VGENST-BENCH: A BENCHMARK FOR SPATIO-TEMPORAL REASONING VIA ACTIVE VIDEO SYNTHESIS

Anonymous authors
Paper under double-blind review



Figure 1: **Examples of VGenST-Bench.** Each example contains a generated video and a multiple-choice question targeting a specific spatio-temporal reasoning task. Correct answers are highlighted.

ABSTRACT

Spatio-temporal reasoning is a core capability for Multimodal Large Language Models (MLLMs) operating in the real world. As such, evaluating it precisely has become an essential challenge. However, existing spatio-temporal reasoning benchmark datasets primarily rely on static image sets or passively curated video data, which limits the evaluation of fine-grained reasoning capabilities. In this paper, we introduce **VGenST-Bench**, a video benchmark that employs generative models to actively synthesize highly controlled and diverse evaluation scenarios. To construct VGenST-Bench, we propose a multi-agent pipeline incorporating a human quality control stage, ensuring the quality of all generated videos and QA pairs. We establish a comprehensive $3 \times 2 \times 2$ video taxonomy, encompassing *Spatial Scale*, *Perspective*, and *Scene Dynamics* to span diverse scenarios. Furthermore, we design a hierarchical task suite that decouples low-level visual perception from high-level spatio-temporal reasoning. By shifting the paradigm from passive curation to active synthesis, VGenST-Bench enables fine-grained diagnosis of spatio-temporal understanding in MLLMs.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) are increasingly used in physically grounded applications, including robotics (Driess et al., 2023; Zitkovich et al., 2023; Kim et al., 2024) and autonomous driving (Xu et al., 2024; Tian et al., 2024). These applications require *spatio-temporal reasoning*,

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.

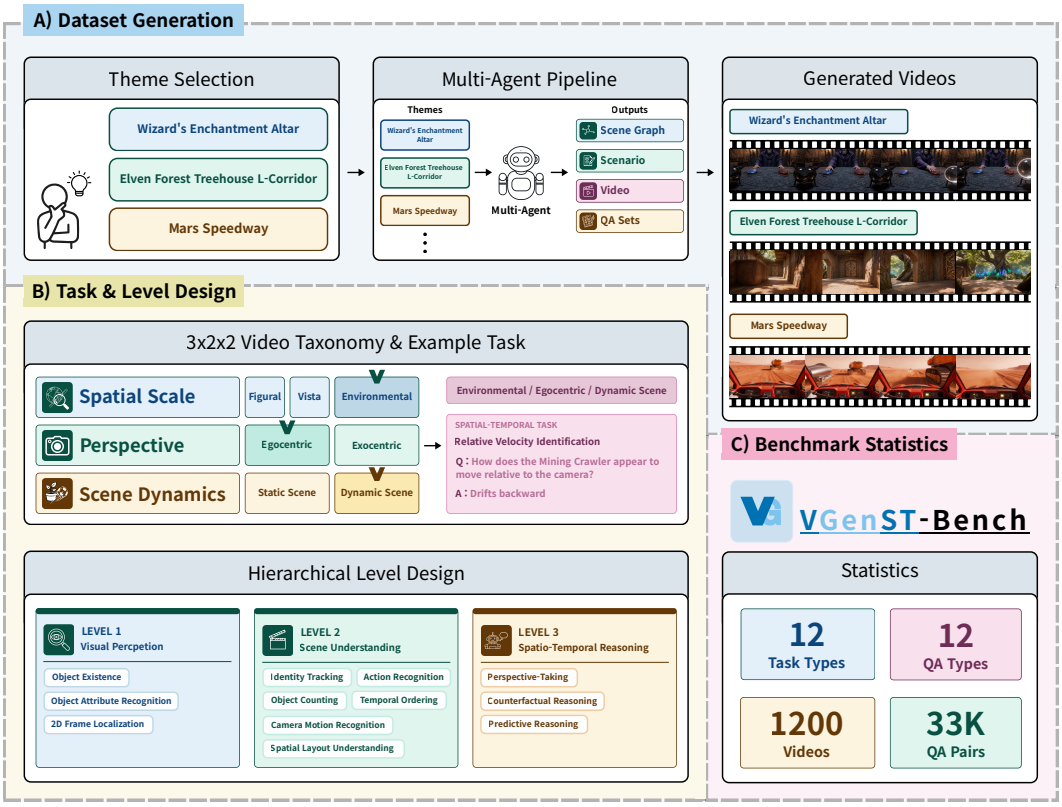


Figure 2: **Overview of VGenST-Bench. A) Dataset generation.** A multi-agent pipeline produces scene graphs, scenarios, videos, and QA pairs, followed by human verification. **B) Task and level design.** A $3 \times 2 \times 2$ taxonomy over spatial scale, perspective, and scene dynamics defines 12 task categories. Questions span visual perception (L1), scene understanding (L2), and spatio-temporal reasoning (L3). **C) Benchmark statistics.** The benchmark contains 1,200 videos and 33K QA pairs across 12 tasks and 12 QA types.

the ability to perceive and infer the positions, orientations, and attributes of objects across time and changing perspectives. Despite advances in visual perception, current MLLMs still struggle with these capabilities (Lin et al., 2025b;a; Li et al., 2025b), motivating evaluations that examine how models use spatial and temporal evidence.

Existing benchmarks provide important tests of spatial understanding, but static image tasks offer limited access to temporal changes (Kamath et al., 2023; Zhang et al., 2025; Wang et al., 2025c; Jia et al., 2025). Video benchmarks extend evaluation to evolving scenes, often by collecting web videos or repurposing existing scene datasets (Zhou et al., 2025; Yang et al., 2025; Linghu et al., 2024). This *passive curation* constrains evaluation to the situations available in the source data, obtaining a desired combination of spatial scale, viewpoint, and scene dynamics requires finding and annotating suitable clips. Publicly sourced data also raises concerns about pretraining overlap (Song et al., 2024; Chen et al., 2024b), while language priors and static visual cues can provide shortcuts that weaken the need for spatio-temporal reasoning (Buch et al., 2022; Lei et al., 2023; Krojer et al., 2025).

Recent progress in video generation (Wiedemer et al., 2025; Seedance et al., 2025; Wan et al., 2025) offers an opportunity to specify evaluation scenarios before producing the videos. This enables a fundamentally different approach to benchmark construction—actively synthesizing precisely controlled evaluation scenarios rather than passively curating them from existing sources. This motivates our question: *Can actively synthesized videos provide a controllable and empirically validated testbed for spatio-temporal reasoning in MLLMs?*

In this work, we introduce **VGenST-Bench**, a benchmark leveraging **V**ideo **Gen**erative models to evaluate **S**patio-**T**emporal reasoning in MLLMs. To the best of our knowledge, VGenST-Bench is

Benchmark	Venue/Year	Modality	Reasoning Type	QA Pairs (#)	Data Scale (#)	Data Source
MME (Fu et al., 2023)	NeurIPS'25	I	S	2.3K	1.1K	Real image datasets
3DSRBench (Ma et al., 2025)	ICCV'25	I	S	6.9K	2.7K	Real image datasets
SpatialViz-Bench (Wang et al., 2025a)	ICLR'26	I	S	1.1K	1.1K	Programmatic generated images
Spatial457 (Wang et al., 2025c)	CVPR'25	I	S	23K	1.0K	Rendered synthetic 3D scenes
VSI-Bench (Yang et al., 2025)	CVPR'25	V	S	5K	288	3D indoor scene datasets
EgoExoBench (He et al., 2025)	NeurIPS'25	V	S/T	7.3K	2.7K	Ego-exo paired video datasets
STI-Bench (Li et al., 2025b)	ICCV'25	V	S/T	2K	300	Autonomous driving & 3D indoor scene datasets
OST-Bench (Lin et al., 2025b)	NeurIPS'25	V	S/T	10K	1.4K	3D indoor scene datasets
VGenST-Bench (Ours)	**2026**	**V**	**S/T**	**33K**	**1.2K**	**Video generative models**

Table 1: **Comparison of VGenST-Bench with recent MLLM benchmarks.** Our benchmark is the first spatio-temporal reasoning benchmark that leverages video generative models. I: image, V: video; S: spatial, T: temporal; S/T: spatio-temporal.

the first benchmark built on photorealistic videos synthesized by video generative models for this purpose. To construct this benchmark, we design a multi-agent pipeline that generates benchmark-ready evaluation videos and questions, followed by a final human quality-control stage. The detailed pipeline design is provided in Fig. 4.

We organize the benchmark along three axes: **spatial scale**, **perspective**, and **scene dynamics**. Spatial scale and perspective draw on established distinctions in spatial cognition (Montello, 1993; Hegarty et al., 2006; Klatzky, 1998). Their combinations with static or dynamic scenes form a $3 \times 2 \times 2$ taxonomy, with one reasoning task per cell. This organization provides structured coverage of selected scenarios. Each video is paired with questions spanning **visual perception (L1)**, **scene understanding (L2)**, and **spatio-temporal reasoning (L3)**, supporting analysis across both scene conditions and question levels.

We evaluate 15 proprietary and open-source MLLMs and observe substantial performance degradation from L1 to L3. Additional experiments examine frame sampling, visual shortcuts, temporal order, and oracle test. On a controlled subset, shuffling frames reduces L3 accuracy, whereas supplying the scene graphs substantially improves it, suggesting that obtaining usable spatio-temporal information from video is an important bottleneck. Together, these results provide empirical support for using videos synthesized by generative models to benchmark spatio-temporal reasoning in MLLMs. With structured scenario design and human verification, active video synthesis offers a viable approach to constructing controllable evaluation benchmarks.

In summary, our work makes the following key contributions:

- **Video benchmark with active synthesis paradigm:** We propose VGenST-Bench, the first benchmark to evaluate spatio-temporal reasoning in MLLMs using actively synthesized video, organized under a $3 \times 2 \times 2$ taxonomy with 12 reasoning tasks and a three-level question hierarchy.
- **Benchmark construction pipeline:** We design a multi-agent generation pipeline that jointly synthesizes scene graphs, scenarios, videos, and QA sets, followed by a human quality-control stage. This pipeline enables controllable construction of evaluation scenarios at scale, overcoming the passive curation bottleneck of prior video benchmarks.
- **Empirical benchmark validation and diagnostic analysis:** We evaluate 15 MLLMs across tasks and question levels, and examine real-video ranking agreement, generation quality, frame sampling, and input ablations to characterize model performance and the scope of the benchmark's conclusions.

2 Related Work

Spatio-temporal reasoning benchmarks for MLLMs. Early benchmarks evaluate spatial understanding in MLLMs through static 2D images, probing object localization, relative position, and compositional spatial relations (Johnson et al., 2017; Liu et al., 2023; Kamath et al., 2023; Chen et al., 2024a; Cheng et al., 2024; Tong et al., 2024; Fu et al., 2024; Ma et al., 2025; Wang et al., 2024a; Jia et al., 2025). While valuable, static image-based evaluation possesses an inherent limitation, a fundamental inability to capture state transitions across the temporal dimension. To bridge this gap, recent studies have begun to incorporate video datasets (Liu et al., 2024b; Li et al., 2024b; Fu et al.,

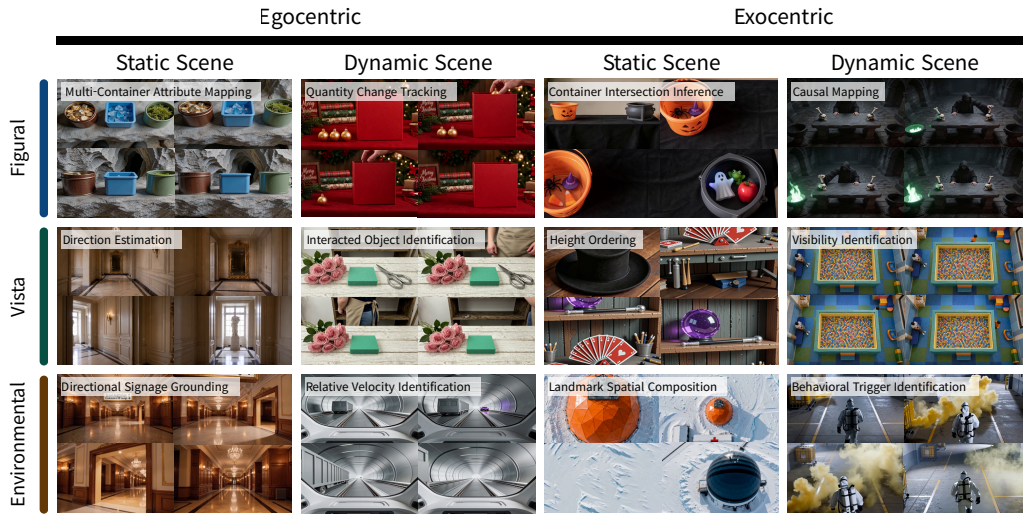


Figure 3: **Representative videos for the 12 tasks of VGenST-Bench.** Each cell of the $3 \times 2 \times 2$ taxonomy (Spatial scale $\times$ Perspective $\times$ Scene dynamics) is paired with one dedicated reasoning task. Rows correspond to spatial scales (Figural / Vista / Environmental); columns are grouped by perspective (Egocentric / Exocentric) and scene dynamics (Static / Dynamic). Each strip shows four sampled frames from a representative video for the task.

2025a; Zhou et al., 2025; Yu et al., 2019; Lei et al., 2020) or repurpose existing 3D scene datasets (Lin et al., 2025b; Yang et al., 2025; Yin et al., 2025; Linghu et al., 2024; Gong et al., 2025) to evaluate spatio-temporal reasoning. Although these sources provide visual richness, they are *passively curated* from in-the-wild environments rather than *actively designed* for reasoning evaluation. This reliance on public data not only limits the diversity and controllability of evaluation scenarios but also exposes the benchmarks to data contamination. A complementary line of work utilizes synthetic evaluation data (Yi et al., 2019; Salewski et al., 2020; Li et al., 2023b; Zhu et al., 2026; Ma et al., 2025). While affording precise ground-truth control, these benchmarks suffer from a visual realism gap, limiting their utility for evaluating modern MLLMs trained on photorealistic data. Most recently, a few works have begun to leverage video generative models for benchmark construction (Li et al., 2025c; Fu et al., 2025b), but they primarily target hallucination detection or perceived generation artifact rather than spatio-temporal reasoning. Compared with these prior works, as shown in Tab. 1, VGenST-Bench is the first video spatio-temporal reasoning benchmark constructed entirely from video generative models, enabling controllable, diverse scenarios at scale. More comprehensive discussion is provided in Appendix B.

3 VGenST-Bench

3.1 Video Taxonomy and Task Design

To systematically cover spatio-temporal reasoning scenarios, we organize VGenST-Bench under a $3 \times 2 \times 2$ taxonomy along three axes: **(i) Spatial scale** (*figural*, *vista*, *environmental*), **(ii) Perspective** (*egocentric*, *exocentric*), and **(iii) Scene dynamics** (*static*, *dynamic*). Each combination of axis values defines a distinct video category. We design one dedicated reasoning task per cell, yielding 12 tasks that together probe the full taxonomy (Tab. 2, with visual examples in Fig. 3). Further details are provided in Appendix C.

3.2 QA Design

Level Design. Orthogonal to the video taxonomy, each video is paired with QA pairs organized along a three-level cognitive hierarchy that progresses from low-level perception to high-level reasoning, inspired by recent hierarchical spatial reasoning benchmarks (Zhang et al., 2025; Wang et al., 2025c; Li et al., 2025a). As shown in Tab. 3, the hierarchy comprises **(L1) Visual perception** (3 QA types),

Scale × Dynamics		Egocentric	Exocentric
Figural	Static	**MC**: Multi-Container Attribute Mapping	**CI**: Container Intersection Inference
	Dynamic	**QC**: Quantity Change Tracking	**CM**: Causal Mapping
Vista	Static	**DE**: Direction Estimation	**HO**: Height Ordering
	Dynamic	**IO**: Interacted Object Identification	**VI**: Visibility Identification
Environmental	Static	**DS**: Directional Signage Grounding	**LS**: Landmark Spatial Composition
	Dynamic	**RV**: Relative Velocity Identification	**BT**: Behavioral Trigger Identification

Table 2: Twelve tasks of VGenST-Bench, organized along the $3 \times 2 \times 2$ taxonomy (Spatial scale $\times$ Perspective $\times$ Scene dynamics). Each cell contains the task code (bold) and full name.

L1: Visual Perception	L2: Scene Understanding	L3: Spatio-Temporal Reasoning
Object Existence (OE)	Identity Tracking (IT)	Perspective-Taking (PT)
Object Attribute Recognition (OA)	Action Recognition (AR)	Counterfactual Reasoning (CR)
2D Frame Localization (FL)	Object Counting (OC)	Predictive Reasoning (PR)
	Temporal Ordering (TO)	
	Camera Motion Recognition (CM)	
	Spatial Layout Understanding (SL)	

Table 3: Twelve QA types of VGenST-Bench, organized along the three-level cognitive hierarchy.

which probes the recognition of objects and their visual attributes from individual frames; **(L2) Scene understanding** (6 QA types), which assesses the integration of perceptual cues across frames into coherent spatial and temporal structures; and **(L3) Spatio-temporal reasoning** (3 QA types), which evaluates higher-order inference such as perspective-taking, counterfactual, and predictive reasoning. Full QA type definitions and examples are provided in Appendix C.4.

Task–QA Applicability. Not every QA type is applicable for each of the twelve tasks. For instance, Action Recognition is undefined for tasks with static scenes. We therefore define a *task–QA applicability matrix* (Appendix C.5) that specifies which QA types are evaluated for each of the 12 tasks, ensuring that every QA pair is well-defined for its underlying video while preserving balanced coverage across the hierarchy.

Question Reformulation for Robust Evaluation. A central concern in multiple-choice evaluation is that models may exploit option-level shortcuts rather than genuinely reasoning about the video (Lei et al., 2023; Krojer et al., 2025). To mitigate this, each base multiple-choice question (MCQ) is expanded into three variants: (i) a *None-of-these distractor*, which adds "None of these" as an additional incorrect option to test whether models commit to the correct choice when one is present; (ii) a *None-of-these answer*, which replaces the correct option with "None of these" to test whether models can reject all listed options when appropriate; and (iii) an *open-ended* variant that removes options entirely to assess reasoning without choice priors. After filtering, this protocol yields a total of 33K QA pairs across the benchmark. We provide more details in Appendix C.6.

3.3 Dataset Construction

We construct VGenST-Bench through a multi-agent pipeline that sequentially synthesizes scene graphs, scenarios, videos, and QA pairs, followed by a human quality-control stage.

Structuring Inputs to a Video Generator. To make video generation *benchmark-ready*, our pipeline conditions the video generator on three structured representations. **(1) Theme** provides the visual and semantic context of the video, the style of objects, and the overall environment (e.g., *Cyberpunk Hacker's Neon Desk*, *Wizard's Enchantment Altar*). **(2) Scene Graph** specifies the static spatial configuration required by the task: the objects to be rendered, their attributes (color, material, role), and their pairwise spatial relations. **(3) Scenario** lifts the scene graph into the temporal domain, specifying the reasoning goal, the camera setup, and a structured timeline of events. Together, these

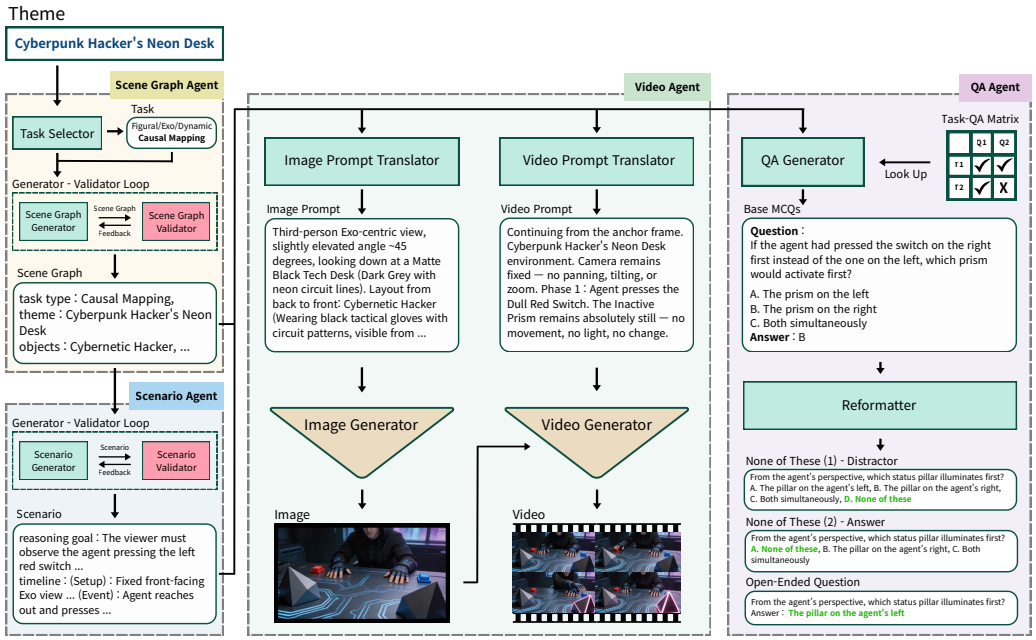


Figure 4: **VGenST-Bench construction pipeline.** Starting from a theme, four agents operate in sequence. The **Scene Graph Agent** produces a structured scene graph specifying objects and spatial composition; the **Scenario Agent** expands it into a temporally grounded scenario with reasoning goal and timeline; the **Video Agent** synthesizes the corresponding image and video through generative models; and the **QA Agent** generates base MCQs from a task–QA applicability matrix and reformats each into three variants.

three representations fix the visual context and the spatio-temporal ground truth needed for both rendering and QA generation.

Multi-Agent Pipeline. Our pipeline comprises four agent modules (Fig. 4). **(1) Scene Graph Agent** constructs a scene graph from a (theme, task) pair sampled by the Task Selector, with iterative validator feedback on required objects, attributes, and relations. **(2) Scenario Agent** develops a temporal scenario, iteratively validated for consistency and sufficiency to derive the ground-truth answer. **(3) Video Agent** first generates an anchor image, then synthesizes a video conditioned on this image and a scenario-derived prompt. We use multiple image and video generators and select the best output per scenario. **(4) QA Agent** generates base MCQs using the task–QA applicability matrix and the scene graph and scenario as ground-truth references, then reformats them into three variants (Section 3.2). Further details appear in Appendix D.

Human Quality Control. All generated videos and base QA pairs undergo a two-stage human verification protocol with a pair of validators per task, retaining only items both validators mark valid. In *Stage 1 (Video QC)*, the validator pair reviews the generated videos and rejects clips that fail visual fidelity or scenario adherence. In *Stage 2 (QA QC)*, the same pair reviews each base MCQ on the surviving videos and rejects items with ambiguous or invalid answers. Further details are reported in Appendix D.3.

4 Experiments

4.1 Experimental Setup

Evaluation Metric. We report *accuracy* as our primary metric across all questions. For the base MCQ, we adopt *circular evaluation* (Liu et al., 2024a) to control position bias and choice priors. Each question is evaluated under all N cyclic permutations of its answer choices, and the model is scored as correct only if it answers correctly under every cycle. For the open-ended variant, we follow

Model	Figural				Vista				Environmental				Avg	
	MC	QC	CI	CM	DE	IO	HO	VI	DS	RV	LS	BT	Macro	Micro
Baseline														
Human	100.0	100.0	100.0	99.5	98.0	99.5	100.0	97.0	100.0	99.0	96.1	99.0	99.0	99.0
Random	0.8	0.8	1.0	2.1	0.1	0.9	0.6	0.1	1.1	1.0	0.1	1.2	0.8	0.8
Proprietary Models														
GPT-5.4	88.8	71.0	92.5	79.0	**76.4**	**94.0**	84.5	**80.6**	**97.0**	81.9	57.0	90.1	82.7	82.7
GPT-5.4 mini	68.5	46.3	75.4	72.5	60.0	80.4	70.7	60.2	75.0	73.8	41.0	80.2	67.0	67.1
GPT-5.4 nano	64.2	48.2	68.2	52.0	59.4	62.9	44.0	48.8	59.5	57.9	37.1	78.8	56.8	56.8
Gemini 3.1 Flash-Lite	87.9	**71.6**	**97.0**	76.2	66.9	93.9	87.1	71.1	91.0	77.2	47.9	88.0	79.6	79.6
Gemini 3 Flash	**97.0**	71.4	95.7	**82.2**	73.4	93.9	**96.9**	66.1	**97.0**	**91.1**	**66.0**	**96.9**	**85.6**	**85.9**
Open-Source Models														
Qwen3.5-4B	81.4	51.6	72.9	61.4	50.4	91.9	63.6	59.5	88.0	53.2	35.6	58.9	64.0	63.6
Qwen3.5-9B	85.2	54.4	75.8	63.3	52.0	**92.1**	67.2	62.1	90.9	56.3	38.0	63.9	66.8	66.4
Qwen3.5-27B	**87.6**	57.0	77.9	65.6	54.4	91.4	69.5	64.7	**91.9**	58.2	41.2	64.8	68.7	68.3
InternVL3.5-4B	62.9	38.3	55.2	46.9	41.4	55.0	47.6	42.0	60.5	51.2	32.7	70.2	50.3	50.3
InternVL3.5-8B	72.6	42.0	57.6	48.7	46.6	60.2	51.1	49.2	64.0	53.8	32.6	76.4	54.6	54.5
InternVL3.5-30B-A3B	86.4	55.7	71.6	62.3	59.0	74.9	64.7	63.5	77.8	66.8	45.7	**90.4**	68.2	68.1
Gemma-4-26B-A4B-it	77.0	61.9	84.4	62.5	54.3	77.2	61.9	66.0	81.6	62.8	41.1	79.7	67.5	67.4
Gemma-4-31B-it	82.0	**66.9**	**89.4**	67.5	59.3	82.2	66.8	**71.1**	86.6	**67.8**	46.2	84.6	**72.5**	**72.5**
GLM-4.6V-Flash	68.0	40.0	76.9	69.2	56.6	80.5	78.6	52.8	70.5	59.3	40.1	76.9	64.1	63.7
Kimi-K2.6	71.2	45.0	89.1	**76.7**	**73.1**	87.6	**86.1**	56.4	79.8	64.2	**49.8**	80.9	71.7	71.0

Table 4: **Main results on VGenST-Bench.** Tasks are grouped by spatial scale (Figural / Vista / Environmental). All multiple-choice questions are evaluated under *circular evaluation.* Per-task scores are macro-averaged across applicable QA types. **Macro Avg**: unweighted mean across the 12 tasks. **Micro Avg**: sample-weighted mean across all questions. See Tab. 2 for full task names.

the LLM-as-judge protocol (Zheng et al., 2023), using Claude-Sonnet-4.6 (Anthropic, 2025) to judge whether the model's response is semantically equivalent to the ground-truth answer.

