equal-weight average (**Avg.**). The best and second-best model results are shown in bold and underlined, respectively.

Model	Frames	Temporal Incompleteness					Spatial Incompleteness						Factual Incompleteness					Avg.
		Pred.	Abd.	Plan.	Ord.	Temp.	Pos-B	Pos-A	B-vs-A	Pair-B	Pair-A	Spa.	SFP	DFP	CFI	OFM	Fact.	
Human Baseline																		
Human Expert	–	87.3	91.2	88.2	92.2	89.7	90.7	90.2	87.7	84.3	86.3	87.8	90.2	85.3	89.2	83.3	87.0	88.2
Commercial Models																		
GPT-5.5	128	63.0	51.2	69.3	**33.9**	54.3	28.3	33.1	22.8	23.6	**26.8**	26.9	72.4	76.4	77.2	74.0	75.0	52.1
Gemini-3.1-Pro-preview	128	**84.1**	**67.5**	**85.0**	22.2	**64.7**	**38.6**	39.0	**23.8**	**32.4**	25.3	**31.8**	73.8	**83.8**	73.4	70.0	75.2	**57.3**
Qwen3.7-Plus	128	78.7	52.8	78.7	22.0	58.1	32.8	**40.6**	22.5	22.8	25.2	28.8	75.6	78.0	76.4	**78.7**	**77.2**	54.7
Open-Source Models (Reasoning)																		
Qwen3.5-397B-A17B	128	71.7	55.1	70.1	18.1	53.7	22.0	25.2	13.4	21.3	16.7	19.7	74.0	77.0	**78.6**	78.0	76.9	50.1
Qwen3.5-27B	128	69.3	51.2	70.1	22.8	53.3	28.8	28.7	19.4	19.0	20.8	23.4	74.6	78.4	77.6	74.6	76.3	51.0
Qwen3-VL-235B-A22B-Thinking	128	65.4	48.8	64.6	22.0	50.2	18.6	24.8	16.5	16.5	19.0	19.1	**76.4**	73.0	65.9	73.2	72.1	47.1
Qwen3-VL-235B-A22B-Thinking	64	63.0	48.8	63.0	24.6	49.9	22.5	28.3	16.5	20.5	20.6	21.7	73.2	67.5	65.9	73.2	69.9	47.2
Qwen3-VL-30B-A3B-Thinking	128	56.7	40.9	52.8	18.1	42.1	18.2	22.0	13.4	18.1	11.1	16.6	66.1	67.5	67.5	67.7	67.2	42.0
Qwen3-VL-30B-A3B-Thinking	64	55.9	41.7	51.2	16.7	41.4	18.2	19.3	13.4	15.7	14.3	16.2	67.7	67.5	67.5	65.4	67.0	41.5
Qwen3-VL-8B-Thinking	128	52.0	37.0	48.0	12.6	37.4	19.0	20.1	13.8	10.3	11.9	15.0	62.2	65.9	72.2	54.4	63.7	38.7
Qwen3-VL-8B-Thinking	64	43.3	44.1	49.6	15.9	38.2	15.4	23.2	12.2	12.6	11.0	14.9	59.1	61.1	61.9	59.1	60.3	37.8
Qwen3-VL-4B-Thinking	128	46.5	30.7	41.7	8.7	31.9	14.6	16.9	7.9	10.2	15.7	13.1	66.1	63.0	59.1	59.1	61.8	35.6
Qwen3-VL-4B-Thinking	64	44.9	33.9	44.9	8.7	33.1	16.5	13.4	10.6	12.6	12.6	13.1	64.6	59.8	58.3	56.7	59.8	35.4
GLM-4.6V-FlashX	50	40.2	21.3	37.0	3.1	25.4	17.7	13.8	10.6	13.4	13.4	13.8	37.0	34.6	48.4	29.1	37.3	25.5
GLM-4.1V-Thinking-FlashX	50	55.9	29.1	44.9	11.1	35.3	18.1	20.1	20.9	17.3	21.3	19.5	58.3	58.3	63.8	50.4	57.7	37.5
Open-Source Models (Instruct)																		
Qwen3-VL-235B-A22B-Instruct	128	67.7	55.9	67.7	17.3	52.2	27.2	27.6	11.8	19.7	12.6	19.8	70.9	69.0	69.8	71.7	70.4	47.4
Qwen3-VL-235B-A22B-Instruct	64	68.5	56.7	67.7	18.3	52.8	27.6	26.8	15.0	22.0	15.9	21.4	68.5	67.5	68.3	70.9	68.8	47.7
Qwen3-VL-30B-A3B-Instruct	128	60.6	48.8	52.8	15.0	44.3	18.5	16.5	8.7	11.0	7.1	12.4	63.8	69.8	69.0	70.1	68.2	41.6
Qwen3-VL-30B-A3B-Instruct	64	60.6	48.8	55.1	15.9	45.1	17.4	18.1	7.5	7.9	7.9	11.8	63.0	64.3	67.5	70.1	66.2	41.0
Qwen3-VL-8B-Instruct	128	57.5	41.7	51.2	15.7	41.5	14.6	18.5	13.0	17.3	11.8	15.0	62.2	54.8	64.3	60.6	60.5	39.0
Qwen3-VL-8B-Instruct	64	56.7	40.2	48.0	17.5	40.6	18.9	20.5	12.6	15.7	16.5	16.9	59.1	55.6	64.3	59.1	59.5	39.0
Qwen3-VL-4B-Instruct	128	40.2	39.4	40.9	17.3	34.4	13.4	15.0	9.1	15.0	14.2	13.3	63.0	61.4	55.1	58.3	59.4	35.7
Qwen3-VL-4B-Instruct	64	35.4	37.8	42.5	19.7	33.9	13.8	15.7	9.8	15.0	14.2	13.7	61.4	62.2	58.3	59.1	60.2	35.9
Gemma4-E2B-Instruct	128	53.5	43.3	44.9	13.4	38.8	16.5	17.7	17.3	19.7	14.2	17.1	62.2	59.8	65.4	55.1	60.6	38.8

Abbreviations: Pred./Abd./Plan./Ord.=Prediction/Abduction/Planning/Ordering; Pos-B/A=Object Position Before/After Movement; B-vs-A=Object Position Before vs. After Movement; Pair-B/A=Pairwise Position Before/After Movement; SFP/DFP=Single-/Dual-Fact Perturbation; CFI=Counterfactual Fact Injection; OFM=Outcome Fact Modification.

Table 2: Evidence-related results ($\times 100$), reported as $\mathbf{D}/\mathbf{J}$ for Evidence Discrimination and Joint Answer–Evidence performance. Dimension scores are macro-averaged over the corresponding tasks, and **Overall** equally averages the three dimensions. Ordering is excluded. The best and second-best model results for each metric are shown in bold and underlined, respectively.

Model	Frames	Temporal Incompleteness				Spatial Incompleteness						Factual Incompleteness					Overall
		Pred.	Abd.	Plan.	Temp.	Pos-B	Pos-A	B-vs-A	Pair-B	Pair-A	Spa.	SFP	DFP	CFI	OFM	Fact.	
Commercial Models																	
GPT-5.5	128	**69.7**/48.0	40.9/27.4	**69.3**/54.3	**59.9**/43.2	58.5/22.1	**57.8**/23.9	**65.7/16.9**	53.5/**18.9**	53.5/**18.5**	**57.8/20.1**	47.2/43.7	52.8/42.5	57.3/**50.4**	44.0/39.9	50.3/44.1	56.0/35.8
Gemini-3.1-Pro-preview	128	64.6/**58.5**	43.7/**36.1**	58.8/**55.6**	55.7/**50.1**	**60.0/27.4**	52.8/**24.3**	60.2/15.1	45.6/14.0	46.7/12.0	53.0/18.6	**59.4/47.5**	**61.4/53.8**	**59.9**/50.0	**58.3/46.8**	**59.7/49.5**	**56.2/39.4**
Qwen3.7-Plus	128	55.1/48.8	35.7/26.6	58.7/49.6	49.8/41.7	55.7/21.2	49.0/22.3	45.0/13.8	51.6/14.6	48.8/13.4	50.0/17.0	41.3/40.1	50.0/43.3	46.0/40.7	44.0/33.9	45.3/39.5	48.4/32.7
Open-Source Models (Reasoning)																	
Qwen3.5-397B-A17B	128	62.2/50.0	38.1/25.0	52.4/44.9	50.9/40.0	51.8/14.4	56.0/15.3	55.6/8.7	**59.1**/14.6	49.6/11.5	54.4/12.9	44.4/38.1	48.4/39.6	47.6/44.7	44.8/38.7	46.3/40.3	50.5/31.1
Qwen3.5-27B	128	50.4/37.8	24.2/18.7	55.5/44.1	43.4/33.5	53.8/21.6	52.4/16.1	54.8/13.7	52.4/13.9	48.8/13.6	52.4/15.8	46.0/41.6	56.5/46.0	52.0/45.9	43.1/37.4	49.4/42.7	48.4/30.7
Qwen3-VL-235B-A22B-Thinking	128	56.7/44.5	42.9/28.6	62.2/45.7	53.9/39.6	42.6/11.5	54.0/17.9	52.6/11.1	51.6/10.2	52.0/13.9	50.6/12.9	44.0/43.7	43.2/40.4	29.7/35.0	45.2/**46.8**	40.5/41.4	48.3/31.3
Qwen3-VL-235B-A22B-Thinking	64	59.1/44.1	**48.0**/27.0	50.8/42.9	52.6/38.0	45.1/12.5	52.4/18.7	53.2/10.7	45.3/11.4	46.0/14.7	48.4/13.6	42.9/42.5	46.0/42.4	43.5/39.8	42.3/45.6	43.7/42.6	48.2/31.4
Qwen3-VL-30B-A3B-Thinking	128	36.2/27.2	30.2/18.7	20.5/12.6	29.0/19.5	35.3/8.9	31.5/8.6	35.9/6.4	35.0/9.1	48.4/6.3	37.2/7.8	23.8/25.8	32.8/28.8	17.5/18.3	33.1/28.2	26.8/25.3	31.0/17.5
Qwen3-VL-30B-A3B-Thinking	64	39.4/30.3	38.5/25.8	26.0/24.8	34.6/27.0	44.8/11.7	39.7/10.0	41.3/7.5	40.9/7.9	51.2/9.1	43.6/9.2	21.4/25.8	28.0/25.6	28.5/28.5	33.1/32.7	27.7/28.1	35.3/21.4
Qwen3-VL-8B-Thinking	128	49.6/33.5	31.7/17.9	35.4/30.7	38.9/27.3	41.5/13.5	50.6/13.6	41.4/8.6	43.7/6.0	51.2/8.3	45.7/10.0	35.7/34.9	47.2/36.8	29.7/37.4	34.0/26.2	36.7/33.8	40.4/23.7
Qwen3-VL-8B-Thinking	64	44.9/29.9	31.0/19.8	31.1/26.0	35.6/25.2	41.7/8.7	37.4/13.9	43.2/6.8	29.5/5.5	44.9/5.5	39.4/8.1	38.9/32.9	49.6/35.2	37.0/32.5	35.9/28.2	40.3/32.2	38.4/21.9
Qwen3-VL-4B-Thinking	128	34.3/22.8	31.7/12.7	32.3/22.8	32.8/19.5	52.4/10.7	49.8/10.7	40.1/4.6	50.8/3.5	**53.9**/12.2	49.4/8.3	25.8/24.2	34.5/24.6	25.4/27.8	27.4/18.1	28.3/23.7	36.8/17.2
Qwen3-VL-4B-Thinking	64	37.4/24.4	26.6/13.5	40.9/22.4	35.0/20.1	46.4/11.8	44.2/9.8	42.9/7.1	44.9/7.5	46.5/8.7	45.0/9.0	29.0/27.4	33.7/23.8	23.8/21.4	30.2/22.2	29.2/23.7	36.4/17.6
GLM-4.6V-FlashX	50	10.6/7.1	7.5/4.8	18.5/10.2	12.2/7.4	34.4/9.3	21.9/6.6	23.9/4.9	39.0/9.8	21.7/6.3	28.2/7.4	14.3/8.3	7.5/6.3	16.7/17.5	10.9/8.1	12.3/10.1	17.6/8.3
GLM-4.1V-Thinking-FlashX	50	1.2/0.0	0.0/0.0	0.0/0.0	0.4/0.0	14.6/3.8	3.8/1.6	4.2/1.6	14.6/4.3	1.2/1.2	7.7/2.5	4.0/3.2	1.6/0.8	2.4/1.2	4.0/1.6	3.0/1.7	3.7/1.4
Open-Source Models (Instruct)																	
Qwen3-VL-235B-A22B-Instruct	128	39.8/30.3	15.5/10.7	44.1/36.6	33.1/25.9	51.4/17.8	45.2/11.8	33.9/6.6	33.9/9.4	52.8/7.9	43.4/10.7	25.4/20.6	23.6/16.0	20.3/14.6	13.3/11.3	20.7/15.6	32.4/17.4
Qwen3-VL-235B-A22B-Instruct	64	46.5/38.2	24.6/13.1	38.2/34.3	36.4/28.5	51.4/16.8	42.8/13.5	42.6/8.7	40.2/10.2	52.4/11.9	45.9/12.2	31.7/26.2	34.8/24.0	26.8/21.1	14.9/11.7	27.1/20.8	36.5/20.5
Qwen3-VL-30B-A3B-Instruct	128	20.9/14.2	12.7/7.5	21.7/12.6	18.4/11.4	43.1/7.9	43.6/8.6	49.2/3.2	36.2/7.9	42.9/3.9	43.0/6.3	13.9/13.1	20.0/18.0	15.0/14.2	15.7/10.5	16.2/14.0	25.9/10.6
Qwen3-VL-30B-A3B-Instruct	64	28.7/21.3	13.9/8.7	26.4/14.6	23.0/14.9	44.8/8.5	41.8/10.4	41.1/2.8	39.0/6.3	40.1/4.0	41.4/6.4	15.9/12.7	25.2/17.6	15.0/14.6	13.7/10.5	17.5/13.9	27.3/11.7
Qwen3-VL-8B-Instruct	128	20.5/15.4	27.0/14.3	11.8/9.8	19.8/13.2	15.0/5.7	11.4/2.0	14.3/2.8	11.8/5.1	23.6/3.5	15.2/3.8	7.5/7.9	10.8/8.0	6.1/2.4	2.8/3.2	6.8/5.4	13.9/7.5
Qwen3-VL-8B-Instruct	64	25.2/20.5	29.8/14.7	17.7/13.8	24.2/16.3	20.2/6.7	7.4/2.4	16.6/3.9	11.4/4.3	28.3/5.1	16.8/4.5	4.8/5.6	10.4/8.8	4.9/3.7	6.9/5.2	6.7/5.8	15.9/8.9
Qwen3-VL-4B-Instruct	128	11.0/6.7	2.8/3.6	13.4/8.7	9.1/6.3	31.8/6.3	32.9/8.6	29.8/4.2	36.6/7.9	42.9/8.3	34.8/7.0	7.9/8.3	6.7/4.8	9.7/4.8	4.4/4.8	7.2/5.7	17.0/6.3
Qwen3-VL-4B-Instruct	64	6.3/3.9	4.8/2.8	17.3/4.7	9.5/3.8	39.7/5.1	38.0/8.2	35.5/3.4	40.6/7.9	39.0/8.7	38.6/6.6	11.9/11.1	7.9/6.7	10.9/5.6	8.9/2.8	9.9/6.6	19.3/5.7
Gemma4-E2B-Instruct	128	24.0/13.4	28.6/15.9	22.0/16.5	24.9/15.3	14.8/4.1	18.9/4.2	19.0/4.6	24.8/5.5	13.8/2.0	18.3/4.1	2.4/4.0	10.7/7.1	3.2/2.8	0.4/2.0	4.2/4.0	15.8/7.8

