VIDEO-HOCA: A DIAGNOSTIC BENCHMARK FOR PHYSICAL ANOMALY REASONING IN VIDEO-LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video-HOCA evaluates how video-language models judge and explain visible physical anomalies. An operational Ontological–Causal taxonomy separates entity-centered from relation-centered violations, and four tasks on the same anomaly collection assess plausibility judgment, category attribution, fine-grained recognition, and mechanism explanation. The benchmark contains 1,439 videos and 3,470 question-answer pairs with human-verified labels and reference answers, plus a control of 250 generator-matched normal/anomalous pairs. Across 20 Instruct-mode models, nine-way category attribution remains difficult (mean Ontological/Causal macro-F1 of 37.9/46.0), and no model reaches the human reference on recognition or explanation. On the matched-generator control, the mean plausibility accuracy of 17 open-weight models is 64.4%, compared with 74.0% on the original real-versus-generated split; every model stays above chance, but the most accurate model on the original split is the least accurate on the control. For the tested model, shuffling frames barely changes plausibility accuracy yet lowers recognition and explanation scores. Annotation agreement and judge-sensitivity analyses characterize the reliability of these observations.

1 INTRODUCTION

Generated videos can depict familiar objects and actions while violating physical expectations (Bai et al., 2025b; Bansal et al., 2025). Assessing such videos requires more than recognizing their visible content. Consider a clip where coffee enters a cup but its level remains unchanged (Figure 4): describing the pouring, identifying the unchanged level, and explaining the expected volume increase are distinct judgments. An evaluation that separates these judgments can identify what a model gets right and where its interpretation fails.

Existing benchmarks address complementary parts of this problem. MVBench and Video-MME evaluate broad video understanding (Li et al., 2024; Fu et al., 2025), while PhysBench covers physical-world understanding across image, video, and text inputs (Chow et al., 2025). IPV-Bench examines impossible events, and VideoHallu and VidHal study hallucinations in synthetic videos and temporal descriptions (Bai et al., 2025b; Li et al., 2025; Choong et al., 2026). Our focus is a shared collection of visible physical violations on which plausibility judgment, category attribution, recognition, and explanation can be examined separately.

The central design challenge is to connect each score to the judgment it measures. Binary tests contrasting real normal videos with generated anomalous videos couple source and label. Taxonomy attribution requires reproducible category boundaries, and open-ended explanations require a scoring protocol that can be checked against human judgments. Different answer spaces also prevent direct interpretation of cross-task score gaps. These considerations motivate matched-source controls, an annotation-agreement study, and alternative-judge and scoring analyses alongside the task results.

We introduce **Video-HOCA**, a diagnostic benchmark for visible, everyday macroscopic physical anomalies. Its operational taxonomy distinguishes **Ontological** violations of an entity's properties, state, or biological capability from **Causal** violations of relations among entities and the environment. Four complementary tasks assess plausibility, category attribution, fine-grained recognition, and physical explanation. In the coffee example, they distinguish a binary judgment from identifying the anomaly and articulating its physical mechanism. We study three questions:

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.

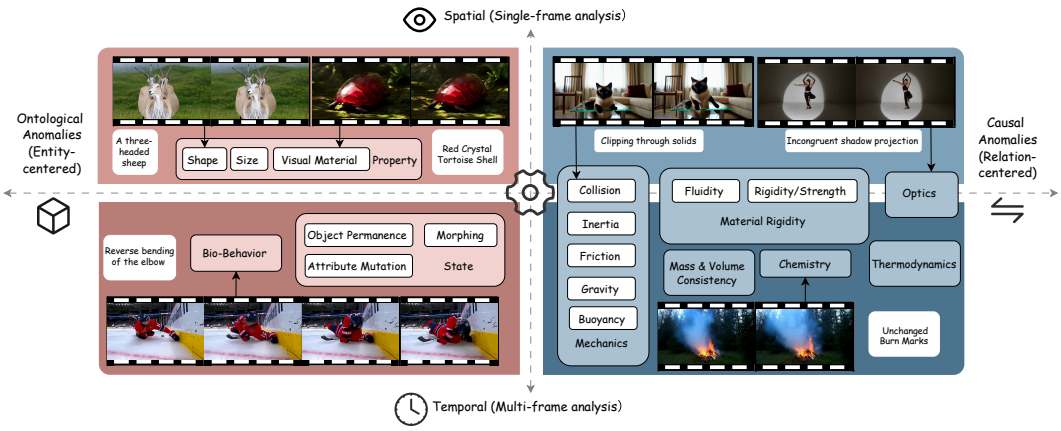


Figure 1: **Video-HOCA overview.** The benchmark organizes physical anomalies into Ontological (entity-centered) and Causal (relation-centered) branches, totaling 9 L2 groups and 18 L3 tags (full hierarchy in Appendix Table 9); spatial and temporal anomaly cases span both branches. Real normal videos provide the original Task I controls.

- **RQ1:** How do models perform on each task under its specified input, label, and scoring conditions?
- **RQ2:** How do anomaly type, temporal information, and decoding configuration affect the observed task profiles?
- **RQ3:** How sensitive are the measurements to video sources, annotation boundaries, judges, and score aggregation?

Video-HOCA combines existing video sources with targeted generation to fill category deficits; human inspection guides prompt refinement and determines which generated videos are retained. The main benchmark contains 1,439 videos and 3,470 QA pairs, and a separate control provides 250 normal/anomalous pairs from matched generators. Across 20 Instruct configurations, nine-way attribution remains difficult (mean Ontological/Causal macro-F1 of 37.9/46.0), and no model reaches the human reference on recognition or explanation. On the matched-generator control, mean open-weight plausibility accuracy is 64.4% versus 74.0% on the original split, and the most accurate model on the original split is the least accurate on the control.

Our contributions are threefold. First, we define an operational taxonomy and a four-task protocol for diagnosing visible physical anomalies (§3.1, §3.3). Second, we document the construction process and assess measurement reliability through a generator-matched plausibility control, an annotation-agreement study, and judge-sensitivity analyses (§3.2, §4.2, §4.4). Third, we report task-specific results for 20 models and analyze how anomaly type, temporal input, and decoding configuration change them (§4.2, §4.3).

2 Related Work

Table 1 compares representative benchmarks by evaluation focus, taxonomy, task format, and video source.

Video understanding and anomaly evaluation. Perception Test, MVBench, MMBench-Video, and Video-MME cover perceptual, temporal, and multimodal video understanding (Patraucean et al., 2023; Li et al., 2024; Fang et al., 2024; Fu et al., 2025); MMVU and Video-MMMU add disciplinary knowledge (Zhao et al., 2025; Hu et al., 2025). VideoHallu evaluates hallucinations in synthetic videos (Li et al., 2025); IPV-Bench examines impossible events (Bai et al., 2025b); VideoPhy-2 and VBench-2.0 evaluate physical commonsense and generation faithfulness (Bansal et al., 2025; Zheng et al., 2025). HumanSAM studies human-centric anomaly classification (Liu et al., 2025).

Physical understanding and controlled hallucination tests. PhysBench spans physical properties, relationships, scenes, and dynamics using image/video/text inputs (Chow et al., 2025). VidHal

Table 1: **Comparison with video benchmarks.** ✓ supported, ◐ partial, ✗ absent. Plausibility judgment: yes/no question to the evaluated model on whether the video is physically plausible; Category attribution: the model assigns an anomaly category; Matched control: same-generator normal counterparts as negatives. Partial: physics is one of several categories in VideoHallu and IPV-Bench, and IPV-Bench's yes/no task asks whether a video is AI-generated; VideoPhy-2 tags physical laws and evaluates video generators. R: real; S: simulated; G: generated.

Benchmark	Physical anomaly	Anomaly taxonomy	Plausibility judgment	Category attribution	Matched control	Source
General video understanding						
MVBench (Li et al., 2024)	✗	✗	✗	✗	✗	R+S
MMBench-Video (Fang et al., 2024)	✗	✗	✗	✗	✗	R
Video-MME (Fu et al., 2025)	✗	✗	✗	✗	✗	R
Synthetic-video and physical-anomaly evaluation						
VideoHallu (Li et al., 2025)	◐	✓	✗	✗	✗	G
IPV-Bench (Bai et al., 2025b)	◐	✓	◐	✗	✗	G+R
VideoPhy-2 (Bansal et al., 2025)	✓	◐	✗	✗	✗	G
Video-HOCA (ours)	✓	✓	✓	✓	✓	G+R

evaluates temporal hallucinations through captions with different hallucination levels and caption ordering (Choong et al., 2026). VidPair-Halluc controls visual backgrounds while varying foreground semantics in paired-video QA (Huang et al., 2026). Video-HOCA instead organizes visible physical violations and their explanations; its same-generator control addresses a different nuisance variable from background-controlled semantic contrasts.

Diagnostic scope. Video-HOCA combines an entity-/relation-centered taxonomy with four complementary tasks on a shared anomaly collection, together with source controls and annotation and judge checks. The taxonomy is an operational rule motivated by work on core physical expectations about objects and their interactions (Spelke, 1990; Baillargeon, 2004). We evaluate models that answer language prompts directly. V-JEPA learns predictive visual representations without a language-answering objective (Bardes et al., 2024); using it here would require a decoder or task head and a separate adaptation protocol. Our findings do not cover all predictive video models.

3 Video-HOCA: Benchmark Design

3.1 Operational Taxonomy

Video-HOCA uses an operational taxonomy: each anomaly is labeled by its *primary violated unit*. **Ontological** anomalies are *entity-centered*: the object or agent is internally inconsistent, unstable, or capable of something its own definition should not allow. **Causal** anomalies are *relation-centered*: otherwise valid entities interact with each other or with the environment in a way that violates a physical regularity. Figure 1 summarizes the two L1 branches and their L2 coverage.

The taxonomy has three levels. **L1** is the Ontological–Causal split used for the main analyses. **L2** contains 9 coarse anomaly groups used for balancing and annotation. **L3** contains the fine-grained anomaly tags used to write options, reference answers, and metadata for individual videos.

L2 groups. The **Ontological** branch contains *Property*, *State*, and *Bio-Behavior*. The **Causal** branch contains *Mechanics*, *Material and Rigidity*, *Mass and Volume Consistency*, *Thermodynamics*, *Chemistry*, and *Optics*. These groups define the annotation categories and are used to balance the dataset and construct the tasks. Full definitions and canonical examples are given in Appendix A.8; the resulting taxonomy distribution is visualized in Figure 6.

Annotation agreement. Three annotators re-labeled 120 stratified anomalous videos using the frozen guideline. They assigned L2 labels directly; L1 labels were induced through the fixed L2-to-L1

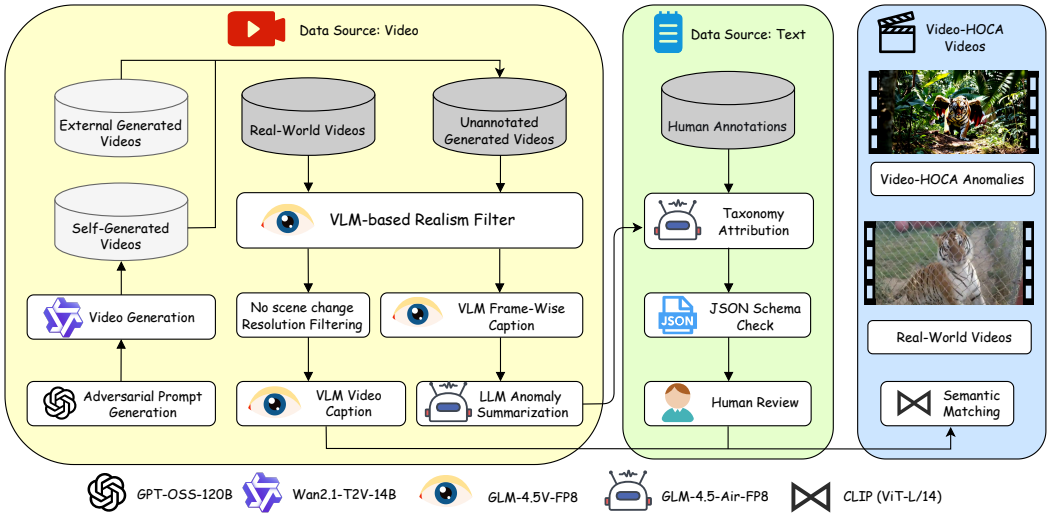


Figure 2: **Data construction and annotation pipeline.** Candidates are collected from existing video sources or generated to supplement underrepresented categories. Model-assisted descriptions and taxonomy assignments undergo human review. Real normal videos are filtered and semantically matched for the original Task I split. The human-guided prompt refinement within targeted generation and the supplementary generated normal controls are detailed in §3.2.

mapping. Fleiss' κ is 0.821 for L1 and 0.708 for L2. Appendix A.3 describes the annotation protocol and the interpretation of these measurements.