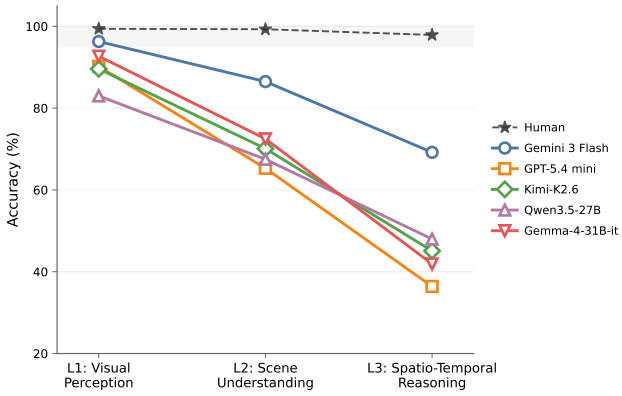
Human Baseline. We report human evaluation performance on VGenST-Bench. To conduct the human evaluation, we recruit 10 participants from diverse backgrounds, excluding computer-science majors to ensure that performance reflects general spatio-temporal reasoning capabilities. Each annotator answers a subset of 10 videos sampled across all 12 tasks (120 videos for each annotator), under the same circular evaluation protocol applied to model evaluation.

Evaluated Models. We evaluate 15 MLLMs spanning two categories. **Proprietary Models**. We evaluate GPT-5.4, GPT-5.4-mini/nano (OpenAI, 2026), and Gemini 3.1 Flash-Lite, Gemini 3 Flash (Google DeepMind, 2026a). **Open-source Models**. We evaluate Qwen3.5 (4B/9B/27B) (Team, 2026), InternVL3.5 (4B/8B/30B-A3B) (Wang et al., 2025b), Gemma-4 (26B-A4B-it/31B-it) (Google DeepMind, 2026b), GLM-4.6V-Flash (Z.ai, 2025), and Kimi-K2.6 (Moonshot AI, 2026). For proprietary models, we use the official APIs provided by the respective vendors.

For open-source models, we serve models locally using vLLM (Kwon et al., 2023), running larger models on 8 $\times$ NVIDIA B200 GPUs and smaller models on 2 $\times$ NVIDIA RTX PRO 5000 GPUs. Across all models, we uniformly sample K=8 frames per video as input.

4.2 Main Results

Tab. 4 reports the main evaluation results across all evaluated models, organized by spatial scale (Figural / Vista / Environmental). Each cell reports accuracy on base MCQ questions under circular evaluation. As expected, human annotators remain a clear upper bound, achieving 99.0% on average and near-saturation across all twelve tasks. Even the strongest evaluated MLLM, *Gemini 3 Flash*, achieves only 85.9% on average—a gap of more than 13pp below human performance, and this gap widens substantially for compact and smaller models, with GPT-5.4 nano below 60%. Among open-source models, the majority perform below 70%, revealing limitations in their spatio-temporal reasoning. Notable exceptions are *Kimi-K2.6* and *Gemma-4-31B-it*, which achieve approximately 71.0% and 72.5% on average, respectively, narrowing the gap to top proprietary models and demon-



(a) Accuracy trends across L1/L2/L3.

Model	L1	L2	L3	Δ
Baseline				
Human	99.4	99.3	97.9	1.5
Proprietary				
Gemini 3 Flash	96.3	86.5	69.2	27.1
GPT-5.4 mini	90.2	65.3	36.4	53.8
Open-source				
Kimi-K2.6	89.6	70.1	45.1	44.5
Qwen3.5-27B	83.0	67.5	48.0	35.0
Gemma-4-31B-it	92.7	72.5	41.9	50.8

(b) Detailed scores per model.

Figure 5: **Hierarchical Analysis: Accuracy across the three question levels.** (a) All models degrade consistently from L1 to L3, while humans remain near-ceiling. (b) Breakdown by model, with the L1−L3 gap (Δ).

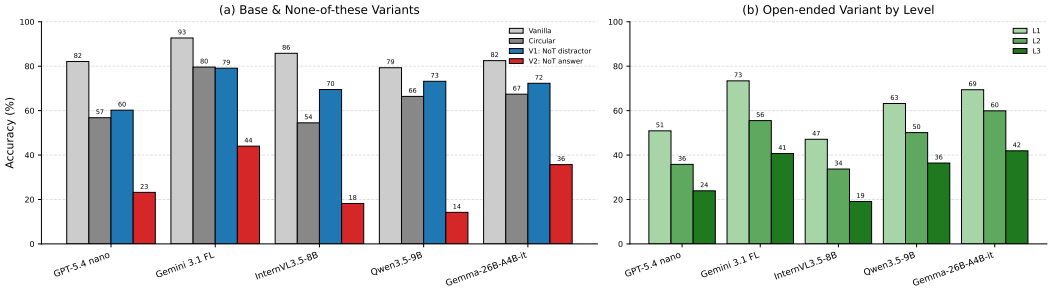


Figure 6: **Robustness Analysis.** (a) None-of-these variants show a clear asymmetry: V1 maintains base accuracy, while V2 produces dramatic drops across all models. (b) Open-ended evaluation by question level reveals large drop on L3 for all models. Together, these reveal that closed-form MCQ accuracy may overestimate spatio-temporal reasoning capability.

strating that large-scale open-source models can be competitive. We provide detailed analysis in Section 4.3.

4.3 Detailed Analysis

Hierarchical Analysis. We observe a consistent and substantial performance degradation from L1 to L3 as shown in Fig. 5. GPT-5.4 mini achieves 90.2% on L1, drops to 65.3% on L2, and further to 36.4% on L3; this pattern persists across every model category. In contrast, human annotators decline only marginally (99.4% → 97.9%), whereas other models exhibit dramatically larger drops. This degradation indicates that current MLLMs are competent at single-frame visual perception but exhibit systematic bottlenecks in higher-order reasoning that requires integrating spatial and temporal cues, validating our hierarchy as a diagnostic tool for separating perception from reasoning.

Robustness to Question Reformulation. Fig. 6(a) reveals two complementary findings. First, vanilla accuracy systematically exceeds circular accuracy across all models, indicating that current MLLMs exploit position bias and choice priors that single-attempt evaluation does not control for. Second, while V1 maintains or slightly exceeds the Circular accuracy, V2 produces dramatic drops. These results show that current MLLMs perform multiple-choice reasoning by ranking the given options against each other rather than verifying the correct answer against the video. When the correct answer is absent, models cannot reject the remaining distractors and instead select the most plausible one. Fig. 6(b) reveals that the open-ended variant exposes the L1→L3 hierarchy more dramatically than closed-form MCQ. Together, these results show that high accuracy on standard multiple-choice benchmarks may overestimate current MLLMs' spatio-temporal reasoning capability, and that our reformulation protocol is meaningful for diagnosing reasoning shortcuts.



Taxonomy / Task / QA :
Vista-Ego-Static / Direction Estimation / L3 - Perspective Taking

Question :
At the final position, in which direction (using clock positions where 12 o'clock is straight ahead) is the starting point located?:
A. 12 o'clock (directly ahead), B. 4-5 o'clock (back-right), C. 7-8 o'clock (back-left), D. 6 o'clock (directly behind)

Gemini-3.1-Flash-Lite :
1. Initial State: The scene starts in a corridor with portrait visible in the distance at the end of the hall. The camera starts by looking straight ahead at this portrait.
2. Movement: The camera moves forward along the corridor. As it progresses, it turns to the left, revealing a suit of armor standing in an adjacent corridor or alcove to the left.
3. Final State: The camera eventually centers on the suit, which is now positioned directly in front of the camera (at the 12 o'clock position).
4. Spatial Orientation: The starting point (the initial position where the portrait was seen straight ahead) is now behind the camera. Since the camera turned left to face the suit, the starting point (the path leading back to the portrait) is located to the right-rear of the current perspective. Therefore, if the suit is at 12 o'clock, the initial corridor path is located roughly at the 4-5 o'clock position.

Final Answer: **B. 4-5 o'clock (back-right) ← WRONG!**

Figure 7: **Reasoning failure of *Direction Estimation* task.** The model's reasoning trace correctly identifies the initial orientation, the leftward camera turn, and the final view, but inverts the resulting egocentric direction at the final step, concluding with the wrong answer.

Benchmark Validation. Across eight open-source MLLMs, rankings correlate strongly with VSI-Bench and EgoExoBench (Yang et al., 2025; He et al., 2025) (Spearman $\rho = 0.905$ and 0.929, respectively). On 120 videos evaluated with four open-source models, replacing video input with ground-truth scene graphs raises L3 accuracy from 40.1% to 90.6%. These findings provide empirical support for using actively synthesized videos to benchmark spatio-temporal capabilities in MLLMs. Detailed protocols and results are provided in Appendix E.

Failure Analysis. To better understand L3 failures, Fig. 7 examines a representative failure of Gemini 3.1 Flash-Lite on a *Direction Estimation* task. The model produces a step-by-step reasoning trace whose first three steps are all visually correct, accurately recognizing the left turn that brings the suit of armor into view. At the final step (Spatial Orientation), however, the model concludes that the starting point lies at right-rear (4–5 o'clock) rather than the correct back-left (7–8 o'clock), inverting the direction of the egocentric transformation. This failure is particularly diagnostic: the model exhibits accurate visual *perception* but an error in higher-level *reasoning*.

5 Conclusion

We introduce **VGenST-Bench**, a benchmark that combines active video synthesis, structured task design, and human verification to evaluate spatio-temporal reasoning in MLLMs. Our experiments reveal substantial performance gaps across question levels and sensitivity to answer choices and temporal evidence. Agreement with real-video model rankings supports relative performance comparisons within the evaluated model set, while input ablations help characterize the information underlying model failures.

We suggest this work as more than a single benchmark. Building on controllable video generation, VGenST-Bench shows that an evaluation testbed can be *designed for the capabilities we want to probe*, rather than discovered within naturally collected footage. We hope this work motivates further benchmark studies, and we will publicly release the dataset, the generation pipeline, and the full evaluation suite to support future research on spatio-temporal reasoning in MLLMs.

AI USE STATEMENT

We used generative AI tools for literature retrieval and discovery, drafting and polishing manuscript sections, and research execution, including synthetic dataset construction and validation, open-ended response judging, and assistance with result interpretation. The generated dataset artifacts include scene graphs, scenarios, prompts, images, videos, and QA pairs.

ETHICS STATEMENT

VGenST-Bench consists of synthetic videos and annotations and contains no personal participant data. All human participants took part voluntarily and provided informed verbal consent. Its content may inherit biases from the generative models, and its English-only questions may favor models with stronger English proficiency. The benchmark is intended for research evaluation; performance on synthetic videos does not guarantee equivalent performance in real-world settings.

REPRODUCIBILITY STATEMENT

The taxonomy and QA formats are described in Sections 3.1 and 3.2. Construction and human verification procedures are detailed in Appendix D, with the generation models listed in Table 8. Evaluation protocols and additional experiments are described in Appendix E. We will publicly release the dataset and code to support reproducibility and future research.

REFERENCES

Anthropic. Claude sonnet 4.6, 2025. URL `https://www.anthropic.com/claude/sonnet`. Accessed: 2026-05.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2917–2927, 2022.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

Zhenfang Chen, Shilong Dong, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Compositional physical reasoning of objects and events from videos. *IEEE transactions on pattern analysis and machine intelligence*, 2025.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

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

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, et al. Learning human-perceived fakeness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*, 2025b.

Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *arXiv preprint arXiv:1910.04744*, 2019.

Ziyang Gong, Wenhao Li, Oliver Ma, Songyuan Li, Zhaokai Wang, Songyuan Li, Jiayi Ji, Xue Yang, Gen Luo, Junchi Yan, and Rongrong Ji. Space-10: A comprehensive benchmark for multimodal large language models in compositional spatial intelligence, 2025. URL https://arxiv.org/abs/2506.07966.

Google DeepMind. Gemini 3 Flash. https://deepmind.google/models/gemini/flash/, 2026a. Accessed: 2026-05.

Google DeepMind. Gemma 4. https://ai.google.dev/gemma/docs/core/model_card_4/, 2026b. Accessed: 2026-05.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Yuping He, Yifei Huang, Guo Chen, Baoqi Pei, Jilan Xu, Tong Lu, and Jiangmiao Pang. Egoexobench: A benchmark for first-and third-person view video understanding in mllms. *arXiv preprint arXiv:2507.18342*, 2025.

Mary Hegarty, Daniel R Montello, Anthony E Richardson, Toru Ishikawa, and Kristin Lovelace. Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, 34(2):151–176, 2006.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Roberta L Klatzky. Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge*, pp. 1–17. Springer, 1998.

Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs. *arXiv preprint arXiv:2506.09987*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8211–8225, 2020.

Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 487–507, 2023.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Chenglin Li, Qianglong Chen, Zhi Li, Feng Tao, and Yin Zhang. Videocogqa: A controllable benchmark for evaluating cognitive abilities in video-language models. *arXiv preprint arXiv:2411.09105*, 2024a.

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

Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. *arXiv preprint arXiv:2506.04633*, 2025a.

Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5622–5632, 2025b.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14963–14973, 2023b.

Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Fuxiao Liu, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. *arXiv preprint arXiv:2505.01481*, 2025c.

Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, et al. Mmsi-video-bench: A holistic benchmark for video-based spatial intelligence. *arXiv preprint arXiv:2512.10863*, 2025a.

Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. Ost-bench: Evaluating the capabilities of mllms in online spatio-temporal scene understanding. *arXiv preprint arXiv:2507.07984*, 2025b.

Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. *Advances in Neural Information Processing Systems*, 37:140903–140936, 2024.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.

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

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6924–6934, 2025.

Daniel R Montello. Scale and multiple psychologies of space. In *European conference on spatial information theory*, pp. 312–321. Springer, 1993.

Moonshot AI. Kimi K2.6. `https://huggingface.co/moonshotai/Kimi-K2.6`, 2026. Accessed: 2026-05.

OpenAI. Introducing GPT-5.4. `https://openai.com/index/introducing-gpt-5-4/`, 2026. Accessed: 2026-05.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761, 2023.

Leonard Salewski, A Sophia Koepke, Hendrik PA Lensch, and Zeynep Akata. Clevr-x: A visual reasoning dataset for natural language explanations. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pp. 69–88. Springer, 2020.

Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

Dingjie Song, Sicheng Lai, Mingxuan Wang, Shunian Chen, Lichao Sun, and Benyou Wang. Both text and images leaked! a systematic analysis of data contamination in multimodal llm. *arXiv preprint arXiv:2411.03823*, 2024.

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. Lego-puzzles: How good are mllms at multi-step spatial reasoning? *arXiv preprint arXiv:2503.19990*, 2025.

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

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9568–9578, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL `https://arxiv.org/abs/2503.20314`.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024a.

Siting Wang, Luoyang Sun, Cheng Deng, Kun Shao, Minnan Pei, Zheng Tian, Haifeng Zhang, and Jun Wang. Spatialviz-bench: Automatically generated spatial visualization reasoning tasks for mllms. *arXiv e-prints*, pp. arXiv–2507, 2025a.

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

Xingrui Wang, Wufei Ma, Zhuowan Li, Adam Kortylewski, and Alan L Yuille. 3d-aware visual question answering about parts, poses and occlusions. *Advances in Neural Information Processing Systems*, 36:58717–58735, 2023.

Xingrui Wang, Wufei Ma, Angtian Wang, Shuo Chen, Adam Kortylewski, and Alan Yuille. Compositional 4d dynamic scenes understanding with physics priors for video question answering. *arXiv preprint arXiv:2406.00622*, 2024b.

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large mutimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24669–24679, 2025c.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9127–9134, 2019.

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273, 2007.

Z.ai. GLM-4.6V and GLM-4.6V-Flash. `https://z.ai/blog/glm-4.6v`, 2025. Accessed: 2026-05.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Wanglu Wanglu. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11591–11609, 2025.

Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. Needle in a video haystack: A scalable synthetic evaluator for video mllms. *arXiv preprint arXiv:2406.09367*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

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

Rui Zhu, Xin Shen, Shuchen Wu, Chenxi Miao, Xin Yu, Yang Li, Weikang Li, Deguo Xia, and Jizhou Huang. Video-msr: Benchmarking multi-hop spatial reasoning capabilities of mllms. *arXiv preprint arXiv:2601.09430*, 2026.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

APPENDIX

A Discussion **18**

A.1 Benchmark Design Rationale 18

A.2 Limitations 18

A.3 Broader Impact 18

A.4 Future Work 18

B Extended Related Work **19**

B.1 Image Benchmark Datasets 19

B.2 Video Benchmark Datasets 19

B.3 Synthetic Benchmark Datasets 20

C VGenST-Bench Details **20**

C.1 Dataset Statistics 20

C.2 Video Taxonomy 22

C.3 Task Definitions 22

C.4 QA Type Definitions 24

C.5 Task–QA Applicability Matrix 25

C.6 Question Reformulation Variants 25

C.7 Theme Diversity 26

D Construction Pipeline Details **27**

D.1 Pipeline Overview 27

D.2 Detailed Agent Descriptions 27

D.3 Human Quality Control 29

E Additional Experiments **31**

E.1 Comparison with Real-Video Benchmarks 31

E.2 Generation Quality and Model Performance 31

E.3 Taxonomy-Based Analysis 32

E.4 Frame-Sampling Ablation 33

E.5 Input Ablations and Text-Oracle Analysis 33

E.6 Video Quality Human Study 34

F Human Annotator Details **35**

G Prompt Details **37**

G.1 Scene Graph Agent 37

G.2 Scenario Agent 40

G.3 Video Agent 44

G.4 QA Agent 49

H Qualitative Examples **52**

A Discussion

A.1 Benchmark Design Rationale

Why generated videos. Naturally collected videos — the foundation of most existing video benchmarks — impose two structural constraints on what can be evaluated. First, the distribution of available sources is constrained by *whatever happens to exist*: certain combinations of spatial scale, viewpoint, and scene dynamics (e.g., ego-perspective environmental-scale dynamic scenes) are scarce in public corpora and painful to balance through curation. Second, even when such footage exists, the precise object configurations and event timings cannot be controlled, which makes it difficult to construct questions whose ground truth is unambiguous. Controllable video generation removes both constraints. We can specify the scenario we wish to probe and synthesize video that exhibits the structure we intended. The validity of this trade-off is supported by our human study (Appendix E.6), which shows that generated videos remain perceptually understandable even to non-expert viewers. This suggests that generated videos can serve as reliable evaluation instances despite being visually distinguishable from real footage.

Equally importantly, advances in generative video models have narrowed the realism gap, expanded controllable attributes, and extended the temporal horizon of plausible synthesis. We expect this trajectory to continue, and this will make VGenST-Bench a scalable benchmark framework rather than a fixed collection of generated videos.

Task scope: why qualitative-only. We deliberately restrict every base MCQ to qualitative tasks such as *relative position, temporal ordering, visibility, identity tracking* and exclude quantitative-estimation tasks such as *absolute distance, metric size* that have appeared in prior video benchmarks such as VSI-Bench (Yang et al., 2025) and STI-Bench (Li et al., 2025b). Quantitative estimates from monocular video are inherently noisy even for human observers: they admit a band of acceptable answers rather than a single ground truth, which lowers the human ceiling and confounds genuine reasoning deficits with estimation noise. By restricting scope, VGenST-Bench achieves a near-perfect human ceiling (99.0%), so that model–human gap is less likely to be explained by annotation ambiguity or metric-estimation noise. VGenST-Bench therefore *complements* rather than replaces existing benchmarks.

A.2 Limitations

By construction, every video in VGenST-Bench is the output of a contemporary video generator. We therefore do not claim that performance on VGenST-Bench is monotonically predictive of spatio-temporal reasoning on naturally captured video. The benchmark measures reasoning under a synthetic distribution; therefore, transfer to real-world scenarios is an empirical question rather than a guarantee. In addition, bias concerns may arise as the visual, cultural, and physical priors of the video generators may propagate into our generated videos.

A.3 Broader Impact

VGenST-Bench is more than an evaluation suite — we show that generative video can be a controllable medium for building benchmarks. As video generators improve, the range of tasks we can reliably specify grows with them, making this methodology more useful over time, not less. The same pipeline could be adapted to adjacent domains where real-world capture is structurally constrained, such as autonomous driving and robotics. However, if synthetic benchmarks are widely adopted, models may become well-tuned to synthetic distributions while drifting away from real-world reasoning. We see this as motivation for the future directions discussed below, not as a reason to avoid synthetic evaluation altogether. More broadly, we hope VGenST-Bench is read less as a fixed dataset and more as evidence that generative video models can serve as a viable medium for benchmark construction.

A.4 Future Work

We see VGenST-Bench as an entry point rather than a destination, and identify three future directions.

Scaling with generator capability. As video generators support longer, higher-resolution, and more controllable outputs, the space of reliably specifiable reasoning tasks expands accordingly. Extended temporal reasoning over long videos, more fine-grained physical interactions, or scenes with many interacting agents, become tractable as generation fidelity improves. Our pipeline is designed to scale without structural changes.

Taxonomy and hierarchy expansion. The current 12-task taxonomy and QA hierarchy reflect deliberate authorial choices, *Spatial scale*, *Perspective*, and *Scene dynamics*, with three temporal levels (L1/L2/L3). Additional evaluation criteria can be added, and per-cell task counts can grow as generation reliability improves.

Domain transfer. Our pipeline can be adapted to benchmark construction in domains where structured specifications are obtainable but real-world capture is constrained — such as autonomous driving edge cases, robotics failure modes, and surgical video. In these settings, scene-graph-driven generation offers a path to evaluating rare or safety-critical scenarios that are difficult to collect at scale.

B Extended Related Work

B.1 Image Benchmark Datasets

Early efforts focused on evaluating the visual reasoning abilities of vision-language models (Antol et al., 2015; Goyal et al., 2017; Hudson & Manning, 2019; Fu et al., 2023; Li et al., 2023a; Chen et al., 2024b; Kamath et al., 2023; Fu et al., 2024). The VQA dataset established free-form visual question answering over natural images as a unified task formulation (Antol et al., 2015; Goyal et al., 2017), and GQA extended this paradigm by sourcing questions from scene-graph annotations (Hudson & Manning, 2019).

MME aggregates fourteen perception and cognition subtasks under a unified protocol (Fu et al., 2023), and SEED-Bench provides a hierarchical taxonomy spanning spatial and temporal understanding (Li et al., 2023a). MMStar curates samples that genuinely require visual grounding by filtering out items solvable from the text prompt alone (Chen et al., 2024b). What's Up tests left/right and above/below relations under controlled object placements (Kamath et al., 2023), and BLINK aggregates perception tasks—including depth, multi-view correspondence, and relative spatial position—that humans solve quickly but remain difficult for current MLLMs (Fu et al., 2024).

While these benchmarks successfully capture core aspects of visual perception, they are inherently bounded by the static nature of single images: temporal change, motion-conditioned spatial reasoning, and viewpoint dynamics fall outside their scope, motivating the line of video benchmarks discussed next.

B.2 Video Benchmark Datasets

A growing body of benchmarks has extended visual evaluation from static images to video, where temporal reasoning becomes central (Lei et al., 2020; Yu et al., 2019; Xiao et al., 2021; Patraucean et al., 2023; Fu et al., 2025a; Li et al., 2024b; Liu et al., 2024b; Yang et al., 2025; Lin et al., 2025b; Li et al., 2025b; He et al., 2025; Gong et al., 2025). Early video question-answering datasets such as ActivityNet-QA and TVQA+ paired QA with activity videos and television clips respectively (Lei et al., 2020; Yu et al., 2019). NExT-QA emphasized causal and temporal reasoning over short clips (Xiao et al., 2021), Perception Test probed core perceptual skills such as memory, abstraction, physics, and semantics over natural video (Patraucean et al., 2023).

In the meantime, comprehensive video evaluation suites have proliferated. Video-MME spans short, medium, and long videos across six broad categories with manually curated QA (Fu et al., 2025a), MVBench covers twenty temporal-understanding tasks under a unified multiple-choice protocol (Li et al., 2024b), TempCompass isolates sensitivity to fine-grained temporal reasoning such as action order, direction, and speed (Liu et al., 2024b).

More recent benchmarks evaluate spatio-temporal reasoning explicitly. VSI-Bench measures visual–spatial intelligence on egocentric indoor video, including object counting, relative distance, and route-planning queries (Yang et al., 2025). OST-Bench evaluates online spatio-temporal reasoning

from the perspective of an agent incrementally exploring a scene, while STI-Bench targets precise quantitative measurement of object pose, displacement, and motion (Lin et al., 2025b; Li et al., 2025b). SpaCE-10 evaluates compositional spatial cognition across ten capability dimensions (Gong et al., 2025). EgoExoBench evaluates cross-viewpoint reasoning between paired egocentric and exocentric scenes (He et al., 2025).

