Supplementary Information

Vision-language models lag human performance on physical dynamics and intent reasoning

Tianjun Gu, Jingyu Gong, Zhizhong Zhang,
Yuan Xie, Lizhuang Ma, Xin Tan,
Athanasios V. Vasilakos

Supplementary Notes

Supplementary Note 1: Benchmark comparison

Supplementary Table 1 compares EscherVerse with representative Theory-of-Mind and spatial-reasoning benchmarks. Direct means that a capability is an explicit organizing target of the benchmark. Partial means that the capability can occur in individual examples but is not the primary target. These labels describe task scope rather than benchmark quality. In particular, goal-conditioned spatial change requires both an observable action goal and a spatial consequence, whereas intent or belief questions without a spatial consequence are counted only under intent/Theory-of-Mind.

Supplementary Note 2: Structure of the six category scores

The category-structure analysis used the six reported category accuracies for all 27 models. Spearman correlations were computed for all 15 category pairs. Intervals were obtained from 5,000 resamples of the observed model rows. We standardized each category across models before principal-component analysis (PCA). Parallel analysis retained a component only when its eigenvalue exceeded the 95th percentile from 5,000 independently permuted matrices of the same dimensions. Leave-one-model-out PCA quantified loading stability. These are analyses of the submitted model panel, not population estimates over all possible vision-language models. They test whether category scores share a performance gradient; they do not establish that the six semantic categories are interchangeable or that any individual question requires every category.

Supplementary Note 3: Necessary-perception controls

We sampled 200 unique benchmark videos without using model outcomes. The sample contained 34 items each from Categories 1 and 2 and 33 items from each remaining category. Answer positions were balanced at 50 per option. Each benchmark question was paired with one control that tested a necessary visual prerequisite in the same uniformly sampled frames. The controls comprised action recognition ($n = 34$), object or actor presence ($n = 67$), orientation ($n = 33$), a simple spatial relation ($n = 33$), and temporal order ($n = 33$). Passing a control is necessary for the corresponding benchmark reasoning step but is not sufficient to solve the original question.

We report Wilson intervals for each marginal accuracy and exact McNemar tests for the paired comparison. The two discordant cells are control-correct/original-wrong and control-wrong/original-correct (Supplementary Tables 5 and 6).

Supplementary Note 4: Same-video category-proxy pairs

The 50-pair panel contains 17 dynamics-versus-intent pairs, 17 camera-centred-versus-actor-goal pairs, and 16 tracking-versus-prediction pairs. Arm A is respectively Category 2, 6, or 1; Arm B is respectively Category 3, 3, or 4. The two questions in each pair use the same source video, but the panel is not a causal matched-pair design because question wording and answer format are not identical within every pair. Four models informed the disclosed post-hoc panel selection. Three additional models were evaluated only after the panel was frozen. For those held-out models, intervals and two-sided sign-flip tests use the 50 within-video accuracy differences. The held-out intervals all include zero, so this panel does not support a fixed difficulty ordering.

Supplementary Note 5: Equal-budget SFT/LoRA ablation

We removed every instruction-tuning row whose video occurred in the evaluation benchmark. From the remaining pool, we constructed three non-overlapping 1,409-example subsets: Intent-focused (Category 3, Human-Centric), Dynamic-only (Category 2, Object-Centric), and a Random mixture. The subsets were exactly matched on question format and normalized gold answer. Each condition used the same Qwen3-VL-4B-Instruct initialization, LoRA

capacity, optimization schedule, frame policy, and two random seeds. Evaluation used disjoint 486-item Intent and 486-item Dynamic panels selected without model outcomes and matched on question format and normalized answer.

The content-specific contrast was the difference-in-differences

$$(A_{\text{IntentSFT,Intent}} - A_{\text{DynamicSFT,Intent}}) - (A_{\text{IntentSFT,Dynamic}} - A_{\text{DynamicSFT,Dynamic}}).$$

We used 10,000 bootstrap replicates, clustering by video within evaluation capability. Seed standard deviations describe variation across the two training runs and are not confidence intervals. The result is limited to this model, LoRA configuration, data budget, and pair of seeds.

The six LoRA training runs were executed concurrently on six slots of an eight-GPU server running Ubuntu 24.04. Each run used one GPU. The recorded training time was 10,031.9–11,641.8 seconds (2.79–3.23 hours) per run. Normalizing the aggregate compute by peak dense BF16 Tensor Core throughput gives approximately 7.93 NVIDIA A100-equivalent GPU-hours. This is a cross-hardware normalization rather than a direct A100 runtime measurement. These figures exclude data preparation and evaluation. Some benchmark results were obtained through external model APIs or retained historical runs for which provider-side hardware and energy use are unavailable; we therefore do not estimate a total carbon footprint for the complete study.

Supplementary Note 6: Quantitative failure-mode audit

We drew 100 errors from each of InternVL3-2B, Qwen2.5-VL-7B-Instruct, and Qwen3-VL-8B-Instruct, proportionally stratified by the six benchmark categories. Two independent raters, blinded to model identity in the labeling interface, assigned one of eight mutually exclusive primary labels. A third rater adjudicated all primary-label disagreements. The 300 model-item errors contained 169 unique benchmark items; 98 items appeared for more than one model. Consequently, the reported Wilson intervals are descriptive and do not assume that all 300 observations are independent benchmark questions. Inverse sampling-probability weights restore each model's source error-pool category composition, and the weighted intervals use the Kish effective sample size. The labels characterize observable answer failures, not latent cognitive mechanisms.

Supplementary Note 7: Independent scope and historical-record audits

The scope audit sampled 20 videos from each benchmark category (120 distinct videos). Two independent model-based authoring runs received 64 uniformly sampled full-span frames but no benchmark question, answer, or category label. Each run proposed three answerable questions per video. Evidence and duplication checks retained 666 questions. Two independent scope raters classified every question as directly in scope, a supporting operation, or outside scope. Absorbed is defined as directly in scope plus supporting operation. The primary estimate requires both raters to mark a question as absorbed. Intervals use 10,000 source-video-clustered bootstrap replicates. This proportion measures alignment with the six-dimensional taxonomy, not task-type representation in the retained benchmark. The separate comparison in Supplementary Note 8 addresses the latter.