5 Analysis

We analyze relationships among benchmark tasks, the effects of model scale, frame budget, and reasoning mode, and the quality of evidence selection under incomplete observations.

5.1 Task Correlation Analysis

We compute task-wise Spearman correlations across the 24 evaluated model configurations and use a low-dimensional partial Spearman analysis to account for broad differences in model capability. For each task pair, we average the accuracies of the remaining 11 tasks to form a pair-specific proxy for overall performance, then rank-transform that proxy and the two target task accuracies. We regress each target rank on the proxy rank with an intercept and compute the Pearson correlation between the residuals. Figure 2 shows that the ordinary correlations are broadly positive and often strong. For example, Prediction and Planning reach $\rho = 0.93$, and most temporal and factual tasks also covary substantially.

After accounting for the pair-specific overall-performance proxy, the matrix becomes more heterogeneous, with both positive and negative coefficients. Prediction–Planning ($\rho = 0.70$) and Single-Fact–Dual-Fact Perturbation ($\rho = 0.72$) remain positively associated, whereas Counterfactual Fact Injection–Outcome Fact Modification is close to zero ($\rho = 0.09$). These descriptive estimates do not establish statistical significance or independence, particularly because configurations from the same model families are related. The contrast between the two panels suggests that aggregate model strength explains part of the broad cross-task agreement, while task-pair relationships vary after accounting for that proxy. This supports evaluation across temporal, spatial, and factual dimensions rather than reliance on one aggregate score.

5.2 Model Scale, Frame Budget, and Reasoning Mode

We compare four Qwen3-VL sizes (4B, 8B, 30B-A3B, and 235B-A22B), two frame budgets (64/128), and Instruct/Thinking variants. Scores follow task-then-dimension macro averaging; Ordering is excluded from evidence metrics. Figure 3 summarizes the 16 configurations.

Model scale. From 4B to 235B, overall accuracy increases by 11.55–11.81 percentage points across the four variant–frame settings. Evidence performance is less monotonic: at 128 frames, 30B Think-

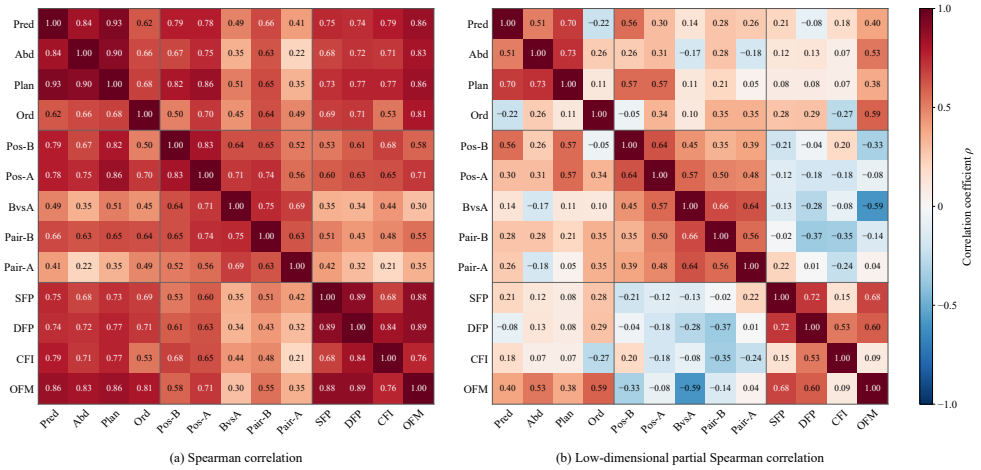


Figure 2: **Task-wise correlations across model configurations.** (a) Spearman correlations of task accuracy. (b) Low-dimensional partial Spearman correlations after residualizing each task pair on the rank of its mean accuracy across the remaining 11 tasks. Results are descriptive over 24 configurations; no significance tests are reported for panel (b).

ing exceeds 8B Thinking in accuracy (41.96% vs. 38.70%) but scores lower on J (17.53 vs. 23.72). Because this family mixes dense and MoE architectures, these are descriptive size comparisons rather than a compute-controlled scaling law.

Frame budget. With each annotated interval $[t_s, t_e]$ fixed, increasing the budget from 64 to 128 frames changes accuracy by only -0.24 to $+0.90$ percentage points across the eight matched configurations. Evidence scores also show mixed changes; for example, 30B Thinking gains 0.45 accuracy points while J drops by 3.91 points. Thus, denser sampling within the available interval provides limited gains over this tested range.

Thinking versus Instruct. At matched model sizes and frame budgets, Thinking improves D and J in all eight comparisons, with mean gains of 15.84 and 11.69 points, respectively, on the $\times 100$ scale. In contrast, its mean accuracy difference is -0.27 percentage points, with higher accuracy observed in only two comparisons. Thinking therefore exhibits stronger evidence discrimination and answer–evidence consistency without a consistent gain in answer accuracy. This distinction supports evaluating evidence alongside answers; it does not establish causal faithfulness or isolate the effect of additional inference-time computation.

5.3 Evidence Quality Analysis

Evidence quality depends on both selecting target observations and rejecting distractors. Table 2 shows that answer accuracy alone does not determine evidence quality. Among the reported 128-frame systems, Gemini-3.1-Pro-preview has the highest overall answer accuracy (57.3%), evidence discrimination score D (56.2), and joint score J (39.4). However, model rankings differ across metrics: Qwen3.7-Plus has higher answer accuracy than GPT-5.5 (54.7% vs. 52.1%) but lower D (48.4 vs. 56.0) and lower J (32.7 vs. 35.8). Correct answers are therefore not always accompanied by stronger evidence performance.

The selection rates illustrate the difficulty of rejecting distractors. Strong models generally select target evidence frequently, but they also retain a substantial fraction of distractors. For example, GPT-5.5 selects 82.1% of target evidence on average while selecting 30.5% of semantically related and 21.6% of irrelevant distractors(Appendix C, Tables 3–5). This asymmetry indicates that detecting question-relevant content is easier than determining which observations are actually diagnostic. Related distractors are typically selected more often than irrelevant ones, indicating that semantic overlap is a harder failure mode than obvious irrelevance.

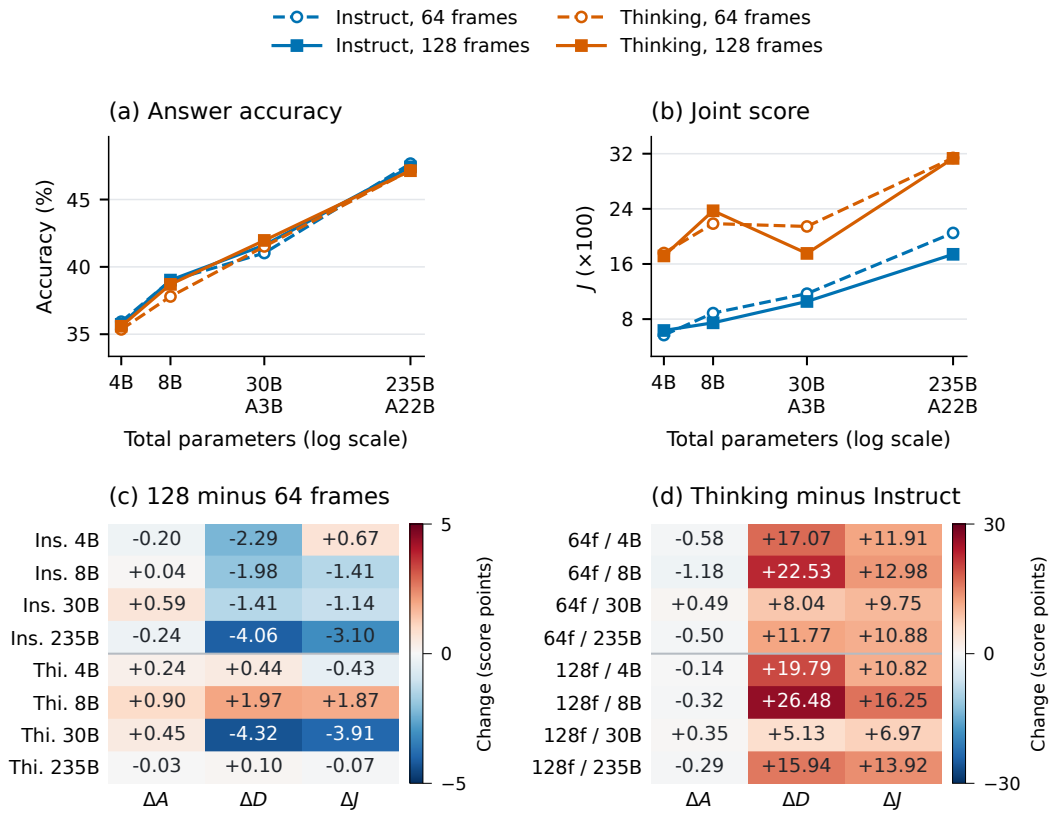


Figure 3: **Model scale, frame budget, and reasoning mode.** (a,b) Overall accuracy A and joint score J across model sizes. (c) Matched frame differences (128 minus 64). (d) Matched variant differences (Thinking minus Instruct). A is a percentage; D and J are reported on the $\times 100$ scale. A3B/A22B indicate activated parameters. Heatmaps use separate, zero-centered color scales and report differences rounded to two decimals. These are descriptive results without confidence intervals.