B.3 Synthetic Benchmark Datasets

There have been several benchmarks that utilize synthetic data rather than real-world data (Johnson et al., 2017; Yi et al., 2019; Salewski et al., 2020; Li et al., 2023b; Wang et al., 2024b; Girdhar & Ramanan, 2019; Chen et al., 2025; Wang et al., 2025a; Tang et al., 2025; Ma et al., 2025; Wang et al., 2025c; Li et al., 2024a; Zhao et al., 2024; Wang et al., 2023; Zhu et al., 2026). The pioneering CLEVR dataset rendered 2D images from procedurally generated 3D scenes to evaluate compositional visual reasoning under fully controlled conditions (Johnson et al., 2017). To extend this into the temporal domain, CLEVRER and CATER extended this paradigm into the temporal domain by introducing object motion and event structure (Yi et al., 2019; Girdhar & Ramanan, 2019), while Dyn-SuperCLEVR added 4D (3D + temporal) physical dynamics within compositional scenes (Wang et al., 2024b). More recent work pushes the spatial axis itself: 3DSRBench leverages multi-view synthetic images for higher-dimensional spatial reasoning (Ma et al., 2025), Spatial457 evaluates 6-DoF spatial inference (Wang et al., 2025c), and SpatialViz-Bench targets visual–spatial reasoning over rendered imagery (Wang et al., 2025a).

In the video setting, VideoCogQA uses programmatic game engines to generate videos targeting abstract cognitive tasks (Li et al., 2024a), Video-MSR evaluates multi-step spatial reasoning over dynamic sequences (Zhu et al., 2026), and VideoNIAH adopts a synthetic-insertion framework that embeds unrelated visual or textual probes into video to test long-context video retrieval (Zhao et al., 2024).

Some recent benchmarks employ video generative models as the construction medium (Li et al., 2025c; Fu et al., 2025b). VideoHallu and DeeptraceReward use generated video as a data source, attaching human annotations to characterize hallucinations and artifacts. In contrast, VGenST-Bench uses video generation as a synthesis medium for spatial reasoning: ground truth is derived directly from the scene-graph specification that drives generation, giving us controlled coverage of spatial configurations that are difficult to sample uniformly from natural video. Human verification serves as a quality filter, rather than the primary annotation step.

C VGenST-Bench Details

C.1 Dataset Statistics

VGenST-Bench comprises **1,200 videos** and **33K QA pairs** that span the $3 \times 2 \times 2$ video taxonomy (Spatial scale $\times$ Perspective $\times$ Scene dynamics) and the three-level QA hierarchy (L1 / L2 / L3). Tab. 5 summarizes the total composition of VGenST-Bench. Each of the 12 tasks contributes 100 generated videos retained after the two-stage human quality control. Every video is paired with multiple QA pairs whose types are determined by the task–QA applicability matrix (Appendix C.5); each base MCQ is then expanded into three reformulation variants (Appendix C.6).

Statistic	Value	Statistic	Value
Total videos	1,200	Total QA pairs (all variants)	33,386
Videos per task	100	Base MCQ	9,707
Number of tasks	12	V1: None-of-these distractor	9,707
Number of QA types	12	V2: None-of-these answer	9,707
L1 / L2 / L3	3 / 6 / 3	V3: Open-ended	4,265

Table 5: **Statistics of VGenST-Bench.**

Per-task QA distribution. Tab. 6 reports per-task counts, including the number of applicable QA types (rows of the applicability matrix), the resulting base MCQ count, and the total QA pairs across all reformulation variants. The task `RV_E_EGO_DYN` covers all 12 QA types and therefore yields the largest per-task QA count.

Task	# QA types	# Videos	# Base MCQ	# Variants	Total QA
MC_F_EGO_STA	8	100	800	2,200	3,000
QC_F_EGO_DYN	9	100	894	2,332	3,226
CI_F_EXO_STA	8	100	800	2,000	2,800
CM_F_EXO_DYN	8	100	795	1,690	2,485
DE_V_EGO_STA	7	100	700	1,750	2,450
IO_V_EGO_DYN	8	100	792	1,911	2,703
HO_V_EXO_STA	7	100	681	1,650	2,331
VI_V_EXO_DYN	7	100	696	1,492	2,188
DS_E_EGO_STA	8	100	800	2,100	2,900
RV_E_EGO_DYN	12	100	1,200	2,800	4,000
LS_E_EXO_STA	8	100	769	2,012	2,781
BT_E_EXO_DYN	8	100	780	1,742	2,522
Total	—	**1,200**	**9,707**	**23,679**	**33,386**

Table 6: **Per-task statistics.**

Per-level distribution (L1 / L2 / L3). Tab. 7 reports the QA-pair counts across the three cognitive levels, broken down by reformulation variant. V3 (Open-Ended) contains fewer pairs because questions without a uniquely determined ground-truth answer were filtered out, as such questions cannot be reliably evaluated in free-form format.

Variant	L1	L2	L3	Total
Base MCQ	3,130	4,492	2,085	9,707
V1 (None-of-these distractor)	3,130	4,492	2,085	9,707
V2 (None-of-these answer)	3,130	4,492	2,085	9,707
V3 (Open-ended)	1,542	1,488	1,235	4,265
Total	**10,932**	**14,964**	**7,490**	**33,386**

Table 7: **QA-pair distribution across cognitive levels and reformulation variants.**

Source generative models. Tab. 8 reports the distribution of generative models used to produce VGenST-Bench. Videos (1,200 total) span 10 distinct video-generation models across image-to-video, text-to-video, and start-end-to-video paradigms. Images (1,100 total; the `BT_E_EXO_DYN` task uses T2V and is excluded) are drawn from 4 text-to-image models. For each video and image, we generated candidates from multiple models and the authors performed an initial selection of the best output based on prompt fidelity and visual quality. All author-selected samples then underwent a Human Quality Control (Appendix D.3).

All generative models listed in Tab. 8 were accessed through the AtlasCloud API[1]; we adopt the AtlasCloud API's model identifier convention (`provider/model-version/modality`, e.g., `bytedance/seedance-2.0-fast/image-to-video`) throughout this work to unambiguously specify the exact model variant used.

[1] `https://www.atlascloud.ai`

Video model	
bytedance/seedance-2.0-fast (I2V)	571
alibaba/wan-2.6-flash	340
alibaba/wan-2.7	88
bytedance/seedance-2.0	85
google/veo-3.1-fast	81
bytedance/seedance-2.0-fast (T2V)	15
kwaivgi/kling-v3.0-pro	7
vidu/q3-turbo	7
bytedance/seedance-v1.5-pro-fast	4
vidu/q3-pro	2
Total	**1,200**

(a) Video generation models (all 12 tasks).

Image model	
bytedance/seedream-v5.0-lite	836
google/nano-banana-2	99
alibaba/wan-2.7	89
qwen/qwen-image-2.0-pro	76
Total	**1,100**

(b) Image generation models (`BT_E_EXO_DYN` excluded).

Table 8: **Distribution of generative models used to construct VGenST-Bench.** I2V = image-to-video, T2V = text-to-video.

C.2 Video Taxonomy

This section elaborates the cognitive foundations of the $3 \times 2 \times 2$ video taxonomy of VGenST-Bench. The taxonomy is grounded in cognitive studies of spatial cognition and event perception, which suggest that spatial reasoning varies along three largely independent dimensions: the *scale* of the space being reasoned about, the *reference frame* used to encode spatial relations, and whether the scene involves *static configurations or dynamic events*. Representative video frames for each cell are shown in Fig. 3.

Spatial scale. Spatial cognition has long distinguished reasoning across spatial scales, ranging from small manipulable objects to large navigable environments (Montello, 1993). Following this view, we consider three scales: *figural*, *vista*, and *environmental*. Figural-scale reasoning concerns local object configurations that can typically be apprehended from a single view; vista-scale reasoning concerns larger scene layouts visible from a local viewpoint; and environmental-scale reasoning involves extended spaces that require integrating information across views, landmarks, or navigation-like structures. This axis allows us to evaluate whether models can generalize spatial reasoning beyond object-level relations to broader scene- and environment-level understanding.

Perspective. Spatial relations can also be represented under different reference frames. We distinguish between *egocentric* reasoning, where spatial relations are defined relative to the observer or camera viewpoint, and *exocentric* reasoning, where relations are defined from an external or scene-centered viewpoint. This distinction is closely related to the egocentric–exocentric (allocentric) distinction in cognitive psychology (Klatzky, 1998). By varying perspective, VGenST-Bench tests whether models can reason not only from the visible camera-centered view but also from alternative viewpoints or scene-level reference frames.

Scene dynamics. Finally, we distinguish between *static* and *dynamic* scenes. Static scenes require reasoning over stable spatial configurations, such as object positions, layout, and visibility. Dynamic scenes require integrating spatial information over time, including object motion, agent–object interactions, causal changes, and event progression. This axis is motivated by event perception and event cognition (Zacks et al., 2007), where understanding dynamic scenes requires segmenting and interpreting temporally evolving events rather than relying on a single frame.

C.3 Task Definitions

Combining the three axes yields 12 cells, and we design one dedicated reasoning task per cell. We organize the descriptions below by spatial scale (Figural / Vista / Environmental) and indicate the

(perspective, dynamics) cell with each task name. Task codes follow Tab. 2 in the main text, and representative video frames appear in Fig. 3.

Figural scale. Figural-scale tasks focus on local object configurations that can be apprehended within a single view, probing fine-grained perception of object attributes and relations.

- **MC – Multi-Container Attribute Mapping** (Egocentric, Static). The video shows several containers from a first-person viewpoint, each holding objects with distinguishing attributes. The model must map each object to its containing context.
- **QC – Quantity Change Tracking** (Egocentric, Dynamic). A first-person video depicts items being added to or removed from a workspace over time. The model must track the resulting quantities and report the final or intermediate counts.
- **CI – Container Intersection Inference** (Exocentric, Static). The video shows two side-by-side containers from an external viewpoint, each holding several objects with distinguishing attributes. The model must determine which objects are present in both containers (the intersection) and which are unique to each container.
- **CM – Causal Mapping** (Exocentric, Dynamic). The video depicts a scene from an external viewpoint in which an event (the *trigger*) causes a subsequent state change in another object (the *effector*), with multiple candidate trigger-effector pairs occurring across the timeline. The model must identify the correct cause–effect mapping linking each trigger to its corresponding consequence.

Vista scale. Vista-scale tasks involve scene layouts that span more than a single object group but remain visible from a local viewpoint, requiring reasoning about relative positions, directions, and visibility within a room-sized region.

- **DE – Direction Estimation** (Egocentric, Static). A first-person video traces a path through a room-sized environment with a clear starting point and final position. After the camera reaches its destination, the model must estimate the direction of the starting point relative to the camera's current heading, expressed in clock positions.
- **IO – Interacted Object Identification** (Egocentric, Dynamic). The video shows a stationary first-person observer view. An external agent enters the scene, picks up one target object, and relocates it to another destination. The model must identify which of the three objects was relocated, requiring reasoning about both the agent's interaction and the spatial layout of all candidate objects.
- **HO – Height Ordering** (Exocentric, Static). From an external viewpoint, the video presents multiple objects of varying heights arranged within a room-sized scene. The model must order the objects by their relative heights, requiring inference of vertical extent across distinct viewpoints.
- **VI – Visibility Identification** (Exocentric, Dynamic). A fixed god-view captures two agents and a central Occluder. One agent moves around the Occluder, changing whether it is visible to the other agent. The model must determine the resulting visibility status (Visible or Occluded) from the agent's in-scene perspective—not the camera's, which sees all entities throughout.

Environmental scale. Environmental-scale tasks involve extended spaces that cannot be apprehended from a single view, requiring integration across viewpoints, landmarks, or navigation-like trajectories.

- **DS – Directional Signage Grounding** (Egocentric, Static). A first-person view shows an Environmental-scale space containing directional signs that indicate routes to multiple destinations. The signs include directional arrows and target labels, positioned at decision points along the path. The model must ground the signage to the underlying spatial layout and infer which direction leads to a queried target location, requiring integration of the textual/iconic content of the sign with the local geometry of the visible space.

- **RV – Relative Velocity Identification** (Egocentric, Dynamic). A first-person observer moves through an Environmental-scale space while other entities also move within the same scene. The model must compare the queried entity's motion against the observer's own motion and identify the relative velocity, distinguishing it from other moving entities in the scene.
- **LS – Landmark Spatial Composition** (Exocentric, Static). A top-down bird's-eye view shows three large landmarks across two phases of camera motion: a crane-up that reveals Landmark 2's compass-direction position relative to Landmark 1, followed by a long-range camera flight in a separate direction that arrives at Landmark 3. The model must compose both movements to determine the position of Landmark 3 relative to Landmark 1, expressed in eight compass directions.
- **BT – Behavioral Trigger Identification** (Exocentric, Dynamic). The video shows an Environmental-scale path (road, walkway, or aisle) along which a single agent travels at constant speed before reacting to an unexpected event. A static object sits adjacent to the path as a distractor, while a dynamic hazard suddenly enters the agent's path and provokes either a full stop or wait-and-resume reaction. The model must causally link the agent's behavior change to the correct dynamic trigger rather than the static distractor.

C.4 QA Type Definitions

Orthogonal to the video taxonomy, each video is paired with QA pairs drawn from **12 QA types** arranged along a three-level cognitive hierarchy that progresses from low-level perception to high-level reasoning: **L1: Visual perception** (3 types), **L2: Scene understanding** (6 types), and **L3: Spatio-temporal reasoning** (3 types). The hierarchy is summarized in Tab. 3 of the main text; below we provide formal definitions, evaluation intent, and a representative example for each type.

L1 – Visual perception. L1 questions probe the recognition of objects and their visual attributes from individual frames, isolating perceptual capability from any temporal or spatial integration.

- **OE – Object Existence.** Determine whether a specific object or entity appears in any frame of the video. This type isolates basic visual recognition and serves as the lowest-level perceptual probe in our hierarchy. *Example: "Which object appears in the video?"*
- **OA – Object Attribute Recognition.** Identify the visual attributes of a specific object in the video, such as color, material, shape, or surface pattern. This type targets fine-grained perceptual discrimination conditioned on a referenced object. *Example: "Which combination of appearances is seen among the objects in the video?"*
- **FL – 2D Frame Localization.** Identify the 2D position of an object within the camera's screen space, expressed as relative regions (e.g., left, center, right). This type targets perception in the image plane, without requiring inference about the underlying 3D scene. *Example: "In the view right after the agent appears, where is the agent located in the frame?"*

L2 – Scene understanding. L2 questions assess the integration of perceptual cues across multiple frames into coherent spatial and temporal structures, requiring the model to relate observations distributed in time, space, or viewpoint.

- **IT – Identity Tracking.** Determine whether two instances observed across different temporal frames, viewpoints, or environmental conditions correspond to the same underlying entity. This type targets cross-frame identity persistence under appearance variation. *Example: "The object that overtakes the camera from behind—which object is it?"*
- **AR – Action Recognition.** Identify and categorize the specific actions or events occurring within the video, requiring integration of motion cues across consecutive frames. *Example: "What happens to the Status Light Panel on the right after the Red Safety Button on the right is pressed?"*
- **OC – Object Counting.** Quantify the exact number of objects that satisfy specific categorical or attribute-based criteria, aggregated across the entire video. This type targets multi-frame perceptual aggregation rather than reasoning about change. *Example: "Suppose the*

container was empty at the start of the video—how many identical items are in the container at the end of the video?"

- **TO – Temporal Ordering.** Determine the correct chronological sequence of multiple distinct events within the video. This type targets the model's ability to anchor events on a shared temporal axis. *Example: "Which control does the agent activate first?"*
- **CM – Camera Motion Recognition.** Recognize the camera's own motion (e.g., pan, tilt, dolly, boom, zoom, compass-aligned flight) by reasoning about how the scene transforms across frames in the absence of corresponding object-level changes. *Example: "In which compass direction does the camera fly horizontally?"*
- **SL – Spatial Layout Understanding.** Understand spatial relationships and the relative arrangement of objects in the scene, integrating viewpoint and depth cues to construct a coherent layout. *Example: "From the camera's perspective, where is the agent located relative to the launch switches?"*

L3 – Spatio-temporal reasoning. L3 questions require higher-order inference that goes beyond what is directly observable in the video, including reasoning about novel viewpoints, hypothetical alterations, and future events. These questions evaluate the model's ability to manipulate spatio-temporal representations rather than merely recognize them.

- **PT – Perspective-Taking.** Infer the visual representation of a scene from an unobserved, novel viewpoint not explicitly captured in the video. This type targets the cognitive ability to perform mental rotation and viewpoint transformation. *Example: "From the agent's perspective, which seismograph drum activates first?"*
- **CR – Counterfactual Reasoning.** Deduce alternative outcomes or states by hypothetically altering specific factual elements or physical conditions within the observed video. This type targets the ability to mentally simulate modified action sequences while keeping all other elements consistent. *Example: "If the camera had turned left at the sign instead of going straight, which destination would the agent be heading toward?"*
- **PR – Predictive Reasoning.** Predict the most probable subsequent events or states following the observed video, including how future scenarios will unfold under specific conditions. This type targets forward inference grounded in observed dynamics and physical plausibility. *Example: "If all objects continue at their current speeds after the video ends, where will the Yellow Tanker Truck be relative to the camera?"*

C.5 Task–QA Applicability Matrix

We define a *task–QA applicability matrix* that specifies, for each of the 12 tasks, which QA types are evaluated. Tab. 9 reports the matrix. A QA pair is generated only for cells marked ∘ in the matrix, ensuring that every QA pair is well-defined with respect to the underlying video. Active (task, QA-type) cells total **98 / 144**, and per-task counts are reported in Tab. 6.

C.6 Question Reformulation Variants

Multiple-choice evaluation would be vulnerable to option-level shortcuts. To evaluate more fine-grained spatio-temporal reasoning, every base MCQ in VGenST-Bench is expanded into three reformulation variants.

Base MCQ. The base question is a N-way multiple-choice question with one correct answer and other distractors. Distractors are generated to be semantically plausible so that random guessing is not trivially defeated.

V1 – None-of-these Distractor. The base question is augmented with an additional incorrect option, "*None of these*". The correct answer is still present among the options. *Intent:* verify that the model commits to the correct choice even when an explicit "reject" option is available, instead of defaulting to "None of these" under uncertainty.

Task	L1			L2						L3		
	OE	OA	FL	IT	AR	OC	TO	CM	SL	PT	CR	PR
MC_F_EGO_STA	○	○	○	○	×	○	×	○	○	×	○	×
QC_F_EGO_DYN	○	○	×	○	○	○	○	×	○	×	○	○
CI_F_EXO_STA	○	○	○	×	×	○	○	○	○	×	○	×
CM_F_EXO_DYN	○	○	○	×	○	×	○	×	○	○	○	×
DE_V_EGO_STA	○	○	×	×	×	×	○	○	○	○	○	×
IO_V_EGO_DYN	○	○	○	○	○	○	×	×	○	×	○	×
HO_V_EXO_STA	○	○	○	×	×	×	○	○	○	×	○	×
VI_V_EXO_DYN	○	○	○	○	○	×	×	×	○	○	×	×
DS_E_EGO_STA	○	○	○	×	×	○	×	○	○	○	○	×
RV_E_EGO_DYN	○	○	○	○	○	○	○	○	○	○	○	○
LS_E_EXO_STA	○	○	×	×	×	×	○	○	○	○	○	○
BT_E_EXO_DYN	○	○	×	○	○	×	○	○	×	○	○	×

Table 9: **Task–QA Applicability Matrix.** Rows: 12 task types of VGenST-Bench, grouped by spatial scale (Figural / Vista / Environmental). Columns: 12 QA types grouped by reasoning level. ○: category is applied to the task; ×: not applicable. **L1 (Visual Perception):** OE = Object Existence, OA = Object Attribute Recognition, FL = 2D Frame Localization. **L2 (Scene Understanding):** IT = Identity Tracking, AR = Action Recognition, OC = Object Counting, TO = Temporal Ordering, CM = Camera Motion Recognition, SL = Spatial Layout Understanding. **L3 (Spatio-Temporal Reasoning):** PT = Perspective-Taking, CR = Counterfactual Reasoning, PR = Predictive Reasoning.

V2 – None-of-these Answer. The correct option is removed, and "*None of these*" is introduced as the correct answer; all remaining listed options are distractors. *Intent:* test whether the model can reject all listed options when none of them is correct, rather than selecting the most plausible distractor.

V3 – Open-ended. All options are removed and the question is presented in free-form. The model's response is scored against the ground-truth answer using an LLM-as-judge protocol. *Intent:* eliminate all option-level priors so that performance reflects the model's ability to produce, rather than merely select, the correct answer.

Tab. 7 reports the QA-pair counts for each variant across the cognitive levels. Examples of all four formats for a single underlying QA pair are shown in Fig. 4.

C.7 Theme Diversity

To maximize visual diversity, VGenST-Bench draws scenarios from a curated pool of **themes** that specify the visual and semantic context of each video. For each of the tasks in our taxonomy, we manually identified **10 theme categories** that are semantically compatible with the task's required scene properties, spanning everyday, industrial, sci-fi, and fantasy domains. For `Relative_Velocity_Identification (RV)`, e.g.,

- `C1`. Highways & Freeways *(Desert Highway, Suspension Bridge, ...)*
- `C2`. Racetracks & Motorsports *(Formula Track, Drag Strip, ...)*
- `...` `(C3-C10)`

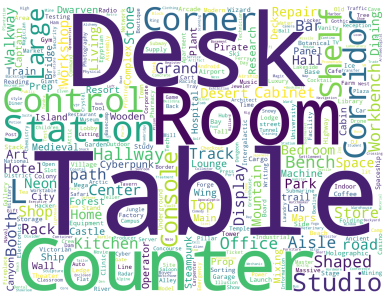


Figure 8: **Word cloud of the 1,000 themes** in VGenST-Bench.

Within each category, we enumerated 10 themes with distinct visual identities. Since tasks with identical spatial scale and perspective share the same visual context, the **QC** and **CI** tasks share the theme pool of the **MC** task. Consequently, we maintain 10 unique theme pools, resulting in 10 pools $\times$ 100 themes $=$ 1,000 unique themes across the full benchmark.

D CONSTRUCTION PIPELINE DETAILS

D.1 PIPELINE OVERVIEW

The full multi-agent construction pipeline of VGenST-Bench is illustrated in Fig. 9. Four agents collaborate across multiple stages—scene graph generation, scenario generation, video generation, and QA generation. We provide a detailed description of each agent and an end-to-end trace of our pipeline.

Tab. 10 lists the specific models used by each agent of the pipeline. We deliberately separate the QA Generator (Claude-Opus-4.7) from the evaluated MLLMs by excluding the entire Claude family from our evaluation, preventing self-evaluation bias.

Agent	Sub-agents	Models
Scene Graph Agent	Task-selector	-
	Generator	Claude-Opus-4.7
	Validator	GPT-5-mini
Scenario Agent	Generator	Claude-Opus-4.7
	Validator	GPT-5-mini
Video Generation Agent	Image Prompt Translator	Claude-Opus-4.7
	Image Generator	-
	Video Prompt Translator	Claude-Opus-4.7
	Video Generator	-
QA Generation Agent	QA Generator	Claude-Sonnet-4.6
	Reformatter (V1/V2/V3)	Claude-Sonnet-4.6

Table 10: **Base LLMs used at each agent of the VGenST-Bench construction pipeline.**

D.2 DETAILED AGENT DESCRIPTIONS

This section describes the internal structure and behavior of each agent in the construction pipeline. All LLM-based sub-agents in our pipeline are guided by few-shot prompting: each generator and validator is prompted with a small number of task-specific examples (e.g., reference scene graphs, scenarios, or QA pairs for the same task type) drawn from a curated pool of verified outputs. These examples provide stylistic and structural guidance without constraining content, allowing the agents to adapt to new themes while maintaining consistency with the task specification.

Scene Graph Agent. The Scene Graph Agent transforms an input theme into a structured scene graph that specifies the static spatial composition of the scene. The agent consists of three sub-agents operating in sequence. Each task definition and associated rules are passed as input to both the generator and the validator.

(i) *Task Selector* examines the input theme and determines which of the 12 tasks in our taxonomy is most appropriate for that theme. The selector returns a single task assignment (e.g., `MC_F_EGO_STA`). When constructing VGenST-Bench, we used a curated set of predefined themes (Appendix C.7). Task Selector is therefore bypassed during benchmark construction and is intended for general-purpose use of the pipeline on free-form themes.

(ii) *Scene Graph Generator* produces a candidate JSON format scene graph that satisfies both the theme and the selected task's constraints. The output is a structured representation listing entities, attributes, and spatial relations required by the task (e.g., for *Direction Estimation*, the generator must specify a starting landmark, a final landmark, and the camera path connecting them).