The historical-record audit is separate from the scope audit. It describes retained human-review handoffs and does not reconstruct the complete generation or filtering pipeline. The workbook and archive are not nested stages; their row-count difference therefore cannot be interpreted as attrition. Counts that cannot be recovered from the retained records are reported as unavailable rather than inferred.

Supplementary Note 8: Task-type representation in the retained question bank

This audit compares newly human-authored reference questions with the questions actually retained in EscherVerse. It addresses fine-grained task representation, separately from the earlier taxonomy-scope audit in Supplementary Note 7.

Question sets and coding. The comparison needs to distinguish what a question is about from what it asks the respondent to do. For example, identifying an object's direction of motion, predicting its next movement, and assessing whether the available views are sufficient to determine its direction all concern motion, but require different judgments. Treating them as a single type would conceal differences in the judgments required.

We therefore describe each question by its content and required operation. Content covers properties, relations, and events such as position, distance, trajectory, and visibility. Operations distinguish observation, comparison, prediction, causal explanation, and assessment of evidence sufficiency, among others. Where needed, we retain subtypes. For visibility, these include occlusion, entry into or exit from the frame, and reflection. Each heatmap row corresponds to a combination identified in at least one question; we do not include arbitrary pairings of labels.

We developed the coding scheme by examining the 200 new questions and a sample of 200 original questions. The resulting codebook contains 17 content categories and 13 operations, with definitions, inclusion and exclusion criteria, and examples. We then applied the same rules to both sets and extended the analysis to the full question bank. This coding scheme was developed specifically for this analysis from the study material. It was not independently prespecified and does not replace the benchmark's six dimensions. Questions may receive multiple labels, but each

content–operation combination must be supported by what the question asks. Mentioning a timestamp, for instance, does not necessarily require comparing states over time.

The 200 new reference questions were written by human annotators based on 60 clips, each from a distinct source video. We sampled 10 clips from each of the six benchmark categories, with three or four questions written per clip. The annotators received the overall task scope and anonymized sampled frames, but no original questions, answers, or clip-level sampling labels. This separate authoring process provides a human-authored reference set, although it does not ensure an unbiased sample of all possible tasks. Thirty clips had been examined in earlier analyses; the questions are new, but the video set is not wholly unseen. Depending on the available material, annotators were given 64 frames for 55 clips and 20 frames for the remaining five. They viewed sampled frames rather than continuous full videos.

For comparison, we sampled 200 existing questions: 100 from videos used for the new questions and 100 from other videos. These subsets allow us to compare questions drawn from the same video sources and from different sources. We also examined all 8,000 original records to determine whether types absent from the smaller sample occurred elsewhere in the bank. The comparison concerns task types: sharing a source video or a label does not establish equivalence between two questions.

Coding was performed on question text using the same codebook. For the full-bank analysis, `gpt-6-sol` received no answers, video input, or original six-category labels. The 8,000 records contain 6,283 unique clip–question groups; identical records shared a coding result, which was mapped back to the original records for frequency reporting. Duplicating a record cannot add a new type, although it can affect its frequency. Only definite assignments contribute to the reported counts; tentative assignments are retained separately. These labels come from an initial model-based classification. They have not undergone a complete human annotation review, and the coding does not verify question answers. In the original bank, 193 records remain flagged because their classification is uncertain, missing, or at a category boundary. We assessed their effect by repeating the comparison after excluding them.

Counting and interpretation. For a given comparison, type representation is the number of distinct reference types with at least one definite assignment in the original set, divided by the number of distinct reference types. Each heatmap row is an observed combination, not an arbitrary pairing of content and operation labels. Cell percentages count questions assigned to that row, with each question counted at most once per row; because coding is multilabel, one question can contribute to multiple rows. Duplicate records can change frequencies but cannot create a new type. White cells indicate no definite assignment and asterisks indicate additional tentative assignments, which do not contribute to shading.

The full original bank has 185 fine-grained types, of which 39 overlap the 45 reference types; their union contains 191 types. The equally sized 200-question original sample contains 74 types, sharing 27 with the reference set, for a two-sample union of 92. Its share of that union is 80.4%, compared with 48.9% for the human-authored reference. These overlapping shares need not sum to 100%. Relative to the complete 191-type inventory, the corresponding shares are 74/191 (38.7%) and 45/191 (23.6%). The original sample therefore has greater observed type diversity at an equal question count, but differences in video composition and question format prevent attributing that advantage solely to the authoring process.

Unidentified combinations and sensitivity. Five of the six combinations not identified concern evidence sufficiency for contact, crossing, orientation, speed and camera position. The sixth is visibility with causal explanation and an unspecified mechanism. The latter is a subtype distinction, not a separate ability; causal explanations involving occlusion are present in the bank. All 14 reference content categories and all nine operations occur even in the 200-question original sample, but a parent category does not establish representation of a specific combination. Supplementary Table 16 gives a concise English translation of one reference question per combination. The examples do not verify the answers.

Excluding all 193 flagged original records leaves 39/45 types represented. Excluding flagged questions from both sets leaves 186 reference questions and 35/41 fine-grained types represented (85.4%). This denominator of 41 denotes remaining fine-grained types, not the coarser combinations in Supplementary Table 13. The same six combinations remain unrepresented. Restricting the reference to 64-frame inputs gives 184 questions and 37/43 types (86.0%); restricting it to the 30 clips not used in the earlier analysis gives 97 questions and 28/32 types (87.5%). These restrictions do not independently validate the labels or codebook. One unresolved original question has a tentative speed–evidence-sufficiency label; accepting it would yield 40/45, but it is not counted as confirmed. The six combinations are therefore not identified under current definite-label coding, rather than proven absent.

Same-clip overlap and limitations. Each of the 186 reference questions without a classification flag was compared with all retained original questions for the same clip, rather than only the 100-question same-video sample. For 28 questions (15.1%), all definite fine-grained types were represented in the original questions for that clip; for 34 (18.3%), at least one was represented. Each sampled clip has one to four unique original questions. These are measures of task-type overlap, not semantic equivalence, answer accuracy, or coverage of all questions a clip could support. Limited question budgets, generation preferences and filtering may all contribute. The audit does not reconstruct the complete candidate-to-retained filtering history, so unidentified types cannot be attributed solely to the generator. Its

results are conditional on sampled videos, human authoring and model-based coding, and do not establish exhaustive coverage of an open-world question space.