The discrepancy is especially pronounced for spatial incompleteness. GPT-5.5 attains a spatial D of 57.8 but a spatial J of only 20.1; similar gaps appear for several other models. Models can therefore distinguish target observations from distractors to some extent while failing to integrate those observations into correct spatial inference. Conversely, Qwen3-VL matched comparisons in Section 5.2 show that Thinking variants consistently improve D and J without a consistent gain in answer accuracy, demonstrating that evidence behavior can change independently of final-answer performance. Together, these findings motivate joint evaluation: target selection measures evidence recall, distractor rejection measures specificity, and J tests whether both align with a correct answer.

6 Conclusion

We introduce IVU-Bench, a benchmark of 13 temporal, spatial, and factual tasks across 2,032 question-answer pairs, with object-level evidence and distractor annotations. Current MLLMs lie far below human performance, particularly on spatial tasks. Cross-task correlations become heterogeneous after accounting for overall model performance; larger models boost answer accuracy, while extra frames yield limited gains. Thinking variants consistently improve evidence discrimination and joint answer–evidence performance without a consistent gain in answer accuracy. Models also select relevant observations while retaining plausible distractors or failing to use evidence for correct spatial inference. Progress demands persistent object-state tracking, integration of observations across time and viewpoints, and stronger links between answers and discriminative visual evidence.Future work will explore models that maintain persistent object states across time and viewpoints, while more explicitly linking their answers to discriminative visual evidence.

REFERENCES

Kirolos Ataallah, Eslam Mohamed Bakr, Mahmoud Ahmed, Chenhui Gou, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. InfiniBench: A Benchmark for Large Multi-Modal Models in Long-Form Movies and TV Shows. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19485–19512, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.984. URL https://aclanthology.org/2025.emnlp-main.984/.

Seunghwan Bang and Hwanjun Song. Reasoning over Video: Evaluating How MLLMs Extract, Integrate, and Reconstruct Spatiotemporal Evidence, April 2026. URL http://arxiv.org/abs/2603.13091. arXiv:2603.13091 [cs.CV].

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. EgoPlan-Bench: Benchmarking Multimodal Large Language Models for Human-Level Planning. *International Journal of Computer Vision*, 134(3):118, February 2026a. ISSN 1573-1405. doi: 10.1007/s11263-025-02676-0. URL https://doi.org/10.1007/s11263-025-02676-0.

Yuefei Chen, Jiang Liu, Xiaodong Lin, and Ruixiang Tang. Distilling Counterfactual Reasoning from Language to Vision: Causal Graph-Guided Post-Training for Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9165–9174, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Chen_Distilling_Counterfactual_Reasoning_from_Language_to_Vision_Causal_Graph-Guided_Post-Training_CVPRF_2026_paper.html.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-STaR: Benchmarking Video-LLMs on Video Spatio-Temporal Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9155–9164, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Cheng_V-STaR_Benchmarking_Video-LLMs_on_Video_Spatio-Temporal_Reasoning_CVPRF_2026_paper.html.

Aditya Chinchure, Sahithya Ravi, Raymond Ng, Vered Shwartz, Boyang Li, and Leonid Sigal. Black Swan: Abductive and Defeasible Video Reasoning in Unpredictable Events. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24201–24210, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chinchure_Black_Swan_Abductive_and_Defeasible_Video_Reasoning_in_Unpredictable_Events_CVPR_2025_paper.html.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding. In *International Conference on Learning Representations*, volume 2025, pp. 97959–98108, May 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. pp. 720–736, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Dima_Damen_Scaling_Egocentric_Vision_ECCV_2018_paper.html.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130(1):33–55, January 2022. ISSN 1573-1405. doi: 10.1007/s11263-021-01531-2. URL https://doi.org/10.1007/s11263-021-01531-2.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui

Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-MME-v2: Towards the Next Stage in Benchmarks for Comprehensive Video Understanding, April 2026. URL http://arxiv.org/abs/2604.05015. arXiv:2604.05015 [cs].

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. EgoTaskQA: Understanding Human Tasks in Egocentric Videos. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3343–3360, December 2022. doi: 10.52202/068431-0242. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/161c94a58ca25bafcaf47893e8233deb-Abstract-Datasets_and_Benchmarks.html.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking Multimodal Large Language Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, Seattle, WA, USA, June 2024a. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.01263. URL https://ieeexplore.ieee.org/document/10658180/.

Jiangtong Li, Li Niu, and Liqing Zhang. From Representation to Reasoning: Towards both Evidence and Commonsense Reasoning for Video Question-Answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21241–21250, June 2022. doi: 10.1109/CVPR52688.2022.02059. URL https://ieeexplore.ieee.org/abstract/document/9878800. ISSN: 2575-7075.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.html.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. VideoVista: A Versatile Benchmark for Video Understanding and Reasoning, June 2024c. URL http://arxiv.org/abs/2406.11303. arXiv:2406.11303 [cs.CV].

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.517. URL https://aclanthology.org/2024.findings-acl.517/.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244, December 2023. doi: 10.52202/075280-2004. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/90ce332aff156b910b002ce4e6880dec-Abstract-Datasets_and_Benchmarks.html.

Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Yunhai Tong, and Haodong Duan. VideoZeroBench: Probing the Limits of Video MLLMs with Spatio-Temporal Evidence Verification, April 2026. URL http://arxiv.org/abs/2604.01569. arXiv:2604.01569 [cs.CV].

Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, Cordelia Schmid, and Tobias Weyand. Minerva: Evaluating Complex Video Reasoning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23968–23978, October 2025. doi: 10.1109/ICCV51701.2025.02222. URL https://ieeexplore.ieee.org/abstract/document/11445745. ISSN: 2380-7504.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception Test: A Diagnostic Benchmark for Multimodal Video Models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42748–42761, December 2023. doi: 10.52202/075280-1852. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8540fba4abdc7f9f7a7b1cc6cd60e409-Abstract-Datasets_and_Benchmarks.html.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A Highly-Detailed Egocentric Video Dataset. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23901–23913, June 2025. doi: 10.1109/CVPR52734.2025.02226. URL https://ieeexplore.ieee.org/abstract/document/11094504. ISSN: 2575-7075.

Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. EgoPlan-Bench2: A Benchmark for Multimodal Large Language Model Planning in Real-World Scenarios. *International Journal of Computer Vision*, 134(5):222, April 2026. ISSN 1573-1405. doi: 10.1007/s11263-026-02826-y. URL https://doi.org/10.1007/s11263-026-02826-y.

Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. SOK-Bench: A Situated Video Reasoning Benchmark with Aligned Open-World Knowledge. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13384–13394, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_SOK-Bench_A_Situated_Video_Reasoning_Benchmark_with_Aligned_Open-World_Knowledge_CVPR_2024_paper.html.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857, December 2024. doi: 10.52202/079017-0907. URL https://proceedings.neurips.cc/paper_files/paper/2024/

hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781, June 2021. doi: 10.1109/CVPR46437.2021.00965. URL https://ieeexplore.ieee.org/abstract/document/9577425. ISSN: 2575-7075.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I Trust Your Answer? Visually Grounded Video Question Answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking Multi-task Long Video Understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, June 2025a. doi: 10.1109/CVPR52734.2025.01278. URL https://ieeexplore.ieee.org/abstract/document/11094860. ISSN: 2575-7075.

Qiji Zhou, YiFan Gong, Guangsheng Bao, Hongjie Qiu, Jinqiang Li, Xiangrong Zhu, Huajian Zhang, and Yue Zhang. Reasoning is All You Need for Video Generalization: A Counterfactual Benchmark with Sub-question Evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2939–2957, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.151. URL https://aclanthology.org/2025.findings-acl.151/.

Yiyang Zhou, Linjie Li, Shi Qiu, Zhengyuan Yang, Yuyang Zhao, Siwei Han, Yangfan He, Kangqi Li, Haonian Ji, Zihao Zhao, Haibo Tong, Lijuan Wang, and Huaxiu Yao. GLIMPSE: Do Large Vision-Language Models Truly Think With Videos or Just Glimpse at Them? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27842–27856, Suzhou, China, November 2025c. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1415. URL https://aclanthology.org/2025.emnlp-main.1415/.

A APPENDIX

This appendix includes supplementary information to support the main paper. Appendix A documents statements regarding AI tool usage, research ethics, and result reproducibility. Appendix B supplements comprehensive descriptive statistics characterizing the video collection, task-level time cost, and evidence distribution of IVU-Bench. Appendix C tabulates complete per-task numerical results for target-evidence selection, distractor-selection rates, and discrimination metrics, which support the evidence-oriented analysis in Section 5. Appendix D provides representative task samples across all 13 subtasks, demonstrating question formats, candidate options, and the annotation format for visual evidence.

A.1 AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript language and improve clarity. The technical content, analyses, figures and tables were independently produced and fully verified by the authors. We take responsibility for all final content of this work.

A.2 ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. All experiments in this study are conducted based on publicly available research datasets, including Ego4D, EPIC-KITCHENS-55, EPIC-KITCHENS-100, and HD-EPIC. These datasets are released for academic research purposes and are used in full compliance with their original licenses, usage terms, and official research guidelines.

We do not collect new data, involve human or animal participants, or conduct any real-world user studies. No personally identifiable information, sensitive personal attributes, or private content is newly extracted, inferred, or exploited beyond the publicly released content of the original datasets. We do not introduce additional privacy, security, or ethical risks through our model training, evaluation, or analysis procedures.

Throughout this work, we maintain transparent, responsible, and ethical usage of existing public resources to ensure research integrity.

A.3 REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results reported in this paper are reproducible.

B ADDITIONAL BENCHMARK STATISTICS

Figure 4 characterizes the composition of the video collection and the distribution of task demands. Most videos are between 2 and 8 minutes long (94 of 127, 74.0%), with the 5–8-minute bin accounting for the largest share (60 videos, 47.2%). Mean task time varies across task types: Temporal Prediction and Temporal Planning take approximately 150 s per example, compared with 299 s for Temporal Ordering and 248 s for Outcome Modification. Pair-B and Pair-A exhibit the largest mean evidence counts, at approximately 10.2–10.3 target, 4.9 related, and 4.2 irrelevant evidence items per example. The recorded evidence counts for Temporal Ordering are zero across all three categories. The normalized task-time ratio is highest for Temporal Ordering (0.73) and Outcome Modification (0.68), and lowest for Prediction and Planning (0.40). These values summarize descriptive patterns in the benchmark and should not be interpreted as inferential or performance comparisons.