3.2 Data Construction and Verification

Existing sources and category coverage. Figure 2 summarizes the construction. We first collect and manually review candidates from existing video sources. Of the 630 retained single-anomaly videos, 473 (75.1%) come from these sources: 375 from Impossible, 57 from VideoPhy-2, and 41 from VBench. Targeted Wan generation supplies the remaining 157 (24.9%). We aim for approximately 35 accepted videos per L3 category; final counts are 31–37.

Human-guided prompt refinement. For underrepresented categories, we supply category definitions to the prompt-generation process using GPT-OSS-120B (OpenAI, 2025) and generate videos with Wan 2.1-T2V-14B (Team Wan et al., 2025). Human inspection of the rendered videos determines whether the intended anomaly and visual quality are suitable. Unsuitable outputs inform prompt revisions; requests for visual clarity and photorealistic or cinematic appearance were added after early outputs lacked these qualities. Once prompts produce usable candidates, we stop refining them and screen videos for inclusion. Appendix A.2 gives the workflow and examples.

Candidate annotation and final verification. Existing descriptions are mapped to the taxonomy with GLM-4.5-Air (GLM-4.5 Team, 2025). For clips without reliable descriptions, GLM-4.5V-FP8 generates frame-wise captions. Temporal summaries of these captions provide candidate anomaly descriptions and labels, while schema checks enforce the annotation format. Human reviewers inspect the videos, correct the scene, anomaly, principle, and taxonomy fields, and determine final labels and reference answers using the interface in Figure 5. Unclear cases are rejected, and clips with multiple salient anomalies are directed to the multi-anomaly split.

Dataset scale and scope. The anomaly-focused set contains 809 videos: 630 single-anomaly clips, 59 additional temporal-grounding clips, and 120 multi-anomaly clips. Adding 630 real normal clips gives 1,439 videos. The 3,470 QA pairs comprise 1,260 for Task I (630 anomalous and 630 normal), 630 each for Tasks II–IV, 200 for temporal grounding, and 120 for multi-anomaly detection. Source datasets and generating models are distinct provenance dimensions: externally collected videos can also originate from Wan or other generators. Our scope is visible, everyday macroscopic anomalies

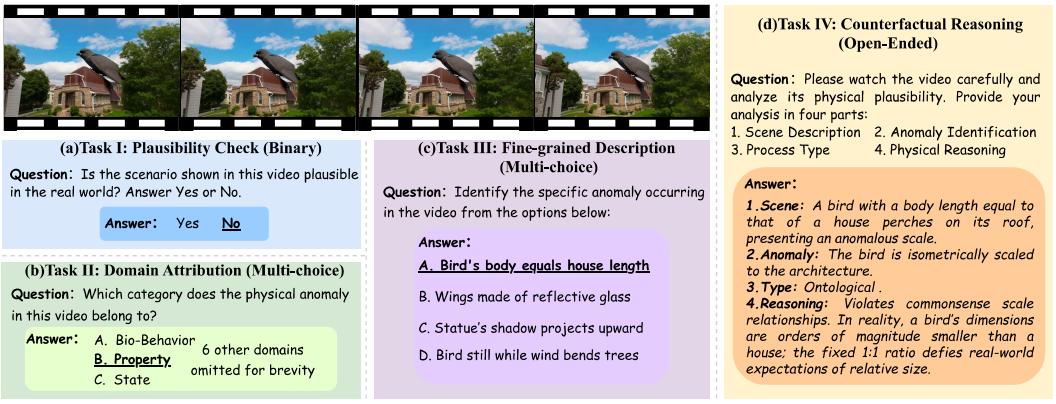


Figure 3: **Four complementary evaluation tasks on a shared video.** The example illustrates (Task I) real-world plausibility judgment, (Task II) anomaly category attribution, (Task III) fine-grained anomaly recognition, and (Task IV) open-ended physical explanation. The tasks use different answer formats and scoring rules.

within the 9-group/18-tag taxonomy, not quantitative physics, hidden-state inference, or long-horizon causal reasoning.

Original real-video controls. Real clips come from Panda-70M (Chen et al., 2024a), with scene, resolution, caption, and realism filtering. GLM-4.5V-FP8 generates whole-video textual descriptions. We encode these descriptions using the text encoder of OpenAI CLIP ViT-L/14 (Radford et al., 2021) and match videos by cosine similarity between description embeddings. Matching uses text, not CLIP frame features. It aligns described scene content but does not control real/generated source differences; the original Task I split couples source and plausibility.

Supplementary generated normal controls. To test this coupling, we generated physically normal counterparts for 272 anomaly candidates from Wan 2.1 or HunyuanVideo. Prompt correction and video-level verification yielded 250 accepted pairs (500 videos): 167 Wan pairs and 83 HunyuanVideo pairs. Both videos in a pair use the same generator, with resolution, frame rate, and duration aligned. This control is reported separately from the main-set video and QA counts; its anomaly side reuses main-set candidates. Both classes are generated, removing the binary source-label coupling; rendering quality and anomaly salience may still differ.

3.3 Tasks and Evaluation

Complementary task definitions. **Task I: Plausibility Judgment.** The exact prompt is *"Is the scenario shown in this video plausible in the real world? Answer Yes or No."* It elicits a binary real-world plausibility judgment on videos labeled for physical plausibility. The wording is broader than a request to name a violated physical law. A correct answer therefore does not establish that the model identified or explained the intended anomaly. We report both the original mixed-source split and the supplementary same-generator control.

Task II: Anomaly Category Attribution. Given an anomalous video, the model assigns it to one of the 9 L2 groups, whose short definitions are included in the prompt. A model may notice that something is wrong but still assign the wrong physical category.

Task III: Fine-grained Anomaly Recognition. This multiple-choice task pairs the ground-truth description with adversarial distractors that GPT-OSS-120B generates from the original annotation and from other taxonomy categories: the alternatives are designed to be plausible descriptions that conflict with the observed event. Human accuracy is 96.0% (Table 2).

Task IV: Open-ended Physical Reasoning. The model gives a structured answer with four parts: scene description, anomaly identification, anomaly type, and physical explanation. The goal is to test

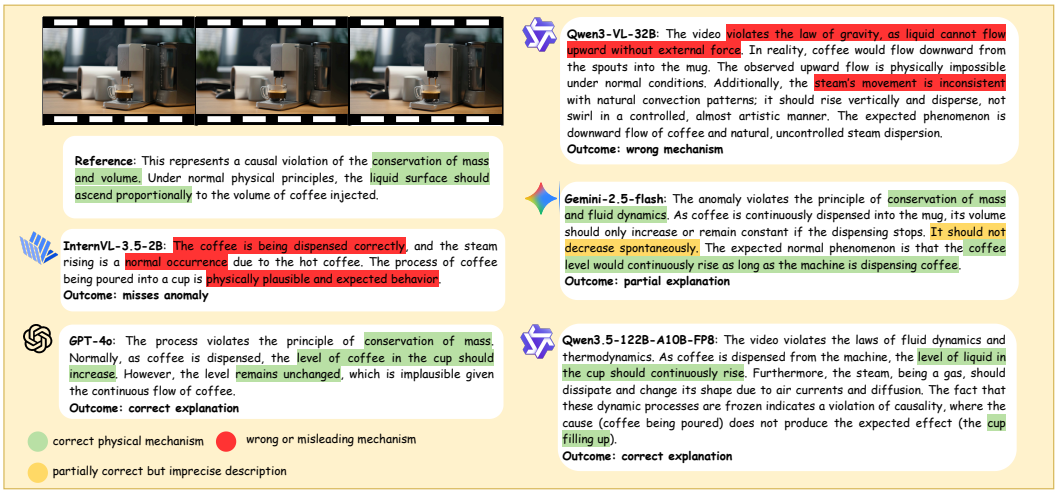


Figure 4: **Task-IV failure modes on a coffee-level anomaly.** All models observe the same video, where coffee is continuously dispensed but the liquid level does not rise. Green marks the correct conservation-based mechanism, red marks wrong or misleading mechanisms, and yellow marks partially correct but imprecise event descriptions. The example shows that models may recognize visible activity while missing or misattributing the physical mechanism.

whether the model can explain the violated physical relation and the expected normal outcome, not only point to the abnormal visual event.

Beyond the four main tasks, Video-HOCA includes two additional probes: *temporal grounding*, which asks for the anomalous time interval, and *multi-anomaly detection*, which asks models to identify concurrent violations in the same video. We use these probes for auxiliary analysis rather than for the main leaderboard. Temporal grounding contributes 200 interval-prediction QA pairs, drawn from 59 newly annotated clips and 141 single-anomaly clips that are reused from the main set with added temporal-interval labels (an annotation example is shown in Figure 7). Multi-anomaly detection uses the 120 multi-anomaly clips with 64-frame input; its exact instruction, definitions, and an example option list are in Appendix A.10. We also run a multi-frame scaling probe that compares the four main tasks under 16-frame and 64-frame inputs.

Task-IV judge protocol. Task IV uses rubric-based evaluation against structured reference answers rather than free-form commonsense judging. GPT-OSS-120B (OpenAI, 2025) scores scene understanding, anomaly identification, anomaly-type alignment, and physical-reasoning quality against the paired reference explanation. Full prompt text and scoring templates are given in Appendix A.14, and judge validation (human–judge and judge–judge) is reported in §4.4.

Metrics. Task I reports ACC and F1 on the original balanced real-normal/generated-anomalous split and ACC on the supplementary generator-matched control. Task II and Task III report macro-F1 on the Ontological and Causal subsets. Macro-F1 averages the label-wise F1 scores in each task, complementing accuracy when prediction performance is uneven across labels. Task IV uses the rubric-based judge score. For compact reporting, we also compute a summary Overall column from Task I F1, Task II/III macro-F1, and the Task IV judge score, normalized via z-scores; the summary is a descriptive index rather than a calibrated reasoning score. It includes the original, source-confounded Task I split. The reference statistics are fixed to the 20 Instruct models for the comparisons in this paper; Appendix A.15 specifies how to compare added models and reports ranking sensitivity under both judges and an alternative weighting. Our conclusions rest on per-task evidence rather than on this index.

Chance reference and comparability. Uniform random answers give expected accuracy of 50% for binary Task I, $1/9 \approx 11.1\%$ for nine-way Task II, and 25% for four-option Task III. These are accuracy references, not the chance values of the reported branch-wise macro-F1 or Task IV rubric

scores. Different answer spaces, candidate options, and metrics preclude inferring an ability gap by directly comparing task scores.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

The main table compares Instruct-mode models under a common decoding protocol. Thinking configurations, which use different sampling settings and token budgets, are evaluated separately. We evaluate 20 zero-shot Instruct configurations in four groups: *Qwen-VL/Qwen3.5* (Qwen2.5-VL-7B (Qwen Team, 2025), Qwen3-VL-2B/8B/32B and Qwen3-VL-30B-A3B (Bai et al., 2025a), Qwen3.5-9B/27B/35B-A3B/122B-A10B-FP8 (Qwen Team, 2026)); *InternVL* (InternVL-2.5-8B (Chen et al., 2024b), InternVL-3.5-2B/8B/14B/38B and InternVL-3.5-30B-A3B (Chen et al., 2024c; Wang et al., 2025a)); *GLM-4.6V* (GLM-4.6V-9B-Flash, GLM-4.6V-106B-FP8 (GLM-V Team, 2025)); and *closed-source* systems (GPT-4o (Hurst et al., 2024), Gemini-2.5-Flash (Comanici et al., 2025), Gemini-3-Flash (Google DeepMind, 2025)). Task IV answers are scored against structured reference explanations by GPT-OSS-120B (OpenAI, 2025); we also report a DeepSeek-V4-Flash (DeepSeek-AI, 2026) judge variant, denoted as Task IV (D).

Decoding and video input. The Instruct-mode main table uses deterministic decoding (temperature $= 0$, no sampling, $1{,}024$-token budget) on all four tasks; Thinking-mode results are kept out of the leaderboard. The Thinking-vs.-Instruct controls in §4.3 use vendor-recommended sampling and an $8{,}192$-token budget (full settings in Appendix A.17). Unless otherwise noted, model inputs use 16 uniformly sampled frames; the multi-frame scaling probe additionally uses 64 frames, and multi-anomaly detection uses 64-frame input. Single-frame and shuffled-frame variants (§4.3) are temporal-confound controls, not part of the leaderboard. We inspected model outputs and judge responses and re-ran missing or failed inference and scoring calls; all tables use the completed records.