Supplementary Tables

Supplementary Table 1: Comparison with representative Theory-of-Mind and spatial-reasoning benchmarks. Direct denotes an explicit evaluation target, whereas Partial denotes a capability that occurs in the task without serving as its primary organizing target. Paired training resource denotes instruction-tuning or post-training data released with the benchmark. Source data are provided as a Source Data file.

a. Scope and setting

Benchmark	Primary target	Input or setting	Dynamic change	Intent/ToM
ToMBench	ToM and social cognition	Text, bilingual MCQ	No	Direct
MMToM-QA	Goals, beliefs and plans	Multimodal household activity	Partial	Direct
EgoToM	Egocentric ToM	Ego4D video	Partial	Direct
MuMA-ToM	Multi-agent ToM	Household video and text	Partial	Direct
SpatialVLM	Metric and 3D space	Images and generated 3D VQA	No	No
VSI-Bench	Video spatial understanding	Real video and navigation QA	Direct	No
Spatial-DISE	Spatial-reasoning taxonomy	Real and synthetic VQA	Direct	No
EscherVerse	Goal-conditioned spatial change	11,328 open-world videos	Direct	Direct

b. Additional capabilities and resources

Benchmark	Goal-conditioned change	Reference frame	Human comparison	Training resource
ToMBench	No	No	Yes	No
MMToM-QA	Partial	No	Yes	No
EgoToM	Partial	Partial	Yes	No
MuMA-ToM	Partial	No	Yes	No
SpatialVLM	No	Partial	No	Yes
VSI-Bench	No	Direct	Yes	No
Spatial-DISE	No	Direct	Yes	Yes
EscherVerse	Direct	Direct	Yes	Yes

Supplementary Table 2: Pairwise Spearman correlations across the six category scores for 27 models. Intervals were obtained by resampling the observed model panel and should not be interpreted as population intervals over all possible models. Source data are provided as a Source Data file.

Pair	ρ	95% interval	Pair	ρ	95% interval
C1–C2	.868	[.704,.946]	C1–C3	.854	[.654,.945]
C1–C4	.864	[.697,.938]	C1–C5	.794	[.539,.919]
C1–C6	.792	[.533,.924]	C2–C3	.950	[.840,.986]
C2–C4	.961	[.870,.990]	C2–C5	.847	[.621,.956]
C2–C6	.928	[.796,.983]	C3–C4	.925	[.797,.976]
C3–C5	.816	[.563,.938]	C3–C6	.864	[.651,.963]
C4–C5	.818	[.556,.951]	C4–C6	.854	[.646,.950]
C5–C6	.753	[.446,.927]			

Supplementary Table 3: Principal-component and parallel analyses of standardized category scores. An observed component was retained when its eigenvalue exceeded the 95th-percentile eigenvalue from random matrices of the same dimensions. Source data are provided as a Source Data file.

Component	Eigenvalue	Variance explained	Parallel threshold	Retained
PC1	5.793	96.56%	2.024	Yes
PC2	0.082	1.37%	1.516	No
PC3	0.072	1.20%	1.217	No
PC4	0.035	0.58%	0.984	No
PC5	0.014	0.24%	0.786	No
PC6	0.003	0.05%	0.590	No

Category	PC1 loading	Leave-one-model-out range
C1	0.979	[0.969,0.983]
C2	0.996	[0.995,0.996]
C3	0.977	[0.970,0.979]
C4	0.991	[0.987,0.993]
C5	0.970	[0.957,0.976]
C6	0.983	[0.979,0.985]

Supplementary Table 4: Accuracy on the 50 same-video question pairs. The first four models informed panel selection and are reported descriptively. The remaining three models were evaluated after the panel was fixed. Intervals and two-sided sign-flip tests use within-video differences. Source data are provided as a Source Data file.

a. Paired accuracies

Model	Analysis role	Arm A	Arm B	B–A
Qwen3-VL-8B-Instruct	panel construction	50.0%	62.0%	+12 points
Qwen2.5-VL-7B-Instruct	panel construction	34.0%	48.0%	+14 points
Qwen3-VL-2B-Instruct	panel construction	40.0%	50.0%	+10 points
InternVL3-2B	panel construction	42.0%	50.0%	+8 points
Qwen3-VL-8B-Thinking	held-out evaluation	46.0%	52.0%	+6 points
Qwen3-VL-4B-Thinking	held-out evaluation	42.0%	50.0%	+8 points
Qwen2.5-VL-32B-Instruct-AWQ	held-out evaluation	52.0%	56.0%	+4 points

b. Held-out uncertainty analysis

Model	95% video-clustered interval	Sign-flip P
Qwen3-VL-8B-Instruct	descriptive	descriptive
Qwen2.5-VL-7B-Instruct	descriptive	descriptive
Qwen3-VL-2B-Instruct	descriptive	descriptive
InternVL3-2B	descriptive	descriptive
Qwen3-VL-8B-Thinking	[-14,+26] points	0.6905
Qwen3-VL-4B-Thinking	[-12,+28] points	0.5709
Qwen2.5-VL-32B-Instruct-AWQ	[-16,+24] points	0.8384

Supplementary Table 5: Complete model-level results for the 200 necessary-perception controls. Brackets are 95% Wilson intervals. Conditional original accuracy is evaluated only among items for which the model passed the control. Discordant counts are used in the exact McNemar test. Source data are provided as a Source Data file.