C EVIDENCE SELECTION AND DISTRACTOR DISCRIMINATION

Evidence metric conventions and abbreviations. All entries are percentages. Table 3 reports the target evidence selection rate, namely how often a model selects the annotated evidence; higher values are better. Table 4 reports the related distractor selection rate, which measures selection of distractors that are semantically related to the target evidence; lower values are better. Table 5 reports the irrelevant distractor selection rate, which measures selection of semantically unrelated



Figure 4: **data-statistic**. (a) Mean task time per example (s). (b) Number of videos in each duration bin. (c) Mean counts of target, related, and irrelevant evidence per example. (d) Mean task time normalized by video duration. Pred., Abd., Plan., and Ord. denote temporal prediction, abduction, planning, and ordering; Pos-B/Pos-A denote object position before/after movement; B-vs-A denotes relative position before versus after movement; Pair-B/Pair-A denote pairwise object position before/after movement; SFP/DFP denote single-/dual-fact perturbation; CFI and OFM denote fact injection and outcome modification, respectively. These are descriptive statistics.

distractors; lower values are better. Tables 6 and 7 report discrimination scores against irrelevant and related distractors, respectively: d_i^U measures rejection of irrelevant distractors, while d_i^R measures rejection of related distractors; higher values indicate stronger discrimination. The `Ord.` column is unavailable because the supplied evidence annotations do not include ordering tasks. We use the following abbreviations throughout these tables: Pred./Abd./Plan./Ord. = Prediction/Abduction/Planning/Ordering; Pos-B/A = Object Position Before/After Movement; B-vs-A = Object Position Before vs. After Movement; Pair-B/A = Pairwise Position Before/After Movement; SFP/DFP = Single-/Dual-Fact Perturbation; CFI = Counterfactual Fact Injection; OFM = Outcome Fact Modification.

Table 3: Target evidence selection rate (%). The best and second-best model results are shown in bold and underlined, respectively. True evidence selection rate; higher values are better.

Model	Frames	Temporal Incompleteness					Spatial Incompleteness						Factual Incompleteness					Avg.
		Pred.	Abd.	Plan.	Ord.	Temp.	Pos-B	Pos-A	B-vs-A	Pair-B	Pair-A	Spa.	SFP	DFP	CFI	OFM	Fact.	
Commercial Models																		
GPT-5.5	128	89.0	79.5	85.8	–	84.8	82.7	80.2	89.8	78.7	78.7	82.0	80.3	81.1	82.5	73.6	79.4	82.1
Gemini-3.1-Pro-preview	128	95.1	87.5	90.0	–	90.9	**87.2**	**90.4**	**90.5**	82.4	**88.0**	**87.7**	87.5	95.0	88.5	**93.7**	91.2	89.9
Qwen3.7-Plus	128	81.9	85.0	84.3	–	83.7	80.2	70.4	74.3	81.1	78.7	77.0	75.6	85.0	75.4	74.4	77.6	79.4
Open-Source Models (Reasoning)																		
Qwen3.5-397B-A17B	128	88.2	85.8	86.6	–	86.9	72.8	74.3	82.3	**85.8**	77.8	78.6	78.7	81.7	80.0	81.6	80.5	82.0
Qwen3.5-27B	128	88.2	85.8	91.3	–	88.5	79.4	76.7	83.4	81.0	77.6	79.6	88.1	94.4	87.9	85.5	89.0	85.7
Qwen3-VL-235B-A22B-Thinking	128	80.3	76.4	84.3	–	80.3	66.0	75.5	83.1	76.4	78.6	75.9	66.9	73.0	60.8	77.6	69.6	75.3
Qwen3-VL-235B-A22B-Thinking	64	83.5	83.5	76.4	–	81.1	67.2	76.3	83.5	70.9	79.4	75.4	70.9	75.4	71.2	74.4	73.0	76.5
Qwen3-VL-30B-A3B-Thinking	128	52.0	59.8	34.6	–	48.8	55.3	49.8	57.1	64.6	74.6	60.3	40.2	54.0	32.0	53.6	44.9	51.3
Qwen3-VL-30B-A3B-Thinking	64	57.5	69.3	46.5	–	57.7	66.8	62.9	70.5	73.2	83.3	71.3	40.9	51.6	45.6	60.0	49.5	59.5
Qwen3-VL-8B-Thinking	128	75.6	78.0	67.7	–	73.8	70.7	74.7	75.1	79.4	84.9	77.0	65.4	82.5	64.0	68.3	70.0	73.6
Qwen3-VL-8B-Thinking	64	74.0	81.9	66.9	–	74.3	74.3	67.2	76.4	72.4	81.9	74.4	66.9	84.9	70.4	72.8	73.8	74.2
Qwen3-VL-4B-Thinking	128	74.8	88.2	71.7	–	78.2	78.0	77.1	70.9	81.1	85.8	78.6	73.2	89.0	68.3	83.2	78.4	78.4
Qwen3-VL-4B-Thinking	64	74.0	86.6	71.7	–	77.4	70.9	70.0	68.9	74.8	81.9	73.3	70.1	83.5	72.2	86.4	78.0	76.3
GLM-4.6V-FlashX	50	20.5	23.6	31.5	–	25.2	53.1	36.7	41.3	63.0	49.6	48.8	24.4	13.4	30.4	20.0	22.0	32.0
GLM-4.1V-Thinking-FlashX	50	2.4	0.0	0.0	–	0.8	21.7	5.1	5.1	19.7	3.1	10.9	4.7	1.6	3.2	4.0	3.4	5.0
Open-Source Models (Instruct)																		
Qwen3-VL-235B-A22B-Instruct	128	88.2	92.1	85.0	–	88.5	69.7	71.1	60.2	67.7	81.9	70.1	70.1	72.2	52.8	57.6	63.2	73.9
Qwen3-VL-235B-A22B-Instruct	64	88.2	95.3	80.3	–	87.9	71.3	69.2	68.9	69.3	79.4	71.6	69.3	81.0	66.4	59.2	69.0	76.2
Qwen3-VL-30B-A3B-Instruct	128	**96.1**	96.9	89.8	–	94.2	75.6	80.2	80.3	74.0	79.5	77.9	81.9	90.5	89.6	85.6	86.9	86.4
Qwen3-VL-30B-A3B-Instruct	64	93.7	94.5	89.8	–	92.7	76.3	76.3	74.4	75.6	79.4	76.4	81.9	90.5	85.6	81.6	84.9	84.6
Qwen3-VL-8B-Instruct	128	48.8	74.0	17.3	–	46.7	21.7	13.4	22.0	20.5	33.1	22.1	14.2	32.5	14.4	14.4	18.9	29.2
Qwen3-VL-8B-Instruct	64	40.2	79.5	25.2	–	48.3	30.7	10.7	26.4	25.2	43.3	27.3	12.6	27.8	10.4	25.6	19.1	31.5
Qwen3-VL-4B-Instruct	128	90.6	93.7	89.8	–	91.3	73.6	78.2	83.1	78.7	80.3	78.8	85.0	92.1	91.3	89.6	89.5	86.5
Qwen3-VL-4B-Instruct	64	92.1	**97.6**	**96.1**	–	**95.3**	83.5	86.2	87.0	81.1	80.3	83.6	**90.6**	**96.1**	**98.4**	92.8	**94.5**	**91.1**
Gemma4-E2B-Instruct	128	33.9	49.6	31.5	–	38.3	23.2	28.5	26.4	33.9	22.8	27.0	7.9	15.7	8.7	6.4	9.7	25.0

Table 4: Related distractor selection rate (%). The best and second-best model results are shown in bold and underlined, respectively. Related distractor selection rate; lower values are better.

Model	Frames	Temporal Incompleteness					Spatial Incompleteness						Factual Incompleteness					Avg.
		Pred.	Abd.	Plan.	Ord.	Temp.	Pos-B	Pos-A	B-vs-A	Pair-B	Pair-A	Spa.	SFP	DFP	CFI	OFM	Fact.	
Commercial Models																		
GPT-5.5	128	22.0	40.5	18.1	–	26.9	30.9	26.3	30.8	28.3	25.2	28.3	40.2	33.3	32.5	39.2	36.3	30.5
Gemini-3.1-Pro-preview	128	36.6	50.6	40.0	–	42.4	31.9	41.1	36.8	36.8	46.7	38.7	41.2	34.2	38.5	43.0	39.2	40.1
Qwen3.7-Plus	128	29.1	57.9	28.3	–	38.5	31.0	23.1	33.3	30.7	31.5	29.9	41.7	46.8	35.7	36.0	40.1	36.2
Open-Source Models (Reasoning)																		
Qwen3.5-397B-A17B	128	27.6	57.9	43.3	–	42.9	25.7	22.7	30.0	25.2	31.0	26.9	42.5	39.2	40.8	43.2	41.4	37.1
Qwen3.5-27B	128	41.7	70.6	37.0	–	49.8	29.7	28.7	35.7	34.1	32.8	32.2	51.6	46.8	46.0	49.2	48.4	43.5
Qwen3-VL-235B-A22B-Thinking	128	26.8	41.3	22.8	–	30.3	26.2	21.9	34.4	25.2	31.7	27.9	26.8	27.2	32.0	35.2	30.3	29.5
Qwen3-VL-235B-A22B-Thinking	64	25.2	42.1	29.9	–	32.4	28.5	27.5	34.0	29.1	38.1	31.4	26.8	33.6	32.0	32.8	31.3	31.7
Qwen3-VL-30B-A3B-Thinking	128	19.7	34.1	15.0	–	22.9	21.0	17.5	23.7	33.1	27.0	24.5	17.3	21.6	14.4	22.4	18.9	22.1
Qwen3-VL-30B-A3B-Thinking	64	18.9	34.9	19.7	–	24.5	24.2	27.1	34.8	39.4	31.7	31.4	27.6	29.6	14.4	25.6	24.3	26.7
Qwen3-VL-8B-Thinking	128	24.4	56.3	37.8	–	39.5	29.8	27.9	34.5	38.9	34.1	33.0	37.0	37.6	40.0	39.0	38.4	37.0
Qwen3-VL-8B-Thinking	64	31.5	55.6	34.6	–	40.6	35.3	35.1	39.1	46.5	37.8	38.8	38.6	37.6	32.0	38.4	36.6	38.7
Qwen3-VL-4B-Thinking	128	40.2	60.3	44.1	–	48.2	29.6	29.1	33.2	33.1	37.8	32.6	53.5	54.8	42.9	55.2	51.6	44.1
Qwen3-VL-4B-Thinking	64	38.6	58.7	37.8	–	45.0	28.1	29.1	28.1	32.3	32.3	30.0	45.7	50.0	47.6	56.0	49.8	41.6
GLM-4.6V-FlashX	50	11.0	15.9	12.6	–	13.2	23.3	14.3	19.8	24.4	32.3	22.8	11.0	4.8	17.6	11.2	11.1	15.7
GLM-4.1V-Thinking-FlashX	50	**0.8**	**0.0**	**0.0**	–	**0.3**	7.5	**0.8**	**1.6**	7.1	**1.6**	**3.7**	**0.8**	**0.0**	**1.6**	**0.0**	**0.6**	**1.5**
Open-Source Models (Instruct)																		
Qwen3-VL-235B-A22B-Instruct	128	53.5	81.0	45.7	–	60.1	21.3	29.5	31.6	33.1	36.2	30.3	45.7	49.6	33.6	44.8	43.4	44.6
Qwen3-VL-235B-A22B-Instruct	64	44.9	73.8	49.6	–	56.1	20.5	32.7	31.2	27.6	32.5	28.9	39.4	46.4	43.2	44.8	43.4	42.8
Qwen3-VL-30B-A3B-Instruct	128	76.4	86.5	74.0	–	79.0	36.0	37.1	35.6	37.8	40.9	37.5	70.1	73.6	75.2	72.0	72.7	63.1
Qwen3-VL-30B-A3B-Instruct	64	62.2	81.7	68.5	–	70.8	33.3	35.9	36.8	37.8	46.0	38.0	68.5	68.8	74.4	70.4	70.5	59.8
Qwen3-VL-8B-Instruct	128	33.1	53.2	6.3	–	30.8	7.5	2.4	11.1	**6.3**	8.7	7.2	7.9	24.8	9.6	12.8	13.8	17.3
Qwen3-VL-8B-Instruct	64	15.7	54.0	7.9	–	25.9	12.2	2.8	11.5	11.8	11.8	10.0	8.7	18.4	5.6	18.4	12.8	16.2
Qwen3-VL-4B-Instruct	128	78.7	92.1	74.8	–	81.9	46.6	48.2	57.3	44.9	33.9	46.2	83.5	86.5	81.0	85.6	84.1	70.7
Qwen3-VL-4B-Instruct	64	85.0	94.4	78.7	–	86.1	49.0	51.8	55.3	36.2	37.8	46.0	81.1	90.5	88.1	85.6	86.3	72.8
Gemma4-E2B-Instruct	128	11.8	23.8	11.8	–	15.8	**6.3**	8.8	6.3	10.2	11.8	8.7	6.3	6.3	5.6	6.4	6.2	10.2

Table 5: Irrelevant distractor selection rate (%). The best and second-best model results are shown in bold and underlined, respectively. Irrelevant distractor selection rate; lower values are better.