4.2 RQ1: TASK PROFILES AND CONTROLLED PLAUSIBILITY

Table 2 reports the four tasks separately. Computed from its rounded values, mean Ontological/Causal macro-F1 is 37.9/46.0 for attribution and 73.3/65.8 for recognition; the best attribution scores per branch are 50.2 and 65.2. No configuration reaches the human reference on Task III (96.0) or Task IV (81.4 under GPT-OSS). Model scale does not order the results within a family: Qwen3.5-27B scores above the larger Qwen3.5-122B-A10B-FP8 on Tasks II–IV. Individual profiles also diverge: Qwen3.5-35B-A3B scores 21.9/1.3 on Task II and 32.7/25.3 on Task III, yet 75.8/61.1 on Task IV. These contrasts depend on answer format and scoring and are best read as task profiles. Figure 4 illustrates distinct explanation errors, such as missing the anomaly or naming an incorrect mechanism; it does not estimate how often each error occurs. **Finding 1:** *the four tasks yield different model profiles; no configuration exceeds 50.2/65.2 Ontological/Causal macro-F1 on nine-way attribution, and recognition and explanation remain below the human reference.*

Source-controlled plausibility. All 17 open-weight models were evaluated on the 250 matched pairs using the same Task I protocol. Accuracy ranges from 59.2% to 69.6%, averaging 64.4%; pair-bootstrap 95% confidence intervals exclude 50% for each model. For the same 17 open-weight models, mean accuracy on the original mixed-source split is 74.0%. Rankings also change: the Spearman correlation between original-split and controlled accuracy is 0.52, and Qwen3.5-122B-A10B-FP8 has the highest original-split accuracy (88.0%) but the lowest controlled accuracy (59.2%). Because the two splits also differ in sample composition, their difference is not an isolated causal estimate of the contribution of source cues. Dynamic-anomaly results span 56.6–68.4%, so the signal is not confined to static cases (Appendix A.1). **Finding 2:** *plausibility discrimination remains above chance without the real-versus-generated contrast, but accuracy is lower and original-split rankings are only partly preserved; this does not establish source-independent physical understanding.*

Table 2: Main Instruct-mode task profiles. Task I reports ACC/F1 (%) on the original mixed-source split, not the matched-source control; Tasks II–III report Ontological/Causal macro-F1 (%); Task IV reports Ontological/Causal rubric scores (0–100), with Task IV (D) using DeepSeek-V4-Flash. Overall standardizes Task I F1, count-weighted Task II/III macro-F1, and the overall Task IV score across the 20 evaluated models; it is not a calibrated physical-reasoning score. Bold and underlining mark the highest and second-highest displayed values in each column (including ties); the Human row is excluded from this marking.

Model	Task I		Task II		Task III		Task IV		Task IV (D)		Overall
	ACC	F1	Onto.	Caus.	Onto.	Caus.	Onto.	Caus.	Onto.	Caus.	
Human	–	–	–	–	96.0		81.4		84.9		–
Open-weight dense models (<7B)											
InternVL-3.5-2B	61.7	69.6	25.0	49.6	70.4	55.6	40.9	47.6	37.2	43.8	38.4
Qwen3-VL-2B	71.7	75.0	34.9	46.6	73.1	67.8	50.2	51.0	44.6	44.6	44.8
Open-weight dense models (≥7B)											
InternVL-2.5-8B	64.6	72.3	42.3	39.5	78.6	63.6	54.3	49.5	51.3	47.1	43.3
Qwen2.5-VL-7B	71.4	76.5	34.3	53.1	67.9	60.9	52.2	54.3	48.1	49.2	46.3
InternVL-3.5-8B	67.8	74.7	40.5	49.4	74.2	67.4	54.1	47.4	51.6	44.4	45.4
Qwen3-VL-8B	75.4	78.2	39.8	49.7	83.7	76.6	64.6	63.7	62.9	59.1	53.1
Qwen3.5-9B	79.8	82.6	34.2	30.9	74.9	67.1	75.3	58.3	71.4	54.8	51.3
GLM-4.6V-9B-Flash	63.6	72.6	38.8	47.3	78.0	67.4	71.0	54.8	68.4	52.3	47.8
InternVL-3.5-14B	73.0	77.6	40.7	59.8	77.7	69.9	63.2	56.0	60.5	53.3	51.5
Qwen3.5-27B	85.6	86.2	47.7	**65.2**	**90.2**	**85.7**	77.5	65.1	75.7	62.2	**63.1**
Qwen3-VL-32B	79.5	82.3	46.0	55.6	84.1	79.0	**78.5**	62.9	75.9	59.3	58.4
InternVL-3.5-38B	75.6	77.5	31.8	54.6	78.3	77.4	69.2	55.7	67.9	52.0	51.5
Open-weight MoE models											
InternVL-3.5-30B-A3B	68.7	74.4	43.8	51.2	82.1	71.2	63.2	48.5	61.2	46.5	48.0
Qwen3-VL-30B-A3B	78.6	80.8	46.4	50.2	81.6	73.4	74.6	64.4	69.8	58.8	56.1
Qwen3.5-35B-A3B	82.9	84.6	21.9	1.3	32.7	25.3	75.8	61.1	74.9	58.2	41.6
GLM-4.6V-106B-FP8	69.8	76.2	31.5	21.9	24.9	18.0	71.0	58.9	70.5	57.3	38.7
Qwen3.5-122B-A10B-FP8	**88.0**	**87.5**	45.0	41.0	85.9	79.2	73.1	61.4	69.0	58.8	58.2
Closed-source systems											
GPT-4o	74.0	75.2	19.5	33.1	59.3	57.1	69.4	63.0	71.0	63.9	44.8
Gemini-2.5-Flash	75.5	77.7	42.8	59.9	82.8	73.0	73.6	65.6	75.6	65.1	55.6
Gemini-3-Flash	84.4	85.5	**50.2**	60.9	85.6	80.9	76.6	**67.6**	**76.4**	**66.8**	62.1

4.3 RQ2: Anomaly Type, Temporal Input, and Model Configuration

Task-conditional anomaly differences. For Qwen3.5-27B, Ontological/Causal scores are 90.2/85.7 on Task III and 77.5/65.1 on Task IV, whereas Task II reverses direction (47.7/65.2). Other configurations differ: Qwen3.5-122B-A10B-FP8 has 45.0/41.0 on Task II. Thus neither branch is uniformly harder across tasks and models. These are within-task comparisons; the differences do not identify an intrinsic causal complexity independent of category composition or annotation ambiguity.

Temporal demand. The single-anomaly set contains 130 static and 124 dynamic Ontological clips, and 128 static and 248 dynamic Causal clips. A dynamic item requires cross-frame evidence; a static item can be judged from a representative frame. Temporal demand and taxonomy branch are therefore associated, and pooled branch differences should not be interpreted as a pure taxonomy effect. Appendix A.13 reports all 20 models and the branch-by-temporal summary. Averaged over models, Ontological/Causal Task IV scores are 66.9/62.8 on static clips and 65.9/55.3 on dynamic clips; this descriptive pattern remains conditional on category composition. **Finding 3:** *neither branch is uniformly harder across tasks and models; the lowest averaged scores occur on dynamic Causal clips, where branch and temporal demand co-vary.*

Input controls. For Qwen3.5-122B-A10B-FP8, shuffling the 16 input frames changes Task I accuracy only from 88.0 to 87.9, whereas using one middle frame reduces it to 80.6. Task III Ontological/Causal scores fall from 85.9/79.2 to 81.8/75.5 under shuffling, and Task IV falls from

73.1/61.4 to 68.5/59.0 (Appendix A.12). **Finding 4:** *in this configuration, frame coverage supports binary judgment, while correct order also matters for recognition and explanation, especially Ontological explanation.*

Decoding controls for Thinking vs. Instruct. Table 8 compares deterministic Instruct, Instruct-Align (the same Instruct checkpoint with Thinking's sampling parameters and token budget), and native Thinking. For GLM-4.6V-106B-FP8, Thinking raises Task III from 24.9/18.0 to 78.9/72.6, whereas Instruct-Align reaches 32.5/26.8. Alignment also does not consistently reproduce Thinking's gains for Qwen3.5-122B-A10B-FP8, and Qwen3-VL-32B varies by task. **Finding 5:** *matching sampling settings and token budget does not consistently reproduce Thinking-mode gains*; these are configuration comparisons, not isolated causal effects of reasoning traces.

4.4 RQ3: Reliability and Sensitivity

Human and alternative-judge checks. On 100 responses from four model configurations, human versus GPT-OSS scores have Pearson $r = 0.896$ and Spearman $\rho = 0.891$. Rescoring 630 responses for each of 20 models with DeepSeek produces 12,600 paired judgments: response-level $r = 0.877$, $\rho = 0.884$, and model-ranking $\rho = 0.970$. The judges agree on 94.2% of model-pair orderings and all pairs separated by more than five points (Appendix A.4). **Finding 6:** *human and judge scores agree on comparative ordering, but not necessarily on absolute calibration or on every individual judgment.*

Rubric wording, length, and model family. Removing two conciseness/brevity instructions changes mean GPT-OSS Task IV scores by about -2.0 points; overall ranking correlation is 0.996 with a maximum one-position shift. Length–score correlation is -0.053 and -0.058 after controlling for model and video effects. The available human-score comparison detects no family effect ($p = 0.73$), but a null result on this sample does not establish the absence of bias. Appendix A.4 reports the checks and their scope.

Generator composition. Removing each of eight major generating sources in turn, together with its matched real controls, gives full-ranking Spearman correlations of 0.991–1.000 and at most two-position changes in the reported sensitivity analysis. This supports ranking stability under those source removals, not immunity to synthetic artifacts. Generator-specific difficulty is also associated with category composition (Appendix A.5).

5 Limitations and Release

Video-HOCA covers visible macroscopic anomalies under an operational taxonomy. It does not test quantitative prediction, unobserved physical states, or long-horizon causal chains. The original Task I split couples source and label; the matched-generator control covers two generators and does not eliminate rendering-quality or salience differences. Tasks II–IV avoid a binary real/generated label contrast but remain subject to candidate-selection and generator effects. Category balancing is not difficulty matching. Human verification corrects proposed annotations but cannot recover all unobserved candidate failures. L2 boundaries remain ambiguous, and the LLM judge remains an imperfect proxy despite agreement checks. Release contents and restrictions are described in Appendix A.16.

6 Conclusion

Video-HOCA separates plausibility, category attribution, fine-grained recognition, and mechanism explanation for visible physical anomalies. Across 20 models, no configuration exceeds 50.2/65.2 macro-F1 on nine-way attribution, recognition and explanation stay below the human reference, and plausibility accuracy is lower on the generator-matched control than on the original split. Frame-order and decoding controls show that these profiles depend on temporal input and model configuration.

AI USE STATEMENT

Video generators and language models assist candidate construction and annotation, and LLM judges score Task IV explanations; these uses are part of the benchmark methodology described in the paper. AI assistants were also used to aid and polish the writing of this manuscript. Model-generated annotations and text were reviewed by the authors; the authors are responsible for the final content, numerical results, and claims.

ETHICS STATEMENT

Video-HOCA relies on human verification of taxonomy labels and reference answers. Annotators review generated and real-world clips with the interface in Appendix A.7, and taxonomy agreement is reported in §3.1. Real-world clips are derived from Panda-70M, and generated clips remain subject to each upstream generator's terms of use. The release is distributed under CC BY 4.0, records the source and license of every clip, and will honor takedown requests (Appendix A.16). The benchmark is a diagnostic evaluation of physical-anomaly reasoning, not a test of deployment safety.

REPRODUCIBILITY STATEMENT

An anonymized artifact accompanies the submission. The benchmark data are available at `https://huggingface.co/datasets/humannobody/video-hoca`, the 250-pair matched-generator control at `https://huggingface.co/datasets/humannobody/task1-250-pairs`, and the evaluation, judging, and table-generation code at `https://anonymous.4open.science/r/video-hoca-86D6`. Appendix A.16 specifies the contents: the dataset or reconstruction scripts for clips that cannot be redistributed, taxonomy labels, four-task question-answer pairs, the unified vLLM evaluation harness and prompt templates, the Task IV judge prompt and rubric (Appendix A.14), and the metric and aggregation code. Decoding settings, frame sampling, and the additional probes are listed in Appendix A.17. Open-weight models were evaluated locally; closed-source models were called through the public APIs available at evaluation time.

REFERENCES

Shuai Bai et al. Qwen3-vl technical report, 2025a.

Zechen Bai, Hai Ci, and Mike Zheng Shou. Impossible videos. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2458–2483. PMLR, 2025b.

Renée Baillargeon. Infants' physical world. *Current directions in psychological science*, 13(3):89–94, 2004.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning

for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024c.

Wey Yeh Choong, Yangyang Guo, and Mohan Kankanhalli. VidHal: Benchmarking temporal hallucinations in vision llms. *Transactions on Machine Learning Research*, 2026.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, pp. 97959–98108, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

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

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

GLM-4.5 Team. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025.

GLM-V Team. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025.

Google DeepMind. Gemini 3 flash model card, December 2025. URL `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf`. Model card.

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

Haojian Huang, Harold Haodong Chen, Meng Luo, Junjia Du, Shanqing Xu, Ziheng Chen, Yanxiang Huang, Yinchuan Li, and Ying-Cong Chen. No place to hide: Benchmarking video hallucination with background-controlled pairs, 2026.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card, 2024.