a. Marginal accuracies

Model	Control correct	Control accuracy	Original correct	Original accuracy	Gap
InternVL3-2B	151/200	75.5% [69.1,80.9]	78/200	39.0% [32.5,45.9]	+36.5 pp
Qwen3-VL-2B-Instruct	81/200	40.5% [33.9,47.4]	76/200	38.0% [31.6,44.9]	+2.5 pp
Qwen3-VL-8B-Instruct	162/200	81.0% [75.0,85.8]	88/200	44.0% [37.3,50.9]	+37.0 pp
Qwen2.5-VL-7B-Instruct	141/200	70.5% [63.8,76.4]	76/200	38.0% [31.6,44.9]	+32.5 pp
Qwen3-VL-8B-Thinking	170/200	85.0% [79.4,89.3]	69/200	34.5% [28.3,41.3]	+50.5 pp
Qwen3-VL-4B-Thinking	172/200	86.0% [80.5,90.1]	78/200	39.0% [32.5,45.9]	+47.0 pp
Qwen2.5-VL-32B-Instruct-AWQ	167/200	83.5% [77.7,88.0]	97/200	48.5% [41.7,55.4]	+35.0 pp

b. Paired outcomes and conditional accuracy

Model	C+O−	C−O+	McNemar P	Original | control pass
InternVL3-2B	92	19	1.12×10^{-12}	39.1%
Qwen3-VL-2B-Instruct	50	45	0.682	38.3%
Qwen3-VL-8B-Instruct	93	19	6.75×10^{-13}	42.6%
Qwen2.5-VL-7B-Instruct	83	18	3.73×10^{-11}	41.1%
Qwen3-VL-8B-Thinking	111	10	1.04×10^{-22}	34.7%
Qwen3-VL-4B-Thinking	103	9	2.31×10^{-21}	40.1%
Qwen2.5-VL-32B-Instruct-AWQ	85	15	4.83×10^{-13}	49.1%

Supplementary Table 6: Perception-control accuracy by prerequisite subtype. Subtype denominators are 34 for action recognition, 67 for object/actor presence, and 33 each for orientation, simple spatial relation, and temporal order. Source data are provided as a Source Data file.

Model	Action	Presence	Orientation	Simple relation	Temporal order
InternVL3-2B	94.1%	92.5%	45.5%	63.6%	63.6%
Qwen3-VL-2B-Instruct	50.0%	37.3%	45.5%	36.4%	36.4%
Qwen3-VL-8B-Instruct	88.2%	95.5%	66.7%	75.8%	63.6%
Qwen2.5-VL-7B-Instruct	85.3%	86.6%	48.5%	51.5%	63.6%
Qwen3-VL-8B-Thinking	94.1%	94.0%	75.8%	87.9%	63.6%
Qwen3-VL-4B-Thinking	91.2%	94.0%	75.8%	93.9%	66.7%
Qwen2.5-VL-32B-Instruct-AWQ	91.2%	97.0%	66.7%	72.7%	75.8%

Supplementary Table 7: Reliability and adjudicated distribution for the 300-error failure-mode audit. Weighted proportions restore each model's source error-pool category composition. Weighted intervals use a Kish effective sample size of 299.3. Positive agreement is the Jaccard overlap of the two raters' pre-adjudication positive sets for each label. Source data are provided as a Source Data file.

a. Adjudicated distribution

Primary label	Final n	Weighted share	95% interval
Benchmark ambiguity	29	9.6%	6.8–13.5%
Protocol or parsing	12	4.1%	2.4–7.0%
Perception or recognition	36	12.1%	8.9–16.3%
Spatiotemporal grounding	86	28.8%	24.0–34.2%
Perspective or reference frame	72	23.7%	19.3–28.9%
Action–goal binding	22	7.4%	5.0–11.0%
Physical prediction or counterfactual	43	14.2%	10.7–18.6%
Other unresolved	0	0.0%	0.0–1.3%

b. Pre-adjudication rater counts

Primary label	Rater 1 n	Rater 2 n	Positive agreement
Benchmark ambiguity	17	24	0.281
Protocol or parsing	11	12	0.917
Perception or recognition	43	44	0.554
Spatiotemporal grounding	90	74	0.691
Perspective or reference frame	67	80	0.670
Action–goal binding	26	23	0.690
Physical prediction or counterfactual	46	43	0.854
Other unresolved	0	0	–

Primary-label raw agreement: 79.3%; Cohen's $\kappa = 0.745$; 114 rows adjudicated. Question/key-problem raw agreement: 89.3%; secondary-label exact agreement: 63.7%. Analysis units: 300 model-item errors; unique benchmark items: 169; repeated across models: 98.

Supplementary Table 8: Scope-status results for 666 independently authored questions from 120 videos. Absorbed combines directly in scope and supporting operation. The primary consensus estimate requires both raters. Intervals are 95% source-video-clustered bootstrap intervals. Source data are provided as a Source Data file.

Decision rule	Count	Rate	95% interval
Rater 1 directly in scope	449	67.4%	63.0–71.6%
Rater 1 absorbed	609	91.4%	88.5–94.1%
Rater 2 directly in scope	354	53.2%	48.8–57.4%
Rater 2 absorbed	518	77.8%	73.0–82.3%
Both raters directly in scope	342	51.4%	47.1–55.6%
Both raters absorbed (primary)	516	77.5%	72.8–82.1%
Both raters outside scope	55	8.3%	–
Taxonomy alignment contested	95	14.3%	–

Supplementary Table 9: Scope-rater agreement and assignment to the six benchmark dimensions. Primary assignments are mutually exclusive and include NONE; multilabel incidence can exceed 666 because a question may invoke more than one dimension. Source data are provided as a Source Data file.

Dimension	Rater 1 primary	Rater 2 primary	Rater 1 multilabel	Rater 2 multilabel
C1 Object permanence and occlusion	32	36	34	36
C2 Dynamic spatial relationships	250	229	383	280
C3 Action and intent-driven reasoning	105	80	139	86
C4 Predictive and counterfactual reasoning	2	2	2	2
C5 Deformation and state transition	86	75	99	81
C6 Egocentric/allocentric reference frames	134	96	151	156
NONE (outside scope)	57	148	–	–
Exact three-status agreement: 69.1%; $\kappa = 0.451$; disagreements: 206.				
Binary absorbed/outside agreement: 85.7%; $\kappa = 0.471$; disagreements: 95.				
Exact primary-dimension agreement: 66.7%; $\kappa = 0.576$.				

Supplementary Table 10: Audit of retained historical human-review records. These sources are non-nested handoffs and cannot be subtracted to estimate full-pipeline attrition. Effective uses the original row when marked correct and a parseable revision when modification was required. Source data are provided as a Source Data file.