Model	Frames	Temporal Incompleteness					Spatial Incompleteness						Factual Incompleteness					Avg.
		Pred.	Abd.	Plan.	Ord.	Temp.	Pos-B	Pos-A	B-vs-A	Pair-B	Pair-A	Spa.	SFP	DFP	CFI	OFM	Fact.	
Commercial Models																		
GPT-5.5	128	16.5	37.0	15.0	–	22.8	18.1	18.6	17.4	22.0	25.2	20.3	26.2	23.0	17.7	20.2	21.8	21.6
Gemini-3.1-Pro-preview	128	24.4	36.2	22.5	–	27.7	22.4	34.0	23.3	36.8	36.0	30.5	15.0	32.9	18.4	28.2	23.6	27.3
Qwen3.7-Plus	128	24.4	40.9	22.8	–	29.4	17.8	19.0	25.0	28.3	28.3	23.7	26.2	23.0	23.4	25.0	24.4	25.8
Open-Source Models (Reasoning)																		
Qwen3.5-397B-A17B	128	24.4	37.0	25.2	–	28.9	16.1	13.8	23.3	28.3	25.4	21.4	26.2	27.2	23.6	31.5	27.1	25.8
Qwen3.5-27B	128	33.9	52.8	34.6	–	40.4	21.3	19.8	21.0	23.0	24.8	22.0	32.8	29.0	25.4	35.8	30.8	31.1
Qwen3-VL-235B-A22B-Thinking	128	20.5	25.2	21.3	–	22.3	20.2	20.9	26.1	24.4	21.4	22.6	20.6	32.0	30.1	31.5	28.5	24.5
Qwen3-VL-235B-A22B-Thinking	64	23.6	28.3	21.3	–	24.4	15.4	19.7	26.1	22.0	28.6	22.4	28.6	24.8	23.6	31.5	27.1	24.6
Qwen3-VL-30B-A3B-Thinking	128	11.8	24.4	13.4	–	16.5	18.6	19.3	19.4	26.0	25.4	21.7	14.3	20.0	15.4	17.7	16.9	18.4
Qwen3-VL-30B-A3B-Thinking	64	17.3	26.0	21.3	–	21.5	19.4	18.6	23.7	25.2	32.5	23.9	10.3	16.8	19.5	29.0	18.9	21.4
Qwen3-VL-8B-Thinking	128	27.6	35.4	26.8	–	29.9	28.4	20.1	32.1	32.5	33.3	29.3	21.4	32.8	28.5	31.1	28.5	29.2
Qwen3-VL-8B-Thinking	64	26.8	45.7	37.0	–	36.5	30.0	23.7	26.9	39.4	36.2	31.2	19.0	32.8	35.0	37.1	31.0	32.9
Qwen3-VL-4B-Thinking	128	40.9	52.0	34.6	–	42.5	21.3	25.3	28.1	27.6	26.0	25.6	41.3	54.0	42.7	56.5	48.6	38.9
Qwen3-VL-4B-Thinking	64	34.6	60.6	23.6	–	39.6	20.5	22.1	23.7	27.6	38.6	26.5	36.5	49.2	48.4	58.1	48.0	38.1
GLM-4.6V-FlashX	50	8.7	16.5	13.4	–	12.9	14.6	15.4	15.0	23.6	23.6	18.4	9.5	7.1	10.6	7.3	8.6	13.3
GLM-4.1V-Thinking-FlashX	50	**1.6**	**0.0**	**0.0**	–	**0.5**	6.7	**1.2**	**0.4**	**3.1**	**2.4**	**2.8**	**0.8**	**0.0**	**0.0**	**0.0**	**0.2**	**1.2**
Open-Source Models (Instruct)																		
Qwen3-VL-235B-A22B-Instruct	128	43.3	71.7	36.2	–	50.4	15.0	22.1	20.6	34.6	22.0	22.9	43.7	48.8	32.5	44.4	42.3	38.5
Qwen3-VL-235B-A22B-Instruct	64	38.6	66.9	34.6	–	46.7	18.9	19.7	20.9	30.7	21.4	22.3	35.7	47.2	35.8	43.5	40.6	36.5
Qwen3-VL-30B-A3B-Instruct	128	74.0	81.9	62.2	–	72.7	29.1	35.6	27.3	37.8	32.3	32.4	65.9	67.2	73.2	67.7	68.5	57.9
Qwen3-VL-30B-A3B-Instruct	64	67.7	78.7	58.3	–	68.2	29.3	32.8	29.6	35.4	32.5	31.9	63.5	61.6	66.7	66.9	64.7	55.0
Qwen3-VL-8B-Instruct	128	23.6	41.7	4.7	–	23.4	**5.9**	2.0	5.1	11.0	10.2	6.9	5.6	19.2	7.3	10.5	10.6	13.6
Qwen3-VL-8B-Instruct	64	14.2	44.9	7.1	–	22.0	9.1	3.9	8.7	15.7	18.1	11.1	7.1	16.8	5.7	19.4	12.2	15.1
Qwen3-VL-4B-Instruct	128	80.3	89.8	78.0	–	82.7	37.0	43.5	49.4	39.4	40.9	42.0	70.6	84.1	82.3	84.7	80.4	68.4
Qwen3-VL-4B-Instruct	64	86.6	91.3	78.7	–	85.6	38.6	45.5	47.8	44.9	44.9	44.3	76.2	85.7	87.1	82.3	82.8	70.9
Gemma4-E2B-Instruct	128	7.9	17.3	7.1	–	10.8	10.6	9.5	8.7	7.9	6.3	8.6	4.8	4.0	5.6	5.6	5.0	8.1

Table 6: Discrimination against irrelevant distractors (d_i^U, %). The best and second-best model results are shown in bold and underlined, respectively. Higher values indicate stronger rejection of irrelevant distractors.

Model	Frames	Temporal Incompleteness					Spatial Incompleteness						Factual Incompleteness					Avg.
		Pred.	Abd.	Plan.	Ord.	Temp.	Pos-B	Pos-A	B-vs-A	Pair-B	Pair-A	Spa.	SFP	DFP	CFI	OFM	Fact.	
Commercial Models																		
GPT-5.5	128	**72.4**	42.9	**70.9**	–	62.1	**64.8**	**61.8**	**72.2**	56.7	53.5	**61.8**	54.0	57.9	64.5	53.2	57.4	60.4
Gemini-3.1-Pro-preview	128	70.7	50.6	67.5	–	**63.0**	**64.8**	56.4	67.0	45.6	52.0	57.1	**72.5**	62.0	**69.7**	**65.4**	**67.4**	**62.5**
Qwen3.7-Plus	128	57.5	44.4	61.4	–	54.4	62.3	51.0	49.0	52.8	50.4	53.1	49.2	61.9	52.4	49.2	53.2	53.6
Open-Source Models (Reasoning)																		
Qwen3.5-397B-A17B	128	63.8	48.4	61.4	–	57.9	56.5	60.5	58.7	57.5	52.4	57.1	52.4	54.4	56.1	50.8	53.4	56.1
Qwen3.5-27B	128	54.3	33.3	56.7	–	48.1	58.0	57.0	62.2	**57.9**	52.8	57.6	55.2	**65.3**	62.3	49.6	58.1	54.6
Qwen3-VL-235B-A22B-Thinking	128	59.8	50.8	63.0	–	57.9	45.6	54.6	56.8	52.0	57.1	53.2	46.8	40.8	30.9	46.8	41.3	50.8
Qwen3-VL-235B-A22B-Thinking	64	59.8	**54.8**	55.1	–	56.6	51.6	56.2	57.2	48.8	50.8	52.9	42.1	50.4	48.0	42.7	45.8	51.8
Qwen3-VL-30B-A3B-Thinking	128	40.2	34.9	21.3	–	32.1	36.5	30.7	38.1	38.6	49.2	38.6	25.4	33.6	17.1	35.5	27.9	32.9
Qwen3-VL-30B-A3B-Thinking	64	40.2	42.9	25.2	–	36.1	47.2	43.9	46.9	48.0	50.8	47.3	30.2	34.4	26.0	31.5	30.5	38.0
Qwen3-VL-8B-Thinking	128	48.0	42.1	40.9	–	43.7	42.1	54.6	42.6	46.8	51.6	47.5	43.7	49.6	35.0	37.7	41.5	44.2
Qwen3-VL-8B-Thinking	64	47.2	35.7	29.9	–	37.6	44.5	43.0	49.2	33.1	45.7	43.1	48.4	52.0	35.0	36.3	42.9	41.2
Qwen3-VL-4B-Thinking	128	33.9	35.7	37.0	–	35.5	56.5	51.8	42.5	53.5	**59.8**	52.8	31.7	34.9	25.0	26.6	29.6	39.3
Qwen3-VL-4B-Thinking	64	39.4	25.4	48.0	–	37.6	50.2	47.8	44.8	47.2	43.3	46.7	33.3	34.1	23.4	29.0	30.0	38.1
GLM-4.6V-FlashX	50	11.8	7.1	18.1	–	12.4	38.7	21.5	26.1	39.4	26.0	30.3	15.1	6.3	20.3	12.9	13.7	18.8
GLM-4.1V-Thinking-FlashX	50	0.8	0.0	0.0	–	0.3	15.0	3.6	4.8	16.5	0.8	8.1	4.0	1.6	3.2	4.0	3.2	3.9
Open-Source Models (Instruct)																		
Qwen3-VL-235B-A22B-Instruct	128	44.9	19.8	48.8	–	37.8	54.5	49.0	39.2	33.1	**59.8**	47.1	26.2	24.0	21.1	13.7	21.3	35.4
Qwen3-VL-235B-A22B-Instruct	64	49.6	27.8	45.7	–	41.0	52.2	49.4	47.6	38.6	57.9	49.2	33.3	34.4	30.9	15.3	28.5	39.6
Qwen3-VL-30B-A3B-Instruct	128	22.0	15.1	27.6	–	21.6	46.6	44.2	53.2	36.2	47.2	45.5	15.9	23.2	16.3	17.7	18.3	28.4
Qwen3-VL-30B-A3B-Instruct	64	26.0	15.1	31.5	–	24.2	46.8	43.4	44.5	40.2	46.8	44.3	18.3	28.8	18.7	15.3	20.3	29.6
Qwen3-VL-8B-Instruct	128	25.2	32.5	12.6	–	23.4	15.8	11.6	17.0	9.4	22.8	15.3	8.7	13.6	7.3	4.0	8.4	15.7
Qwen3-VL-8B-Instruct	64	26.0	34.1	18.1	–	26.1	21.7	6.8	17.8	9.4	25.2	16.2	5.6	11.2	4.9	6.5	7.0	16.4
Qwen3-VL-4B-Instruct	128	10.2	4.0	11.8	–	8.7	36.8	35.4	33.8	39.4	39.4	36.9	14.3	7.9	8.9	4.8	9.0	18.2
Qwen3-VL-4B-Instruct	64	5.5	6.3	17.3	–	9.7	45.1	41.4	39.3	36.2	35.4	39.5	14.3	10.3	11.3	10.5	11.6	20.3
Gemma4-E2B-Instruct	128	26.0	31.7	24.4	–	27.4	12.7	18.7	17.9	26.0	16.5	18.4	3.2	11.9	3.2	0.8	4.8	16.8

Table 7: Discrimination against related distractors (d_i^R, %). The best and second-best model results are shown in bold and underlined, respectively. Higher values indicate stronger rejection of related distractors.