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

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. In *Advances in Neural Information Processing Systems*, volume 38, pp. 84651–84683, 2025.

Chang Liu, Yunfan Ye, Fan Zhang, Qingyang Zhou, Yuchuan Luo, and Zhiping Cai. Humansam: Classifying human-centric forgery videos in human spatial, appearance, and motion anomaly. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14028–14038, 2025.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023. Datasets and Benchmarks Track.

Qwen Team. Qwen2.5-VL, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990.

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

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, 2025a.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-R1: Post-training large vision language model for temporal video grounding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 92893–92927, 2025b.

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

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

A Appendix

A.1 Source-Controlled Plausibility Evaluation

We selected 272 anomalous candidates from the main set and generated normal counterparts using the corresponding Wan 2.1 or HunyuanVideo model. Iterative prompt correction and human inspection retained 250 normal/anomalous pairs, with resolution, frame rate, and duration aligned within each pair. The accepted pairs comprise 167 Wan and 83 HunyuanVideo pairs. Each of the 17 open-weight models receives the same Task I instruction as in the original evaluation. Both sides are generated; thus a real-versus-generated decision rule cannot directly solve the labels. Other rendering differences can still correlate with plausibility.

Table 3: Available Task I source-control summaries. Ranges are across the 17 open-weight models, not confidence intervals. The original-split mean is a separate reference; sample composition is not held fixed across splits.

Evaluation condition	Mean ACC (%)	Model ACC range (%)
Original mixed-source split (same 17 models)	74.0	–
Matched-generator control (250 pairs)	64.4	59.2–69.6
Wan control (167 pairs)	–	60.5–69.8
HunyuanVideo control (83 pairs)	–	51.8–75.3
Dynamic-anomaly control	–	56.6–68.4

Table 4 reports the per-model controlled accuracies. Pair-bootstrap intervals use 10,000 resamples of the 250 pairs (seed 20260728). Accuracy intervals exclude 50% for every model. Anomaly-class F1 is reported separately and is not tested against a universal 50% chance threshold. The aggregates support above-chance plausibility judgment under this protocol, not a causal decomposition of source and physics cues. An item-matched comparison to the original real controls would be needed to isolate changes in the normal counterpart while fixing the anomaly set.

Table 4: Per-model Task I accuracy on the 250 same-generator pairs. Ctrl. ACC and F1 are on the matched pairs; the 95% CI is the pair bootstrap for accuracy. Wan and Hunyuan are accuracies on the 167 and 83 pairs from each generator. Dynamic is accuracy on the 144 dynamic-anomaly pairs.

Model	**Ctrl. ACC**	**95% CI**	**F1**	**Wan**	**Hunyuan**	**Dynamic**
InternVL-3.5-2B	61.0	[57.6, 64.4]	49.6	65.0	53.0	56.9
Qwen3-VL-2B	61.0	[57.4, 64.6]	58.1	65.6	51.8	56.6
InternVL-2.5-8B	61.6	[58.6, 64.6]	48.9	62.9	59.0	58.7
Qwen2.5-VL-7B	63.0	[59.6, 66.4]	56.3	63.5	62.0	61.1
InternVL-3.5-8B	61.4	[58.4, 64.6]	50.9	63.5	57.2	59.4
Qwen3-VL-8B	68.0	[64.4, 71.6]	66.2	69.8	64.5	68.1
Qwen3.5-9B	68.4	[65.0, 71.8]	64.9	68.9	67.5	67.7
GLM-4.6V-9B-Flash	60.6	[57.8, 63.4]	41.5	62.9	56.0	56.9
InternVL-3.5-14B	64.6	[61.2, 68.0]	59.5	65.0	63.9	62.2
Qwen3.5-27B	66.6	[63.2, 70.0]	69.7	66.2	67.5	65.3
Qwen3-VL-32B	69.6	[66.0, 73.2]	66.8	69.5	69.9	68.4
InternVL-3.5-38B	68.2	[64.8, 71.6]	67.1	64.7	75.3	66.0
InternVL-3.5-30B-A3B	63.4	[60.0, 66.8]	55.5	64.1	62.0	62.5
Qwen3-VL-30B-A3B	67.2	[63.6, 70.8]	65.5	68.6	64.5	64.2
Qwen3.5-35B-A3B	67.2	[63.4, 70.8]	66.9	68.0	65.7	65.3
GLM-4.6V-106B-FP8	64.2	[61.2, 67.4]	50.7	65.9	60.8	62.2
Qwen3.5-122B-A10B-FP8	59.2	[56.2, 62.2]	68.2	60.5	56.6	59.0

A.2 Construction Workflow and Prompt Records

Source-first supplementation. The retained single-anomaly set contains 375 videos from Impossible, 57 from VideoPhy-2, 41 from VBench, and 157 from targeted Wan generation. These acquisition-source counts differ from generator-family counts because externally collected videos can originate from the same generating model. Targeted generation is used for category deficits after external collection, with an acceptance target of approximately 35 videos per L3 category. Accepted counts of 31–37 do not imply an equal number of generation attempts or equal category difficulty.

Human feedback and stopping conditions. Category definitions condition candidate prompt generation. Rendered videos are inspected for the requested anomaly, realism, clarity, and unrelated failures. Observed failures inform prompt revisions and another generation attempt. Requests such as "4K" or a photorealistic/cinematic appearance are visual-style instructions introduced to improve candidate quality; they do not guarantee the resolution or realism of the output. Prompt refinement stops once usable candidates can be produced. Video screening continues toward the category coverage target. Final annotations are based on the visible video rather than copied automatically from the requested category or any prompt-side explanation.

What the retained text records establish. The available prompt records include short candidates, expanded descriptions, and category-specific supplements. They document candidate construction, not a complete approval log or a one-to-one list of accepted videos. The examples below illustrate generation targets; their presence in the archive does not establish that the corresponding video was retained. Early candidates or revised text can differ from the final prompts used for accepted videos.

Table 5: Examples from retained candidate-prompt records. These are exact prompt excerpts/examples, not verified prompt-to-benchmark-video matches.

Target	Candidate generation prompt
Shape	A human hand with seven fingers, each having three extra joints.
Rigidity	A 4K cinematic shot of a modern kitchen where a heavy granite island begins to sag and bend in the middle like a sheet of soft rubber, caused by the weight of a single ceramic plate.
Visual material	A high-definition 4K shot of a white swan gliding on a pond; its feathers are composed entirely of sharp, translucent glass needles that scintillate and refract sunlight with every movement.

Selection limitations. Feedback can improve generation yield for selected categories while also favoring failures that are easy to render or recognize. Uniform target counts do not remove this selection effect. Shared stylistic wording may also affect generated appearance. Multi-source collection and video-level verification limit reliance on any one prompt stream but do not prove independence from prompt style or equal difficulty across categories.

A.3 Taxonomy Agreement and Boundary Analysis

Three annotators assigned L2 labels to a stratified subset of 120 videos. The pooled nine-way Fleiss' κ is 0.708. Mapping the L2 labels to the Ontological/Causal branches gives an L1 coefficient of 0.821; this is not a separate direct-labeling experiment. Agreement is summarized for each level rather than used as evidence that the taxonomy is uniquely determined.

Material behavior concerns how a substance deforms or flows, whereas mass/volume consistency concerns amount changes during an interaction. Chemistry and thermodynamics can co-occur in a visible event, so the intended dominant violation may remain uncertain without additional evidence. Aggregate agreement does not by itself quantify these individual boundaries.

A.4 Judge, Rubric, and Response-Length Sensitivity

Table 2 and the rubric decomposition below use the same Task IV scores.

Table 6 reports per-model rubric scores under both judges. Component maxima differ (25/25/15/35), so raw component values must not be compared as if they were on a common scale. Cross-task scores also do not measure the probability that the same response identifies an anomaly yet fails to explain it; that would require a paired item-level analysis.

Across 12,600 paired judgments, response-level Pearson/Spearman correlations are 0.877/0.884, and model-ranking Spearman correlation is 0.970. Per-model Pearson correlations range from 0.840 to 0.904. Within-family ranking agreement is 1.000 for InternVL and 0.983 for Qwen. Judges agree on 94.2% of model pairs and on all pairs separated by more than five score points. These statistics characterize agreement, not correctness against physical truth on every sample.

Comparison with human scores. On the available human-scored sample, judge-minus-human family differences are Qwen $+0.96$, GLM $+0.54$, and InternVL -1.84 points; bootstrap confidence intervals include zero, and a permutation test gives $p = 0.73$. GPT-4o has a negative quality-controlled residual in this comparison. The sample does not establish the absence of model-family bias.

Length sensitivity. Over 12,600 responses, length–score Pearson correlation is -0.053 and -0.058 after controlling for model and video effects. Controlling for mean answer length leaves model ranking unchanged in the reported analysis. Conditioning on human score gives partial correlation -0.013. On paired answers for the same video, the judge follows the human preference in 89.6% of cases when the preferred answer is shorter and 87.3% when it is longer. These checks do not show a positive verbosity advantage on the observed responses; they do not rule out all stylistic effects.

Rubric wording. A neutral rubric removes only the two instructions about conciseness and not penalizing brevity, keeping dimensions, point values, and output format fixed. All 630 responses from each of the 20 models were re-scored. Original-versus-neutral Task IV Pearson/Spearman correlations are 0.983/0.976 under GPT-OSS and 0.993/0.994 under DeepSeek. The reported Overall correlation is 0.998/0.996 with at most one rank shift; the top six remain fixed. Mean GPT-OSS Task IV scores decrease by about two points. Absolute calibration is therefore not identical despite similar rankings.

Table 6: Full Task IV rubric decomposition under both judges. Scene, Anomaly, Process and Reasoning are worth 25, 25, 15 and 35 points and sum to the 0–100 Task IV score. r and MAE are computed over the 630 per-video paired judgements of that model. Process is the Ontological-vs-Causal label match, which is why it is all-or-nothing.

Model	**GPT-OSS-120B judge**					**DeepSeek-V4-Flash judge**					**Agreement**	
	Scene	Anom.	Proc.	Reas.	**Total**	Scene	Anom.	Proc.	Reas.	**Total**	r	MAE
Open-weight dense models (<7B)												
InternVL-3.5-2B	18.31	7.07	9.74	9.81	**44.92**	16.27	7.64	9.33	7.90	**41.14**	0.860	10.21
Qwen3-VL-2B	19.30	9.40	9.76	12.25	**50.71**	17.40	9.06	9.58	8.56	**44.60**	0.879	10.75
Open-weight dense models (≥7B)												
InternVL-2.5-8B	17.98	9.87	9.90	13.68	**51.43**	17.13	9.93	9.64	12.08	**48.79**	0.904	9.62
Qwen2.5-VL-7B	19.05	9.98	10.14	14.24	**53.42**	17.14	10.18	10.01	11.40	**48.73**	0.882	10.29
InternVL-3.5-8B	18.43	8.93	8.83	13.87	**50.06**	17.32	9.70	8.48	11.81	**47.31**	0.889	9.65
Qwen3-VL-8B	20.13	13.73	10.21	19.97	**64.05**	19.02	13.69	10.19	17.78	**60.67**	0.840	11.03
Qwen3.5-9B	20.34	14.35	8.93	21.55	**65.17**	19.37	14.39	8.90	18.87	**61.53**	0.862	10.59
GLM-4.6V-9B-Flash	20.30	12.52	9.48	19.06	**61.36**	19.21	13.22	9.45	16.87	**58.76**	0.869	10.44
InternVL-3.5-14B	19.34	11.79	10.50	17.28	**58.91**	18.12	12.23	10.32	15.55	**56.22**	0.872	10.32
Qwen3.5-27B	20.55	15.52	10.74	23.25	**70.06**	20.03	15.52	10.88	21.18	**67.61**	0.852	10.21
Qwen3-VL-32B	21.00	15.72	9.62	22.85	**69.18**	20.08	15.48	9.61	20.82	**65.99**	0.848	9.90
InternVL-3.5-38B	20.14	12.62	9.10	19.26	**61.12**	18.80	13.25	8.98	17.40	**58.43**	0.882	9.50
Open-weight MoE models												
InternVL-3.5-30B-A3B	18.90	10.33	9.38	15.82	**54.43**	18.36	10.92	8.94	14.16	**52.39**	0.882	9.15
Qwen3-VL-30B-A3B	20.62	14.99	10.83	22.10	**68.55**	19.46	14.27	10.70	18.80	**63.22**	0.864	9.60
Qwen3.5-35B-A3B	20.34	14.58	10.02	22.08	**67.02**	19.85	14.84	10.03	20.18	**64.90**	0.873	9.37
GLM-4.6V-106B-FP8	20.79	13.53	9.95	19.54	**63.81**	20.00	14.23	9.92	18.49	**62.64**	0.854	10.19
Qwen3.5-122B-A10B-FP8	20.27	14.27	10.02	21.57	**66.13**	19.47	14.16	10.05	19.21	**62.89**	0.875	9.73
Closed-source systems												
GPT-4o	19.62	14.54	10.60	20.80	**65.56**	19.50	15.60	10.56	21.10	**66.75**	0.901	8.37
Gemini-2.5-Flash	20.74	14.86	10.95	22.26	**68.80**	20.53	16.02	10.86	21.94	**69.35**	0.869	9.11
Gemini-3-Flash	21.16	16.21	10.48	23.40	**71.25**	20.94	16.52	10.55	22.70	**70.71**	0.855	9.16
Pooled over all 12,600 model×video paired judgements: $r = 0.877$, $\rho = 0.884$, MAE $= 9.86$ pts												

A.5 Generator-Composition Sensitivity

The source audit identifies 19 generating-source labels in the 630-video single-anomaly set. Among eight sources contributing at least 20 videos, mean Task IV scores range from 56.0 (HunyuanVideo) to 67.4 (Pyramid Flow). After controlling for model and L3 category, the mean absolute source deviation decreases from 3.26 to 1.68 points. This indicates that source and anomaly composition are associated; raw source averages should not be read as direct generator-quality comparisons.