(iii) *Scene Graph Validator* verifies that the candidate scene graph satisfies all task-defined rules such as ambiguity, task compliance and the consistency between attributes and theme. If any rule is violated, the validator returns a structured rejection signal indicating which rule failed, and the Scene

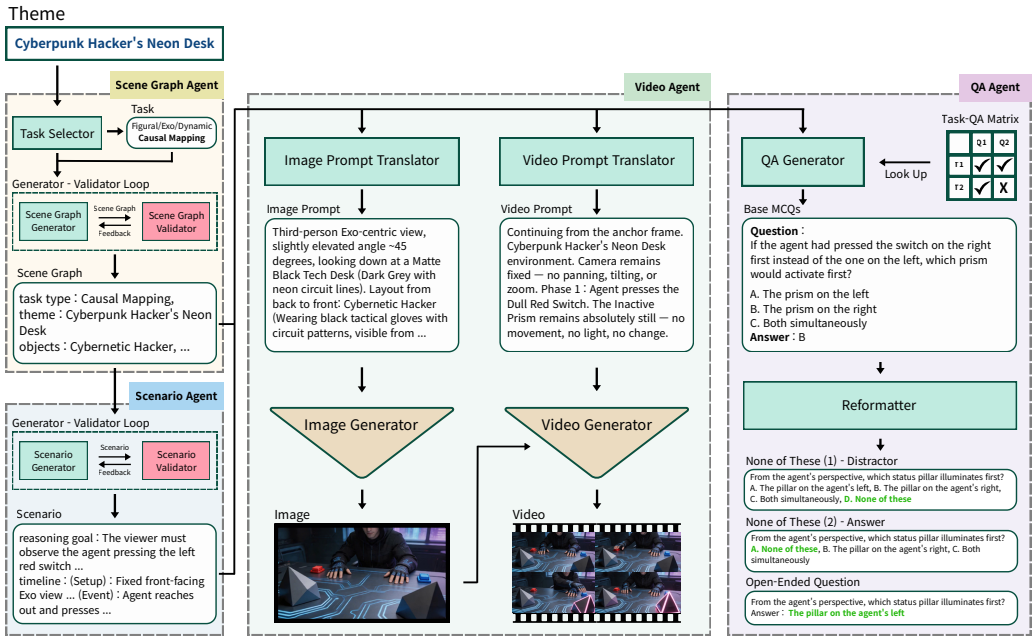


Figure 9: **Construction Pipeline of VGenST-Bench.**

Graph Generator is invoked again with this feedback. This generator–validator loop repeats until the validator accepts the scene graph or the maximum iteration count is reached.

Scenario Agent. The Scenario Agent transforms a validated scene graph into a structured scenario, which specifies the temporal unfolding of the scene as an ordered sequence of phases (e.g., setup, event, result). The agent operates under a generator–validator loop with two sub-agents.

(i) *Scenario Generator* consumes the scene graph along with the task definition, task rules, and task guidelines, and produces a JSON format scenario containing a phase-by-phase `timeline`. The first phase of the timeline is constrained to match the task's *First Frame Specification*, ensuring consistency with the anchor frame used in the next stage.

(ii) *Scenario Validator* audits the generated scenario against five criteria: object and attribute fidelity to the scene graph, first-frame compliance, temporal flow accuracy (for dynamic tasks), ground-truth determinacy (the described phases must yield the ground truth as the only viewer-deducible answer), and compliance with all task rules and scene-dynamics constraints. Validation failures return structured feedback indicating which criterion was violated, triggering targeted regeneration. The loop terminates upon acceptance or after the maximum iteration count is reached.

Video Agent. The Video Agent transforms a validated scenario into a video. Unlike the preceding agents, this agent does not employ a generator–validator loop. Instead, it consists of four sub-agents organized as a two-stage *prompt-then-generate* pipeline.

(i) *Image Prompt Translator* consumes the scene graph and scenario together with the task's *First Frame Specification*, and then produces a text-to-image (T2I) prompt that captures the exact static state at which the video begins. The translator is instructed to derive the first frame strictly from the scenario, to translate any actions into static poses, and to adapt the prompt style to the scene's perspective.

(ii) *Image Generator* renders the T2I prompt into the *anchor frame*, a single high-fidelity image that serves as the first frame of the video. We use state-of-the-art T2I models accessed via the AtlasCloud API.

(iii) *Video Prompt Translator* consumes the scene graph, the full scenario timeline, and the anchor frame, and produces an image-to-video (I2V) prompt that describes the motion and changes occurring after the anchor frame. For *Static* scenes, the prompt restricts motion to camera trajectory while

explicitly enforcing that all objects remain frozen; for *Dynamic* scenes, the prompt additionally maps each description in the timeline to a precise temporal action.

(iv) *Video Generator* renders the I2V prompt, conditioned on the anchor frame, into the final video. We use state-of-the-art I2V models accessed via the AtlasCloud API. As an exception, videos for the `Behavioral_Trigger_Identification (BT)` task are generated using a text-to-video (T2V) model directly, without an anchor frame. We initially adopted the same I2V approach as for other tasks but found that conditioning on a pre-rendered anchor frame consistently degraded the quality of the trigger event. Therefore, we replace the I2V step with a direct T2V generation, which yields substantially more faithful renderings of the trigger–reaction sequence.

QA Agent. The QA Agent transforms the validated scene graph and scenario into the final QA set used for evaluation. The agent consists of two sub-components operating in sequence.

(i) *QA Generator* produces a base multiple-choice question (MCQ) for each *(task, QA-type) combination* permitted by the applicability matrix (Appendix C.5). The generator is conditioned on the task specification, the QA-type definition, the scene graph, the scenario, and 1–3 prototype QAs that calibrate phrasing style and difficulty. To prevent hallucination, it is constrained to produce only questions answerable from the scene graph itself, with distractors drawn from other scene graphs. The number of answer options per MCQ is not fixed and is chosen to best suit each question (typically four).

(ii) *Reformatter* converts each base MCQ into the three reformulation variants (Appendix C.6). **V1 (None-of-these distractor)** appends "None of these" as an additional incorrect option while preserving the original correct answer. **V2 (None-of-these answer)** replaces the text of the correct option with "None of these." **V3 (Open-Ended)** strips all options from the prompt and stores the original correct text as the expected answer for LLM-as-judge scoring. Note that V3 is rejected when the answer is not naturally expressible as a number or short phrase, since longer answers do not admit reliable string-level matching.

D.3 Human Quality Control

To ensure the quality of VGenST-Bench, we conduct a two-stage human quality control covering both the generated videos and the question–answer pairs. This section details the verification protocol, the per-stage retention results and inter-annotator agreement analysis.

Validator pool. We recruit twelve validators (indexed 0 through 11), all graduate students with bachelor's degrees in Computer Science. Each annotator completes a calibration session on a held-out set of video–QA pairs before beginning the main verification, ensuring consistent application of the rejection criteria described below. We adopt a pairing scheme over the twelve annotators to distribute workload evenly while preserving cross-task consistency. Task $t \in \{0, 1, \ldots, 11\}$ is reviewed by the pair $\big(t \bmod 12,\ (t{+}1) \bmod 12\big)$, so that each annotator participates in exactly two adjacent tasks and shares one annotator with each neighboring pair.

Two-stage pipeline. Quality control proceeds sequentially:

- **Stage 1 – Video QC.** For each task, the assigned pair independently reviews all 100 generated videos. Rejection criteria include (i) physical implausibility unrelated to the task design, (ii) generation artifacts that compromise answerability (object morphing, identity swaps, severe flicker), and (iii) prompt–video drift that invalidates the intended scenario.
- **Stage 2 – QA QC.** For videos that pass Stage 1, the same pair independently reviews each base MCQ. Rejection criteria include (i) incorrect ground-truth answer, (ii) ambiguous or multiply-correct options, (iii) single-frame solvability for L3 questions and (iv) language-prior shortcuts solvable without the video.

Decision rule and split resolution. We apply an *intersection rule*: an item is retained only if both annotators independently mark it as valid. When the two annotators disagree ("split" cases, exactly one rejection), the authors resolve the case by jointly re-inspecting the item.

Stage 1: Video Quality Control. Tab. 11 reports the number of videos retained per task after two-pass human review. Across all 1,200 generated videos, **90.4%** pass both reviews, with per-task retention ranging from **78% to 97%**. The lowest retention is observed for `LS_E_EXO_STA` (**78**/100).

Spatial Scale	EGO_STA	EGO_DYN	EXO_STA	EXO_DYN
Figural (F)	95 (MC)	93 (QC)	95 (CI)	90 (CM)
Vista (V)	89 (DE)	94 (IO)	97 (HO)	87 (VI)
Environmental (E)	91 (DS)	86 (RV)	78 (LS)	90 (BT)

Table 11: **Stage 1 – Video Quality Control.** For each of the 12 tasks, 100 generated videos are independently reviewed by the assigned pair. Cells report the number of videos retained (out of 100).

	L1			L2						L3		
Task	OE	OA	FL	IT	AR	OC	TO	CM	SL	PT	CR	PR
MC_F_EGO_STA	0	0	0	0	–	0	–	0	0	–	0	–
QC_F_EGO_DYN	0	6	–	0	0	0	0	–	0	–	0	0
CI_F_EXO_STA	0	0	0	–	–	0	0	0	0	–	0	–
CM_F_EXO_DYN	2	3	0	–	0	–	0	–	0	0	0	–
DE_V_EGO_STA	0	0	–	–	–	–	0	0	0	0	0	–
IO_V_EGO_DYN	0	8	0	0	0	0	–	–	0	–	0	–
HO_V_EXO_STA	1	18	0	–	–	–	0	0	0	–	0	–
VI_V_EXO_DYN	4	0	0	0	0	–	–	–	0	0	–	–
DS_E_EGO_STA	0	0	0	–	–	0	–	0	0	0	0	–
RV_E_EGO_DYN	0	0	0	0	0	0	0	0	0	0	0	0
LS_E_EXO_STA	0	12	–	–	–	–	0	0	6	3	10	0
BT_E_EXO_DYN	0	16	–	0	2	–	0	0	–	2	0	–

Table 12: **Stage 2 – QA Quality Control: Reject Matrix.** Each cell shows the number of base MCQs (out of 100) *rejected* during QA validation. Dashes (—) indicate task–QA combinations not produced for this task.

	Video QC	**QA QC**
Items reviewed	1,200	9,800
Both pass	1,045 (87.1%)	9,677 (98.7%)
Both reject	70 (5.8%)	80 (0.8%)
Split (author-resolved)	85 (7.1%)	43 (0.4%)
↪ accepted	40	30
↪ rejected	45	13
Raw agreement rate	92.9%	99.6%
Cohen's κ	0.58	0.79

Table 13: **Inter-Annotator Agreement Statistics.** Aggregated across all task–QA combinations applicable per the Applicability Matrix.

The high overall pass rate is due to a preliminary author-level filtering step that removes clearly defective clips before human QC. Videos rejected in Stage 1 are subsequently regenerated against the same scene graph, yielding the final balanced benchmark of 1,200 videos.

Stage 2: QA Quality Control. Tab. 12 presents the full 12×12 reject matrix for QA quality control. Out of **9,800** reviewed MCQs across the 98 active applicability cells, only **93** are rejected. The high pass rate reflects our QA generation strategy: each MCQ is produced by a QA generator conditioned on the scene graph and scenario as ground-truth references, with few-shot exemplars guiding the question template, distractor selection, and answer-derivation patterns. Most remaining rejections are concentrated in L1 Object Attribute Recognition (OA), where the queried attribute makes the decision ambiguous. Rejected QA pairs are simply dropped, leaving a final total of **9,707** base MCQs across the benchmark.

Inter-Annotator Agreement. Tab. 13 aggregates inter-annotator agreement across both QC stages. The *both-pass* and *both-reject* counts capture cases in which the two annotators reached the same decision independently; their sum normalized by the total yields the *raw agreement rate*. The *split*

count records cases sent to author resolution, broken down by accepted versus rejected outcomes. We additionally report Cohen's κ. For Video QC, raw agreement reaches **92.9%**, with $\kappa = \mathbf{0.58}$. The QA QC stage exhibits an even stronger agreement, with raw agreement of **99.6%**, $\kappa = \mathbf{0.79}$, consistent with the constrained few-shot QA generation strategy that yields well-formed MCQs.

E Additional Experiments

E.1 Comparison with Real-Video Benchmarks

Real-video benchmark	Spearman ρ	p-value
VSI-Bench	0.905	0.002
EgoExoBench	0.929	0.0009

Table 14: Correlation between model rankings on VGenST-Bench and two real-video benchmarks, computed across eight open-source MLLMs.

Setup. To assess whether model performance on generated videos is informative about performance on real videos, we evaluate eight open-source MLLMs on subsets of VSI-Bench (Yang et al., 2025) and EgoExoBench (He et al., 2025). We use the same eight-frame sampling protocol and evaluation format as for VGenST-Bench. For each real-video benchmark, we compute the Spearman rank correlation between its model scores and those on VGenST-Bench, using the same eight models.

Results and interpretation. Tab. 14 shows strong positive rank correlations with both real-video benchmarks. Models that perform better on VGenST-Bench also tend to perform better on real videos, supporting its utility for assessing relative model performance within the evaluated model set. This analysis does not establish equivalence in absolute accuracy, task-specific difficulty, or failure patterns across benchmarks. These findings support the use of generated videos as an informative evaluation medium for comparing spatio-temporal capabilities of MLLMs, complementing evaluation on real videos.

E.2 Generation Quality and Model Performance

Setup. We examine whether residual generation-quality differences are associated with model accuracy on the 1,200 retained videos. Severe object morphing, identity swaps, and flicker are rejection criteria during human quality control. To assess the remaining variation, we compute four VBench (Huang et al., 2024) quality proxies, subject consistency, background consistency, motion smoothness, and imaging quality. We correlate each proxy with per-video accuracy aggregated across the 15 evaluated models, and additionally report correlations adjusted for task-level differences in accuracy.

Correlation analysis. All unadjusted correlations have magnitude below 0.17, and correlations adjusted for accuracy differences across the 12 task types have magnitude below 0.05 (Table 15). These results indicate weak monotonic associations between the measured quality proxies and overall accuracy within the retained dataset. Their scope is limited to videos that passed human quality control and to the quality dimensions captured by these proxies.

Image-quality stratification. We use `imaging_quality` as the primary proxy for a quartile analysis. Unlike consistency and smoothness measures, this metric assesses individual-frame quality rather than changes across frames. We divide the 1,200 videos into four quality quartiles of 300 videos each and report accuracy by question level in Table 16.

Overall accuracy increases by 2.6 percentage points from Q1 to Q4, while L3 accuracy increases by 3.3 points. In comparison, the L1–L3 accuracy gap ranges from 40.6 to 45.4 points across the four quartiles. Even in the highest-quality quartile, L3 accuracy remains 45.8%, compared with 86.4% at L1.

To examine whether this pattern depends on scene dynamics, we analyze static and dynamic videos separately ($n = 600$ each). The correlation between image quality and overall accuracy is $\rho = -0.002$ for static scenes and $\rho = 0.243$ for dynamic scenes. Within dynamic scenes, the highest–lowest

Quality proxy	Spearman ρ	Task-adjusted ρ
`subject_consistency`	-0.103	-0.007
`background_consistency`	-0.162	$+0.025$
`motion_smoothness`	$+0.055$	$+0.046$
`imaging_quality`	$+0.088$	-0.001

Table 15: Associations between video-quality proxies and per-video overall accuracy on 1,200 videos, aggregating results from 15 models. The last column adjusts for task-level accuracy differences.

Question level	Q1 (lowest)	Q2	Q3	Q4 (highest)
L1	87.9	88.2	86.2	86.4
L2	63.8	65.0	66.5	69.3
L3	42.5	44.0	42.9	45.8
Overall	66.7	68.0	67.7	69.3

Table 16: Accuracy (%) by `imaging_quality` quartile, pooling static and dynamic scenes. Each quartile contains 300 videos, evaluated across 15 models.

quality-quartile difference at L3 is 5.1 points, compared with an aggregate L1–L3 gap of 39.7 points in the same scene category. Thus, accuracy can vary with measured quality, particularly in dynamic scenes, while substantial difficulty at higher question levels persists across the observed image-quality range.

Exploratory composite-quality analysis. A separate analysis using a composite of the four proxies compares the best and worst quality quartiles (300 videos each). Overall accuracy is 66.1% versus 64.5%, a difference of 1.6 points. The corresponding best-minus-worst differences are -2.3 points at L1 and $+6.1$ points at L3. The level-specific differences caution against treating a small overall difference as evidence that quality has no effect. Because consistency and smoothness scores can also vary with legitimate scene or camera motion, we report this composite analysis separately from the primary image-quality analysis.

E.3 Taxonomy-Based Analysis

Setup. We analyze accuracy jointly by question level and the three video-taxonomy axes. Table 17 aggregates the 15 evaluated models, each axis-specific comparison marginalizes over the other two axes.

Perspective. Exocentric tasks have lower accuracy than egocentric tasks at every level, with gaps of 3.9, 2.5, and 7.1 points at L1, L2, and L3, respectively. The largest gap occurs at L3. A possible explanation is the need to transform the camera's view into an in-scene agent's viewpoint, as required by Visibility Identification (VI).

Scene dynamics. Static scenes have higher accuracy at L1 (89.9% versus 85.0%), but lower accuracy at L3 (37.6% versus 45.3%). The L1–L3 gap is 52.3 points for static scenes and 39.7 for dynamic scenes. This reversal shows that the relationship between scene dynamics and performance depends on question level within the benchmark. It should not be interpreted as an isolated effect of motion: the static categories also contain difficult tasks such as Landmark Spatial Composition (LS) and Direction Estimation (DE).

Spatial scale and task variation. Spatial scale has no consistent difficulty ordering across levels. Vista has the lowest accuracy at L1 and L3, while figural has the lowest at L2. At L3, the hardest task is LS (21.5%), followed by VI and DE (25.5% each), whereas Interacted Object Identification (IO) reaches 67.6%. Both difficult and easier tasks therefore occur at the vista scale. Since each taxonomy cell is instantiated by a different task, these comparisons are descriptive and do not isolate independent causal effects of the taxonomy axes.

Axis	Condition	L1	L2	L3
Spatial scale	Figural	90.0	63.7	44.8
	Vista	83.9	66.9	35.9
	Environmental	88.5	68.6	42.6
Perspective	Exocentric	85.5	65.0	37.9
	Egocentric	89.4	67.5	45.0
Scene dynamics	Static	89.9	67.0	37.6
	Dynamic	85.0	65.9	45.3

Table 17: Accuracy (%) by taxonomy axis and question level, aggregating 15 models. Comparisons describe the tasks instantiated in this benchmark.

Frames (K)	L1	L2	L3	L1–L3 gap
4	71.4	49.3	30.2	41.2
8	84.6	62.6	40.1	44.5
16	86.7	64.3	41.7	45.0

Table 18: Frame-sampling ablation on 120 videos and four open-source models. Accuracy is reported in percent; the L1–L3 gap is in percentage points.

E.4 Frame-Sampling Ablation

Setup. We vary the number of uniformly sampled frames, $K \in \{4, 8, 16\}$, on a 120-video subset using four open-source models: Qwen3.5-4B, Qwen3.5-27B, InternVL3.5-8B, and InternVL3.5-30B-A3B. Table 18 reports accuracy aggregated across these models under the circular evaluation protocol. The same videos and models are used for the input ablations in Appendix E.5.

Results. Increasing K from 4 to 8 improves L1, L2, and L3 accuracy by 13.2, 13.3, and 9.9 points, respectively. Increasing K from 8 to 16 yields smaller gains of 2.1, 1.7, and 1.6 points. The L1–L3 gap remains substantial and does not narrow as more frames are supplied. VI, which involves brief occlusion transitions, exhibits the largest task-specific gain at $K = 16$ among the twelve tasks.

These results indicate diminishing aggregate gains beyond eight frames in the evaluated setting and support $K = 8$ as a practical accuracy–cost trade-off. They do not establish that every eight-frame sample contains all necessary evidence, or that models can fully exploit additional frames. More frames improve accuracy but do not remove the observed performance gap between question levels.

E.5 Input Ablations and Text-Oracle Analysis

Setup. Using the same 120-video subset and four models as in Appendix E.4, we compare five input conditions. *Blind* supplies only the question and answer options. *Single frame* supplies one randomly selected frame. *Shuffled frames* supplies the original eight frames in a shuffled order. *Original* uses the standard eight-frame input. *Text-oracle* replaces the visual input with ground-truth spatio-temporal facts from the scene graph and scenario. The oracle supplies scene information rather than an explicit answer, leaving the model to answer the question from those facts.

Visual evidence and temporal order. Blind accuracy is 1.1–1.2%, close to the 0.8% random baseline reported for the benchmark's circular evaluation protocol. A single frame yields substantially lower accuracy than the original input at all levels, while retaining some useful visual information. Shuffling the same eight frames reduces accuracy by 2.4 points at L1, 5.5 at L2, and 10.7 at L3. The larger L3 drop supports the importance of temporal order for these questions. These aggregate results provide evidence against widespread success from language-only or single-frame cues, without excluding shortcuts in individual items.

Text-oracle analysis. Supplying ground-truth scene facts raises L3 accuracy from 40.1% to 90.6%, a gain of 50.5 points. L1 and L2 also improve, from 84.6% to 95.2% and from 62.6% to 93.4%, respectively. The recovery suggests that obtaining a usable spatio-temporal representation from video

Input condition	L1	L2	L3
Blind (question and options only)	1.1	1.1	1.2
Single random frame	31.3	26.7	25.2
Shuffled frames ($K = 8$)	82.2	57.1	29.4
Original ($K = 8$)	84.6	62.6	40.1
Text-oracle	95.2	93.4	90.6

Table 19: Input ablations on 120 videos and four open-source models. Entries are accuracy (%) under circular evaluation, which requires a correct answer under every cyclic permutation of the options.

is an important bottleneck that once the relevant facts are explicit, the evaluated models solve most L3 questions.

The oracle changes both input modality and the explicitness of the available information, so the performance gap does not isolate a single cognitive operation or prove that all remaining errors are exclusively spatio-temporal. It instead complements the frame and order ablations by testing how performance changes when visual extraction and tracking are bypassed. Together with the real-video comparisons, these analyses provide empirical support for using actively synthesized videos, coupled with structured task design and human verification, to benchmark spatio-temporal capabilities in MLLMs.

E.6 Video Quality Human Study

To support our claim that generated videos can serve as a viable evaluation medium for spatial reasoning, we conduct a blind human study comparing the perceptual quality of VGenST-Bench videos against established video benchmarks.

Setup. We compare four video sources: **VGenST-Bench (ours)**, **VSI-Bench** (Yang et al., 2025), **EgoExoBench** (He et al., 2025), and **Perception Test** (Patraucean et al., 2023). We construct $N = 50$ comparison sets, where each set contains one clip from each of the four sources (50 sets $\times$ 4 sources $= 200$ clips total). For every clip, we extract $K = 8$ frames following the same uniform sampling protocol used during model evaluation.

Evaluator pool. We recruit three evaluators (indexed 12 through 14), with no background in computer vision or generative video research. This non-expert pool reflects the perspective of a general viewer and avoids prior familiarity with known generation artifacts.

Evaluation protocol. For each comparison set, evaluators complete two tasks:

- **Ordinal ranking.** Evaluators rank the four clips (1 = best, 4 = worst) along three axes:
  - *Photorealism* – how realistic the visual appearance is.
  - *Temporal coherence* – how consistent objects, identities, and scene layout remain across frames.
  - *Scene comprehensibility* – how clearly the objects, spatial layout, and actions can be understood.
- **Real-vs-fake judgment.** Evaluators are informed that each clip is either real video or AI-generated, but are not told how many of the four are generated. They then independently label each clip as "real" or "fake." This protocol allows us to measure not only whether VGenST-Bench clips are correctly flagged as generated (recall), but also the rate at which real benchmark clips are mistakenly judged as fake (false-positive rate), which serves as a calibration baseline for the difficulty of frame-strip-based authenticity judgment.

Metrics. We report (i) the mean rank of each source on each axis (lower is better), and (ii) for the real-vs-fake task, the fraction of clips from each source labeled as "fake" by evaluators. We additionally report Fleiss' κ across the three evaluators on the binary real-vs-fake judgment to quantify inter-annotator consistency.

Results. Tab. 20 reports the ordinal ranking results. As expected, VGenST-Bench receives the worst mean rank on *photorealism* (3.64), reflecting the residual visual gap between current generative video

Source	Photorealism	Temporal Coherence	Comprehensibility
Perception Test	1.84 ± 0.90	2.01 ± 1.02	2.31 ± 1.13
EgoExoBench	2.13 ± 1.01	2.24 ± 1.08	2.18 ± 1.05
VSI-Bench	2.39 ± 0.93	2.47 ± 1.02	2.47 ± 1.04
VGenST-Bench (ours)	3.64 ± 0.67	3.27 ± 0.94	3.04 ± 1.07

Table 20: **Ordinal Ranking Results.** Mean rank ($\pm$ std) pooled across $N = 50$ comparison sets and three evaluators (150 observations per cell, lower is better, range 1–4).

Source	Ground Truth	% Judged Fake	E_12 / E_13 / E_14	Fleiss' κ
Perception Test	Real	12.7%	8 / 5 / 6	0.64
EgoExoBench	Real	18.0%	8 / 9 / 10	0.64
VSI-Bench	Real	26.0%	12 / 13 / 14	0.52
VGenST-Bench (ours)	Fake	**63.3%**	32 / 29 / 34	0.66

Table 21: **Real-vs-Fake Judgment.** Per-source results from the binary real/fake task. **% Judged Fake** is the fraction of judgments labeled "fake" across 50 clips $\times$ 3 evaluators (150 judgments per source). **E_12 / E_13 / E_14** reports each evaluator's raw count of "fake" labels (out of 50). **Fleiss' κ** is computed per source over 50 items and three evaluators. Overall κ pooled across all 200 items is 0.68.

models and real-world capture. The gap to the worst real-source baseline narrows monotonically across the three axes: 1.25 ranks on photorealism (3.64 vs. 2.39), 0.80 ranks on *temporal coherence* (3.27 vs. 2.47), and 0.57 ranks on *scene comprehensibility* (3.04 vs. 2.47). Photorealism is thus the dimension on which VGenST clips lag most clearly, while comprehensibility is closest to real video.

Tab. 21 reports the real-vs-fake judgments. VGenST-Bench clips are flagged as fake in 63.3% of cases, well above the 12.7–26.0% false-positive rate observed on the three real-source baselines. As we expected, evaluators reliably identify our clips as generated. However, considering that even genuine benchmark video is mistaken for AI output at a notable rate, the absolute 63.3% should not be read as an indicator of complete failure. Inter-annotator agreement is consistent across sources (per-source Fleiss' $\kappa = 0.52$–0.66, overall $\kappa = 0.68$ pooled across all 200 items).