Model	Frames	Temporal Incompleteness					Spatial Incompleteness						Factual Incompleteness					Avg.
		Pred.	Abd.	Plan.	Ord.	Temp.	Pos-B	Pos-A	B-vs-A	Pair-B	Pair-A	Spa.	SFP	DFP	CFI	OFM	Fact.	
Commercial Models																		
GPT-5.5	128	**66.9**	38.9	**67.7**	–	**57.8**	52.1	**53.8**	**59.1**	50.4	**53.5**	**53.8**	40.5	47.6	**50.0**	34.7	43.2	**51.6**
Gemini-3.1-Pro-preview	128	58.5	36.7	50.0	–	48.4	**55.2**	49.2	53.3	45.6	41.3	48.9	**46.2**	**60.8**	**50.0**	**51.3**	**52.1**	49.8
Qwen3.7-Plus	128	52.8	27.0	55.9	–	45.2	49.2	47.0	41.0	50.4	47.2	47.0	33.3	38.1	39.5	38.7	37.4	43.2
Open-Source Models (Reasoning)																		
Qwen3.5-397B-A17B	128	60.6	27.8	43.3	–	43.9	47.1	51.4	52.4	**60.6**	46.8	51.7	36.5	42.4	39.0	38.7	39.2	44.9
Qwen3.5-27B	128	46.5	15.1	54.3	–	38.6	49.6	47.8	47.5	46.8	44.8	47.3	36.8	47.6	41.8	36.6	40.7	42.2
Qwen3-VL-235B-A22B-Thinking	128	53.5	34.9	61.4	–	50.0	39.7	53.4	48.4	51.2	46.8	47.9	41.3	45.6	28.5	43.5	39.7	45.9
Qwen3-VL-235B-A22B-Thinking	64	58.3	**41.3**	46.5	–	48.7	38.5	48.6	49.2	41.7	41.3	43.9	43.7	41.6	39.0	41.9	41.6	44.7
Qwen3-VL-30B-A3B-Thinking	128	32.3	25.4	19.7	–	25.8	34.1	32.3	33.7	31.5	47.6	35.8	22.2	32.0	17.9	30.6	25.7	29.1
Qwen3-VL-30B-A3B-Thinking	64	38.6	34.1	26.8	–	33.2	42.4	35.5	35.7	33.9	51.6	39.8	12.7	21.6	30.9	34.7	25.0	32.6
Qwen3-VL-8B-Thinking	128	51.2	21.4	29.9	–	34.2	40.9	46.6	40.2	40.5	50.8	43.8	27.8	44.8	24.4	30.3	31.8	36.6
Qwen3-VL-8B-Thinking	64	42.5	26.2	32.3	–	33.7	38.9	31.9	37.3	26.0	44.1	35.6	29.4	47.2	39.0	35.5	37.8	35.7
Qwen3-VL-4B-Thinking	128	34.6	27.8	27.6	–	30.0	48.2	47.8	37.7	48.0	48.0	46.0	19.8	34.1	25.8	28.2	27.0	34.3
Qwen3-VL-4B-Thinking	64	35.4	27.8	33.9	–	32.4	42.7	40.6	40.9	42.5	49.6	43.3	24.6	33.3	24.2	31.5	28.4	34.7
GLM-4.6V-FlashX	50	9.4	7.9	18.9	–	12.1	30.0	22.3	21.8	38.6	17.3	26.0	13.5	8.7	13.0	8.9	11.0	16.4
GLM-4.1V-Thinking-FlashX	50	1.6	0.0	0.0	–	0.5	14.2	4.0	3.6	12.6	1.6	7.2	4.0	1.6	1.6	4.0	2.8	3.5
Open-Source Models (Instruct)																		
Qwen3-VL-235B-A22B-Instruct	128	34.6	11.1	39.4	–	28.4	48.2	41.4	28.5	34.6	45.7	39.7	24.6	23.2	19.5	12.9	20.1	29.4
Qwen3-VL-235B-A22B-Instruct	64	43.3	21.4	30.7	–	31.8	50.6	36.3	37.7	41.7	46.8	42.6	30.2	35.2	22.8	14.5	25.7	33.4
Qwen3-VL-30B-A3B-Instruct	128	19.7	10.3	15.7	–	15.3	39.5	43.0	45.2	36.2	38.6	40.5	11.9	16.8	13.8	13.7	14.1	23.3
Qwen3-VL-30B-A3B-Instruct	64	31.5	12.7	21.3	–	21.8	42.8	40.2	37.7	37.8	33.3	38.4	13.5	21.6	11.4	12.1	14.6	24.9
Qwen3-VL-8B-Instruct	128	15.7	21.4	11.0	–	16.1	14.2	11.2	11.5	14.2	24.4	15.1	6.3	8.0	4.9	1.6	5.2	12.1
Qwen3-VL-8B-Instruct	64	24.4	25.4	17.3	–	22.4	18.6	8.0	15.4	13.4	31.5	17.4	4.0	9.6	4.9	7.3	6.4	15.4
Qwen3-VL-4B-Instruct	128	11.8	1.6	15.0	–	9.5	26.9	30.3	25.8	33.9	46.5	32.7	1.6	5.6	10.5	4.0	5.4	15.8
Qwen3-VL-4B-Instruct	64	7.1	3.2	17.3	–	9.2	34.4	34.7	31.8	44.9	42.5	37.6	9.5	5.6	10.5	7.3	8.2	18.3
Gemma4-E2B-Instruct	128	22.0	25.4	19.7	–	22.4	17.0	19.1	20.2	23.6	11.0	18.2	1.6	9.5	3.2	0.0	3.6	14.7

D BENCHMARK TASK EXAMPLES

In this section, we present two videos, each with 16 question samples. Green boxes denote positive evidence that supports the correct answer; yellow boxes mark semantically related distractor evidence which does not support the correct answer; red boxes represent semantically irrelevant distractor evidence. These examples provide a concrete view of the benchmark's input–output structure. Each item pairs sampled egocentric video frames with a task-specific question and seven candidate answers. Answering these questions requires interpreting visual evidence within its temporal and scene context, rather than relying solely on object recognition in an isolated frame. The examples thus clarify how video observations are translated into structured reasoning tasks. The frames fed into the model are not overlaid with bounding boxes; instead, the model receives bounding-box coordinates paired with the corresponding image indices. A reference format is shown below:

```
- evidence_id=0; input_frame_index=ids1; bbox_normalized=
                                    (x=x1, y=y1, width=w1, height=h1)
- evidence_id=1; input_frame_index=ids2; bbox_normalized=
                                    (x=x2, y=y2, width=w2, height=h2)
- evidence_id=2; input_frame_index=ids3; bbox_normalized=
                                    (x=x3, y=y3, width=w3, height=h3)
```

Below we present examples annotated with the three types of bounding boxes for demonstration.

Temporal Prediction

Based on the currently observed video, what is the most likely next action?



Answer
"A": "Take the plastic sealed bag from the drawer, place the grilled fish into the bag, then seal the bag and place the tomato salad next to it",
"B": "Take the plastic sealed bag from the drawer, put it in the grilled fish fillet and tomato salad, and then place the bag on the countertop",
"C": "Put the grilled fish and tomato salad in the plastic bag, take the plastic bag from the drawer, then seal the bag and put it in the refrigerator",
"D": "Take the plastic sealed bag from the drawer, put in the grilled fish fillet and tomato salad, then seal the bag and put it in the refrigerator",
"E": "Take the plastic sealed bag from the drawer, put in the grilled fish fillet and tomato salad, and then open the bag for later use",
"F": "Take out a glass, pour in juice, and drink the juice",
"G": "Take the plastic sealed bag from the drawer, put it in the grilled fish fillet and tomato salad, and throw it in the trash can",
"H": "Take out the plastic bag from the drawer, take out the bread slices, and apply jam"

Temporal Abduction

A process is missing between the two video segments. Determine what most likely happened in between.



Answer
"A": "Teared off a long roll of paper, cleaned the cutting board with a tissue, and threw the paper into the trash can",
"B": "Brush the cutting board with a brush, tear off a long roll of paper to absorb water from the sink, and clean the kitchen utensils and frying pan after unclogging",
"C": "Cut carrots, stir egg mixture, and bake small pancakes",
"D": "Pick up the seasoning from the table, put two bottles of seasoning into the cabinet, and close the cabinet door",
"E": "Use your hands to clear the water channel of the sink, tear off a long roll of paper to suck up the water in the sink, and clean the kitchen utensils and frying pans after dredging",
"F": "Tear off a long roll of paper to suck water from the sink, use your hands to clear the water channel of the sink, and then clean the kitchen utensils and frying pan after clearing it",
"G": "Take the plastic sealed bag from the drawer and put it in the grilled fish fillet and tomato salad",
"H": "Use your hands to unclog the water channel in the sink, tear off a cloth to suck up the water in the sink, and clean the kitchen utensils and frying pans after unclogging"

Temporal Planning

To complete the task, what action should be performed next?



Answer
"A": "Take the plastic sealed bag from the drawer, put in the grilled fish fillet and tomato salad, and then open the bag for later use",
"B": "Put the grilled fish and tomato salad in the plastic bag, take the plastic bag from the drawer, then seal the bag and put it in the refrigerator",
"C": "Take the plastic sealed bag from the drawer, place the grilled fish into the bag, then seal the bag and place the tomato salad next to it",
"D": "Take the plastic sealed bag from the drawer, put it in the grilled fish fillet and tomato salad, and throw it in the trash can",
"E": "Take the plastic sealed bag from the drawer, put it in the grilled fish fillet and tomato salad, and then place the bag on the countertop",
"F": "Take out the plastic bag from the drawer, take out the bread slices, and apply jam",
"G": "Take out a glass, pour in juice, and drink the juice",
"H": "Take the plastic sealed bag from the drawer, put in the grilled fish fillet and tomato salad, then seal the bag and put it in the refrigerator"

Temporal Ordering

Which sequence is correctly ordered according to the temporal order?



Answer
"A": "2->5->3->4->1",
"B": "1->4->3->2->5",
"C": "3->2->1->4->5",
"D": "4->5->1->3->2",
"E": "5->2->1->4->3",
"F": "3->2->5->4->1",
"G": "1->2->4->3->5",
"H": "5->4->3->2->1"

Spatial Object Position Before Movement

Taking the center of [faucet] as the spatial center, the direction from the center of [faucet] to the center of [electric kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. Before [Sealed bags for storing food] is moved, where is [Sealed bags for storing food] located relative to [faucet]?



Answer
"A": "The target object is located front-right of the reference object and is lower than the reference object.",
"B": "The target object is located front-left of the reference object and is at the same height as the reference object.",
"C": "The target object is located in front of the reference object and is higher than the reference object.",
"D": "The target object is located behind the reference object and is higher than the reference object.",
"E": "The target object is located back-right of the reference object and is higher than the reference object.",
"F": "The target object is located on the left side of the reference object and is lower than the reference object.",
"G": "The target object is located on the right side of the reference object and is at the same height as the reference object.",
"H": "The target object is located back-left of the reference object and is at the same height as the reference object."

Spatial Object Position Before Movement

Taking the center of [faucet] as the spatial center, the direction from the center of [faucet] to the center of [electric kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. Before [Small iron pot] is moved, where is [Small iron pot] located relative to [faucet]?