The leave-one-generator-out analysis removes a major source and its semantically matched real normal controls, preserving the original Task I class balance. Recomputed four-task rankings have Spearman correlation 0.9910–1.0000 with the full ranking, maximum displacement two positions, and unchanged top two models across the eight removals. The top three remain the same in seven cases; removing Wan exchanges positions three and four between closely scored models. Stability to these removals does not resolve the real/generated confound, which is addressed separately by the matched-generator Task I control.

Table 7: Sensitivity of Video-HOCA to the choice of video generator, over the 630 single-anomaly videos and all 20 evaluated models. "Resid." is the generator's mean Task-IV deviation after removing model and Level-3 anomaly-class effects, which is necessary because generators were used for class-specific gap filling. "Within" is the Spearman correlation between the model ranking computed on that generator alone and on the full set. "LOO" drops the generator (and its matched real videos) and recomputes the full four-task z-score leaderboard.

Generator	**Videos**	**Difficulty (model mean)**				**LOO leaderboard**		**Within**
		T-I	T-III	T-IV	Resid.	ρ	max shift	ρ
Wan 2.1	177	59.8	67.6	59.33	-1.39	0.9910	2	0.923
HunyuanVideo	95	47.4	67.2	55.97	-1.89	0.9985	1	0.956
Hailuo	79	65.4	74.7	66.35	+1.49	0.9925	2	0.976
Kling	58	64.0	75.5	62.29	+1.57	0.9970	1	0.857
Sora	53	61.1	69.6	65.05	+3.51	0.9940	2	0.899
Mochi	44	56.9	72.6	63.59	+0.79	1.0000	0	0.911
Luma	28	67.7	63.0	60.72	-1.60	1.0000	0	0.952
Pyramid Flow	26	66.9	75.4	67.45	+1.22	0.9955	1	0.856

Task IV averages span 56.0–67.4 (SD 3.43). Mean absolute source deviation falls from 3.26 to 1.68 points after controlling for anomaly class. Task I class accuracies average 58.7% on generated anomalous clips and 90.4% on matched real controls; 19/20 models score lower on anomalous clips. This class asymmetry alone does not identify which visual cues drive answers.

A.6 Detailed Instruct/Thinking Comparison

The table compares Instruct, Instruct-Align, and native Thinking for each model. Instruct-Align retains the Instruct checkpoint and matches the corresponding Thinking sampling parameters and output budget. All diagnostic Overall values use the fixed normalization from the finalized 16-frame main table. These comparisons do not independently isolate training, architecture, or reasoning-trace effects.

Table 8: Comparison of Instruct, Instruct-Align, and Thinking decoding on Video-HOCA. Instruct-Align keeps the Instruct checkpoint and matches Thinking's sampling budget. Task I reports ACC/F1 (%); Task II and Task III report Ontological/Causal macro-F1 (%); Task IV reports rubric-based Ontological/Causal judge scores (0–100). Diag. Overall uses the same fixed main-table z-score normalization as Table 2, computed from Task I F1, count-weighted Task II/III macro-F1, and overall Task IV judge score.

Model	**Setting**	**Task I**		**Task II**		**Task III**		**Task IV**		**Diag. Overall**
		ACC	**F1**	**Onto.**	**Caus.**	**Onto.**	**Caus.**	**Onto.**	**Caus.**	
GLM-4.6V-106B-FP8	Instruct	69.8	76.2	31.5	21.9	24.9	18.0	71.0	58.9	38.7
GLM-4.6V-106B-FP8	Instruct-Align	68.4	75.1	36.2	36.9	32.5	26.8	73.6	60.7	42.4
GLM-4.6V-106B-FP8	Thinking	68.5	75.7	45.7	59.8	78.9	72.6	78.3	62.0	54.5
Qwen3.5-122B-A10B-FP8	Instruct	88.0	87.5	45.0	41.0	85.9	79.2	73.1	61.4	58.2
Qwen3.5-122B-A10B-FP8	Instruct-Align	86.7	87.7	44.6	35.4	88.6	80.3	68.0	61.5	57.1
Qwen3.5-122B-A10B-FP8	Thinking	83.7	86.0	52.4	50.9	86.6	82.6	78.6	62.4	60.7
Qwen3-VL-32B	Instruct	79.5	82.3	46.0	55.6	84.1	79.0	78.5	62.9	58.4
Qwen3-VL-32B	Instruct-Align	78.7	81.5	44.8	52.6	84.4	76.7	80.5	61.9	57.4
Qwen3-VL-32B	Thinking	74.8	79.4	50.9	58.8	86.9	79.3	76.6	61.9	57.6

A.7 Annotation Interface and Dataset Visualizations

This subsection collects the supporting visualizations referenced in the main paper: the human review tool used by annotators (Figure 5), the taxonomy distribution (Figure 6), and an example of the temporal-grounding annotation (Figure 7).

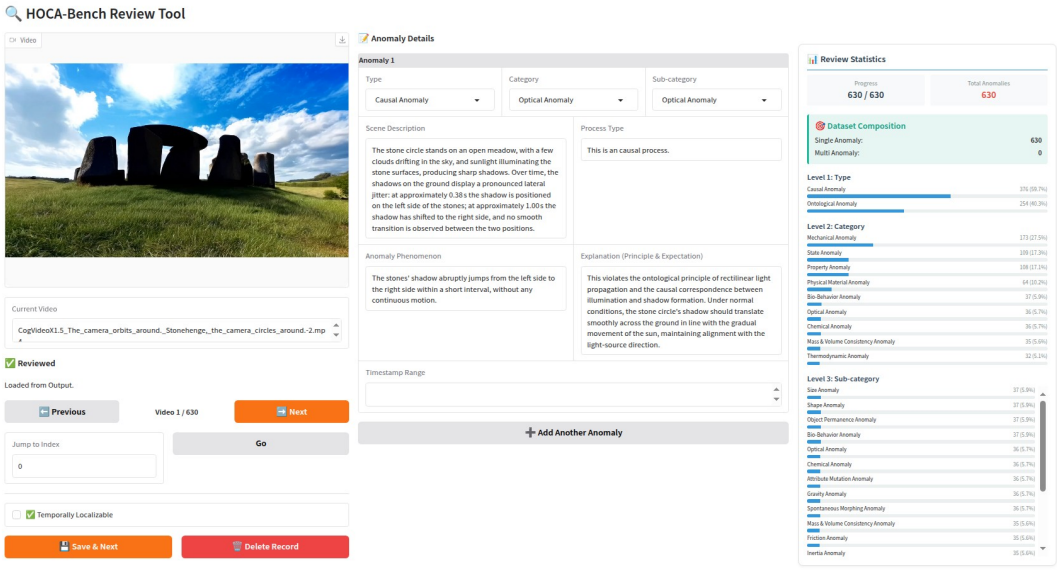


Figure 5: **Video-HOCA annotation interface.** The browser-based review tool used by annotators. It pairs a video player with editable fields for taxonomy type, category, sub-category, scene description, anomaly phenomenon, principle/expectation, and an optional temporal-localization range. The right panel tracks per-category dataset progress.



Figure 6: **Taxonomy distribution.** The sunburst chart visualizes the Ontological–Causal branches and their constituent categories.

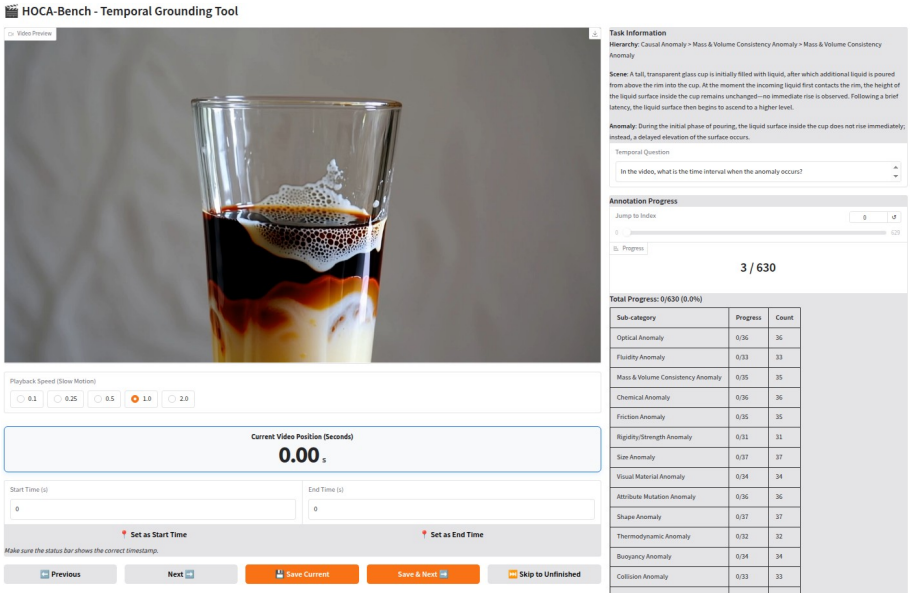


Figure 7: **Temporal annotation example.** Annotated time intervals for the temporal-grounding probe. The probe localizes when anomalous events occur and is auxiliary to the main Ontological–Causal claims.

A.8 Detailed Taxonomy Guidelines

Our annotation pipeline follows the same operational principle used in the main paper: the final label is determined by the *primary violated unit*, not by the most visually salient event alone. The guideline organizes labels by the Ontological/Causal distinction and nine L2 subcategories. In the agreement experiment, annotators selected L2 directly and L1 was derived by the fixed mapping, as described in §3.1. Samples without a stable primary violated unit are marked for adjudication and are not used as canonical examples.

For the L1 Ontological-vs.-Causal decision, the practical rule is:

- assign **Ontological** if the dominant violation concerns an entity's own identity, intrinsic property, state stability, or biological capability;
- assign **Causal** if the dominant violation concerns an interaction law, contact relation, force law, or environment-mediated transition between otherwise valid entities.

Adjudication rule. Canonical examples are selected only when the primary violated unit is stable under the written guideline. If an item can be reasonably explained by multiple rule layers, annotators do not force it into a canonical category. They first identify the dominant violation; if no dominant reading is stable, the item is flagged for adjudication or excluded from the scored canonical set. This policy is meant to keep the benchmark focused on physical anomaly reasoning rather than on ambiguous taxonomy edge cases.

Taxonomy attribution prompt. The following prompt is used after a video has been summarized into a short physical-realism analysis. It is given both to the LLM annotation pipeline and to human annotators as the written classification guideline. L2/L3 names are unified with Table 9; the decision procedure is entity-centered vs. relation-centered, matching the operational rule in §3.1.

```
You are an expert video anomaly analyst. Your task is to analyze a video
based on a provided "Analysis Summary" and classify the primary physical
anomaly according to the Video-HOCA taxonomy.

### 1. Taxonomy Structure
```

Table 9: Full Video-HOCA taxonomy. L1 is the Ontological–Causal split; L2 is the 9 anomaly groups; L3 lists the fine-grained tags actually used as labels. For the five L2 groups with no further subdivision (Bio-Behavior, Mass and Volume Consistency, Thermodynamics, Chemistry, Optics), the L3 tag is identical to the L2 name. Canonical examples are selected to have a stable primary violated unit.