Observed quantity	Count
Human-delivery workbook rows	3,569
Marked correct without modification	1,518
Marked as requiring modification	2,020
Blank/untrusted status	31
Effective usable decision rows	3,290
Effective unique video–question keys	3,123
Effective unique videos	2,577
Repeated effective keys	152
Keys with conflicting effective answers	49
Effective keys strictly mapped to final benchmark	1,967
Keys with at least one human answer equal to final	1,461
Separate manual-review archive rows	771
Archive rows strictly mapped to final benchmark	770
Mapped rows with same answer as final	310
Mapped rows with a different answer from final	460

Supplementary Table 11: Per-run and aggregate accuracy for the equal-budget SFT/LoRA ablation. The Base row is a single frozen evaluation of the common initialization. SD is the standard deviation across the two training seeds. Source data are provided as a Source Data file.

a. Per-run accuracies

Condition	Run	Intent accuracy	Dynamic accuracy
Base	frozen base	50.21%	37.65%
Intent-focused	seed 20260814	55.14%	39.09%
Intent-focused	seed 20260815	59.88%	39.09%
Intent-focused	mean (SD)	57.51% (3.35)	39.09% (0.00)
Dynamic-only	seed 20260814	53.70%	42.80%
Dynamic-only	seed 20260815	55.35%	44.44%
Dynamic-only	mean (SD)	54.53% (1.16)	43.62% (1.16)
Random mixture	seed 20260814	52.88%	42.80%
Random mixture	seed 20260815	51.44%	40.74%
Random mixture	mean (SD)	52.16% (1.02)	41.77% (1.45)

b. Gains over the frozen base

Condition	Run	Intent gain	Dynamic gain
Base	frozen base	–	–
Intent-focused	seed 20260814	+4.94 pp	+1.44 pp
Intent-focused	seed 20260815	+9.67 pp	+1.44 pp
Intent-focused	mean (SD)	+7.30 pp	+1.44 pp
Dynamic-only	seed 20260814	+3.50 pp	+5.14 pp
Dynamic-only	seed 20260815	+5.14 pp	+6.79 pp
Dynamic-only	mean (SD)	+4.32 pp	+5.97 pp
Random mixture	seed 20260814	+2.67 pp	+5.14 pp
Random mixture	seed 20260815	+1.23 pp	+3.09 pp
Random mixture	mean (SD)	+1.95 pp	+4.12 pp

Difference-in-differences: 7.51 pp; bootstrap estimate: 7.59 pp; 95% video-clustered interval: 3.10–12.17 pp; two-sided bootstrap $P = 0.0018$.

Supplementary Table 12: Data construction and training configuration for the equal-budget SFT/LoRA ablation. The three training subsets have zero pairwise row overlap and exclude all benchmark videos. Source data are provided as a Source Data file.

Component	Setting	Value
Source audit	Valid instruction-tuning rows	35,963
Source audit	Rows after benchmark-video exclusion	24,398
Eligible pool	Intent-focused rows before matching	4,944
Eligible pool	Dynamic-only rows before matching	1,584
Frozen subsets	Examples per condition	1,409
Frozen subsets	Question-type counts per condition	52 fill-in, 20 multiple-select, 1,133 single-choice, 204 true/false
Frozen subsets	Unique videos: Intent/Dynamic/Random	1,399 / 1,180 / 1,310
Evaluation	Intent/Dynamic items	486 / 486
Evaluation	Unique videos: Intent/Dynamic/overall	421 / 361 / 782
Model	Initialization	Qwen3-VL-4B-Instruct
Optimization	Epochs; batch; accumulation	1; 1; 8
Optimization	Learning rate; scheduler; warmup	2×10^{-4}; cosine; 0.03
Precision/input	Precision; video frames	bfloat16; 8
LoRA	Rank; alpha; dropout	16; 32; 0.05
LoRA	Target modules	q_proj, k_proj, v_proj, o_proj
LoRA	Visual encoder	Frozen
Seeds	Training seeds	20260814, 20260815
Bootstrap	Replicates; cluster	10,000; video within capability

Supplementary Table 13: Types observed in the 200 human-authored reference questions and represented in the full original bank. Only definite assignments are counted. Source data are provided as a Source Data file.

Level of comparison	New questions	Represented in original bank
Content category	14	14 (100%)
Required operation	9	9 (100%)
Content–operation combination	41	38 (92.7%)
Combination including subtype	45	39 (86.7%)

Supplementary Table 14: Task diversity in equally sized sets of 200 questions. Shares refer to the union of the two samples, not all possible tasks. Source data are provided as a Source Data file.

Measure	New questions	Original sample
Number of questions	200	200
Content categories observed	14	17
Operations observed	9	13
Fine-grained types observed	45	74
Share of the two-sample union (92 types)	48.9%	80.4%