Answer
"A": "The target object is located front-left of the reference object and is lower than the reference object.",
"B": "The target object is located behind the reference object and is at the same height as the reference object.",
"C": "The target object is located on the left side of the reference object and is higher than the reference object.",
"D": "The target object is located back-right of the reference object and is lower than the reference object.",
"E": "The target object is located on the right side of the reference object and is lower than the reference object.",
"F": "The target object is located front-right of the reference object and is at the same height as the reference object.",
"G": "The target object is located front-left of the reference object and is at the same height as the reference object.",
"H": "The target object is located back-left of the reference object and is higher than the reference object."

Spatial Object Position After Movement

Taking the center of [faucet] as the spatial center, the direction from the center of [faucet] to the center of [electric kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. After [Sealed bags for storing food] is moved, where is [Sealed bags for storing food] located relative to [faucet]?



Answer
"A": "The target object is located on the right side of the reference object and is at the same height as the reference object.",
"B": "The target object is located back-left of the reference object and is lower than the reference object.",
"C": "The target object is located front-right of the reference object and is lower than the reference object.",
"D": "The target object is located behind the reference object and is higher than the reference object.",
"E": "The target object is located in front of the reference object and is higher than the reference object.",
"F": "The target object is located front-left of the reference object and is at the same height as the reference object.",
"G": "The target object is located back-left of the reference object and is higher than the reference object.",
"H": "The target object is located on the left side of the reference object and is lower than the reference object."

Spatial Object Position After Movement

Taking the center of [faucet] as the spatial center, the direction from the center of [faucet] to the center of [electric kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. After [Small iron pot] is moved, where is [Small iron pot] located relative to [faucet]?



Answer
"A": "The target object is located behind the reference object and is at the same height as the reference object.",
"B": "The target object is located back-right of the reference object and is at the same height as the reference object.",
"C": "The target object is located in front of the reference object and is at the same height as the reference object.",
"D": "The target object is located front-right of the reference object and is higher than the reference object.",
"E": "The target object is located back-left of the reference object and is lower than the reference object.",
"F": "The target object is located on the left side of the reference object and is higher than the reference object.",
"G": "The target object is located front-left of the reference object and is at the same height as the reference object.",
"H": "The target object is located on the right side of the reference object and is lower than the reference object."

Spatial Relative Position Before and After Object Movement

Taking the center of [faucet] as the spatial center, the direction from the center of [faucet] to the center of [electric kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. Where is [Sealed bags for storing food] before it is moved relative to [Sealed bags for storing food] after it is moved?



Answer
"A": "The target object is located in front of the reference object and is higher than the reference object.",
"B": "The target object is located front-right of the reference object and is lower than the reference object.",
"C": "The target object is located behind the reference object and is higher than the reference object.",
"D": "The target object is located on the right side of the reference object and is at the same height as the reference object.",
"E": "The target object is located on the left side of the reference object and is lower than the reference object.",
"F": "The target object is located front-left of the reference object and is higher than the reference object.",
"G": "The target object is located back-left of the reference object and is at the same height as the reference object.",
"H": "The target object is located back-right of the reference object and is higher than the reference object."

Spatial Relative Position Before and After Object Movement

Taking the center of [faucet] as the spatial center, the direction from the center of [faucet] to the center of [electric kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. Where is [Small iron pot] before it is moved relative to [Small iron pot] after it is moved?



Answer
"A": "The target object is located front-right of the reference object and is lower than the reference object.",
"B": "The target object is located back-right of the reference object and is at the same height as the reference object.",
"C": "The target object is located on the left side of the reference object and is higher than the reference object.",
"D": "The target object is located front-left of the reference object and is at the same height as the reference object.",
"E": "The target object is located on the right side of the reference object and is at the same height as the reference object.",
"F": "The target object is located behind the reference object and is lower than the reference object.",
"G": "The target object is located in front of the reference object and is at the same height as the reference object.",
"H": "The target object is located back-left of the reference object and is higher than the reference object."

Spatial Relative Position of Two Objects Before Movement

Taking the center of [faucet] as the spatial center, the direction from the center of [faucet] to the center of [electric kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. Where is [Sealed bags for storing food] before it is moved relative to [Small iron pot] before it is moved?



Answer
"A": "The target object is located front-left of the reference object and is at the same height as the reference object.",
"B": "The target object is located front-right of the reference object and is lower than the reference object.",
"C": "The target object is located behind the reference object and is higher than the reference object.",
"D": "The target object is located on the left side of the reference object and is lower than the reference object.",
"E": "The target object is located back-right of the reference object and is higher than the reference object.",
"F": "The target object is located back-left of the reference object and is at the same height as the reference object.",
"G": "The target object is located in front of the reference object and is higher than the reference object.",
"H": "The target object is located on the right side of the reference object and is at the same height as the reference object."

Spatial Relative Position of Two Objects Before Movement

Taking the center of [faucet] as the spatial center, the direction from the center of [faucet] to the center of [electric kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. Where is [Sealed bags for storing food] after it is moved relative to [Small iron pot] after it is moved?



Answer
"A": "The target object is located on the right side of the reference object and is at the same height as the reference object.",
"B": "The target object is located front-right of the reference object and is higher than the reference object.",
"C": "The target object is located on the left side of the reference object and is lower than the reference object.",
"D": "The target object is located behind the reference object and is higher than the reference object.",
"E": "The target object is located back-right of the reference object and is lower than the reference object.",
"F": "The target object is located front-left of the reference object and is at the same height as the reference object.",
"G": "The target object is located in front of the reference object and is higher than the reference object.",
"H": "The target object is located back-left of the reference object and is lower than the reference object."

Factual Single Fact Perturbation

If during [Prepare food], [There are no plastic bags inside the drawer], in the modified factual scenario, what would the camera wearer most likely do next from the moment of the factual change?



Answer
"A": "Tap the drawer and try shaking the plastic bag out of the gap, then put it into the grilled fish fillet and tomato salad",
"B": "Wrap the grilled fish fillet and tomato salad with plastic wrap, open the refrigerator, and place them in the refrigerator compartment",
"C": "Open the refrigerator, pick up the plate containing food, and directly throw the pot into the refrigerator to freeze",
"D": "Repeatedly open and close the drawer, keep flipping inside to find the plastic bag, and then put it into the grilled fish fillet and tomato salad",
"E": "Open the refrigerator, pick up the plate containing food, and directly put the grilled fish fillet and tomato salad into the refrigerator for freezing",
"F": "Go to the storage cabinet to find an alternative fresh-keeping bag, take out the fresh-keeping bag, and put it into the grilled fish fillet and tomato salad",
"G": "Go to the storage cabinet to find an alternative fresh-keeping bag, take out the fresh-keeping bag, and put it into the toast",
"H": "Take a plastic sealed bag from the drawer, put it in grilled fish, tomato salad, and then seal the bag with plastic"

Factual Dual Fact Perturbation

If during [Prepare food], [There are no plastic bags inside the drawer], and [The freezer layer of the refrigerator is malfunctioning and unable to freeze], in the modified factual scenario, what would the camera wearer most likely do next from the moment of the factual changes?



Answer
"A": "Find a fresh-keeping box to store grilled fish fillet and tomato salad. Put them in the refrigerator compartment first, and then cover the box tightly",
"B": "Find a fresh-keeping box to store grilled fish fillet and tomato salad, without covering the box, and place them in the refrigerator compartment",
"C": "Find a fresh-keeping box to store grilled fish fillet, leave tomato salad in place, cover the box tightly, and place it in the refrigerator compartment",
"D": "Store grilled fish fillet and tomato salad in a fresh-keeping bag, seal the bag tightly, and place it in the refrigerator compartment",
"E": "Find a fresh-keeping box to store grilled fish fillet and tomato salad, cover the box tightly, and place it in the refrigerator compartment",
"F": "Find a fresh-keeping box to store grilled fish fillet and tomato salad, cover the box tightly, and place it on the countertop next to the refrigerator compartment",
"G": "Find a fresh-keeping box to store grilled fish fillet and tomato salad, cover the box tightly, put it in the refrigerator compartment, and take it out for later use",
"H": "Find a fresh-keeping box to store grilled fish fillet and tomato salad, cover the box tightly, and place it in the freezer compartment of the refrigerator"

Factual Fact Injection

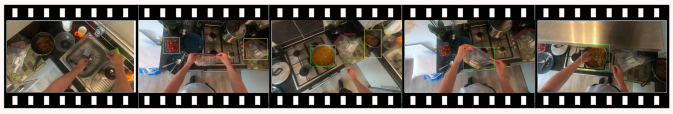
If during [Put food into plastic bags], [The small iron pot suddenly boils], in the modified factual scenario, what would the camera wearer most likely do next in this scenario?



Answer
"A": "Put down what's in your hand, turn down the heating function on the stove, then pick up the plate and pour tomato salad into the plastic bag",
"B": "Put down what you are holding, turn off the heating function on the stove, then pick up the plate and pour the grilled fish fillet into the plastic bag",
"C": "Put down the things in your hand, turn off the heating function on the stove, then pick up the plate, pour the tomato salad into the plastic bag, and seal the bag tightly",
"D": "Put down the things in your hand, turn off the heating function on the stove, then pick up the plate and pour the tomato salad next to the plastic bag",
"E": "Put down the things in your hand, first pick up the plate and pour tomato salad into the plastic bag, then turn off the heating function on the stove",
"F": "Put down what you are holding, turn off the heating function on the stove, then pick up the bowl and pour tomato salad into the plastic bag",
"G": "Put down what you are holding, turn off the heating function on the stove, then pick up the plate and pour tomato salad into the plastic bag",
"H": "Put down the things in your hand, turn off the heating function on the stove, then pick up the plate and pour the tomato salad in the plastic bag back onto the plate"

Factual Outcome Modification

For [Prepare food], if [Put the plastic sealed bag into the refrigerator] is changed to [Put the plastic sealed bag into the cabinet], after modifying the outcome fact, which steps should have been performed or not performed before the outcome?



Answer
"A": "Seal the bag, then open the refrigerator, put it in the plastic sealed bag, and stir the food in the small iron pot",
"B": "Seal the bag, place the plastic sealed bag next to the cabinet, and then stir the food in the small iron pot",
"C": "Take the plastic sealed bag from the drawer, put it in the grilled fish fillet, tomato salad, seal it and put it in the refrigerator for freezing",
"D": "Seal the bag, then stir the food in the small iron pot first, and then open the cabinet to put the plastic sealed bag in",
"E": "Seal the bag, then open the cabinet, put it in the plastic bag, and stir the food in the small iron pot",
"F": "Seal the bag, then open the cabinet, put it into the food preservation box, and stir the food in the small iron pot",
"G": "Seal the bag, then open the cabinet and place it in the plastic bag. Then, transfer the food from the small iron pot to a plate",
"H": "Seal the bag, then open the cabinet, put it in the plastic sealed bag, and continue to observe the food in the small iron pot"

Temporal Prediction

Based on the currently observed video, what is the most likely next action?



Answer
"A": "Pick up the pizza box, tear open the pizza box, and throw the pizza box into the trash can",
"B": "Open the refrigerator door, take out the milk, and place the milk on the table",
"C": "Remove the grill from the oven, place it on the table, and tear open the pizza box",
"D": "Pick up tomatoes, peel them, and use a slicer to slice potatoes",
"E": "Pick up the onion, peel it with a knife, and use a slicer to slice the onion",
"F": "Take out tomatoes and potatoes from the refrigerator, and place them on the cutting board",
"G": "Pick up a small knife, cut the grill, and place the grill on the table",
"H": "Pick up the trash can, lift the lid, and discard the grill"

Temporal Abduction

A process is missing between the two video segments. Determine what most likely happened in between.



Answer
"A": "Take out the pizza box from the refrigerator, place it on the table, and take out the grill from the oven",
"B": "Tear open the pizza box, place the pizza on the table, and throw the pizza box into the trash can",
"C": "Close the cabinet door, take out the plate, and pour water into the plate",
"D": "Take out a cutting board from the cabinet, take out tomatoes, take out potatoes, and place them on the cutting board",
"E": "Tear off the pizza cling film, place the pizza on the grill, and throw the cling film into the trash can",
"F": "Take out the cutting board, place the grill, and discard the pizza on the grill",
"G": "Move ham to another pizza, spread shredded onions on the surface of the pizza. Move ham to another pizza, spread shredded onions on the surface of the pizza",
"H": "Pick up a small knife, cut onions, and throw onions into the trash can"

Temporal Planning

Based on the currently observed video, what is the most likely next action?