L2 group	L3 tag	Primary violation	Canonical example
Ontological anomalies (entity-centered violations)			
Property	Shape	Geometrically impossible or contradictory shape.	Three-headed goat.
	Size	Object scale violates physical bounds.	Tiny elephant on a dinner plate.
	Visual Material	Surface appearance inconsistent with entity identity.	Turtle with a ruby-like shell; transparent wood.
State	Object Permanence	Spontaneous appearance or disappearance of the whole entity.	Chair appears out of thin air.
	Morphing	Shape or size changes without external cause.	Paper grows longer by itself.
	Attribute Mutation	A visual attribute is temporally inconsistent.	Butterfly wings change from yellow to white.
Bio-Behavior	Bio-Behavior	A biological agent violates its own anatomy, motor ability, or species-level capability.	Human joints bend backward; eyeballs detach from sockets and swing.
Causal anomalies (relation-centered violations)			
Mechanics	Gravity	Universal gravitation fails (e.g., upward fall, no pull).	Mug floats off the table; ball rolls uphill.
	Buoyancy	Archimedes' principle is violated.	Heavy stone floats on water.
	Friction	Surface resistance is incorrect.	Object slides indefinitely on grass.
	Collision	Impenetrability or momentum transfer fails.	Solid objects pass through each other.
	Inertia	Newton's first law is violated.	Object stops or reverses without external force.
Material and Rigidity	Rigidity/Strength	A solid loses rigidity or deforms abnormally.	Glass bends like rubber.
	Fluidity	Flow, viscosity, or surface tension behaves incorrectly.	Square water droplets; liquid behaves like a solid.
Mass and Volume Consistency	Mass and Volume Consistency	Matter or volume changes during interaction without a plausible cause.	Water volume increases while pouring; cutting an object increases total volume.
Thermodynamics	Thermodynamics	Temperature-dependent state changes fail or occur under the wrong condition.	Ice remains solid in boiling water; liquid boils without a heat source.
Chemistry	Chemistry	Combustion or chemical transformation fails under ordinary reaction conditions.	Paper does not burn in flame; wood turns to ash without heat.
Optics	Optics	Light transport, shadow, or reflection violates geometric constraints.	Shadow points toward the light source; mirror shows an impossible reflection.

```
- Ontological anomaly (entity-centered): the dominant violation lies in
  an entity's own identity, intrinsic property, state, or biological
  capability, regardless of any interaction the entity participates in.
    * Property:     Shape, Size, Visual Material.
    * State:        Object Permanence, Morphing,
                    Attribute Mutation.
    * Bio-Behavior: anatomy / motor / species-level capability
                    violations of a biological agent.
- Causal anomaly (relation-centered): the dominant violation lies in a
  physical relation between otherwise valid entities, or between an
  entity and the environment.
    * Mechanics:                      Gravity, Buoyancy, Friction,
                                      Collision, Inertia.
    * Material and Rigidity:          Rigidity, Fluidity.
    * Mass and Volume Consistency:    matter / volume changes during
                                      interaction without a plausible cause.
    * Thermodynamics:                 temperature-dependent state changes
                                      that fail or occur under the wrong
                                      condition.
    * Chemistry:                      combustion or chemical transformation
                                      that fails under ordinary reaction
                                      conditions.
    * Optics:                         light transport, shadow, or
                                      reflection violates geometric
                                      constraints.

### 2. Decision Procedure
Step 1. Decide L1 (Ontological vs Causal) by identifying the
        *primary violated unit*:
          - If the violation is entity-centered (intrinsic to one
            entity's identity, property, state, or biological
            capability), assign Ontological.
          - If the violation is relation-centered (between entities,
            or between an entity and the environment), assign Causal.
          - The presence of interaction does NOT automatically imply
            Causal. A clip can show two entities interacting while
            the dominant violation is still intrinsic to one of them
            (e.g., a person whose joints bend backward while walking:
            the contact is normal, the bio-behavior violation is
            intrinsic).
Step 2. Within the chosen L1 branch, assign the L2 group whose written
        definition the violation matches most precisely. If two groups
        apply equally well, mark the item for adjudication rather than
        forcing a label.
Step 3. Within L2, assign the most specific L3 sub-category from the
        list in Section 1.

### 3. Annotation Principles
1. Visual honesty: describe exactly what is shown, regardless of how
   impossible it is.
2. Setup vs observation: scene_description states the stage BEFORE the
   anomaly; anomaly_phenomenon describes the "wrong" event.
3. Dual-part reasoning ("Principle + Expectation" template) for
   normal_phenomenon. Examples:
   - Gravity: violates the Law of Universal Gravitation. Normally, an
     unsupported object undergoes vertical downward displacement.
   - Collision: violates impenetrability / momentum transfer. Normally,
     the subject should stop or rebound on contact rather than clip
     through.
   - Object Permanence: violates object constancy. Normally,
     macroscopic objects persist unless a specific physical process
```

```
     intervenes.
   - Rigidity: violates the property of solid rigidity. Normally, a
     solid resists deformation and keeps structural integrity under
     contact.

### 4. Few-Shot Examples
Example 1
Summary: A mug sits on a table, then suddenly flies up into the air.
Output:
{
  "type": "Causal",
  "l2_group": "Mechanics",
  "l3_tag": "Gravity",
  "scene_description": "A wooden table with a ceramic mug placed in
                        the center.",
  "anomaly_phenomenon": "The mug lifts off the table and floats
                         upwards without any external force.",
  "primary_violated_unit": "the gravitational relation between the
                            mug and the Earth",
  "normal_phenomenon": "The mug should remain stationary on the
                        table; once unsupported it should fall under
                        gravity, not rise."
}

Example 2
Summary: A person is walking, and their shirt suddenly changes from
blue to red.
Output:
{
  "type": "Ontological",
  "l2_group": "State",
  "l3_tag": "Attribute Mutation",
  "scene_description": "A person wearing a blue shirt is walking
                        across a room.",
  "anomaly_phenomenon": "The shirt's color suddenly changes from blue
                         to red without any lighting change.",
  "primary_violated_unit": "the temporal stability of the shirt's
                            color attribute",
  "normal_phenomenon": "An object's color should remain temporally
                        consistent unless acted upon by external light
                        or chemical changes."
}

### 5. Current Task
Analyze the following summary and produce the JSON described above.

Summary:
[INSERT SUMMARY TEXT HERE]

Output (JSON only, no extra text):
```

A.9 TASK-III DISTRACTOR GENERATION PROMPT

Task III pairs the ground-truth anomaly description with three adversarial distractors. We use GPT-OSS-120B (OpenAI, 2025) as the distractor generator, conditioned on the ground-truth (L1, L2, L3) annotation and on the Video-HOCA taxonomy in Table 9. The prompt deliberately samples one distractor from the same L1 branch (a different L3 under the same Ontological–Causal branch), one from the opposite L1 branch, and one "adversarial" distractor that sounds physical but contradicts the actual video event. The released review tool (Figure 5) is used by human reviewers to keep, swap, or edit each generated distractor before the multiple-choice item is finalized.

```
You are an expert physics question generator for a Video-HOCA multi-choice
question.
Read the English ground-truth scene and anomaly below, then generate
3 plausible "distractor" descriptions. The correct option is the Observed
Anomaly; keep it as a factual visual description of the event
(e.g., "The cup floats upwards spontaneously.").

### Taxonomy Definitions
{VIDEO_HOCA_TAXONOMY_DEFINITIONS}        // see Table tab:taxonomy-def for
                                         // the full L2/L3 hierarchy

### Input (Ground Truth)
- Scene Setup:      "{scene}"
- Observed Anomaly: "{phenomenon}"
- Classification:   {l1} -> {l2} -> {l3}

### Distractor Generation (3 false options)
- Distractor 1 (Hard, same L1 branch): generate a description of a
  different anomaly type that belongs to the SAME L1 branch
  (Ontological or Causal).
   - Logic: if GT is Gravity (Causal), generate a Collision description
     for the same objects in the scene.
   - Example: "The cup passes through the table directly."
- Distractor 2 (Medium, different L1 branch): generate a description from
  the OTHER L1 branch.
   - Logic: if GT is Gravity (Causal), generate a Morphing description
     (Ontological).
   - Example: "The cup suddenly turns into a bird."
- Distractor 3 (Adversarial, logical/normalcy): generate a description
  that sounds physical but is either
    (a) a slightly wrong version of the actual event (e.g., "The cup
        falls sideways"); or
    (b) a non-anomalous outcome that contradicts the video (e.g., "The
        cup shatters upon impact" when it actually floated).
   - Goal: a valid English sentence that clearly describes a different
     outcome than what happened in the video.

### Output (JSON)
{
  "scene_description":        "...",   // scene setup
  "correct_option":           "...",   // observed anomaly
  "distractor_1_hard":        "...",   // same L1, different L3
  "distractor_1_type":        "...",   // L3 category for distractor 1
  "distractor_2_medium":      "...",   // different L1
  "distractor_2_type":        "...",   // L3 category for distractor 2
  "distractor_3_adversarial": "..."    // adversarial description
}
```

A.10 Multi-Anomaly Inference Prompt

The probe uses the following instruction and 18 definitions, followed by a video-specific option list. M001 illustrates the option format. Letters refer to the options for that video rather than fixed taxonomy IDs.

```
Analyze the video carefully. Based on the physical anomaly definitions
   provided below, select ALL applicable anomalies observed in the
   video from the options.

### Physical Anomaly Definitions:
1. Shape Anomaly: Geometrically impossible or contradictory shapes
   violating basic geometric common sense.
```

```
2. Size Anomaly: Object scale violates physical or commonsense
   constraints relative to its environment.
3. Visual Material Anomaly: Incorrect surface appearance including
   texture, color, or material type contradictory to the object's
   nature.
4. Object Permanence Anomaly: Spontaneous appearance or disappearance
   of a whole object without a physical process.
5. Spontaneous Morphing Anomaly: Shape or size changes without any
   external cause or interaction.
6. Attribute Mutation Anomaly: Unjustified temporal inconsistency of
   visual attributes like material or color.
7. Bio-Behavior Anomaly: Organisms performing biologically impossible
   actions or anatomical violations for their species.
8. Gravity Anomaly: Violation of gravitational laws, such as objects
   falling upwards or floating without support.
9. Buoyancy Anomaly: Violation of buoyancy laws where density logic is
   reversed.
10. Friction Anomaly: Incorrect surface resistance where objects slide
   indefinitely or stop without friction.
11. Collision Anomaly: Interaction logic errors including clipping
   (passing through solids) or missing collisions.
12. Inertia Anomaly: Violation of inertia where objects stop or change
   trajectory suddenly without force.
13. Rigidity/Strength Anomaly: Solid materials behaving like liquids
   or exhibiting abnormal deformation during interaction.
14. Fluidity Anomaly: Violation of flow, viscosity, or surface tension
   logic.
15. Mass & Volume Consistency Anomaly: Inconsistency of mass or volume
   during interaction (matter is not conserved).
16. Thermodynamic Anomaly: Incorrect state change or thermal behavior
   under temperature changes.
17. Chemical Anomaly: Unrealistic fire behavior, combustion, or
   chemical transformations.
18. Optical Anomaly: Errors in light transport, shadows, or
   reflections that contradict the light source.

Options:
A. Friction Anomaly
B. Fluidity Anomaly
C. Optical Anomaly
D. Mass & Volume Consistency Anomaly
E. Inertia Anomaly
F. Collision Anomaly
Select all that apply (e.g., A, C).
```

A.11 Additional Diagnostic Probes

Table 10 reports temporal grounding and multi-anomaly detection; Table 11 reports frame-count sensitivity. Multi-anomaly numbers in (b) use the original 64-frame stress-test setting with Thinking mode where supported, so they are reported as a stress test rather than part of the Instruct-only main leaderboard.

Temporal grounding. The two models in Table 10(a) show the gap between a general-purpose Video-LLM and a temporal-localization specialist. Time-R1-7B (Wang et al., 2025b), which is RL-trained for temporal localization, reaches mIoU 30.3 and R@0.5 15.5, with more than 3× the mIoU of vanilla Qwen2.5-VL-7B (9.0). These two systems differ in training and specialization, so the comparison does not isolate a causal effect of RL. Absolute performance leaves room for improvement in this setting.

Multi-anomaly detection. Table 10(b) reports F1 scores in the 43–57 range and exact-match accuracy of 1.7–15.0. These results indicate difficulty identifying the complete set of concurrent violations. Because the answer space differs from the single-anomaly tasks, the probe is reported separately from the main leaderboard.

Multi-frame scaling. Table 11 uses the same deterministic Instruct-mode decoding as the main table (temperature $= 0$, no Thinking mode); the only change is the frame budget, going from 16 to 64 uniformly sampled frames. The effects are task- and model-specific. Gemini-2.5-Flash declines on Task I (ACC $75.5 \rightarrow 70.1$) while Task II/III change little. GLM-4.6V-106B-FP8 declines on Task II Causal ($21.9 \rightarrow 2.3$) and Task III Ontological ($24.9 \rightarrow 17.3$). Increasing frame count therefore does not uniformly improve the reported task scores. Diagnostic Overall values use the same fixed reference statistics as the finalized 16-frame main table.

Table 10: Additional diagnostic probes. (a) Temporal grounding reports mean IoU and recall at IoU $= 0.5$. (b) Multi-anomaly detection reports F1 and exact-match accuracy under the original 64-frame stress-test setting.