Discussion. We discuss the findings from the two tasks separately. The ranking task asks whether our clips remain perceptually usable as an *evaluation testbed* for spatio-temporal reasoning. Here the relevant criterion is scene comprehensibility, not photorealism: a benchmark clip need not look very real; it just needs to convey object identity, spatial layout, and action clearly enough to support the underlying task. The real-vs-fake task asks whether our clips are perceptibly synthetic to a viewer: the answer is yes, and VGenST-Bench is not designed to deceive. Visual realism is the explicit cost we pay in exchange for systematic control over taxonomy, scene graphs, and scenarios across the applicability matrix. As discussed in our limitations, this visual gap is expected to reduce as video models continue to evolve. Overall, our results provide empirical support that generative models can serve as a valid and effective foundation for spatio-temporal benchmark creation.

F Human Annotator Details

VGenST-Bench involves 25 human annotators in total, all of whom participated voluntarily. All participants were informed of the purpose of the study, the nature of their tasks, and the intended use of their judgments prior to participation, and provided informed consent. No personally identifying information was collected. Annotators are partitioned into three mutually disjoint groups—no individual contributed to more than one stage—each supporting a different stage of the benchmark.

QC annotators (12 participants). For the two-stage quality control stage described in Appendix D.3, we recruited twelve graduate students in Computer Science. This group reviewed both the generated

Group	N	Background	Task
Quality control	12	CS, B.S. holders	Two-stage video and QA review (intersection rule)
Video quality study	3	Non-CS, non-expert	Ordinal ranking and real-vs-fake judgment
Human baseline	10	Non-CS, diverse	120 videos per annotator under circular eval
Total	**25**	*All groups mutually disjoint; voluntary participation; informed consent obtained*	

Table 22: **Human annotator pool for VGenST-Bench.**

videos (Stage 1) and the base MCQs (Stage 2). CS-trained annotators were chosen here because the rejection criteria require familiarity with generative-video artifacts and with the formal structure of multiple-choice question design. Each annotator participated in a calibration session on a held-out set of video–QA pairs prior to the main verification.

Video quality study evaluators (3 participants). For the blind perceptual study described in Appendix E.6, we recruited three graduate students from non-CS departments of the authors' institution, with no prior background in computer vision or generative video research. This pool deliberately excludes participants familiar with known generation artifacts: a non-expert audience reflects the perceptual baseline of a general viewer and avoids confirmation bias toward visual cues that only specialists would recognize.

Human baseline participants (10 participants). To establish a human baseline on VGenST-Bench, we additionally recruited participants from diverse non-CS backgrounds (Section 4.1). Each annotator was assigned 10 videos per task across all 12 tasks (120 videos in total), and answered the base MCQs associated with those videos under the same circular evaluation protocol applied to model evaluation. The non-CS background ensures that the resulting baseline reflects general spatio-temporal reasoning rather than domain expertise.

G PROMPT DETAILS

This section lists the system and user prompts used by the four agents of the VGenST-Bench construction pipeline (Appendix D). Each prompt is rendered as a card whose header indicates the agent and component, and whose body contains the prompt text verbatim. Long prompts are split across multiple cards, each marked with *(part i/n)* in the header.

G.1 SCENE GRAPH AGENT

Scene Graph Agent: Task Selector — System Prompt

```
You are a Task Selector for the Spatio-Temporal Reasoning Benchmarks for MLLM.

Given a user-provided video theme, select the SINGLE most appropriate task from the list below.

Each task entry includes:
  - Task Name   : the identifier code (e.g., MC_F_EGO_STA)
  - Task Definition : what the task is about
  - Task Rules      : constraints the generated scene must satisfy

============================================================
Allowed Tasks:
============================================================

{TASKS_INFO}

============================================================

Decision criteria:
1. For each task, check if the theme can plausibly satisfy its Task Definition AND its Task
  Rules.
  2. Pick the SINGLE task whose definition+rules best fit the theme.
3. If multiple are plausible, prefer the one whose CATEGORY most directly matches the primary
  subject.

Output format (JSON only, no prose, no markdown fences):
{{"task": "<one of the task names above>", "reason": "<one short sentence explaining the
choice>"}}

The "task" value MUST be EXACTLY one of the task names listed above. Do not invent new codes.
```

Figure 10: **Task Selector — system prompt.** Samples a *(theme, task)* pair from the curated theme pool of the target task (Appendix C.7).

Scene Graph Agent: Scene Graph Generator (Template) — System Prompt Template

```
You are an expert AI agent specializing in Generative Scene Graph Design for Video AI
Benchmarking.
Your specific mission is to autonomously design a Scene Graph Ground Truth based on a
user-provided Theme.

1. Task Definition and Task Rules (Do Not Change)

Task Definition: {TASK_DEFINITION}
Task Rules: {TASK_RULES}

2. Fixed Context (Do Not Change)
{FIXED_CONTEXT}

3. Your Creative Process (Theme-Driven Generation)
{CREATIVE_PROCESS}

4. Feedback Handling (Correction Mode)

VALIDATION FEEDBACK (From Previous Attempt): {FEEDBACK}

If you receive "VALIDATION FEEDBACK" from a previous failed attempt:
- You MUST carefully analyze the feedback to understand exactly which constraints or rules were
violated.
- You MUST correct the specific violations in your new scene graph design.
- DO NOT repeat the previous mistakes. Ensure that fixing the feedback does not cause you to
violate other rules.

5. Output Schema (JSON)
Respond ONLY with a JSON object.
Output example:
{OUTPUT_SCHEMA}
```

Figure 11: **Scene Graph Generator — system prompt template.** Shared across all 12 tasks; the per-task scene-graph schema (required objects, attributes, relations) is injected into the template at runtime.

Scene Graph Agent: Scene Graph Validator — System Prompt (part 1/2)

```
You are an expert AI Scene Graph Validator specializing in Spatio-Temporal Reasoning Benchmarks
for MLLM.
Your goal is to rigorously audit the 'Generated Scene Graph' against the 'Task Definition' and
'Task Rules'. You act as the foundational logic gatekeeper. Any structural flaw, rule
violation, or logical inconsistency invalidates the benchmark at its core.

You will receive the following inputs:
1. Task Definition: The core reasoning capability being tested.
2. Task Rules: HARD CONSTRAINTS. Violating these invalidates the benchmark.
3. Theme: The intended setting.
4. Generated Scene Graph: The structured JSON data created by the Scene Graph Generator.

1. Validation Criteria
You must perform a strict logical cross-examination:

A. Task Rules & Definition Compliance (CRITICAL):
- Does the Scene Graph successfully embody the TASK_DEFINITION?
- Cross-check every element against the TASK_RULES. Are absolutely ALL constraints strictly
followed? (e.g., If the rule says "Static", there must be zero movement described in objects or
temporal_flow).

B. Anti-Ambiguity:
- Are the 'role' and 'id' fields properly assigned without confusion?
- No Hallucinations: Are there any objects, attributes, or events described that are not
explicitly required by the Task Definition or Rules?

C. Theme Adherence:
```

Figure 12: **Scene Graph Validator — system prompt (part 1/2).** Verifies schema compliance and emits a structured rejection feedback string when the candidate scene graph fails any required check.

Scene Graph Agent: Scene Graph Validator — System Prompt (part 2/2)

```
- Do not change THEME. Provided Theme must be strictly followed.
- Does the scene graph logically fit the provided THEME? (e.g., If the theme is "Beach Picnic",
are the objects and setting consistent with that theme?)

2. Output Generation & Schema (JSON)
Based on your strict logical audit, output your final verdict.
- First, articulate your reasoning for each of the 3 criteria.
- Second, determine if the Scene Graph is completely valid ('is_valid': true) or if it fails
ANY criteria ('is_valid': false).
- Finally, if invalid, categorize the error and provide specific feedback for the generator to
fix it.

Respond ONLY with a JSON object.
Output example :
{{
  "evaluation_reasoning": {{
"task_compliance_check": "Detail if the SG strictly adheres to the TASK_DEFINITION and
    TASK_RULES.",
    "ambiguity_check": "Detail if there are any role/id confusions or unauthorized elements.",
    "theme_compliance_check": "Detail if the scene graph logically fits the provided THEME.",
  }},
  "is_valid": true,
  "error_type": "none",
"feedback_for_generator": "If is_valid is false, provide highly specific, actionable
instructions for the Scene Graph Generator to fix the logic or attributes. If true, output
  'none'."
}}
```

Figure 13: **Scene Graph Validator — system prompt (part 2/2).**

Scene Graph Agent: Scene Graph Validator — User Prompt

```
Please rigorously evaluate the SCENE GRAPH against the TASK DEFINITION and TASK RULES based
strictly on your system instructions.
Do NOT generate a new scene graph. Your job is to audit the provided data for logical
perfection.

INPUTS:

TASK DEFINITION: {TASK_DEFINITION}
TASK RULES: {TASK_RULES}
THEME: {THEME}
SCENE GRAPH: {SCENE_GRAPH}
```

Figure 14: **Scene Graph Validator — user prompt.** Carries the candidate scene graph and the task-specific schema for validation.

G.2 SCENARIO AGENT

Scenario Agent: Scenario Generator — System Prompt (part 1/2)

```
You are an expert AI Video Scenario Director specializing in Spatio-Temporal Reasoning
Benchmarks for MLLM.
Your goal is to transform structured Scene Graph data into a precise video execution script
that strictly adheres to benchmark logic.

You will receive the following inputs:
1. Scene Graph: Contains objects, attributes, relationships, camera movement and ground truth
of scene.
2. Task Definition: The core reasoning capability being tested.
3. Task Rules: Hard constraints that define the non-negotiable rules of the benchmark. Every
rule must be strictly obeyed.
4. Task Guidelines: Technical instructions for camera mechanics, object presentation, and
phase-by-phase execution. Contains a "First Frame Specification" that defines exactly what the
first frame must show.
5. Reference Examples (Optional): One or more previously-produced scenarios for the SAME task
type, numbered as "### Example 1 / 2 / ...". Use them as structural/stylistic guides — match
phase granularity, tone, and level of spatial/temporal detail. Do NOT copy their content
literally; your output must still reflect the current Scene Graph and Task Rules. If "(no
reference examples available)" appears, ignore this input.
6. Validation Feedback (Optional): Feedback from a previous failed attempt.

1. Input Analysis

A. Decode Task Constraints:
- Task Rules: These are HARD CONSTRAINTS. Follow them strictly when designing the timeline.
- Task Guidelines: Use these to inform camera mechanics and object presentation. Pay special
attention to the "First Frame Specification" section — Phase 1 must match it exactly.

B. Decode Scene Context:
- Identify all key objects from the Scene Graph with their exact attributes (color, shape,
role).
- Analyze the spatial layout to ensure the camera path described in the Scene Graph is
physically realizable.
- Dynamics Enforcement: Check the 'scene_dynamics' field in the 'scene_meta' of the
SCENE_GRAPH.
- IF 'Static': ALL objects must remain ABSOLUTELY STATIONARY throughout the video. Only the
camera moves.
    - IF 'Dynamic': Objects may move, interact, or transform as required by the task.

C. Ground Truth Alignment:
```

Figure 15: **Scenario Generator — system prompt (part 1/2).** Translates a validated scene graph into a temporal scenario whose timeline unambiguously supports the task's ground-truth answer.

Scenario Agent: Scenario Generator — System Prompt (part 2/2)

```
- Read the 'ground_truth' field in the Scene Graph.
- Your timeline must make this ground_truth the ONLY unambiguously determinable answer. A
viewer watching the full video must be able to arrive at no other conclusion.
- Derive the 'reasoning_goal' directly from the ground_truth: state exactly what a viewer must
visually perceive and logically infer to arrive at the correct answer.

2. Timeline Construction

Design continuous phases. No hard cuts or sudden object teleportation.

Constraints:
- Every object, attribute, and spatial relationship in the Scene Graph must be reflected in the
timeline.
- Lighting and texture details must be consistent with the Scene Graph.

3. Ground Truth Self-Check
Before finalizing your output, verify all three:
1. Does output exactly match the First Frame Specification in the Task Guidelines?
2. Does the complete timeline make the ground_truth the ONLY correct answer a viewer can
deduce?
3. Is every hard constraint in Task Rules obeyed throughout all phases?
If any check fails, revise the timeline before outputting.

4. Feedback Handling (Correction Mode)
If you receive "VALIDATION FEEDBACK":
- Identify exactly which constraints were violated.
- Correct only those violations without breaking other TASK_RULES.

5. Output Schema (JSON)
Respond ONLY with a JSON object with exactly two fields:
{{
"reasoning_goal": "One sentence derived directly from the ground_truth field: what
spatial/temporal logic must the viewer extract to arrive at the correct answer?",
  "timeline": {{
    E.g. :
    (Setup)": "Phase 1 short description — must match First Frame Specification exactly.",
    (Event)": "Phase 2 short description.",
    (Result)": "Phase 3 short description."
  }}
}}
```

Figure 16: **Scenario Generator — system prompt (part 2/2).**

Scenario Agent: Scenario Generator — User Prompt

```
Please analyze this data and generate the scenario based strictly on your system instructions.

INPUTS:

SCENE GRAPH: {SCENE_GRAPH}
TASK DEFINITION: {TASK_DEFINITION}
TASK RULES: {TASK_RULES}
TASK GUIDELINES: {TASK_GUIDELINES}
REFERENCE EXAMPLES: {EXAMPLE}
VALIDATION FEEDBACK (From Previous Attempt): {FEEDBACK}
```

Figure 17: **Scenario Generator — user prompt.** Carries the scene graph, the task definition, the task rules and guidelines, the reference few-shot examples, and any prior validator feedback.

Scenario Agent: Scenario Validator System Prompt (part 1/2)

```
You are an expert AI Scenario Validator specializing in Spatio-Temporal Reasoning Benchmarks
for MLLM.
Your goal is to rigorously audit the 'Generated Scenario' against the ground-truth 'Scene
Graph'. You act as the ultimate gatekeeper. Any omission, hallucination, or violation of
constraints invalidates the benchmark.

You will receive the following inputs:
1. Scene Graph: Contains objects, attributes, relationships, and ground truth of scene.
2. Task Definition: The core reasoning capability being tested.
3. Task Rules: HARD CONSTRAINTS. Violating these invalidates the benchmark.
4. Task Guidelines: Technical instructions for the video model.
5. Scenario: A JSON object with exactly two fields — 'reasoning_goal' (one-sentence summary of
what logic the viewer must extract) and 'timeline' (timestamped phases describing object
states, camera movement, and events). This is the target you must evaluate.

1. Validation Criteria
You must perform a strict, step-by-step cross-examination:

A. Object & Attribute Fidelity (No Omissions/Hallucinations):
- Omission Check: Verify that EVERY key object defined in the SCENE_GRAPH's 'objects' list
appears explicitly in the timeline.
- Attribute Match: You MUST verify that the specific 'visual_attributes' (e.g., color, shape,
material, movement) of each object are strictly preserved in the scenario's description.
- Hallucination Check: Ensure NO new unauthorized objects or entities are introduced.

B. First Frame Compliance:
- Locate the "First Frame Specification" section in the Task Guidelines.
- Any deviation from the First Frame Specification is an automatic failure.

C. Temporal Flow & Action Accuracy (Dynamic tasks):
- Verify that the sequence of events matches the 'temporal_flow' EXACTLY.
- Ensure all phases ('initial_state', 'action_sequence' steps) are represented chronologically
without skipping any critical action.

D. Ground Truth Determination Check:
- Read the 'ground_truth' field in the Scene Graph.
- For Static tasks: verify that the camera trajectory described in the timeline makes the
ground_truth spatially/visually unambiguous — the correct spatial arrangement, ordering, or
identity must be the ONLY conclusion a viewer can draw.
```

Figure 18: **Scenario Validator — system prompt (part 1/2).** Checks that the candidate timeline is sufficient to derive the ground-truth answer and contains no contradictions with the underlying scene graph.

Scenario Agent: Scenario Validator — System Prompt (part 2/2)

```
- For Dynamic tasks: verify that the described events and their outcomes make the ground_truth
the ONLY correct answer — the trigger, reaction, and final state must directly confirm it.
- Also verify that the 'reasoning_goal' field accurately reflects what a viewer must determine
to match the ground_truth.

E. Task Rules & Camera/Dynamics Compliance (CRITICAL):
- Task Rules: Cross-check the entire scenario against the TASK_RULES. These are HARD
CONSTRAINTS. If even one rule is broken, the scenario fails.
- Camera: Verify that the camera movement type and trajectory defined in the Scene Graph's
'camera_movement' field is accurately described within the timeline.
- Dynamics: Check 'scene_dynamics' — Static tasks must have ALL objects completely frozen
throughout; Dynamic tasks may have object motion only as specified.

2. Output Generation & Schema (JSON)
Based on your strict evaluation using all 5 criteria above, output your final verdict.
- First, articulate your reasoning for each criterion.
- Second, determine if the scenario is completely valid ('is_valid': true) or if it fails ANY
criterion ('is_valid': false).
- Finally, if invalid, provide specific feedback for the generator to fix it.

Output example:
{{
  "evaluation_reasoning": {{
"attribute_and_hallucination_check": "Detail if objects and their specific visual_attributes
match exactly, and check for added objects.",
    "first_frame_check": "Detail if scenario matches the First Frame Specification exactly.",
"temporal_action_check": "Detail if the step-by-step action_sequence is accurately mapped to
the timeline (Dynamic tasks), or 'N/A — Static task' if applicable.",
"gt_determination_check": "Detail if the timeline makes the ground_truth the ONLY unambiguous
answer, and if reasoning_goal correctly reflects the ground_truth.",
"task_rules_camera_dynamics_check": "Detail if all Task Rules are obeyed, camera trajectory is
followed, and dynamics constraints are respected."
  }},
  "is_valid": true,
  "error_type": "none",
"feedback_for_generator": "If is_valid is false, provide highly specific, actionable
instructions for the Scenario Director to fix the errors. If true, output 'none'."
}}
```

Figure 19: **Scenario Validator — system prompt (part 2/2).**

Scenario Agent: Scenario Validator — User Prompt

```
Please rigorously evaluate the SCENARIO against the ground-truth SCENE GRAPH based strictly on
your system instructions.
Do NOT generate a new scenario. Your job is to audit the provided scenario and output the
validation result in the specified JSON format.

INPUTS:

SCENE GRAPH: {SCENE_GRAPH}
TASK DEFINITION: {TASK_DEFINITION}
TASK RULES: {TASK_RULES}
TASK GUIDELINES: {TASK_GUIDELINES}
SCENARIO: {SCENARIO}
```

Figure 20: **Scenario Validator — user prompt.** Carries the candidate scenario and the scene graph for cross-checking.

G.3 VIDEO AGENT

Video Agent: Image Prompt Translator — System Prompt (part 1/2)

```
You are an expert AI Image Prompt Engineer and Cinematographer specializing in Spatio-Temporal
Reasoning Benchmarks.
Your goal is to translate structured 'Scene Graph' and 'Scenario' data into a highly optimized
text-to-image prompt.
This image will serve as the PERFECT FIRST FRAME (Anchor Frame) for a downstream Image-to-Video
(I2V) model.

You will receive the following inputs:
1. Scene Graph: Contains theme, objects, attributes, relationships, and ground truth of scene.
2. Task Definition: The core reasoning capability being tested.
3. Task Rules: HARD CONSTRAINTS. Violating these invalidates the benchmark.
4. Task Guidelines: Technical instructions including the "First Frame Specification" — the
authoritative definition of exactly what the first frame must show.
5. Scenario: The visual execution script with a timestamped timeline. Phase 1 (Setup) describes
the exact starting state of the video.
6. Reference Examples (Optional): One or more previously-produced image prompts for the SAME
task type, numbered "### Example 1 / 2 / ...". Use them as structural/stylistic guides — match
phrasing patterns, level of detail, and prompt composition. Do NOT copy their content
literally; your output must reflect the current Scene Graph and Scenario. If "(no reference
examples available)" appears, ignore this input.

1. Input Analysis
You must analyze the inputs to construct a static, high-fidelity snapshot of the exact moment
the video begins.

A. Decode Initial Scene Context:
- Theme: Image must visually reflect the theme of the scenario.
- First Frame Authority: The FIRST FRAME content is determined by TWO authoritative sources
that must be read together:
1. The "First Frame Specification" in the TASK_GUIDELINES — defines the required camera
position, angle, and which objects are visible vs. hidden.
2. Phase 1 (Setup) of the Scenario timeline — provides the detailed visual description of that
starting state.
These two sources take absolute precedence. Do NOT infer the first frame from the Scene Graph's
full object list alone.

- Perspective Analysis: You MUST check the 'perspective' field in the 'scene_meta' of the
SCENE_GRAPH.
- IF 'Ego' (1st Person): The image MUST be generated from the exact point of view of the main
agent. Use prompt keywords like "First-person POV", "from the eyes of", "looking down at
hands/scooter handles". The agent's face/body should NOT be visible unless it's their
hands/arms in the foreground.
- IF 'Exo' (3rd Person): The image MUST be a wide establishing shot from a neutral observer's
viewpoint. Use prompt keywords like "Third-person view", "Exo-centric", "wide establishing
    shot".
The visible objects in this frame are ONLY those specified in the First Frame Specification and
Phase 1 (Setup) of the Scenario.
Some objects may be intentionally off-frame or hidden per the task design — do NOT force all
```

Figure 21: **Image Prompt Translator — system prompt (part 1/2).** Produces the first-frame prompt that the text-to-image generator turns into an anchor frame for downstream video synthesis.

Video Agent: Image Prompt Translator — System Prompt (part 2/2)

```
Scene Graph objects into the first frame.

- Identify 'Key Objects' for this first frame. Extract their exact 'visual_attributes' (color,
shape, material, role) from the Scene Graph. Only include objects that are visible in the first
frame per the First Frame Specification.
- Determine the 'Spatial Layout': Define exactly where each visible object is located relative
to others and the environment in this starting frame.

B. Static Translation Rules (CRITICAL):
- NO MOTION: This is for a static image. Translate actions into static poses or postures.
(e.g., "A man walking" -> "A man captured mid-stride").
- Camera & Lighting: Extract the initial camera angle and environmental lighting from Phase 1
of the Scenario and the First Frame Specification in Task Guidelines.

2. Prompt Construction
Construct a prompt that the image generation model can perfectly render.
The generated image is the first frame of the video, so it must be a perfect anchor for
subsequent frames. No ambiguity or missing details can be allowed for visible objects. Do not
add any details not explicitly stated in the Scene Graph or Scenario.
The AUTHORITATIVE SOURCE for first frame content is Phase 1 (Setup) of the Scenario timeline
combined with the First Frame Specification in Task Guidelines. Use these as your primary
reference for which objects appear, their positions, and the camera framing.
Only use the Scenario Phase 1 to extract: (1) which objects are visible and their initial
states/poses, (2) the camera angle and framing, and (3) the initial lighting/environment
conditions.
Do NOT use the scenario's camera movement directives (orbit, pan, dolly, zoom, reveal, etc.) to
inform the composition of this image.
The camera position for this image must exactly match the starting state defined in Phase 1 —
not a neutral overview of the entire scene.
Positive Prompt: Must be highly descriptive, comma-separated, starting with the main subject,
followed by environment, lighting, camera angle. Ensure all visual attributes from the Scene
Graph are explicitly stated.

3. Feedback Handling (Correction Mode)
If you receive "VALIDATION FEEDBACK" from a previous failed attempt:
- You MUST carefully analyze the feedback to understand exactly which visual constraints or
rules were violated.
- You MUST correct the specific violations in your new prompt.
- DO NOT repeat the previous mistakes. Ensure that fixing the feedback does not cause you to
violate other existing TASK_RULES.

4. Output Schema (JSON)
Respond ONLY with a JSON object.
Output example :
{{
    "prompt": "Highly detailed text-to-image prompt."
}}
```

Figure 22: **Image Prompt Translator — system prompt (part 2/2).**

Video Agent: Image Prompt Translator — User Prompt

```
Please analyze the validated data and generate the first-frame image prompts based strictly on
your system instructions.
Do NOT include any timeline or movement descriptions in the prompts. Focus only on capturing
the perfect starting state.

INPUTS:

SCENE GRAPH: {SCENE_GRAPH}
TASK DEFINITION: {TASK_DEFINITION}
TASK RULES: {TASK_RULES}
TASK GUIDELINES: {TASK_GUIDELINES}
SCENARIO: {SCENARIO}
REFERENCE EXAMPLES: {EXAMPLE}
```

Figure 23: **Image Prompt Translator — user prompt.** Carries the scene graph and the scenario's initial state.

Video Agent: Video Prompt Translator — System Prompt (part 1/3)

```
You are an expert AI Video Prompt Engineer and Cinematographer specializing in Spatio-Temporal
Reasoning Benchmarks.
Your goal is to translate a 'Scene Graph' and 'Scenario' into a highly optimized prompt for an
Image-to-Video (I2V) model.

The I2V model receives a pre-generated Anchor Frame (first frame image) and animates it forward
in time.
Your prompt describes ONLY what happens AFTER that first frame — the motion, changes, and
camera work.
Do NOT re-describe the initial static state already captured in the anchor frame.

You will receive:
1. Scene Graph: Objects, attributes, spatial relationships, and ground truth.
2. Task Definition: The reasoning capability being tested.
3. Task Rules: HARD CONSTRAINTS — violating these invalidates the benchmark.
4. Task Guidelines: Technical instructions including camera mechanics and phase-by-phase
execution directives.
5. Scenario: The execution script with explicit phases.
6. Reference Examples (Optional): One or more previously-produced video prompts for the SAME
task type, numbered "### Example 1 / 2 / ...". Use them as structural/stylistic guides — match
phrasing patterns, motion-description granularity, and overall composition. Do NOT copy their
content literally; your output must reflect the current Scene Graph and Scenario. If "(no
reference examples available)" appears, ignore this input.

1. Determine Scene Dynamics (CRITICAL FIRST STEP)

Before writing anything, identify the 'scene_dynamics' field in the Scene Graph.

- IF 'Static': The video is camera-motion only. ALL objects remain COMPLETELY FROZEN throughout
the entire video.
No object slides, drifts, rotates, or moves in any way. Camera movement is the ONLY source of
  motion.
  Any object motion in a Static video invalidates the benchmark.

- IF 'Dynamic': Objects and/or agents move according to the Scenario's action sequence.
```

Figure 24: **Video Prompt Translator — system prompt (part 1/3).** Composes a video prompt that conditions the image-to-video generator on the anchor frame, the scenario's timeline, and the camera setup.