Answer
"A": "Open the refrigerator door, take out the milk, and place the milk on the table",
"B": "Take out tomatoes and potatoes from the refrigerator, and place them on the cutting board",
"C": "Pick up the pizza box, tear open the pizza box, and throw the pizza box into the trash can",
"D": "Remove the grill from the oven, place it on the table, and tear open the pizza box",
"E": "Pick up a small knife, cut the grill, and place the grill on the table",
"F": "Pick up the trash can, lift the lid, and discard the grill",
"G": "Pick up tomatoes, peel them, and use a slicer to slice potatoes",
"H": "Pick up the onion, peel it with a knife, and use a slicer to slice the onion"

Temporal Ordering

Which sequence is correctly ordered according to the temporal order?



Answer
"A": "4->2->3->1->5",
"B": "3->2->1->4->5",
"C": "5->4->3->2->1",
"D": "2->5->3->4->1",
"E": "1->4->3->2->5",
"F": "3->1->4->5->2",
"G": "1->3->5->4->2",
"H": "4->2->3->5->1"

Spatial Object Position Before Movement

Taking the center of [microwave oven] as the spatial center, the direction from the center of [microwave oven] to the center of [White hot water kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. Before [Green cutting board] is moved, where is [Green cutting board] located relative to [microwave oven]?



Answer

"A": "The target object is located front-left of the reference object and is higher than the reference object.",
"B": "The target object is located back-right of the reference object and is higher than the reference object.",
"C": "The target object is located front-right of the reference object and is at the same height as the reference object.",
"D": "The target object is located behind the reference object and is at the same height as the reference object.",
"E": "The target object is located front-right of the reference object and is lower than the reference object.",
"F": "The target object is located back-left of the reference object and is lower than the reference object.",
"G": "The target object is located on the left side of the reference object and is lower than the reference object.",
"H": "The target object is located on the right side of the reference object and is higher than the reference object."

Spatial Object Position Before Movement

Taking the center of [microwave oven] as the spatial center, the direction from the center of [microwave oven] to the center of [White hot water kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. Before [cherry tomato] is moved, where is [cherry tomato] located relative to [microwave oven]?



Answer

"A": "The target object is located back-right of the reference object and is higher than the reference object.",
"B": "The target object is located front-left of the reference object and is higher than the reference object.",
"C": "The target object is located on the left side of the reference object and is lower than the reference object.",
"D": "The target object is located front-right of the reference object and is at the same height as the reference object.",
"E": "The target object is located back-left of the reference object and is lower than the reference object.",
"F": "The target object is located on the right side of the reference object and is higher than the reference object.",
"G": "The target object is located in front of the reference object and is higher than the reference object.",
"H": "The target object is located behind the reference object and is at the same height as the reference object."

Spatial Object Position After Movement

Taking the center of [microwave oven] as the spatial center, the direction from the center of [microwave oven] to the center of [White hot water kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. After [Green cutting board] is moved, where is [Green cutting board] located relative to [microwave oven]?



Answer

"A": "The target object is located front-right of the reference object and is lower than the reference object.",
"B": "The target object is located on the left side of the reference object and is higher than the reference object.",
"C": "The target object is located back-right of the reference object and is at the same height as the reference object.",
"D": "The target object is located behind the reference object and is lower than the reference object.",
"E": "The target object is located on the right side of the reference object and is at the same height as the reference object.",
"F": "The target object is located front-left of the reference object and is at the same height as the reference object.",
"G": "The target object is located back-left of the reference object and is higher than the reference object.",
"H": "The target object is located front-right of the reference object and is higher than the reference object."

Spatial Object Position After Movement

Taking the center of [microwave oven] as the spatial center, the direction from the center of [microwave oven] to the center of [White hot water kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. After [cherry tomato] is moved, where is [cherry tomato] located relative to [microwave oven]?



Answer

"A": "The target object is located front-left of the reference object and is higher than the reference object.",
"B": "The target object is located behind the reference object and is at the same height as the reference object.",
"C": "The target object is located front-right of the reference object and is at the same height as the reference object.",
"D": "The target object is located front-right of the reference object and is lower than the reference object.",
"E": "The target object is located on the right side of the reference object and is higher than the reference object.",
"F": "The target object is located back-left of the reference object and is lower than the reference object.",
"G": "The target object is located on the left side of the reference object and is lower than the reference object.",
"H": "The target object is located back-right of the reference object and is higher than the reference object."

Spatial Relative Position Before and After Object Movement

Taking the center of [microwave oven] as the spatial center, the direction from the center of [microwave oven] to the center of [White hot water kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. Where is [Green cutting board] before it is moved relative to [Green cutting board] after it is moved?



Answer
"A": "The target object is located back-left of the reference object and is higher than the reference object.",
"B": "The target object is located front-right of the reference object and is at the same height as the reference object.",
"C": "The target object is located behind the reference object and is at the same height as the reference object.",
"D": "The target object is located on the right side of the reference object and is lower than the reference object.",
"E": "The target object is located front-left of the reference object and is higher than the reference object.",
"F": "The target object is located front-left of the reference object and is lower than the reference object.",
"G": "The target object is located on the left side of the reference object and is higher than the reference object.",
"H": "The target object is located back-right of the reference object and is lower than the reference object."

Spatial Relative Position Before and After Object Movement

Taking the center of [microwave oven] as the spatial center, the direction from the center of [microwave oven] to the center of [White hot water kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. Where is [cherry tomato] before it is moved relative to [cherry tomato] after it is moved?



Answer
"A": "The target object is located on the left side of the reference object and is at the same height as the reference object.",
"B": "The target object is located on the right side of the reference object and is lower than the reference object.",
"C": "The target object is located back-left of the reference object and is at the same height as the reference object.",
"D": "The target object is located behind the reference object and is higher than the reference object.",
"E": "The target object is located front-right of the reference object and is higher than the reference object.",
"F": "The target object is located front-left of the reference object and is lower than the reference object.",
"G": "The target object is located back-right of the reference object and is lower than the reference object.",
"H": "The target object is located on the left side of the reference object and is lower than the reference object."

Spatial Relative Position of Two Objects Before Movement

Taking the center of [microwave oven] as the spatial center, the direction from the center of [microwave oven] to the center of [White hot water kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. Where is [Green cutting board] before it is moved relative to [cherry tomato] before it is moved?



Answer
"A": "The target object is located back-right of the reference object and is lower than the reference object.",
"B": "The target object is located back-left of the reference object and is higher than the reference object.",
"C": "The target object is located behind the reference object and is at the same height as the reference object.",
"D": "The target object is located on the left side of the reference object and is higher than the reference object.",
"E": "The target object is located behind the reference object and is lower than the reference object.",
"F": "The target object is located front-left of the reference object and is lower than the reference object.",
"G": "The target object is located front-right of the reference object and is at the same height as the reference object.",
"H": "The target object is located on the right side of the reference object and is lower than the reference object."

Spatial Relative Position of Two Objects Before Movement

Taking the center of [microwave oven] as the spatial center, the direction from the center of [microwave oven] to the center of [White hot water kettle] is defined as the forward direction. The vertical direction is defined with the ground direction as downward, and the supporting surface of an object is used as its relative height reference. Where is [Green cutting board] after it is moved relative to [cherry tomato] after it is moved?



Answer
"A": "The target object is located back-right of the reference object and is at the same height as the reference object.",
"B": "The target object is located on the left side of the reference object and is higher than the reference object.",
"C": "The target object is located behind the reference object and is at the same height as the reference object.",
"D": "The target object is located on the right side of the reference object and is at the same height as the reference object.",
"E": "The target object is located back-left of the reference object and is higher than the reference object.",
"F": "The target object is located front-right of the reference object and is lower than the reference object.",
"G": "The target object is located behind the reference object and is lower than the reference object.",
"H": "The target object is located front-left of the reference object and is at the same height as the reference object."

Factual Single Fact Perturbation

If during [Processing pizza], [There are no onions in the basket], in the modified factual scenario, what would the camera wearer most likely do next from the moment of the factual change?



Answer
"A": "Pick up a small knife, cut the pizza box, and throw the pizza box into the trash can",
"B": "Pick up the potatoes directly, peel them, and use a slicer to cut the potatoes",
"C": "Pick up the pizza box, tear open the pizza box, and remove the onions from the pizza box",
"D": "Pick up the slicer, cut the grill, and place the grill on the cutting board",
"E": "Move the grill position, wipe the grill, and remove surface dust",
"F": "Take out onions from the basket, place them on the cutting board, and peel the onion skin",
"G": "Open the refrigerator, search for alternative vegetables, and place them on the cutting board for processing",
"H": "Pick up the trash can, clean up the garbage on the ground, and close the lid of the trash can"

Factual Dual Fact Perturbation

If during [Preparing to process pizza], [There is no grill inside the oven], and [There are no onions in the basket], in the modified factual scenario, what would the camera wearer most likely do next from the moment of the factual changes?



Answer
"A": "Open the window for ventilation, pick up a cloth to wipe away desktop stains, and clean the cloth",
"B": "Open the cabinet, find the baking tray, take out the baking tray and place it on the desktop",
"C": "Open the refrigerator, search for alternative vegetables, and place them on the cutting board",
"D": "Find substitute items such as baking pans, select other vegetables from the refrigerator, and place them in the sink",
"E": "Pick up the pizza box, tear open the pizza box, and place the pizza directly on the table",
"F": "Remove the grill from the oven, take onions from the basket, and place them on the cutting board",
"G": "Pick up tomatoes and mushrooms, place them on a cutting board, and directly use a slicer to slice them",
"H": "Find substitute items such as baking pans, select other vegetables from the refrigerator, and place them on the cutting board"

Factual Fact Injection

If during [cut onions], [Suddenly a phone call came], in the modified factual scenario, what would the camera wearer most likely do next in this scenario?



Answer
"A": "Put down the knife and onion, dry your hands, and then pick up your phone to make a call",
"B": "Put down the knife and onion, dry your hands, then pick up the pizza and leave the workbench",
"C": "Put down the knife and onion, don't wipe your hands, then pick up your phone to answer the call",
"D": "Put down the knife and onion, dry your hands, and then pick up your phone to answer the call",
"E": "Put down the knife and onion, dry your hands, then pick up your phone and put it to your ear",
"F": "Put down the knife and onion, dry your hands, then pick up the onion and continue cutting",
"G": "Put down the knife and onion, dry your hands, and then pick up your phone to check your messages",
"H": "Put down the knife, pick up the onion and put it aside, then pick up your phone to answer the call"

Factual Outcome Modification

For [Preparing to process pizza], if [Pizza is placed on an iron rack] is changed to [Pizza is placed directly on the table], after modifying the outcome fact, which steps should have been performed or not performed before the outcome?



Answer
"A": "Open the pizza packaging, do not move the iron rack, place the pizza on the table, and then use the cutting board and dish",
"B": "Open the pizza packaging, take out the iron rack, place the pizza on the table, and then bring the cutting board and dishes",
"C": "Open the pizza packaging, do not take the iron rack, place the pizza on the table, and then do not take the cutting board or dish",
"D": "Open the pizza packaging, do not take the iron rack, place the pizza on the table, then bring the dishes first, and then bring the cutting board",
"E": "Open the pizza packaging, do not take the iron rack, place the pizza on the table, and then use the cutting board and seasoning",
"F": "Open the pizza packaging, do not take the iron rack, place the pizza on the table, and then use the cutting board and vegetables",
"G": "Open the pizza packaging, do not take the iron rack, place the pizza on the table, and then take the iron rack and vegetables",
"H": "Open the pizza packaging, do not take the iron rack, place the pizza on the cutting board, and then bring the cutting board and dishes"