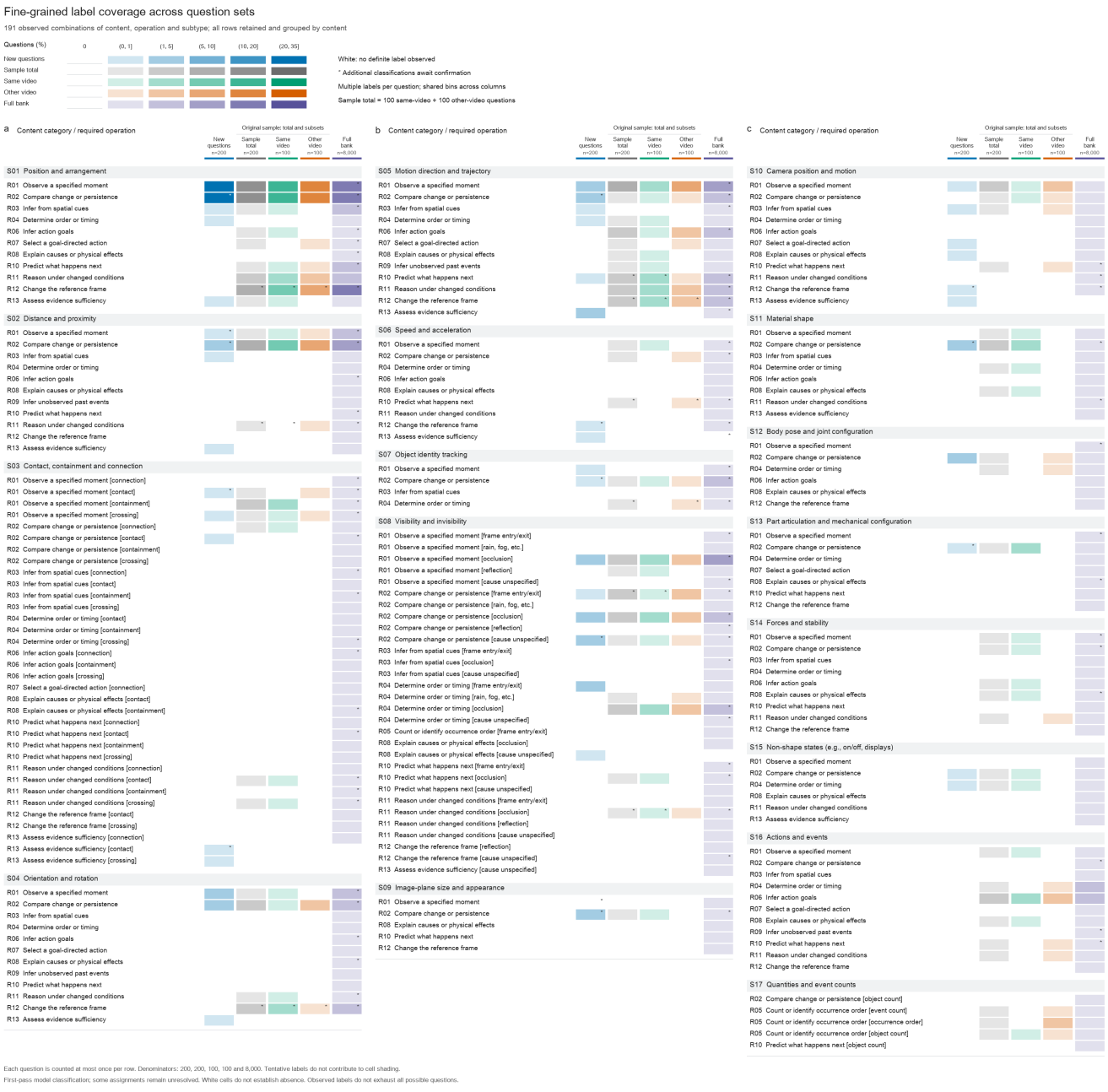
Supplementary Table 15: Sensitivity of fine-grained reference-type representation in the full original bank. The 41-type denominator after excluding flagged questions is distinct from the coarser 41 combinations in Supplementary Table 13. Source data are provided as a Source Data file.

Comparison	Reference questions	Types represented
All questions; definite labels	200	39/45 (86.7%)
Exclude 193 flagged original records	200	39/45 (86.7%)
Exclude flagged questions from both sets	186	35/41 (85.4%)
Restrict reference to 64-frame inputs	184	37/43 (86.0%)
Restrict reference to 30 additional clips	97	28/32 (87.5%)

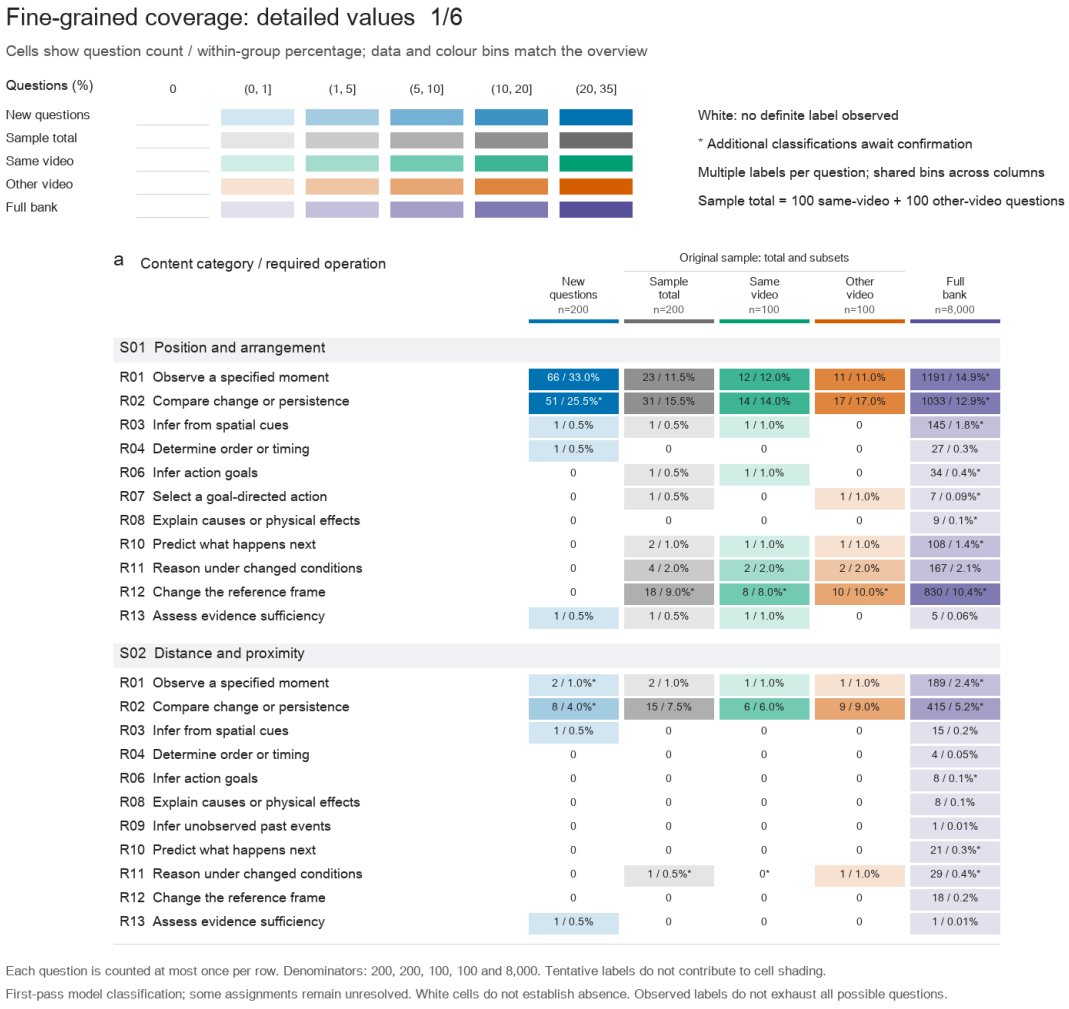
Supplementary Table 16: Reference questions illustrating six combinations not identified under the current definite-label coding. English text is a concise translation; source identifiers link the examples to coding records. The examples do not independently verify answers. Source data are provided as a Source Data file.