(a) Temporal grounding

Model	mIoU	R@0.5
Time-R1-7B(Wang et al., 2025b)	30.3	15.5
Qwen2.5-VL-7B	9.0	6.0

(b) Multi-anomaly detection

Model	F1	Acc.
InternVL-3.5-38B	43.3	1.7
Qwen3-VL-32B	57.0	12.5
GLM-4.6V-106B-FP8	47.0	15.0

Table 11: Multi-frame evaluation comparison on Video-HOCA. Task I reports ACC/F1 (%); Task II and Task III report Ontological/Causal macro-F1 (%); Task IV reports Ontological/Causal judge scores (0–100). Overall uses the Task-wise means and standard deviations computed from the 16f main-table models only; both 16f and 64f rows report 50 plus 10 times the average task-wise z-score.

Model	Input	Task I		Task II		Task III		Task IV		Overall
		ACC	F1	Ontological	Causal	Ontological	Causal	Ontological	Causal	
Gemini-2.5-Flash	16f	75.5	77.7	42.8	59.9	82.8	73.0	73.6	65.6	55.6
	64f	70.1	70.4	40.4	59.3	82.9	71.7	76.2	63.4	51.4 (-4.2)
GLM-4.6V-106B	16f	69.8	76.2	31.5	21.9	24.9	18.0	71.0	58.9	38.7
	64f	69.2	75.8	28.2	2.3	17.3	16.7	71.9	59.5	35.2 (-3.5)
Qwen3-VL-8B	16f	75.4	78.2	39.8	49.7	83.7	76.6	64.6	63.7	53.1
	64f	75.5	78.1	39.7	49.7	83.4	77.0	65.3	64.5	53.3 (+0.2)
Qwen3-VL-32B	16f	79.5	82.3	46.0	55.6	84.1	79.0	78.5	62.9	58.4
	64f	79.4	82.2	47.1	55.7	84.1	79.2	70.2	63.8	57.9 (-0.6)

A.12 Temporal Frame-Order Ablation

Table 12 compares original 16-frame input, a single middle frame, and shuffled 16-frame input for Qwen3.5-122B-A10B-FP8. Binary plausibility accuracy is nearly unchanged by shuffling (88.0 to 87.9) but drops to 80.6 with one frame. Shuffling reduces Task III by 4.1/3.7 points and Task IV by 4.6/2.4 points on the Ontological/Causal subsets. Using a single frame produces a larger Ontological Task IV loss (15.1 points), while the Causal loss is the same (2.4 points). These results distinguish access to multiple frames from access to their correct order: coverage is important for coarse judgment, and temporal order also affects recognition and open-ended explanation. The comparison concerns one model and these specific input manipulations.

Table 12: Temporal frame-order ablation on Video-HOCA. The experiment compares the original 16-frame input, a single middle-frame input, and a shuffled 16-frame input. Task I reports ACC/F1 (%); Task II and Task III report Ontological–Causal scores (%); Task IV reports rubric-based Ontological–Causal judge scores (0–100). Bold marks the best temporal setting in each column.

Model	Temporal setting	Frames	Task I		Task II		Task III		Task IV	
			ACC	F1	Onto.	Caus.	Onto.	Caus.	Onto.	Caus.
Qwen3.5-122B-A10B-FP8	Original order	16f	**88.0**	**87.5**	**45.0**	**41.0**	**85.9**	**79.2**	**73.1**	**61.4**
Qwen3.5-122B-A10B-FP8	Middle frame	1f	80.6	82.0	39.5	40.2	77.9	70.7	58.0	59.0
Qwen3.5-122B-A10B-FP8	Shuffled order	16f	87.9	87.2	44.4	36.8	81.8	75.5	68.5	59.0

A.13 Full Static/Dynamic Results

Table 14 reports static/dynamic results for the 20 Instruct-mode models on 258 static and 372 dynamic anomalous clips. The table is descriptive and does not by itself isolate temporal demand from anomaly-category composition.

Table 13: Static/dynamic analysis by anomaly branch. Reported model-averaged task scores within each branch and temporal group. Static Causal scores remain below static Ontological scores in this comparison; category composition is not controlled by this stratification.

Branch	Temporal type	Videos	Task III	Task IV
Ontological	Static	130	74.6	66.9
	Dynamic	124	71.6	65.9
Causal	Static	128	72.6	62.8
	Dynamic	248	62.2	55.3
Static gap (Onto. – Caus.)	–	–	+2.0	+4.1

Table 14: Static vs. dynamic subset analysis on Video-HOCA. The subset labels come from `static_dynamic_labels.jsonl`. Only anomalous videos are included: Task I reports anomaly recall (the fraction answered No), not balanced accuracy. An always-No predictor would score 100%. Task II/III report macro-F1 (%), and Task IV reports the rubric-based judge score (0–100).

Model	Subset	Videos	T-I recall	Task II	Task III	Task IV
Labeled videos: Static=258, Dynamic=372						
Open-weight dense models (<7B)						
InternVL-3.5-2B	Static	258	48.1	32.0	66.8	49.9
	Dynamic	372	27.2	31.6	57.6	41.4
Qwen3-VL-2B	Static	258	70.9	41.5	76.8	56.1
	Dynamic	372	49.7	27.1	64.9	47.0
Open-weight dense models (≥7B)						
InternVL-2.5-8B	Static	258	51.6	33.4	79.9	58.1
	Dynamic	372	26.9	28.4	62.4	46.8
Qwen2.5-VL-7B	Static	258	62.4	44.5	74.4	60.4
	Dynamic	372	41.4	33.2	56.5	48.6
InternVL-3.5-8B	Static	258	54.7	40.9	76.5	55.0
	Dynamic	372	30.4	31.6	66.0	46.6
Qwen3-VL-8B	Static	258	72.9	42.0	86.4	66.2
	Dynamic	372	55.6	33.3	74.8	62.6
Qwen3.5-9B	Static	258	70.5	21.7	71.2	68.8
	Dynamic	372	59.4	25.3	69.8	62.7
GLM-4.6V-9B-Flash	Static	258	45.0	40.4	78.6	61.1
	Dynamic	372	20.7	32.1	66.9	61.5
InternVL-3.5-14B	Static	258	64.7	42.7	79.6	61.4
	Dynamic	372	44.4	39.2	68.9	57.2
Qwen3.5-27B	Static	258	85.3	45.7	89.4	72.1

Model	Subset	Videos	T-I recall	Task II	Task III	Task IV
	Dynamic	372	79.0	48.8	86.0	68.7
Qwen3-VL-32B	Static	258	73.6	45.3	85.8	74.8
	Dynamic	372	57.0	36.4	77.9	65.3
InternVL-3.5-38B	Static	258	74.8	38.6	81.8	67.5
	Dynamic	372	61.8	34.4	74.9	56.7
Open-weight MoE models						
InternVL-3.5-30B-A3B	Static	258	59.3	39.0	83.1	60.0
	Dynamic	372	37.1	36.7	70.7	50.6
Qwen3-VL-30B-A3B	Static	258	74.8	41.3	83.3	76.8
	Dynamic	372	61.3	33.6	72.4	62.8
Qwen3.5-35B-A3B	Static	258	77.9	1.1	33.0	70.3
	Dynamic	372	66.7	7.7	24.9	64.8
GLM-4.6V-106B-FP8	Static	258	53.9	15.7	22.8	64.6
	Dynamic	372	34.4	18.1	19.8	63.3
Qwen3.5-122B-A10B-FP8	Static	258	94.6	35.1	85.2	64.4
	Dynamic	372	90.1	30.4	79.7	67.3
Closed-source systems						
GPT-4o	Static	258	74.8	22.5	58.9	67.3
	Dynamic	372	65.9	21.4	57.3	64.4
Gemini-2.5-Flash	Static	258	74.4	39.8	78.3	70.1
	Dynamic	372	59.1	43.9	75.7	67.9
Gemini-3-Flash	Static	258	82.9	40.9	85.9	72.7
	Dynamic	372	73.1	47.7	80.6	70.3

A.14 TASK-IV JUDGE RUBRIC

Task IV is evaluated against a structured reference answer. For each question, the reference specifies (i) scene description, (ii) anomaly identification, (iii) anomaly type (Ontological vs. Causal), and (iv) physical explanation. The judge prompt asks GPT-OSS-120B (OpenAI, 2025) to score the model's response against these four components using a fixed rubric on a 0–100 scale. The reference specifies the expected content for each scored component.

The paired human–judge comparison covers 100 Task-IV responses and gives Pearson 0.896 and Spearman 0.891. Separately, the Human row in Table 2 scores 81.4 under GPT-OSS-120B and 84.9 under DeepSeek-V4-Flash. These results describe agreement and the scoring of human answers, respectively; alternative-judge analyses appear in Appendix A.4.

Field-name mapping. The verbatim prompt below uses `scene_score`, `anomaly_score`, `process_score`, and `reasoning_score` as the four rubric fields, which correspond respectively to the main-text components *scene description*, *anomaly identification*, *anomaly type*, and *physical explanation*.