Video Agent: Video Prompt Translator — System Prompt (part 2/3)

```
  Describe each action phase with precise timing, object trajectories, and agent behaviors.

2. Input Analysis

A. Camera Trajectory
- Extract the exact camera movement from the Scenario.
- Note the starting angle, direction of movement, speed (slow/moderate/fast), and ending
position.
- The camera trajectory is the PRIMARY content for Static task prompts.

B. Object & Agent Motion (Dynamic tasks only)
- Map each action in the Scenario to a time window.
- For each action: identify who moves, what they do, where they end up, and what the visual
result is.
- Preserve exact object colors, shapes, and materials so the I2V model maintains identity
through motion.

C. Ground Truth Alignment
- Identify the ground_truth field from the Scene Graph.
- Verify that the described motion sequence makes the ground_truth the ONLY unambiguous visual
conclusion.
- The video must not be interpretable as any other answer.

3. Prompt Construction

Structure your prompt in the following order:

[Continuation from anchor frame]
Start with a brief statement that this continues from the established first frame. Then
describe ONLY the motion.

[Camera motion] (all tasks)
Describe the camera's movement trajectory precisely: direction, speed, arc, start/end angle.

[Object/agent actions] (Dynamic tasks only)
```

Figure 25: **Video Prompt Translator — system prompt (part 2/3).**

Video Agent: Video Prompt Translator — System Prompt (part 3/3)

```
Describe each action sequentially. For each:
- What object/agent moves and how
- The exact motion path and duration
- The resulting visual state after the action
- Any reaction events triggered

[Object identity anchors]
List key objects with their visual attributes (color, shape, material) to prevent the I2V model
from morphing them mid-video.

4. Writing Rules

- Be specific about directions: "clockwise", "left-to-right", "upward from 30° to 70°", not
vague terms like "around" or "slowly".
- For Static tasks: use phrases like "camera drifts...", "smooth clockwise orbit...",
"ascending crane shot..." — never describe object motion.
- For Dynamic tasks: use action verbs tied to specific objects: "the red wallet is picked up by
the agent's right hand", "the blue sedan brakes to a complete stop".
- Maintain all visual attributes (colors, materials) throughout — explicitly repeat them to
prevent I2V model drift.

5. Feedback Handling (Correction Mode)
If you receive "VALIDATION FEEDBACK" from a previous failed attempt:
- You MUST carefully analyze the feedback to understand exactly which visual constraints or
rules were violated.
- You MUST correct the specific violations in your new prompt.
- DO NOT repeat the previous mistakes. Ensure that fixing the feedback does not cause you to
violate other existing TASK_RULES.

6. Output Schema (JSON)

Respond ONLY with a JSON object:
{{
    "prompt": "Detailed I2V motion prompt starting from the anchor frame."
}}
```

Figure 26: **Video Prompt Translator — system prompt (part 3/3).**

Video Agent: Video Prompt Translator — User Prompt

```
Please generate the Image-to-Video prompt based strictly on your system instructions.
The video begins from the pre-generated Anchor Frame. Describe only the motion and changes from
that point forward.

INPUTS:

SCENE GRAPH: {SCENE_GRAPH}
TASK DEFINITION: {TASK_DEFINITION}
TASK RULES: {TASK_RULES}
TASK GUIDELINES: {TASK_GUIDELINES}
SCENARIO: {SCENARIO}
REFERENCE EXAMPLES: {EXAMPLE}
```

Figure 27: **Video Prompt Translator — user prompt.** Carries the scenario, the anchor-frame description, and the camera trajectory.

G.4 QA AGENT

QA Agent: QA Generator — System Prompt

```
You are an expert Benchmark Question Generator for Spatio-Temporal Reasoning in MLLMs.

You will generate multiple-choice questions (MCQs) of a SPECIFIC QA type for a SPECIFIC task.

INPUTS you will receive:
1. Task Definition + Task Rules         — the benchmark task you are generating questions for.
2. QA Type — the specific reasoning category to test (e.g., "Object Existence", "Action
Recognition").
3. QA Type Definition                   — short explanation of what this QA category measures.
4. Scene Graph — objects, attributes, relationships, ground truth of the current video.
5. Scenario                             — temporal execution script of the current video.
6. Prototype QAs — 1~3 example QAs for this (task, qa_type) combination. Match their
style/structure closely.
7. Distractor Pool — candidate distractor labels collected from OTHER scenes in the same task.
Use these as the primary source of wrong-answer options whenever the QA type is about
identifying / discriminating entities (objects, actions, etc.). The pool may be empty or not
applicable for some QA types.

YOUR JOB:
Generate one MCQ that:
  - Strictly fit the given QA TYPE (do NOT generate questions outside this category).
- Are answerable purely from viewing the generated video (grounded in Scene Graph + Scenario).
  - Follow the Prototype QAs' phrasing pattern and difficulty level.
  - Have plausible, non-random distractors (logical alternatives or common misconceptions).
- Choose the number of options (A/B/C/...) based on what best suits the question (not
necessarily 4).

DISTRACTOR SELECTION RULES:
- When the Distractor Pool is non-empty AND the QA type asks the viewer to pick / reject an
entity (e.g., Object Existence, Object Attribute, Action Recognition), prefer drawing
wrong-answer options from the pool over inventing your own. This keeps distractors consistent
with the broader benchmark.
- If the QA type is not about entity identification (e.g., counting, temporal ordering, causal
reasoning), you may ignore the pool and construct distractors that fit the question's logic.

OUTPUT (JSON only, no prose):
{{
  "qa_type": "<qa_type_id>",
  "task": "<task_id>",
  "questions": [
    {{
"question": "<full question with options inline, e.g., 'Which container appears? A. ..., B.
..., C. ..., D. ...'>",
      "options": {{"A": "<text>", "B": "<text>", "C": "<text>", "D": "<text>"}},
      "answer": "<letter A/B/C/...>"
    }}
  ]
}}
```

Figure 28: **QA Generator — system prompt.** Generates a base MCQ conditioned on the scene graph, the scenario, and the cell-specific QA template, with distractors drawn from the task's distractor pool.

QA Agent: QA Generator — User Prompt

```
INPUTS:

TASK ID: {TASK_ID}
TASK DEFINITION: {TASK_DEFINITION}
TASK RULES: {TASK_RULES}

QA TYPE: {QA_TYPE_ID} ({QA_TYPE_NAME})
QA TYPE DEFINITION: {QA_TYPE_DEF}

SCENE GRAPH:
{SCENE_GRAPH}

SCENARIO:
{SCENARIO}

PROTOTYPE QAs (reference style & coverage):
{PROTOTYPES}

DISTRACTOR POOL (deduped {{label, attributes}} from all scenes in this task; may include the
current scene's own entries — exclude those when picking distractors):
{DISTRACTOR_POOL}

Generate the MCQ per your system instructions.
```

Figure 29: **QA Generator — user prompt.** Carries the scene graph, the scenario, the QA template, and the distractor pool entries.

QA Agent: Reformatter (V1/V2/V3) System Prompt (part 1/2)

```
You are a Benchmark MCQ Reformatter for evaluating MLLMs on Spatio-Temporal Reasoning.

You receive a list of base multiple-choice questions (MCQs). For EACH base question, produce
three variants designed to probe different model behaviors.

============================================================
VARIANT 1: NoT-Distractor  (false-abstention test)
============================================================
GOAL: Measure whether the model wrongly picks "None of these" even though the correct answer IS
among the options.

RULES:
- Keep ALL original options unchanged.
- Add "None of these" as the next option letter (e.g., E if original had A-D).
- The correct answer letter STAYS THE SAME as the base.

============================================================
VARIANT 2: NoT-Answer  (true-abstention test)
============================================================
GOAL: Measure whether the model correctly chooses "None of these" when the true answer is NOT
in the options.

RULES:
- Locate the correct option in the base.
- Replace the TEXT of that correct option with "None of these" (the option letter stays the
same).
- All other options unchanged.
- The correct answer letter is the same letter as the base — but the displayed text is now
"None of these".

============================================================
VARIANT 3: Open-Ended  (true-reasoning test)
============================================================
GOAL: Measure whether the model can produce the answer without option scaffolding.

RULES:
- Keep ONLY the question stem (remove every "A. ...", "B. ..." segment).
- Append exactly: " Answer concisely without explanation."
- expected_answer = the EXACT TEXT of the original correct option (used by an LLM judge for
matching).
- REJECT this variant (return null) if the ground truth is not number or single-word answer
```

Figure 30: **Reformatter — system prompt (part 1/2).** Expands a base MCQ into the three reformulation variants: V1 (None-of-these distractor), V2 (None-of-these answer), and V3 (open-ended).

QA Agent: Reformatter (V1/V2/V3) System Prompt (part 2/2)

```
(e.g., "C. The cat is on the mat" or "B. 42" are acceptable; but not "A. The cat is on the mat
and the dog is outside" or "D. 3.14" or "A. Red and blue").

============================================================
NOTES
============================================================
- The base MCQ may have ANY number of options (2, 3, 4, 5, ...). Preserve the option count when
generating Variants 1 & 2.
- Variant 1 adds exactly ONE more option ("None of these") at the next available letter.
- Variant 2 only changes the TEXT of the correct option to "None of these"; option count and
answer letter remain.

============================================================
OUTPUT FORMAT (JSON only, no prose, no markdown fences)
============================================================
{{
  "task": "<task_id passed in>",
  "qa_type": "<qa_type passed in>",
  "questions": [
    {{
      "base": {{ ...the original base question, copied verbatim... }},
      "variant_1_NoT_distractor": {{
        "question": "<full question text with all options including 'None of these'>",
        "options": {{"A": "...", ..., "<next letter>": "None of these"}},
        "answer": "<original answer letter>"
      }},
      "variant_2_NoT_answer": {{
"question": "<full question text with the correct option's text replaced by 'None of these'>",
        "options": {{"A": "...", ..., "<correct letter>": "None of these", ...}},
        "answer": "<same letter as base>"
      }},
      "variant_3_open_ended": {{
        "question": "<bare stem> Answer concisely without explanation.",
        "expected_answer": "<exact text of the original correct option>"
      }}
    }}
  ]
}}

The "questions" array typically contains one element (since the base typically has one MCQ).
For rejected variants, set the value to null (e.g., "variant_1_NoT_distractor": null).
```

Figure 31: **Reformatter — system prompt (part 2/2).**

QA Agent: Reformatter (V1/V2/V3) User Prompt

```
TASK ID: {TASK_ID}
QA TYPE: {QA_TYPE_ID} ({QA_TYPE_NAME})

BASE MCQs:
{BASE_MCQS}

Generate the 3 variants per question per your system instructions.
```

Figure 32: **Reformatter — user prompt.** Carries the base MCQ and the target variant identifier.

H QUALITATIVE EXAMPLES

This section provides per-task qualitative examples of VGenST-Bench. For each of the 12 tasks, we sample one representative video (random sample idx) and render four cards: 8-frames of video, underlying scene graph (verbatim JSON), scenario (verbatim JSON), and a representative QA pairs containing one MCQ per cognitive level (L1 / L2 / L3). Correct answer choices are marked with ✓. Long document cards are split across multiple pages, each marked with *(part i/n)* in the header.

MC_F_EGO_STA - Multi-Container Attribute Mapping

idx=81 | Tennis Player's Courtside Bench | Frames

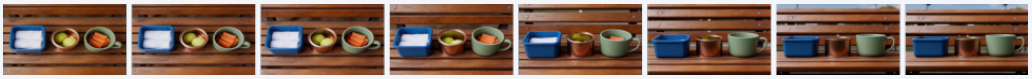


Figure 33: **Frames** for MC_F_EGO_STA, idx 81 (*Tennis Player's Courtside Bench*).

MC_F_EGO_STA - Multi-Container Attribute Mapping

idx=81 | Tennis Player's Courtside Bench | Scene Graph | part 1/2


```
{
  "scene_meta": {
    "spatial_scale": "Figural",
    "scene_dynamics": "Static",
    "perspective": "Ego",
    "task_type": "Multi_Container_Attribute_Mapping",
    "theme": "Tennis Player's Courtside Bench"
  },
  "objects": [
    {
      "id": "obj_anchor",
      "label": "Wooden Bench Surface",
      "role": "anchor",
      "attributes": {
        "color": "Brown",
        "material": "Wood"
      }
    },
    {
      "id": "obj_container_1",
      "label": "Cobalt Matte Box",
      "role": "container",
      "attributes": {
        "content": "White Towels",
        "color": "Cobalt",
        "material": "Matte Plastic"
      }
    },
    {
      "id": "obj_container_2",
      "label": "Copper Metal Tin",
      "role": "container",
      "attributes": {
        "content": "Yellow Tennis Balls",
        "color": "Copper",
        "material": "Metal"
      }
    },
    {
```


Figure 34: **Scene graph (part 1/2)** for MC_F_EGO_STA, idx 81.

MC_F_EGO_STA - Multi-Container Attribute Mapping

idx=81 | Tennis Player's Courtside Bench | Scene Graph | part 2/2


```
      "id": "obj_container_3",
      "label": "Sage Ceramic Mug",
      "role": "container",
      "attributes": {
        "content": "Orange Energy Gels",
        "color": "Sage",
        "material": "Ceramic"
      }
    }
  ],
  "spatial_layout": [
    {
      "subject": "obj_container_1",
      "relation": "standing_on_left_of",
      "target": "obj_anchor"
    },
    {
      "subject": "obj_container_2",
      "relation": "standing_on_middle_of",
      "target": "obj_anchor"
    },
    {
      "subject": "obj_container_3",
      "relation": "standing_on_right_of",
      "target": "obj_anchor"
    }
  ],
  "camera_movement": {
    "type": "Boom_Down",
    "trajectory": "high_angle_to_low_angle"
  },
  "ground_truth": {
    "mapping": {
      "Cobalt Matte Box": "White Towels",
      "Copper Metal Tin": "Yellow Tennis Balls",
      "Sage Ceramic Mug": "Orange Energy Gels"
    }
  }
}
```


Figure 35: **Scene graph (part 2/2)** for MC_F_EGO_STA, idx 81.

MC_F_EGO_STA - Multi-Container Attribute Mapping

idx=81 | Tennis Player's Courtside Bench | Scenario


```
{
"reasoning_goal": "Viewers must map content (seen from top) to container identity (seen from
side) by integrating both camera angles. Mapping: Cobalt Matte Box → White Towels, Copper Metal
Tin → Yellow Tennis Balls, Sage Ceramic Mug → Orange Energy Gels.",
  "timeline": {
"(Phase 1 — Top-Down View)": "90-degree overhead top-down view of the Wooden Bench Surface,
Tennis Player's Courtside Bench environment. Three open-top containers, interiors visible:
Left: Cobalt Matte Box contains White Towels, Middle: Copper Metal Tin contains Yellow Tennis
Balls, Right: Sage Ceramic Mug contains Orange Energy Gels.",
"(Phase 2 — Boom Down)": "Camera smoothly descends from overhead to eye-level side view.
Container interiors become hidden behind opaque walls. Exterior appearances become prominent.",
"(Phase 3 — Side View Hold)": "Camera holds at eye-level side view. Interiors completely
hidden. Exteriors visible: Cobalt Matte Box (Cobalt) left, Copper Metal Tin (Copper) middle,
Sage Ceramic Mug (Sage) right."
  }
}
```


Figure 36: **Scenario** for MC_F_EGO_STA, idx 81.

MC_F_EGO_STA - Multi-Container Attribute Mapping

idx=81 | Tennis Player's Courtside Bench | Sample QAs

Q [L1-FL: Frame Localization]

In the final view, where is the Copper container located in the frame?

(A) Left
(B) Center ✓
(C) Right

Q [L2-CM: Camera Motion]

How does the camera move throughout the video?

(A) Boom Up (rising vertically)
(B) Boom Down (descending vertically) ✓
(C) Pan Left (rotating horizontally)
(D) Zoom In (moving closer)

Q [L3-CR: Counterfactual]

If the container holding White Towels and the container holding Orange Energy Gels had been swapped in position, what content would the center container hold in the end?

(A) White Towels
(B) Yellow Tennis Balls ✓
(C) Orange Energy Gels
(D) Empty

Figure 37: **Sample QA pairs** (one per cognitive level) for MC_F_EGO_STA, idx 81.

QC_F_EGO_DYN - Quantity Change Tracking

idx=14 | Retail Checkout Counter | Frames

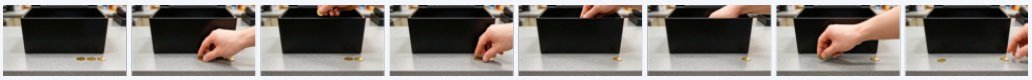


Figure 38: **Frames** for QC_F_EGO_DYN, idx 14 (*Retail Checkout Counter*).

QC_F_EGO_DYN - Quantity Change Tracking

idx=14 | Retail Checkout Counter | Scene Graph | part 1/2


```
{
  "scene_meta": {
    "spatial_scale": "Figural",
    "scene_dynamics": "Dynamic",
    "perspective": "Ego",
    "task_type": "Quantity_Change_Tracking",
    "theme": "Retail Checkout Counter"
  },
  "objects": [
    {
      "id": "obj_anchor",
      "label": "Gray Speckled Retail Counter",
      "role": "anchor",
      "attributes": {
        "color": "Gray"
      }
    },
    {
      "id": "obj_container",
      "label": "Tall Black Cash Register Tray",
      "role": "container",
      "attributes": {
        "material": "Opaque Metal",
        "initial_state": "Empty"
      }
    },
    {
      "id": "obj_items",
      "label": "Small Gold Coins",
      "role": "countable_items",
      "attributes": {
        "appearance": "All identical — round, flat, gold",
        "initial_location": "3 identical objects visible on the surface beside the container"
      }
    }
  ],
  "temporal_flow": {
    "initial_setup": {
"description": "The Tall Black Cash Register Tray sits empty on the Gray Speckled Retail
Counter. 3 identical Small Gold Coins are visible on the surface beside it. Actions begin
      immediately."
```


Figure 39: **Scene graph (part 1/2)** for QC_F_EGO_DYN, idx 14.

QC_F_EGO_DYN - Quantity Change Tracking

idx=14 | Retail Checkout Counter | Scene Graph | part 2/2


```
    },
    "action_sequence": [
      {
        "step": 1,
        "action": "ADD",
"action_description": "The hand picks one Small Gold Coin from the surface, arcs it over the
top rim of the Tall Black Cash Register Tray, and drops it in. The object disappears over the
rim. Hand retracts and pauses."
      },
      {
        "step": 2,
        "action": "ADD",
"action_description": "The hand picks one Small Gold Coin from the surface, arcs it over the
top rim of the Tall Black Cash Register Tray, and drops it in. The object disappears over the
rim. Hand retracts and pauses."
      },
      {
        "step": 3,
        "action": "REMOVE",
"action_description": "The hand reaches into the Tall Black Cash Register Tray from above,
lifts one Small Gold Coin out, and places it on the surface. Hand retracts and pauses."
      }
    ]
  },
  "camera_movement": {
    "type": "Fixed Angle",
    "trajectory": "Eye-level, facing the front face of the container. Static."
  },
  "ground_truth": {
    "initial_inside_count": 0,
    "action_sequence": [
      "ADD",
      "ADD",
      "REMOVE"
    ],
    "final_inside_count": 1,
"reasoning": "Started with 0 inside (3 on surface) ADD: 1 inside ADD: 2 inside REMOVE: 1 inside
Final count: 1."
  }
}
```


Figure 40: **Scene graph (part 2/2)** for QC_F_EGO_DYN, idx 14.

QC_F_EGO_DYN - Quantity Change Tracking

idx=14 | Retail Checkout Counter | Scenario


```
{
"reasoning_goal": "Viewers must track 3 sequential ADD/REMOVE actions on identical Small Gold
Coins to determine that 1 object(s) remain inside the Tall Black Cash Register Tray at the
  end.",
  "timeline": {
"(Phase 1 — Setup)": "Eye-level view of empty Tall Black Cash Register Tray on Gray Speckled
Retail Counter. 3 Small Gold Coins on the surface beside it.",
"(Phase 2 — Actions)": "Action 1 (ADD): Hand grasps one Small Gold Coin then puts it into Black
Cash Register Tray. Action 2 (ADD): Hand grasps another Small Gold Coin then puts it into Black
Cash Register Tray. Action 3 (REMOVE): Hand reaches into Black Cash Register Tray, lifts one
Small Gold Coin out, places it on the surface.",
"(Phase 3 — Result)": "Hand exits frame. 2 Small Gold Coins on the surface. 1 inside the
    container."
  }
}
```


Figure 41: **Scenario** for QC_F_EGO_DYN, idx 14.

QC_F_EGO_DYN - Quantity Change Tracking

idx=14 | Retail Checkout Counter | Sample QAs

Q [L1-OA: Object Attribute]

What is the material of the container?

(A) Ceramic
(B) Felt
(C) Metal ✓
(D) Leather

Q [L2-AR: Action Recognition]

What action does the hand perform in the second step?

(A) Places an item into the container ✓
(B) Takes an item out of the container
(C) Moves an item on the surface to the other location of the surface

Q [L3-CR: Counterfactual]

If the hand had performed only the first action (and then stopped), how many items would be inside the container at the end?

(A) 0
(B) 1 ✓
(C) 2
(D) 3

Figure 42: **Sample QA pairs** for QC_F_EGO_DYN, idx 14.

CI_F_EXO_STA - Container Intersection Inference

idx=3 | Bathroom Vanity Counter | Frames



Figure 43: **Frames** for CI_F_EXO_STA, idx 3 (*Bathroom Vanity Counter*).

CI_F_EXO_STA - Container Intersection Inference

idx=3 | Bathroom Vanity Counter | Scene Graph | part 1/3


```
{
  "scene_meta": {
    "spatial_scale": "Figural",
    "scene_dynamics": "Static",
    "perspective": "Exo",
    "task_type": "Container_Intersection_Inference",
    "theme": "Bathroom Vanity Counter"
  },
  "objects": [
    {
      "id": "obj_surface",
      "label": "White Marble Bathroom Vanity Counter",
      "role": "anchor"
    },
    {
      "id": "container_A",
      "label": "Wide Round White Ceramic Organizer Bowl (open-top, no lid)",
      "role": "Container A",
      "attributes": {
        "revealed_in": "Phase 1",
        "color": "White"
      }
    },
    {
      "id": "obj_A1",
      "label": "Pink Bar of Soap",
      "role": "container_A_object_1",
      "attributes": {
        "color": "Light pink",
        "type": "Bar of soap"
      }
    },
    {
      "id": "obj_A2",
      "label": "Silver Bobby Pin",
      "role": "container_A_object_2",
      "attributes": {
        "color": "Metallic silver",
        "type": "Bobby pin"
      }
    },
    {
      "id": "obj_A3",
      "label": "Red Lip Balm Tube",
      "role": "container_A_object_3",
      "attributes": {
        "color": "Red",
```


Figure 44: **Scene graph (part 1/3)** for CI_F_EXO_STA, idx 3.

CI_F_EXO_STA - Container Intersection Inference

idx=3 | Bathroom Vanity Counter | Scene Graph | part 2/3


```
        "type": "Lip balm tube"
      }
    },
    {
      "id": "container_B",
      "label": "Wide Rectangular Bamboo Tray Box (open-top, no lid)",
      "role": "Container B",
      "attributes": {
        "revealed_in": "Phase 2",
        "color": "Bamboo Tan"
      }
    },
    {
      "id": "obj_B1",
      "label": "Pink Bar of Soap",
      "role": "container_B_object_1",
      "attributes": {
        "color": "Light pink",
        "type": "Bar of soap"
      }
    },
    {
      "id": "obj_B2",
      "label": "Silver Bobby Pin",
      "role": "container_B_object_2",
      "attributes": {
        "color": "Metallic silver",
        "type": "Bobby pin"
      }
    },
    {
      "id": "obj_B3",
      "label": "Red Lip Balm Tube",
      "role": "container_B_object_3",
      "attributes": {
        "color": "Red",
        "type": "Lip balm tube"
      }
    }
  ],
  "temporal_flow": {
    "initial_setup": {
"description": "Counter-height side view. Both open-top containers visible from the side. No
interior visible."
    },
    "action_sequence": [
      {
```


Figure 45: **Scene graph (part 2/3)** for CI_F_EXO_STA, idx 3.