Combination and source	Reference question (concise translation)
Contact / evidence sufficiency `vq-052-q03`	Between approximately 26 and 33 seconds, does the helicopter's position relative to the grass change? Are these views sufficient to establish that it has touched down?
Crossing / evidence sufficiency `vq-056-q02`	After the turn, how does the distant white car's position change relative to the passage and within the image? Is this sufficient to establish that it crossed the passage?
Orientation / evidence sufficiency `vq-046-q04`	Which direction within the image is the car facing at approximately 8.8 and 13 seconds? Is this sufficient to establish that it made a U-turn between the two views?
Speed / evidence sufficiency `vq-005-q02`	What happens to the two motorcycles after the camera turns towards the stands? Is this sufficient to establish that they stopped?
Camera position / evidence sufficiency `vq-050-q03`	Between 19.125 and 20.719 seconds, the pylon shifts from the left towards the centre of the image. Does its changing position relative to the surrounding city establish that the pylon itself moved?
Visibility / causal explanation / unspecified mechanism `vq-034-q04`	Why is the distant village harder to see in its entirety near the end than at approximately 24 seconds?

Supplementary Figures



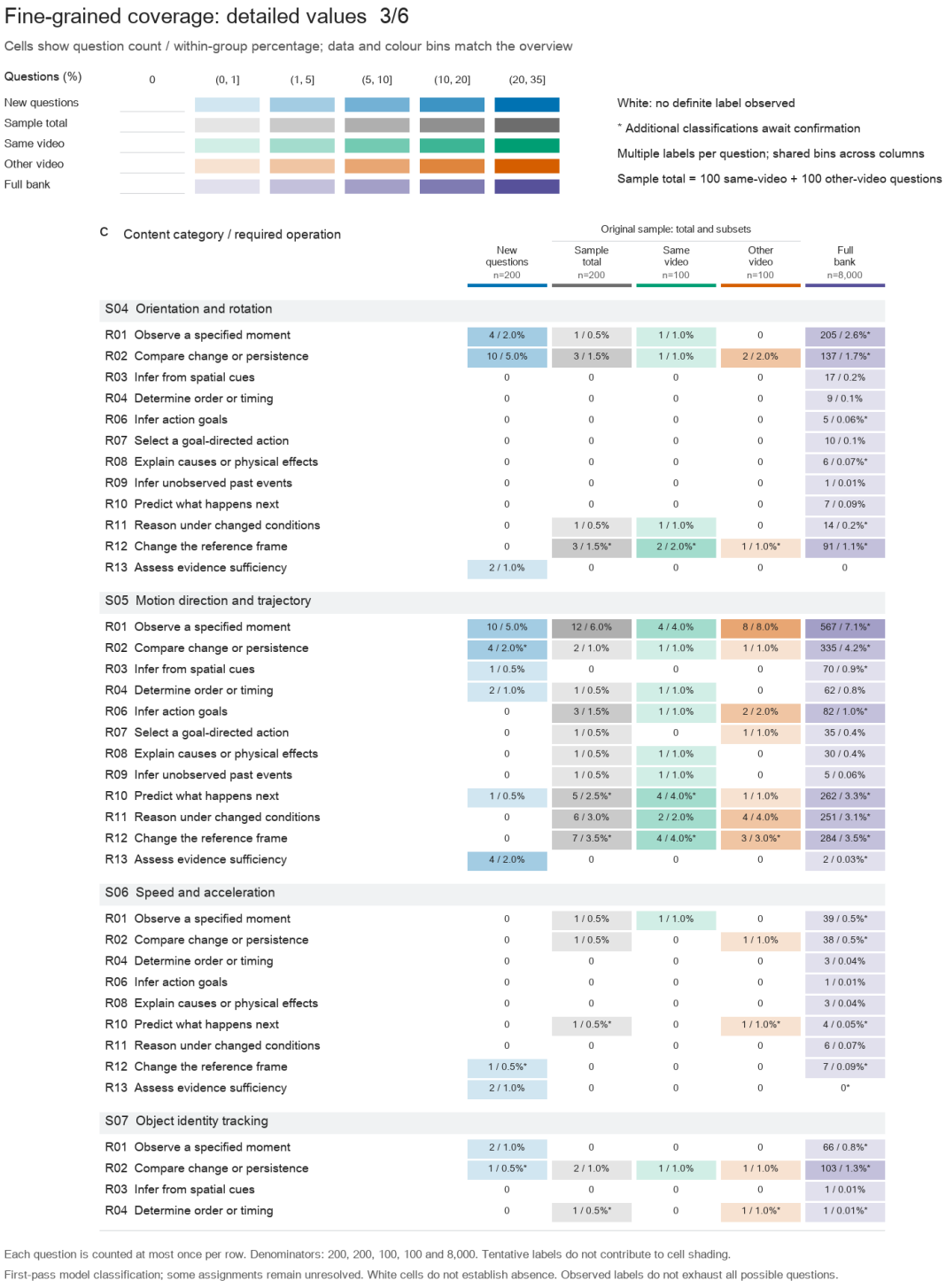
Supplementary Figure 1: Fine-grained task distributions across the human-authored reference questions, original samples and full original bank. All 191 observed combinations are grouped by content. White cells indicate no definite assignment; asterisks indicate additional assignments awaiting confirmation. Questions may receive multiple labels. Tentative assignments do not contribute to shading. Source data are provided as a Source Data file.