Verbatim judge prompt. The exact prompt template sent to the judge model (with reference answer `gt_text` and model output `prediction_text`) is:

```
You are an expert physics professor and visual reasoning judge. Your task
is to evaluate a model's physical reasoning performance based on a video's
ground-truth (GT) and the model's prediction.

### Evaluation Rubric (Total 100 pts)

**Significant guideline: physical and semantic equivalence.**
The evaluation should prioritize accurate identification of the
underlying physical violation over exact linguistic matching.
- Granularity and summarization: GT references are often highly
  detailed. Predictions that provide a concise summary capturing the
  core physical event should be rewarded. If the prediction identifies
  the primary entity and the central action, it should receive full
```

```
  credit for scene and anomaly identification. Ignore missing
  environmental details (background, lighting, minor textures) and do
  not penalize brevity. Think like an intelligent evaluator, not a
  keyword matcher.
- Object-class flexibility: using functionally related terms (e.g.,
  "paintbrush" instead of "pencil") is acceptable if the visual
  features are ambiguous or similar.
- Physical essence vs. academic terms: focus on common sense and
  intuition. Full credit requires correctly identifying the violated
  physical law/common sense AND describing the normal expectation
  (e.g., "It violates the law of inertia because the object moves on
  its own; normally, it should stay still."). Academic terminology is
  NOT required for full marks.
   - Note on generic terms: if the model only says "it violates
     physical common sense" without explaining the specific phenomenon
     (e.g., "object should stay still"), deduct partial points
     (-10 pts). The explanation must relate to the specific event.

**Rubric summary**
1. Scene (25 pts):
   - Entity identification (10 pts): award 10 pts if the primary
     subject (e.g., "lamp", "cat", "dog") is correctly identified,
     even if specific sub-parts (e.g., "lamp shade") are omitted.
   - Action/behavior (10 pts): award 10 pts if the core motion or
     event is captured. Be lenient: generic terms like "moving" or
     "changing" are acceptable summaries for more specific actions
     like "rotating" or "morphing".
   - Outcome/context (5 pts): award 5 pts for describing the final
     state or environmental details.
   - Full-credit logic: if the prediction is a concise and accurate
     summary of "who did what", award the full 25 pts. Do not
     penalize for brevity compared to a highly detailed GT.
2. Anomaly (25 pts): physical essence of the violation. If the
   prediction identifies that a physical impossibility occurred to the
   correct object, award at least 15-20 pts even if the description is
   simple (e.g., "moving by itself" vs. "rotating independently").
3. Process (15 pts): exact label match (Causal/Ontological).
4. Reasoning (35 pts):
   - Full credit: correctly explains what scientific common
     sense/physical law was violated AND describes what the normal
     behavior of the event should be. Academic terms or common-sense
     descriptions are BOTH acceptable.
   - Partial deduction (-10): if the explanation is tautological
     (e.g., just says "violates physics" without explaining what
     behavior is wrong and what is normal).
   - Note: do NOT penalize for "imprecise terminology" if the logic
     is sound.

### Few-Shot Examples for Grading

Example 1: Perfect alignment with full reasoning (full marks)
- Ground-truth:
  { "scene": "A cat sits on a windowsill. Its tail extends to an
              extreme length.",
    "anomaly": "The cat's tail extends to an impossible length.",
    "process": "Ontological",
    "reasoning": "Violates biological regularities. Normally, a cat's
                  tail has a fixed skeletal structure and cannot grow
                  indefinitely." }
- Model prediction:
  "A cat sits by the window. Its tail is incredibly long, like a
```

```
   snake. This is Ontological. It violates biological common sense
   because a tail cannot grow to such an extreme length. Normally, a
   cat's tail has a fixed structure and fixed length."
- Judge output:
  { "breakdown": { "scene_score": 25, "anomaly_score": 25,
                   "process_score": 15, "reasoning_score": 35 },
    "justification": "The model correctly captures the scene and tail
                      anomaly. Its reasoning identifies the biological
                      violation (15 pts) and explicitly describes the
                      normal state (20 pts), providing a complete
                      explanation." }

Example 2: Generic violation and missing normal state (low reasoning)
- Ground-truth:
  { "scene": "A baseball bat hits a glass bottle on a wooden floor.
              The bottle remains stationary after impact.",
    "anomaly": "The bottle remains still while shards fly outwards.",
    "process": "Causal",
    "reasoning": "Violates momentum conservation. Normally, the bottle
                  should move due to impact." }
- Model prediction:
  "A bat hits a bottle. The bottle doesn't move. Causal. This violates
   common sense."
- Judge output:
  { "breakdown": { "scene_score": 25, "anomaly_score": 25,
                   "process_score": 15, "reasoning_score": 5 },
    "justification": "Scene and anomaly are correct. For reasoning, it
                      only provides a generic 'violates common sense'
                      without event-specific violation content (5/15)
                      and fails to describe what should have happened
                      normally (0/20). Total reasoning: 5/35." }

Example 3: Precise identification but flawed reasoning (zero reasoning)
- Ground-truth:
  { "scene": "An apple on a table rotates around its axis without
              contact.",
    "anomaly": "Apple spins spontaneously without external force.",
    "process": "Ontological",
    "reasoning": "Violates the conservation of angular momentum.
                  Normally, a stationary object remains still unless
                  acted upon by a torque." }
- Model prediction:
  "The apple starts spinning by itself. Ontological. It violates
   Gravity."
- Judge output:
  { "breakdown": { "scene_score": 25, "anomaly_score": 25,
                   "process_score": 15, "reasoning_score": 0 },
    "justification": "Reasoning score: 0/35. Citing 'Gravity' as the
                      reason for rotation is physically incorrect. It
                      fails BOTH the violation identification and the
                      normal-state components because the premise is
                      flawed." }

Example 4: Label mismatch but perfect reasoning (reasoning full)
- Ground-truth:
  { "scene": "Cat food in a bowl increases while being eaten.",
    "anomaly": "Cat food quantity increases during consumption.",
    "process": "Causal",
    "reasoning": "Violates conservation of mass. Normally, food
                  quantity should decrease when eaten." }
- Model prediction:
```

```
  "A cat is eating. The food multiplies. This is Ontological. It
   violates the law of conservation of matter because the food
   quantity increases during the eating process. Normally, the amount
   of food should decrease as it is consumed."
- Judge output:
  { "breakdown": { "scene_score": 25, "anomaly_score": 25,
                   "process_score": 0, "reasoning_score": 35 },
    "justification": "Process: 0 (mismatch). Reasoning: 35/35. It
                      correctly identifies the violation (15/15) and
                      describes the normal expectation (20/20) in
                      separate, clear components." }

Example 5: Partial scene and generic anomaly (partial marks)
- Ground-truth:
  { "scene": "A table lamp's shade rotates independently while the
              base stays still.",
    "anomaly": "Lampshade rotates without external force.",
    "process": "Causal",
    "reasoning": "Violates inertia. Normally a shade doesn't rotate
                  on its own." }
- Model prediction:
  "A lamp moves around. Ontological. Violates common sense, objects
   should stay still."
- Judge output:
  { "breakdown": { "scene_score": 20, "anomaly_score": 15,
                   "process_score": 0, "reasoning_score": 10 },
    "justification": "Scene: 20/25. Correctly identifies entity
                      'lamp' (10) and action 'moves' (10), but misses
                      specific details. Anomaly: 15/25. Captures the
                      essence of self-motion but is very generic.
                      Process: 0 (mismatch). Reasoning: 10/35.
                      Identifies the violation generally (5/15) but
                      fails to describe the specific normal state of
                      a lamp (5/20), getting only partial credit." }

### Input Data
- Ground-truth reference:
{gt_text}

- Model's prediction:
{prediction_text}

### Output Format
Provide your evaluation in JSON. You may include reasoning before the
JSON block, but the final response must contain a valid JSON object
wrapped in a ```json ... ``` code block:

{
  "breakdown": {
    "scene_score":     0,
    "anomaly_score":   0,
    "process_score":   0,
    "reasoning_score": 0
  },
  "justification": "Detailed explanation for each score point based on
                    the rubric."
}
```

A.15 Aggregation and Ranking Sensitivity

All configurations use the completed inference and scoring records. The same model set supplies the normalization statistics and both judge variants in Table 15.

Overall standardizes Task I F1, count-weighted Ontological/Causal macro-F1 for Tasks II–III, and the Task IV score across the 20 models, averages the four z-scores, and displays $50 + 10\bar{z}$. It depends on the reference model set and the original Task I source composition. We do not interpret it as an absolute physical-reasoning scale. Table 15 reports ranks from the finalized results under the GPT-OSS and DeepSeek judges, using either z-score Overall or a fixed 0.2/0.2/0.2/0.4 weighted average. Rank changes are relative to the GPT-OSS z-score column.

For a task vector $s = (s_1, s_2, s_3, s_4)$, the reported index is $50 + \frac{10}{4}\sum_{t=1}^{4}(s_t - \mu_t)/\sigma_t$. The reference pool is the same 20 Instruct configurations, excluding the Human row and auxiliary decoding/frame settings. The GPT-OSS reference (μ_t, σ_t) values, rounded to three decimals, are $(78.351, 4.909)$, $(42.737, 11.273)$, $(68.837, 16.445)$, and $(61.297, 7.646)$ for Tasks I–IV. Both the 64-frame and decoding comparisons reuse these fixed statistics. The alternative-judge ranking uses its own Task IV reference statistics on the same 20 models.

A new model can be compared on this index only by keeping the same reference statistics and evaluation protocol. Re-estimating them on an expanded model pool defines a new index version and changes old scores. The raw weighted alternative, $0.2s_1 + 0.2s_2 + 0.2s_3 + 0.4s_4$, does not depend on model-pool means or standard deviations, although it still depends on task definitions and weights. We report per-task scores and this alternative alongside z-score rankings rather than treating the index as an absolute ability scale.

Table 15: Ranking sensitivity under alternative Task IV judges and aggregation rules. Each cell reports score (rank); arrows show rank changes relative to the GPT-OSS z-score setting. The z-score variants standardize Task I F1, count-weighted Task II/III macro-F1, and the selected Task IV judge score over the same main-table models. The weighted variants use raw scores with weights 0.2/0.2/0.2/0.4 for Tasks I–IV.

Model	GPT-OSS z-score	DeepSeek z-score	GPT-OSS weighted	DeepSeek weighted
Qwen3.5-27B	63.1 (1)	62.9 (1) ·	74.4 (1) ·	73.4 (1) ·
Gemini-3-Flash	62.1 (2)	62.4 (2) ·	73.5 (2) ·	73.3 (2) ·
Qwen3-VL-32B	58.4 (3)	58.0 (3) ·	70.7 (3) ·	69.4 (3) ·
Qwen3.5-122B-A10B-FP8	58.2 (4)	57.9 (4) ·	68.9 (5) ↓1	67.6 (5) ↓1
Qwen3-VL-30B-A3B	56.1 (5)	55.1 (6) ↓1	68.7 (6) ↓1	66.5 (6) ↓1
Gemini-2.5-Flash	55.6 (6)	56.3 (5) ↑1	69.0 (4) ↑2	69.3 (4) ↑2
Qwen3-VL-8B	53.1 (7)	52.8 (7) ·	66.3 (7) ·	64.9 (7) ·
InternVL-3.5-14B	51.5 (8)	51.6 (8) ·	64.1 (9) ↓1	63.0 (9) ↓1
InternVL-3.5-38B	51.5 (9)	51.5 (9) ·	64.6 (8) ↑1	63.5 (8) ↑1
Qwen3.5-9B	51.3 (10)	50.9 (10) ·	63.1 (10) ·	61.6 (10) ·
InternVL-3.5-30B-A3B	48.0 (11)	48.4 (11) ·	61.4 (12) ↓1	60.6 (12) ↓1
GLM-4.6V-9B-Flash	47.8 (12)	47.8 (12) ·	62.2 (11) ↑1	61.1 (11) ↑1
Qwen2.5-VL-7B	46.3 (13)	46.0 (13) ·	58.5 (13) ·	56.6 (15) ↓2
InternVL-3.5-8B	45.4 (14)	45.7 (15) ↓1	58.2 (15) ↓1	57.1 (14) ·
Qwen3-VL-2B	44.8 (15)	44.1 (16) ↓1	57.6 (16) ↓1	55.2 (17) ↓2
GPT-4o	44.8 (16)	45.8 (14) ↑2	58.4 (14) ↑2	58.8 (13) ↑3
InternVL-2.5-8B	43.3 (17)	43.7 (17) ·	57.1 (17) ·	56.0 (16) ↑1
Qwen3.5-35B-A3B	41.6 (18)	41.5 (18) ·	51.3 (19) ↓1	50.5 (19) ↓1
GLM-4.6V-106B-FP8	38.7 (19)	39.0 (19) ·	50.1 (20) ↓1	49.6 (20) ↓1
InternVL-3.5-2B	38.4 (20)	38.6 (20) ·	52.1 (18) ↑2	50.6 (18) ↑2

A.16 Artifact and Release Details

All open-weight models were evaluated locally on a single node with 8× NVIDIA H20 96 GB GPUs using a unified vLLM-based inference stack. Closed-source models (GPT-4o, Gemini-2.5-Flash, Gemini-3-Flash) were evaluated through their public APIs available at the time of submission.

The data are hosted at `https://huggingface.co/datasets/humannobody/video-hoca` (main benchmark) and `https://huggingface.co/datasets/humannob`

ody/task1-250-pairs (matched-generator Task I control); the code is hosted at https://anonymous.4open.science/r/video-hoca-86D6. The anonymized artifact package released with the paper contains the *benchmark* and the *evaluation code* required to reproduce Table 2: (i) the Video-HOCA dataset, comprising videos (or URLs with reconstruction scripts for clips that cannot be redistributed directly), per-clip taxonomy labels, the four-task QA pairs, and split files for the static/dynamic and temporal-grounding subsets; (ii) the unified vLLM-based evaluation harness with the prompt templates and decoding configurations used in the paper; (iii) the Task-IV judge prompt and rubric templates from Appendix A.14; and (iv) the metric and aggregation code that turns model outputs into the Task I/II/III/IV scores and the z-score and weighted overall columns reported in Appendix A.15. We do *not* ship raw per-model inference outputs or per-response judge scores in this release; users reproduce those by running the provided harness against their chosen model endpoints.

License and takedown policy. The Video-HOCA release, including the dataset, annotations, taxonomy guideline, judge rubric, evaluation harness, and metric code, is distributed under the Creative Commons Attribution 4.0 International license (CC BY 4.0). Real-world clips for the plausibility task are derived from Panda-70M (Chen et al., 2024a) via the curation pipeline of §3.2; generated anomaly clips are produced by, or reused from, multiple video generators, and each upstream generator's terms of use continue to apply to the underlying visual content. The release manifest records the source and original license or terms of use for every clip. We will honor takedown requests from any upstream provider, video generator, or original creator who objects to the inclusion of a particular clip, and the corresponding record will be removed from the dataset and from the manifest in the next release version.

A.17 Additional Experimental Details

All experiments are zero-shot with fixed prompts per task family. For closed-source models we use the public APIs available at evaluation time (OpenAI for GPT-4o (Hurst et al., 2024), Google for Gemini-2.5-Flash (Comanici et al., 2025)); for open-weight models we use a unified vLLM-based inference stack. Task II and Task III predictions are parsed with a single-choice extraction protocol, and Task IV uses the rubric-based judge protocol described in Appendix A.14.

Main-table decoding uses deterministic Instruct-mode settings for all four tasks: temperature $= 0$, no sampling, and max_new_tokens=1024. The Thinking-vs.-Instruct controls in §4.3 use an 8,192-token output budget for the aligned Instruct and Thinking settings. The Qwen3-VL and Qwen3.5 Thinking settings use temperature $= 1.0$, top-$p = 0.95$, top-$k = 20$; GLM-4.6V-106B uses temperature $= 0.8$ and top-$p = 0.6$. Both Task IV judges, GPT-OSS-120B and DeepSeek-V4-Flash, run locally with vLLM using deterministic decoding (temperature $= 0$) and a 4,096-token output budget.

Main-table video inputs use 16 uniformly sampled frames. The temporal-grounding probe uses 200 interval-prediction QA pairs (covering 59 newly annotated clips and 141 single-anomaly clips reused from the main set with added temporal-interval labels) and reports mIoU and R@0.5. The multi-anomaly probe uses 120 clips with concurrent anomalies and 64-frame input, reporting F1 and exact-match accuracy. The multi-frame scaling probe compares 16-frame and 64-frame inputs on the four main tasks. The full-video / middle-frame / shuffled-frame controls in §4.3 use 16 uniformly sampled frames, a single middle frame, and 16 frames with their order randomly permuted, respectively.