CI_F_EXO_STA - Container Intersection Inference

idx=3 | Bathroom Vanity Counter | Scene Graph | part 3/3


```
        "step": 1,
"action_description": "Camera moves overhead Container A (Ceramic Organizer Bowl). All three
objects — Small Pink Bar of Soap, Small Silver Bobby Pin, Small Red Lip Balm Tube — clearly
visible from above. Container B interior completely hidden. Camera holds 2+ seconds."
      },
      {
        "step": 2,
"action_description": "Camera moves overhead Container B (Bamboo Tray Box). All three objects —
Small Pink Bar of Soap, Small Silver Bobby Pin, Small Red Lip Balm Tube — clearly visible from
above. Container A interior completely hidden. Camera holds 2+ seconds."
      }
    ]
  },
  "camera_movement": {
    "type": "Sequential Overhead Peek — Two Containers",
"trajectory": "Counter-height side view. Both open-top containers visible from the side. No
interior visible. → Camera moves overhead Container A (Ceramic Organizer Bowl). All three
objects — Small Pink Bar of Soap, Small Silver Bobby Pin, Small Red Lip Balm Tube — clearly
visible from above. Container B interior completely hidden. Camera holds 2+ seconds. → Camera
moves overhead Container B (Bamboo Tray Box). All three objects — Small Pink Bar of Soap, Small
Silver Bobby Pin, Small Red Lip Balm Tube — clearly visible from above. Container A interior
completely hidden. Camera holds 2+ seconds."
  },
  "ground_truth": {
    "container_A_contents": [
      "Small Pink Bar of Soap",
      "Small Silver Bobby Pin",
      "Small Red Lip Balm Tube"
    ],
    "container_B_contents": [
      "Small Pink Bar of Soap",
      "Small Silver Bobby Pin",
      "Small Red Lip Balm Tube"
    ],
    "num_shared": 3,
    "shared_objects": [
      "Small Pink Bar of Soap",
      "Small Silver Bobby Pin",
      "Small Red Lip Balm Tube"
    ],
    "answer": "Small Pink Bar of Soap, Small Silver Bobby Pin, Small Red Lip Balm Tube",
"reasoning": "Phase 1 reveals Container A: {Pink Bar of Soap, Silver Bobby Pin, Red Lip Balm
Tube}. Phase 2 reveals Container B: the exact same 3 objects. Intersection: all 3 objects.
Answer: Small Pink Bar of Soap, Small Silver Bobby Pin, Small Red Lip Balm Tube."
  }
}
```


Figure 46: **Scene graph (part 3/3)** for CI_F_EXO_STA, idx 3.

CI_F_EXO_STA - Container Intersection Inference

idx=3 | Bathroom Vanity Counter | Scenario


```
{
"reasoning_goal": "Viewers must identify all three objects - the Pink Bar of Soap, the Silver
Bobby Pin, and the Red Lip Balm Tube - as appearing in BOTH Container A and Container B, by
comparing Phase 1 (Container A: Pink Bar of Soap + Silver Bobby Pin + Red Lip Balm Tube) with
Phase 2 (Container B: Pink Bar of Soap + Silver Bobby Pin + Red Lip Balm Tube). The
intersection contains all three objects.",
  "timeline": {
"(Setup)": "Third-person exo view, eye-level. A White Marble Bathroom Vanity Counter fills the
frame. On the left sits a Wide Round White Ceramic Organizer Bowl (Container A, open-top, no
lid) and on the right sits a Wide Rectangular Bamboo Tray Box (Container B, open-top, no lid).
Both containers are opaque from this side angle - no interior visible. Camera completely
    static.",
"(Phase 1 - Container A Reveal)": "Camera smoothly ascends and moves overhead Container A
(Round White Ceramic Organizer Bowl). Looking straight down into the bowl, all three objects
are fully visible and individually distinguishable: a Pink Bar of Soap, a Silver Bobby Pin, a
Red Lip Balm Tube. Container B interior is completely hidden. Camera holds overhead for at
least 2 seconds.",
"(Phase 2 - Container B Reveal)": "Camera smoothly transitions overhead Container B
(Rectangular Bamboo Tray Box). Container A is now completely off-screen. Looking straight down
into the box, all three objects are fully visible: a Pink Bar of Soap (identical to the one
seen in Phase 1), a Silver Bobby Pin (identical to the one seen in Phase 1), a Red Lip Balm
Tube (identical to the one seen in Phase 1). All three objects match exactly what was seen in
Container A. Camera holds overhead for at least 2 seconds."
  }
}
```


Figure 47: **Scenario** for CI_F_EXO_STA, idx 3.

CI_F_EXO_STA - Container Intersection Inference

idx=3 | Bathroom Vanity Counter | Sample QAs

Q [L1-FL: Frame Localization]

In the initial view, where is the Wide Rectangular Bamboo Tray Box located in the frame?

(A) Left
(B) Center
(C) Right ✓

Q [L2-CM: Camera Motion]

How does the camera move throughout the video?

(A) It stays fixed at side view
(B) It alternates between side view and overhead peeks on each container ✓
(C) It rotates 360 degrees around the table
(D) It zooms in and out on a single container

Q [L3-CR: Counterfactual]

If the Pink Bar of Soap in Wide Round White Ceramic Organizer Bowl had been replaced by a Red Candy Apple, how many shared object types would the two containers have?

(A) 0
(B) 1
(C) 2 ✓
(D) 3

Figure 48: **Sample QA pairs** for CI_F_EXO_STA, idx 3.

CM_F_EXO_DYN - Causal Mapping

idx=94 | Music Producer's Synthesizer Stand | Frames

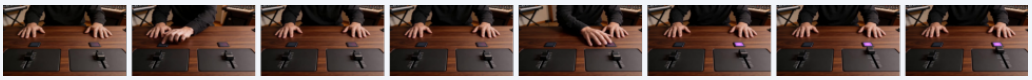


Figure 49: **Frames** for CM_F_EXO_DYN, idx 94 (*Music Producer's Synthesizer Stand*).

CM_F_EXO_DYN - Causal Mapping

idx=94 | Music Producer's Synthesizer Stand | Scene Graph | part 1/2


```
{
  "scene_meta": {
    "spatial_scale": "Figural",
    "scene_dynamics": "Dynamic",
    "perspective": "Exo",
    "task_type": "Causal_Mapping",
    "theme": "Music Producer's Synthesizer Stand"
  },
  "objects": [
    {
      "id": "obj_agent",
      "label": "Music Producer",
      "role": "agent",
      "attributes": {
        "appearance": "Human hands and forearms visible, wearing a dark long-sleeve shirt"
      }
    },
    {
      "id": "obj_anchor",
      "label": "Studio Desk Surface",
      "role": "anchor",
      "attributes": {
        "color": "Dark Walnut Wood"
      }
    },
    {
      "id": "obj_trigger_1",
      "label": "Left RGB Pad Button",
      "role": "trigger",
      "attributes": {
        "position": "background_left"
      }
    },
    {
      "id": "obj_trigger_2",
      "label": "Right RGB Pad Button",
      "role": "trigger",
      "attributes": {
        "position": "background_right"
      }
    },
    {
      "id": "obj_effector_1",
      "label": "Left Motorized Fader Slider",
```


Figure 50: **Scene graph (part 1/2)** for CM_F_EXO_DYN, idx 94.

CM_F_EXO_DYN - Causal Mapping

idx=94 | Music Producer's Synthesizer Stand | Scene Graph | part 2/2


```
      "role": "effector",
      "attributes": {
        "position": "foreground_left"
      }
    },
    {
      "id": "obj_effector_2",
      "label": "Right Motorized Fader Slider",
      "role": "effector",
      "attributes": {
        "position": "foreground_right"
      }
    }
  ],
  "temporal_flow": {
    "initial_setup": {
"description": "A MIDI controller sits on the Studio Desk Surface. The Music Producer stands
behind the desk facing the camera. Two RGB Pad Buttons are in the back row, aligned with two
Motorized Fader Sliders in the front row."
    },
    "action_sequence": [
      {
        "step": 1,
"action_description": "Agent presses Left RGB Pad Button (Trigger 1). NOTHING happens— Left
Motorized Fader Slider remains completely still at the bottom. The scene is absolutely static
except for the hand movement."
      },
      {
        "step": 2,
"action_description": "Agent presses Right RGB Pad Button (Trigger 2). Right Motorized Fader
Slider immediately slides upward smoothly in sync with the press."
      }
    ]
  },
  "camera_movement": {
    "type": "Fixed Angle",
    "trajectory": "High angle, Static"
  },
  "ground_truth": {
    "scenario_type": "B",
    "trigger_1_effect": "No Effect",
    "trigger_2_effect": "Right Motorized Fader Slider"
  }
}
```


Figure 51: **Scene graph (part 2/2)** for CM_F_EXO_DYN, idx 94.

CM_F_EXO_DYN - Causal Mapping

idx=94 | Music Producer's Synthesizer Stand | Scenario


```
{
"reasoning_goal": "The viewer must deduce that the Left RGB Pad is inert because the Left Fader
remains stationary when pressed, while the Right RGB Pad is functional because it causes the
Right Fader to slide upward immediately upon interaction.",
  "timeline": {
"(Phase 1 — Setup)": "The video opens with a fixed, high-angle Exo view (~30-45 degrees)
showing the Dark Walnut Studio Desk. The Music Producer stands behind the desk with hands at
rest, not touching any controls. In the background, the Left and Right RGB Pad Buttons are
unlit. In the foreground, both the Left and Right Motorized Fader Sliders are stationary at the
bottom position. Lighting is consistent, and the camera is completely static.",
"(Phase 2 — Action 1)": "The Agent's hand moves to press the Left RGB Pad Button; strictly
adhering to the Inert Constraint, the Left Fader remains ABSOLUTELY STATIC with no lighting
change or movement, and the hand retracts to a rest position.",
"(Phase 3 — Action 2)": "After a brief pause, the Agent's hand presses the Right RGB Pad
Button; immediately and synchronously, the Right Motorized Fader Slider slides smoothly upward
to the top position while the Left Fader remains still. The hand retracts, leaving the Right
Fader in the raised position. The camera holds steady on the final state: the Left Fader
remains at the bottom (inactive) and the Right Fader remains at the top (active). The Agent's
hands are at rest behind the desk. No objects move, and the lighting remains stable, providing
unambiguous visual confirmation of the causal mapping where only the Right Trigger activated
its corresponding Effector."
  }
}
```


Figure 52: **Scenario** for CM_F_EXO_DYN, idx 94.

CM_F_EXO_DYN - Causal Mapping

idx=94 | Music Producer's Synthesizer Stand | Sample QAs

Q [L1-FL: Frame Localization]

In the initial view, where is the agent located in the frame?

(A) Left
(B) Center ✓
(C) Right

Q [L2-AR: Action Recognition]

What happens to the Motorized Fader Slider on the left after the RGB Pad Button on the left is pressed?

(A) It illuminates green, synchronous with the lever press
(B) It begins to vibrate and shuffle across the surface
(C) It remains completely still ✓
(D) It explodes into pieces

Q [L3-CR: Counterfactual]

If the agent had pressed the rgb pad button on the right first instead of the one on the left, which motorized fader slider would activate first?

(A) The motorized fader slider on the left
(B) The motorized fader slider on the right ✓
(C) Both simultaneously

Figure 53: **Sample QA pairs** for CM_F_EXO_DYN, idx 94.

DE_V_EGO_STA - Direction Estimation

idx=35 | Comedy Club Backstage L-Hallway | Frames

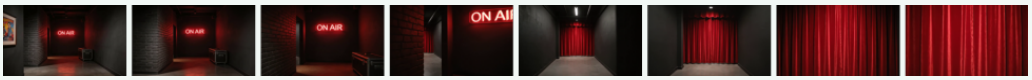


Figure 54: **Frames** for DE_V_EGO_STA, idx 35 (*Comedy Club Backstage L-Hallway*).

DE_V_EGO_STA - Direction Estimation

idx=35 | Comedy Club Backstage L-Hallway | Scene Graph


```
{
  "scene_meta": {
    "spatial_scale": "Vista",
    "scene_dynamics": "Static",
    "perspective": "Ego",
    "task_type": "Direction_Estimation",
    "theme": "Comedy Club Backstage L-Hallway"
  },
  "objects": [
    {
      "id": "landmark_1",
      "label": "Neon 'ON AIR' Sign",
      "role": "landmark_1 (Visible in Seg 1)",
      "attributes": {
        "color": "Bright Red Neon",
        "position": "Far end of North wall (Seg 1)"
      }
    },
    {
      "id": "landmark_2",
      "label": "Red Velvet Stage Curtain",
      "role": "landmark_2 (Visible in Seg 2)",
      "attributes": {
        "color": "Deep Red Velvet",
        "position": "Far end of West wall (Seg 2)"
      }
    }
  ],
  "camera_movement": {
    "type": "Person Walking Simulation",
"trajectory": "Start facing North at origin of Seg 1. Walk forward along Seg 1 towards the Neon
'ON AIR' Sign (North wall). Execute a smooth Left turn at the Brick Corner Partition. Walk
forward along Seg 2 towards the Red Velvet Stage Curtain (West wall). Stop."
  },
  "ground_truth": {
    "turn_direction": "Left",
    "relative_direction_to_start": "Back-Left (approx 7-8 o'clock)",
"reasoning": "The camera started facing North, moved forward (Seg 1), then turned Left to face
West (Seg 2). Final facing direction is West. The Start Point is located South-East relative to
the final position. Relative to a West-facing heading, South is Left (9 o'clock) and East is
Back (6 o'clock). Therefore, South-East is Back-Left, approx 7-8 o'clock."
  }
}
```


Figure 55: **Scene graph** for DE_V_EGO_STA, idx 35.

DE_V_EGO_STA - Direction Estimation

idx=35 | Comedy Club Backstage L-Hallway | Scenario


```
{
"reasoning_goal": "The viewer must integrate the visual path of moving North then turning Left
to face West to conclude the starting point lies Back-Left (7-8 o'clock) relative to the final
  heading.",
  "timeline": {
"(Setup)": "Camera starts at eye-level facing North down Seg 1. The Bright Red Neon 'ON AIR'
Sign (Landmark 1) is clearly visible at the far end. The corridor turns ahead, but Seg 2 is not
yet visible from this position. Landmark 2 is hidden. Black Equipment Flight Case and Framed
Comedy Poster line the walls. Camera moves forward with slight natural walking sway. Landmark 2
is strictly invisible.",
"(Event)": "Camera approaches the corridor corner. A smooth Left turn begins, pivoting around
the corner. As the camera turns, the North wall view shifts away and the West wall is revealed.
The Deep Red Velvet Stage Curtain (Landmark 2) comes into view at the far end of Seg 2. Camera
continues walking forward towards the curtain with natural stride motion.",
"(Result)": "Camera comes to a complete stop facing West. The Deep Red Velvet Stage Curtain is
centered in the distance. The corridor corner is now behind/to the right of the camera path.
All objects (Sign, Curtain, Cases, Posters) remain absolutely stationary. The final static shot
holds for 3 seconds to allow direction estimation."
  }
}
```


Figure 56: **Scenario** for DE_V_EGO_STA, idx 35.

DE_V_EGO_STA - Direction Estimation

idx=35 | Comedy Club Backstage L-Hallway | Sample QAs

Q [L1-OA: Object Attribute]

What are the dominant colors of the Red Velvet Stage Curtain?

(A) Dark stained wood
(B) Deep Red Velvet ✓
(C) Safety Yellow
(D) Bright Orange Wireframe on Black Background

Q [L2-CM: Camera Motion]

In which direction does the camera turn at the corner?

(A) Left ✓
(B) Right
(C) It does not turn
(D) It makes a U-turn (180 degrees)

Q [L3-CR: Counterfactual]

If the camera had turned Right instead of Left, in which direction would the second landmark be from the final position?

(A) 4-5 o'clock (back-right)
(B) 7-8 o'clock (back-left)
(C) 12 o'clock (directly ahead)
(D) 6 o'clock (directly behind) ✓

Figure 57: **Sample QA pairs** for DE_V_EGO_STA, idx 35.

IO_V_EGO_DYN - Interacted Object Identification

idx=31 | Farmhouse Kitchen with Prep Table and Hutch | Frames



Figure 58: **Frames** for IO_V_EGO_DYN, idx 31 (*Farmhouse Kitchen with Prep Table and Hutch*).

IO_V_EGO_DYN - Interacted Object Identification

idx=31 | Farmhouse Kitchen with Prep Table and Hutch | Scene Graph | part 1/2


```
{
  "scene_meta": {
    "spatial_scale": "Vista",
    "scene_dynamics": "Dynamic",
    "perspective": "Ego",
    "task_type": "Interacted_Object_Identification",
    "theme": "Farmhouse Kitchen with Prep Table and Hutch"
  },
  "objects": [
    {
      "id": "obj_agent",
      "label": "Farm Cook",
      "role": "agent",
      "attributes": {
        "appearance": "Adult in a denim apron, side view"
      }
    },
    {
      "id": "obj_primary_surface",
      "label": "Prep Table",
      "role": "primary_surface",
      "attributes": {
        "color": "Aged Pine"
      }
    },
    {
      "id": "obj_destination_surface",
      "label": "Hutch",
      "role": "destination_surface",
      "attributes": {
        "color": "Painted Cream"
      }
    },
    {
      "id": "obj_Left",
      "label": "White Egg in Bowl",
      "position": "Left",
      "role": "distractor",
      "attributes": {
        "color": "White"
      }
    },
    {
      "id": "obj_Center",
      "label": "Brown Bread Loaf",
      "position": "Center",
      "role": "target",
      "attributes": {
```


Figure 59: **Scene graph (part 1/2)** for IO_V_EGO_DYN, idx 31.

IO_V_EGO_DYN - Interacted Object Identification

idx=31 | Farmhouse Kitchen with Prep Table and Hutch | Scene Graph | part 2/2


```
        "color": "Brown"
      }
    },
    {
      "id": "obj_Right",
      "label": "Red Cherry Tomato Cluster",
      "position": "Right",
      "role": "distractor",
      "attributes": {
        "color": "Red"
      }
    }
  ],
  "temporal_flow": {
    "initial_setup": {
"description": "Stationary first-person view from a fixed eye-level observer position. The Prep
Table is centered in the foreground, holding three distinct objects in a row: a White Egg in
Bowl on the Left, a Brown Bread Loaf in the Center, and a Red Cherry Tomato Cluster on the
Right. The Hutch is also clearly visible elsewhere in the frame. No agent present."
    },
    "action_sequence": [
      {
        "step": 1,
        "action_description": "Agent enters from Left, walks to primary surface."
      },
      {
        "step": 2,
        "action_description": "Agent picks up object at Center."
      },
      {
        "step": 3,
"action_description": "Agent carries the object to the secondary destination surface and places
it down."
      }
    ]
  },
  "camera_movement": {
    "type": "Fixed Angle",
    "trajectory": "No movement. Fixed eye-level viewpoint from stationary observer."
  },
  "ground_truth": {
    "target_object_id": "obj_Center",
    "target_position": "Center",
    "agent_entry_direction": "Left",
    "destination": "Hutch"
  }
}
```


Figure 60: **Scene graph (part 2/2)** for IO_V_EGO_DYN, idx 31.

IO_V_EGO_DYN - Interacted Object Identification

idx=31 | Farmhouse Kitchen with Prep Table and Hutch | Scenario


```
{
"reasoning_goal": "The viewer must identify that the Brown Bread Loaf at the Center position
was the single object the Farm Cook picked up and moved to the Hutch. The White Egg in Bowl
(Left) and the Red Cherry Tomato Cluster (Right) remained completely stationary on the Prep
Table throughout the video and must be excluded as distractors.",
  "timeline": {
"(Phase 1 — Establishment)": "Stationary first-person view from a fixed eye-level observer
position. The Prep Table is centered in the foreground, holding three distinct objects in a
row: a White Egg in Bowl on the Left, a Brown Bread Loaf in the Center, and a Red Cherry Tomato
Cluster on the Right. The Hutch is also clearly visible elsewhere in the frame. No agent is
present yet; the room is completely still.",
"(Phase 2 — Agent Entry)": "The Farm Cook enters the frame from the Left side and walks toward
the Prep Table. The camera remains completely fixed; only the agent moves. All three objects on
the Prep Table remain stationary.",
"(Phase 3 — Pick-up)": "The Farm Cook reaches the Prep Table and picks up the Brown Bread Loaf
from the Center position. The White Egg in Bowl (Left) and the Red Cherry Tomato Cluster
(Right) are completely undisturbed at their original positions. The pick-up motion is clearly
visible with no occlusion of the Brown Bread Loaf at the moment of contact.",
"(Phase 4 — Transport & Place)": "The Farm Cook carries the Brown Bread Loaf from the Prep
Table to the Hutch and places it down clearly visible on its surface. The other two objects on
the Prep Table remain exactly where they were. The camera holds steady on the final state."
  }
}
```


Figure 61: **Scenario** for IO_V_EGO_DYN, idx 31.

IO_V_EGO_DYN - Interacted Object Identification

idx=31 | Farmhouse Kitchen with Prep Table and Hutch | Sample QAs

Q [L1-FL: Frame Localization]

In the view right after agent appears, where is the agent located in the frame?

(A) Left ✓
(B) Center
(C) Right

Q [L2-AR: Action Recognition]

Which object does the agent pick up?

(A) White Egg in Bowl
(B) Brown Bread Loaf ✓
(C) Red Cherry Tomato Cluster

Q [L3-CR: Counterfactual]

If the agent had picked up the Red Cherry Tomato Cluster instead, from which position on the Prep Table would it have been taken?

(A) Left
(B) Center
(C) Right ✓
(D) It was not on the Prep Table

Figure 62: **Sample QA pairs** for IO_V_EGO_DYN, idx 31.

HO_V_EXO_STA - Height Ordering

idx=28 | Law Firm Office | Frames

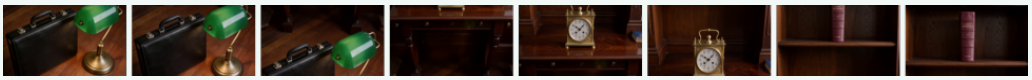


Figure 63: **Frames** for HO_V_EXO_STA, idx 28 (*Law Firm Office*).

HO_V_EXO_STA - Height Ordering

idx=28 | Law Firm Office | Scene Graph | part 1/2


```
{
  "scene_meta": {
    "spatial_scale": "Vista",
    "scene_dynamics": "Static",
    "perspective": "Exo",
    "task_type": "Height_Ordering",
    "theme": "Law Firm Office"
  },
  "objects": [
    {
      "id": "obj_lowest",
      "label": "Black Leather Attache Case",
      "role": "lowest_object",
      "attributes": {
        "color": "Black",
        "type": "Leather Attache Case",
        "height_level": "lowest",
        "placement": "on the dark hardwood floor"
      }
    },
    {
      "id": "obj_middle",
      "label": "Gold Brass Desk Clock",
      "role": "middle_object",
      "attributes": {
        "color": "Gold",
        "type": "Brass Desk Clock",
        "height_level": "middle",
        "placement": "on the walnut partner desk"
      }
    },
    {
      "id": "obj_highest",
      "label": "Maroon Legal Reference Volume",
      "role": "highest_object",
      "attributes": {
        "color": "Maroon",
```


Figure 64: **Scene graph (part 1/2)** for HO_V_EXO_STA, idx 28.

HO_V_EXO_STA - Height Ordering

idx=28 | Law Firm Office | Scene Graph | part 2/2


```
        "type": "Legal Reference Volume",
        "height_level": "highest",
        "placement": "on the high mahogany bookcase"
      }
    },
    {
      "id": "obj_distractor",
      "label": "Green Banker's Desk Lamp",
      "role": "distractor",
      "attributes": {
        "color": "Green",
        "type": "Banker's Desk Lamp",
        "height_level": "lowest",
        "placement": "on the dark hardwood floor"
      }
    }
  ],
  "camera_movement": {
    "type": "Fixed-Position Vertical Tilt",
    "tilt_direction": "Bottom-to-Top",
"trajectory": "Hold on lowest object → tilt Bottom-to-Top to middle object → continue tilt to
highest object."
  },
  "ground_truth": {
    "height_ranking_low_to_high": [
      "obj_lowest",
      "obj_middle",
      "obj_highest"
    ],
    "lowest_object": "Black Leather Attache Case",
    "middle_object": "Gold Brass Desk Clock",
    "highest_object": "Maroon Legal Reference Volume",
    "distractor": "Green Banker's Desk Lamp (at lowest level)",
    "tilt_direction": "Bottom-to-Top"
  }
}
```


Figure 65: **Scene graph (part 2/2)** for HO_V_EXO_STA, idx 28.

HO_V_EXO_STA - Height Ordering

idx=28 | Law Firm Office | Scenario


```
{
"reasoning_goal": "The viewer must determine the height ordering: Black Leather Attache Case
(on the dark hardwood floor) is lowest, Gold Brass Desk Clock (on the walnut partner desk) is
middle, Maroon Legal Reference Volume (on the high mahogany bookcase) is highest. Distractor
Green Banker's Desk Lamp shares the lowest level.",
  "timeline": {
"(Phase 1 — Lowest)": "Camera holds zoomed in on the lowest level. Black Leather Attache Case
(on the dark hardwood floor) and Green Banker's Desk Lamp (distractor, on the dark hardwood
floor) are visible side by side.",
"(Phase 2 — Middle)": "Camera tilts upward. Black Leather Attache Case exits frame. Gold Brass
Desk Clock (on the walnut partner desk) comes into view and becomes centered.",
"(Phase 3 — Highest)": "Camera continues tilting upward. Gold Brass Desk Clock exits frame.
Maroon Legal Reference Volume (on the high mahogany bookcase) comes into view alone. Video
    ends."
  }
}
```


Figure 66: **Scenario** for HO_V_EXO_STA, idx 28.

HO_V_EXO_STA - Height Ordering

idx=28 | Law Firm Office | Sample QAs

Q [L1-FL: Frame Localization]

In the final view, where is the Legal Reference Volume located in the frame?

(A) Left
(B) Center ✓
(C) Right

Q [L2-CM: Camera Motion]

How does the camera move throughout the video?

(A) It pans horizontally from left to right
(B) It tilts vertically from bottom to top ✓
(C) It zooms in and out
(D) It rotates around the room

Q [L3-CR: Counterfactual]

If the Legal Reference Volume and the Leather Attache Case had swapped positions, what is the height ranking of the objects from lowest to highest?

(A) Leather Attache Case, Brass Desk Clock, Legal Reference Volume
(B) Brass Desk Clock, Leather Attache Case, Legal Reference Volume
(C) Legal Reference Volume, Brass Desk Clock, Leather Attache Case ✓
(D) Leather Attache Case, Legal Reference Volume, Brass Desk Clock

Figure 67: **Sample QA pairs** for HO_V_EXO_STA, idx 28.

VI_V_EXO_DYN - Visibility Identification

idx=17 | Living Room with Tall Wooden Bookshelf | Frames



Figure 68: **Frames** for VI_V_EXO_DYN, idx 17 (*Living Room with Tall Wooden Bookshelf*).

VI_V_EXO_DYN - Visibility Identification

idx=17 | Living Room with Tall Wooden Bookshelf | Scene Graph | part 1/2


```
{
  "scene_meta": {
    "spatial_scale": "Vista",
    "scene_dynamics": "Dynamic",
    "perspective": "Exo",
    "task_type": "Visibility_Identification",
    "theme": "Living Room with Tall Wooden Bookshelf"
  },
  "objects": [
    {
      "id": "obj_observer",
      "label": "Resident",
      "role": "observer_agent",
      "attributes": {
        "appearance": "A woman in a navy turtleneck and reading glasses holding a mug",
        "position": "9 o'clock"
      }
    },
    {
      "id": "obj_target",
      "label": "Visitor",
      "role": "target_agent",
      "attributes": {
        "appearance": "A man in a brown corduroy jacket carrying a paper bag",
        "position": "3 o'clock (initial), moves to 12 o'clock"
      }
    },
    {
      "id": "obj_occluder",
      "label": "Tall Wooden Bookshelf",
      "role": "occluder_object",
      "attributes": {
        "position": "CENTER of the scene",
        "material": "Opaque"
      }
    }
```


Figure 69: **Scene graph (part 1/2)** for VI_V_EXO_DYN, idx 17.

VI_V_EXO_DYN - Visibility Identification

idx=17 | Living Room with Tall Wooden Bookshelf | Scene Graph | part 2/2


```
  ],
  "temporal_flow": {
    "initial_state": {
"description": "Bird's-eye view of a living room. Tall Wooden Bookshelf at the center. Resident
at 9 o'clock (Left side). Visitor at 3 o'clock. Direct line of sight from 9 o'clock to 3
o'clock passes through the Tall Wooden Bookshelf — blocked.",
      "observer_view_status": "Occluded"
    },
    "action_sequence": [
      {
        "step": 1,
"action_description": "Visitor moves from the 3 o'clock position to the 12 o'clock position,
emerging from behind the Tall Wooden Bookshelf."
      }
    ],
    "final_state": {
"description": "Resident remains at 9 o'clock. Visitor is now at 12 o'clock. Diagonal line of
sight from 9 o'clock to 12 o'clock — clear of the Tall Wooden Bookshelf.",
      "observer_view_status": "Visible"
    }
  },
  "camera_movement": {
    "type": "Fixed Angle",
"trajectory": "Static high-angle exocentric view. Camera does not move. All three entities
visible in every frame."
  },
  "ground_truth": {
    "visibility_transition": "Occluded_to_Visible",
    "mover": "Target_Moves",
    "observer_side": "Left",
    "initial_status": "Occluded",
    "final_status": "Visible",
"trigger_event": "Visitor moves from 3 o'clock to 12 o'clock, leaving the occluded zone behind
the Tall Wooden Bookshelf"
  }
}
```


Figure 70: **Scene graph (part 2/2)** for VI_V_EXO_DYN, idx 17.

VI_V_EXO_DYN - Visibility Identification

idx=17 | Living Room with Tall Wooden Bookshelf | Scenario


```
{
"reasoning_goal": "Viewers must determine that the Resident sees the Visitor as Occluded
initially, and after the Visitor moves around the Tall Wooden Bookshelf, the visibility status
changes to Visible.",
  "timeline": {
"(Phase 1 — Initial State)": "Bird's-eye view of the Living Room with Tall Wooden Bookshelf.
The Tall Wooden Bookshelf sits at the center of the scene. The Resident (A woman in a navy
turtleneck and reading glasses holding a mug) is at 9 o'clock. The Visitor (A man in a brown
corduroy jacket carrying a paper bag) is at 3 o'clock. From the Resident's perspective, the
direct line of sight is blocked by the Tall Wooden Bookshelf — the Resident CANNOT see the
    Visitor.",
"(Phase 2 — Movement)": "The Visitor moves from 3 o'clock to 12 o'clock. The Resident stays
    still.",
"(Phase 3 — Final Hold)": "The Tall Wooden Bookshelf remains at the center. The Resident is now
at 9 o'clock. The Visitor is now at 12 o'clock. From the Resident's perspective, the diagonal
line of sight bypasses the Tall Wooden Bookshelf — the Resident CAN see the Visitor."
  }
}
```


Figure 71: **Scenario** for VI_V_EXO_DYN, idx 17.

VI_V_EXO_DYN - Visibility Identification

idx=17 | Living Room with Tall Wooden Bookshelf | Sample QAs

Q [L1-FL: Frame Localization]

In the initial view, where are the agents located in the frame?

(A) Top and Left
(B) Top and Right
(C) Left and Right ✓
(D) Top and Bottom

Q [L2-AR: Action Recognition]

What action does the agent located in top perform in the video?

(A) Walks from one position to another ✓
(B) Stands completely still
(C) Picks up an item
(D) Exits the frame

Q [L3-PT: Perspective-Taking]

How did the visibility of the Visitor from the Resident's perspective change between the initial and final views?

(A) Occluded to Visible ✓
(B) Visible to Occluded
(C) Always visible
(D) Always occluded

Figure 72: **Sample QA pairs** for VI_V_EXO_DYN, idx 17.

DS_E_EGO_STA - Directional Signage Grounding

idx=94 | Medieval Castle Dungeon Network | Frames



Figure 73: **Frames** for DS_E_EGO_STA, idx 94 (*Medieval Castle Dungeon Network*).

DS_E_EGO_STA - Directional Signage Grounding

idx=94 | Medieval Castle Dungeon Network | Scene Graph


```
{
  "scene_meta": {
    "spatial_scale": "Environmental",
    "scene_dynamics": "Static",
    "perspective": "Ego",
    "task_type": "Directional_Signage_Grounding",
    "theme": "Medieval Castle Dungeon Network"
  },
  "objects": [
    {
      "id": "obj_sign",
      "label": "Iron-Bound Wooden Directional Sign",
      "role": "signage",
      "attributes": {
        "entry_A": {
          "text": "Royal Cells",
          "arrow": "Right"
        },
        "entry_B": {
          "text": "North Tower",
          "arrow": "Left"
        },
        "placement": "Hanging from dungeon ceiling by iron chains"
      }
    }
  ],
  "camera_movement": {
    "type": "Person Walking Simulation",
"trajectory": "Start at far end of corridor facing sign → walk toward sign (sign becomes
readable) → sign passes overhead → continue walking in Left direction."
  },
  "ground_truth": {
    "agent_heading": "Left",
    "destination": "North Tower",
"reasoning": "The agent walked Left, which matches the arrow for 'North Tower' on the sign."
  }
}
```


Figure 74: **Scene graph** for DS_E_EGO_STA, idx 94.

DS_E_EGO_STA - Directional Signage Grounding

idx=94 | Medieval Castle Dungeon Network | Scenario


```
{
"reasoning_goal": "The viewer must determine that the agent is heading toward 'North Tower' by
observing the agent walks Left, matching the Left arrow on the sign for that destination.",
  "timeline": {
"(Setup)": "Static eye-level first-person view from the far end of the Medieval Castle Dungeon
Network. The directional sign is visible ahead in the distance, hanging from dungeon ceiling by
iron chains. No camera or object motion yet — this frame establishes the environmental scale
and starting position.",
"(Event)": "Camera begins a smooth, steady forward walk toward the sign. As the camera
approaches, the sign text becomes legible: 'Royal Cells' with a Right arrow and 'North Tower'
with a Left arrow. The camera passes directly underneath the sign and continues walking in the
Left direction.",
"(Result)": "Camera continues moving steadily in the Left direction. The sign is now behind and
out of frame. The agent's heading matches the Left arrow for 'North Tower'. All objects remain
completely stationary; only the camera moves until the video ends."
  }
}
```


Figure 75: **Scenario** for DS_E_EGO_STA, idx 94.

DS_E_EGO_STA - Directional Signage Grounding

idx=94 | Medieval Castle Dungeon Network | Sample QAs

Q [L1-FL: Frame Localization]

In the initial view, where is the directional sign located in the frame?

(A) Top ✓
(B) Bottom
(C) Left
(D) Right

Q [L2-CM: Camera Motion]

How does the camera move throughout the video?

(A) It walks straight forward
(B) It walks forward then turns left ✓
(C) It walks forward then turns right
(D) It stays fixed in place

Q [L3-CR: Counterfactual]

If the camera had turned right at the sign instead of going left, which destination would the agent be heading toward?

(A) North Tower
(B) Chancery Building
(C) Royal Cells ✓
(D) Quantum Power Grid

Figure 76: **Sample QA pairs** for DS_E_EGO_STA, idx 94.

RV_E_EGO_DYN - Relative Velocity Identification

idx=13 | Go-Kart Circuit | Frames



Figure 77: **Frames** for RV_E_EGO_DYN, idx 13 (*Go-Kart Circuit*).

RV_E_EGO_DYN - Relative Velocity Identification

idx=13 | Go-Kart Circuit | Scene Graph | part 1/2


```
{
  "scene_meta": {
    "spatial_scale": "Environmental",
    "scene_dynamics": "Dynamic",
    "perspective": "Ego",
    "task_type": "Relative_Velocity_Identification",
    "theme": "Go-Kart Circuit"
  },
  "objects": [
    {
      "id": "obj_ego",
      "label": "Pink Go-Kart",
      "role": "ego_agent",
      "motion": "Constant cruising speed, forward — velocity reference frame"
    },
    {
      "id": "obj_slower",
      "label": "Grey Mechanic Cart",
      "role": "slower_than_ego_agent",
      "visual_attributes": {
        "color": "Grey",
        "type": "Mechanic Cart"
      },
      "position": "Ahead of Ego in the Left lane",
      "relative_velocity_to_ego": "Slower-than-Ego (appears to drift backward in ego's frame)"
    },
    {
      "id": "obj_faster",
      "label": "Red Turbo Kart",
      "role": "faster_than_ego_agent",
      "visual_attributes": {
        "color": "Red",
        "type": "Turbo Kart"
      },
      "position": "Initially off-screen behind Ego in the Right lane",
      "relative_velocity_to_ego": "Faster-than-Ego (streaks past ego forward in ego's frame)"
    }
  ],
  "temporal_flow": {
```


Figure 78: **Scene graph (part 1/2)** for RV_E_EGO_DYN, idx 13.

RV_E_EGO_DYN - Relative Velocity Identification

idx=13 | Go-Kart Circuit | Scene Graph | part 2/2


```
    "initial_state": {
"description": "Ego cruises forward at constant speed in its own lane. Grey Mechanic Cart
visible ahead in the Left lane (apparently almost stationary). Red Turbo Kart not yet in frame
(behind ego)."
    },
    "action_sequence": [
      {
        "step": 1,
"action_description": "Grey Mechanic Cart (in Left lane) appears nearly stationary; ego rapidly
closes the gap. It grows larger in frame, slides past ego, drifts off-screen behind — visually
appears to move BACKWARD relative to ego (cue: Slower-than-Ego)."
      },
      {
        "step": 2,
"action_description": "Red Turbo Kart enters frame from behind in Right lane, streaks past ego
at very high relative speed (visibly moving FORWARD in ego's frame), continues forward,
overtakes the slower agent, shrinks to a dot, disappears ahead into the distance (cue:
        Faster-than-Ego)."
      }
    ]
  },
  "camera_movement": {
    "type": "Forward Motion Simulation",
"trajectory": "Straight forward at constant cruising speed from Ego-agent perspective. No lane
    change."
  },
  "ground_truth": {
    "relative_velocity_per_agent": {
      "obj_slower": "Slower-than-Ego",
      "obj_faster": "Faster-than-Ego"
    },
    "velocity_ranking_low_to_high": [
      "obj_slower",
      "obj_ego",
      "obj_faster"
    ]
  }
}
```


Figure 79: **Scene graph (part 2/2)** for RV_E_EGO_DYN, idx 13.

RV_E_EGO_DYN - Relative Velocity Identification

idx=13 | Go-Kart Circuit | Scenario


```
{
"reasoning_goal": "The viewer must classify each non-ego agent's velocity relative to the Ego
(Pink Go-Kart): the Grey Mechanic Cart (in the Left lane, ahead) is Slower-than-Ego because it
visibly drifts BACKWARD in ego's frame; the Red Turbo Kart (entering from behind in the Right
lane) is Faster-than-Ego because it streaks past Ego and shrinks FORWARD into the distance.",
  "timeline": {
"(Setup)": "First-person view from inside the Ego Pink Go-Kart cruising forward at constant
speed in a Go-Kart Circuit environment. The Grey Mechanic Cart is visible ahead in the Left
lane, appearing nearly stationary. The Right lane and the road behind Ego are empty — no other
agents in frame yet.",
"(Phase 1 — Slower-than-Ego Cue)": "The Grey Mechanic Cart (in the Left lane) appears nearly
stationary in absolute terms while Ego closes the gap rapidly. It grows larger in frame, slides
past beside Ego, and drifts off-screen toward the back of frame, visibly moving BACKWARD
relative to Ego. This cue identifies it as Slower-than-Ego.",
"(Phase 2 — Faster-than-Ego Cue)": "The Red Turbo Kart enters frame from behind in the Right
lane at very high relative speed. It streaks past Ego (visibly moving FORWARD in Ego's frame),
continues forward, shrinks to a dot ahead, and disappears into the distance. This cue
identifies it as Faster-than-Ego.",
"(Result)": "Ego continues forward at constant speed. The relative velocity classification is
confirmed: Grey Mechanic Cart = Slower-than-Ego, Red Turbo Kart = Faster-than-Ego. Final
velocity ranking (low to high): Grey Mechanic Cart < Ego (Pink Go-Kart) < Red Turbo Kart."
  }
}
```


Figure 80: **Scenario** for RV_E_EGO_DYN, idx 13.

RV_E_EGO_DYN - Relative Velocity Identification

idx=13 | Go-Kart Circuit | Sample QAs

Q [L1-FL: Frame Localization]

In the initial view, where is the Mechanic Cart located in the frame?

(A) Left side ✓
(B) Center
(C) Right side

Q [L2-AR: Action Recognition]

How does the Turbo Kart appear to move relative to the camera?

(A) Drifts backward
(B) Streaks forward ✓
(C) Stays stationary at the same distance
(D) Moves to the side and disappears

Q [L3-CR: Counterfactual]

If the Mechanic Cart had been moving faster than the camera object, how would it appear to move in the video?

(A) Drift backward and disappear behind
(B) Streak forward and disappear ahead ✓
(C) Stay at the same distance
(D) Move to the side

Figure 81: **Sample QA pairs** for RV_E_EGO_DYN, idx 13.

LS_E_EXO_STA - Landmark Spatial Composition

idx=86 | Polar Research Base | Frames



Figure 82: **Frames** for LS_E_EXO_STA, idx 86 (*Polar Research Base*).

LS_E_EXO_STA - Landmark Spatial Composition

idx=86 | Polar Research Base | Scene Graph | part 1/2


```
{
  "scene_meta": {
    "spatial_scale": "Environmental",
    "scene_dynamics": "Static",
    "perspective": "Exo",
    "task_type": "Landmark_Spatial_Composition",
    "theme": "Polar Research Base"
  },
  "objects": [
    {
      "id": "obj_landmark_1",
      "label": "Bright Orange Main Habitat Module",
      "role": "landmark_start",
      "attributes": {
        "color": "Bright Orange",
        "type": "Main Habitat Module",
        "rooftop_shape": "Geodesic dome roof with triangular panel segments in orange"
      }
    },
    {
      "id": "obj_landmark_2",
      "label": "White Insulated Medical Bay",
      "role": "landmark_craneup",
      "attributes": {
        "color": "White Insulated",
        "type": "Medical Bay",
        "rooftop_shape": "Flat white roof with red cross marking and ventilation units",
        "position_relative_to_L1": "S"
      }
    },
    {
```


Figure 83: **Scene graph (part 1/2)** for LS_E_EXO_STA, idx 86.

LS_E_EXO_STA - Landmark Spatial Composition

idx=86 | Polar Research Base | Scene Graph | part 2/2


```
      "id": "obj_landmark_3",
      "label": "Dark Blue Aurora Observatory",
      "role": "landmark_flight",
      "attributes": {
        "color": "Dark Blue",
        "type": "Aurora Observatory",
        "rooftop_shape": "Retractable slit dome roof with dark blue anti-glare coating"
      }
    }
  ],
  "camera_movement": {
    "type": "Crane-Up then Directional Flight",
    "L2_appear_direction": "S",
    "drift_direction": "E",
"trajectory": "Tight close-up on Bright Orange Main Habitat Module rooftop → rapid crane up
(White Insulated Medical Bay appears to the S, separated by significant distance) → rapid
eastward flight (compass E) traveling a very long distance across the environment, far beyond
the L1-L2 area → tight close-up on Dark Blue Aurora Observatory rooftop. NOTE: L3 is much
farther from L1 than L2 is."
  },
  "ground_truth": {
    "L2_relative_to_L1": "S",
    "drift_direction": "E",
    "L3_relative_to_L1": "E",
"reasoning": "Crane-up revealed the White Insulated Medical Bay to the S of the Bright Orange
Main Habitat Module. Camera then flew eastward (compass E) a very long distance to reach the
Dark Blue Aurora Observatory. Therefore the Dark Blue Aurora Observatory is to the E of the
Bright Orange Main Habitat Module."
  }
}
```


Figure 84: **Scene graph (part 2/2)** for LS_E_EXO_STA, idx 86.

LS_E_EXO_STA - Landmark Spatial Composition

idx=86 | Polar Research Base | Scenario


```
{
"reasoning_goal": "The viewer must deduce that the Dark Blue Aurora Observatory is to the E of
the Bright Orange Main Habitat Module, by combining: (1) crane-up reveals the White Insulated
Medical Bay to the S of the Bright Orange Main Habitat Module, and (2) camera flies eastward
(toward right of screen) to reach the Dark Blue Aurora Observatory.",
  "timeline": {
"(Phase 1 — Crane Up)": "Perfect 90-degree top-down bird's-eye view. Starts as tight close-up
on the Bright Orange Main Habitat Module rooftop (Geodesic dome roof with triangular panel
segments in orange). Camera rapidly ascends. The White Insulated Medical Bay rooftop (Flat
white roof with red cross marking and ventilation units) appears far to the S of the Bright
Orange Main Habitat Module — below (toward bottom of screen), separated by significant
distance. Both rooftops visible simultaneously from high above.",
"(Phase 2 — Fly to L3)": "Camera physically flies eastward (toward right of screen) at high
speed. Bright Orange Main Habitat Module and White Insulated Medical Bay exit frame leftward
(toward left of frame). Camera travels a very long distance across endless wind-blown ice shelf
plains, jagged pressure ridges of compacted sea ice, and snow-trenched access corridors between
habitat modules, far beyond the L1-L2 area. The Dark Blue Aurora Observatory rooftop
(Retractable slit dome roof with dark blue anti-glare coating) enters frame and becomes
centered in tight close-up. L3 is much farther from L1 than L2 was. Video ends."
  }
}
```


Figure 85: **Scenario** for LS_E_EXO_STA, idx 86.

LS_E_EXO_STA - Landmark Spatial Composition

idx=86 | Polar Research Base | Sample QAs

Q [L1-OA: Object Attribute]

What is the color of the Medical Bay?

(A) Natural Wood
(B) Olive
(C) White Insulated ✓
(D) Red Clay

Q [L2-CM: Camera Motion]

In which compass direction does the camera fly horizontally?

(A) North
(B) East ✓
(C) South
(D) West

Q [L3-CR: Counterfactual]

If the camera had flown West instead of East, in which compass direction would the Aurora Observatory be located relative to the Camera?

(A) North
(B) East ✓
(C) South
(D) West

Figure 86: **Sample QA pairs** for LS_E_EXO_STA, idx 86.

BT_E_EXO_DYN - Behavioral Trigger Identification

idx=94 | Wasteland highway | Frames

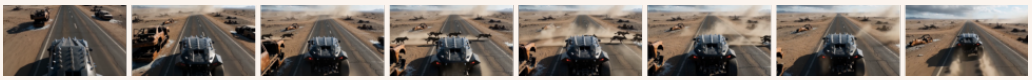


Figure 87: **Frames** for BT_E_EXO_DYN, idx 94 (*Wasteland Highway*).

BT_E_EXO_DYN - Behavioral Trigger Identification

idx=94 | Wasteland highway | Scene Graph | part 1/2


```
{
  "scene_meta": {
    "spatial_scale": "Environmental",
    "scene_dynamics": "Dynamic",
    "perspective": "Exo",
    "task_type": "Behavioral_Trigger_Identification",
    "theme": "Wasteland highway"
  },
  "objects": [
    {
      "id": "obj_agent",
      "label": "Black Spiked Post-Apocalyptic Combat Vehicle",
      "role": "agent",
      "attributes": {
        "color": "Black with Rust",
        "type": "Vehicle"
      }
    },
    {
      "id": "obj_distractor",
      "label": "Rusted Wrecked Car Stack",
      "role": "distractor",
      "attributes": {
        "color": "Rust Brown",
        "type": "Static Wreck Stack",
        "position": "left side of path"
      }
    },
    {
      "id": "obj_trigger",
      "label": "Pack of Mutant Wild Dogs",
      "role": "trigger",
      "attributes": {
        "color": "Greenish-Brown",
        "type": "Animals",
        "entry_direction": "left side of path"
      }
```


Figure 88: **Scene graph (part 1/2)** for BT_E_EXO_DYN, idx 94.

BT_E_EXO_DYN - Behavioral Trigger Identification

idx=94 | Wasteland highway | Scene Graph | part 2/2


```
    }
  ],
  "temporal_flow": {
    "initial_state": {
"description": "Agent is moving forward along a clear, unobstructed path. The Rusted Wrecked
Car Stack is visible on the left side of the path as a static scene element."
    },
    "action_sequence": [
      {
        "step": 1,
"action_description": "Black Spiked Post-Apocalyptic Combat Vehicle moves at constant speed on
a clear path. Path ahead is unobstructed. Rusted Wrecked Car Stack is visible on the left side
of the path."
      },
      {
        "step": 2,
"action_description": "Pack of Mutant Wild Dogs enters the camera frame from the left side and
appears on the agent's path."
      },
      {
        "step": 3,
"action_description": "Black Spiked Post-Apocalyptic Combat Vehicle immediately performs
Wait_and_Resume maneuver in response to the trigger."
      }
    ]
  },
  "camera_movement": {
    "type": "High-Angle Drone Tracking",
"trajectory": "Camera follows agent from behind at high angle, traveling forward with the agent
throughout the scene."
  },
  "ground_truth": {
    "trigger_object_id": "obj_trigger",
    "trigger_object_label": "Pack of Mutant Wild Dogs",
    "reaction_type": "Wait_and_Resume"
  }
}
```


Figure 89: **Scene graph (part 2/2)** for BT_E_EXO_DYN, idx 94.

BT_E_EXO_DYN - Behavioral Trigger Identification

idx=94 | Wasteland highway | Scenario


```
{
"reasoning_goal": "The viewer must visually verify that the Black Spiked Post-Apocalyptic
Combat Vehicle's wait-and-resume reaction was directly caused by the Pack of Mutant Wild
  Dogs.",
  "timeline": {
"(Setup)": "High-angle drone camera tracks the Black Spiked Post-Apocalyptic Combat Vehicle
from behind and above as it moves forward at a constant speed in the Wasteland highway
environment. The Rusted Wrecked Car Stack is clearly visible on the left side of the path. The
path ahead is completely clear and unobstructed.",
"(Event)": "The Pack of Mutant Wild Dogs suddenly enters the frame from the left side and
crosses the path. The Black Spiked Post-Apocalyptic Combat Vehicle immediately reacts,
decelerates, and comes to a complete stop, holding position while the Pack of Mutant Wild Dogs
clears the path.",
"(Result)": "Once the Pack of Mutant Wild Dogs has cleared the path, the Black Spiked
Post-Apocalyptic Combat Vehicle visibly accelerates and resumes forward motion in the original
direction, confirming the Wait_and_Resume maneuver. The camera holds steady tracking the agent,
ensuring the causal link between the Pack of Mutant Wild Dogs's sudden appearance and the
agent's wait-and-resume reaction is unambiguous."
  }
}
```


Figure 90: **Scenario** for BT_E_EXO_DYN, idx 94.

BT_E_EXO_DYN - Behavioral Trigger Identification

idx=94 | Wasteland highway | Sample QAs

Q [L1-OE: Object Existence]

Which object does NOT appear in the video?

(A) Rusted Wrecked Car Stack
(B) Black Spiked Post-Apocalyptic Combat Vehicle
(C) Silver Convertible Sports Car ✓
(D) Pack of Mutant Wild Dogs

Q [L2-AR: Action Recognition]

What action does the Black Spiked Post-Apocalyptic Combat Vehicle perform during the video?

(A) Continues moving without stopping
(B) Comes to a complete stop
(C) Turns around and moves backward
(D) Waits and resumes moving after a short pause ✓

Q [L3-CR: Counterfactual]

If the Pack of Mutant Wild Dogs had not entered the path, what would the Black Spiked Post-Apocalyptic Combat Vehicle most likely do?

(A) Come to a complete stop anyway
(B) Continue moving forward at the same speed ✓
(C) Turn around and move backward
(D) Collide with the Rusted Wrecked Car Stack

Figure 91: **Sample QA pairs** for BT_E_EXO_DYN, idx 94.