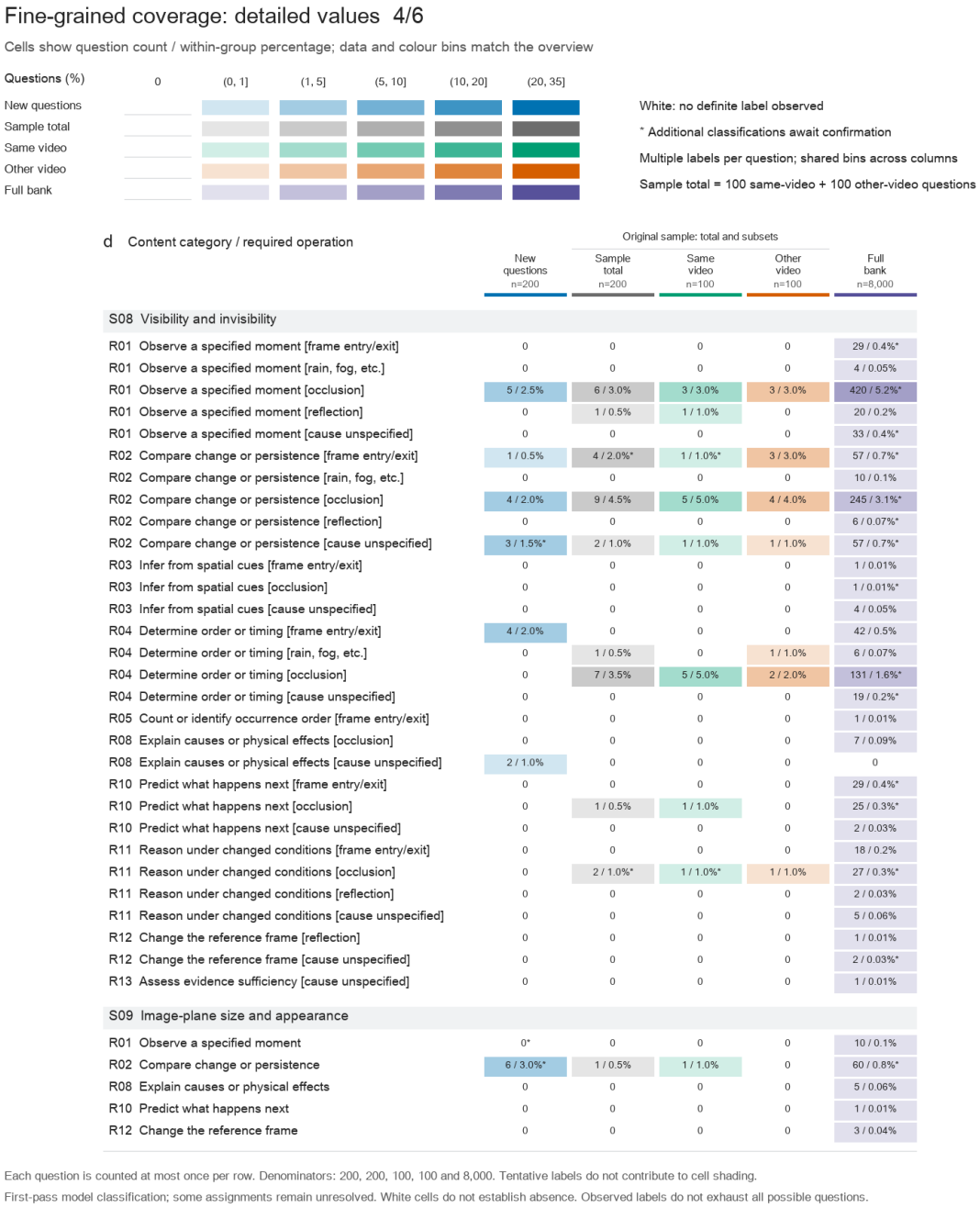
Supplementary Figure 2: Detailed task distributions (1/6). Cells show counts and within-group percentages for the corresponding rows of Supplementary Figure 1. White cells indicate no definite assignment; asterisks indicate additional assignments awaiting confirmation. Questions may receive multiple labels. Tentative assignments do not contribute to shading. Source data are provided as a Source Data file.



Supplementary Figure 3: Detailed task distributions (2/6). Cells show counts and within-group percentages for the corresponding rows of Supplementary Figure 1. White cells indicate no definite assignment; asterisks indicate additional assignments awaiting confirmation. Questions may receive multiple labels. Tentative assignments do not contribute to shading. Source data are provided as a Source Data file.



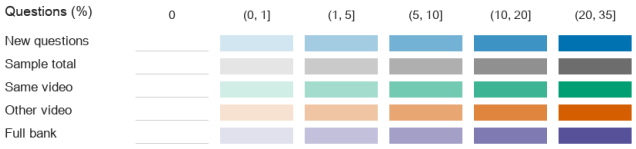
Supplementary Figure 4: Detailed task distributions (3/6). Cells show counts and within-group percentages for the corresponding rows of Supplementary Figure 1. White cells indicate no definite assignment; asterisks indicate additional assignments awaiting confirmation. Questions may receive multiple labels. Tentative assignments do not contribute to shading. Source data are provided as a Source Data file.



Supplementary Figure 5: Detailed task distributions (4/6). Cells show counts and within-group percentages for the corresponding rows of Supplementary Figure 1. White cells indicate no definite assignment; asterisks indicate additional assignments awaiting confirmation. Questions may receive multiple labels. Tentative assignments do not contribute to shading. Source data are provided as a Source Data file.

Fine-grained coverage: detailed values 5/6

Cells show question count / within-group percentage; data and colour bins match the overview



White: no definite label observed

* Additional classifications await confirmation

Multiple labels per question; shared bins across columns

Sample total = 100 same-video + 100 other-video questions

e Content category / required operation

	Original sample: total and subsets				
	New questions n=200	Sample total n=200	Same video n=100	Other video n=100	Full bank n=8,000
S10 Camera position and motion					
R01 Observe a specified moment	1 / 0.5%	4 / 2.0%	1 / 1.0%	3 / 3.0%	30 / 0.4%
R02 Compare change or persistence	0	2 / 1.0%	1 / 1.0%	1 / 1.0%	44 / 0.5%
R03 Infer from spatial cues	1 / 0.5%	1 / 0.5%	0	1 / 1.0%	9 / 0.1%
R04 Determine order or timing	0	0	0	0	4 / 0.05%
R06 Infer action goals	0	0	0	0	3 / 0.04%
R07 Select a goal-directed action	1 / 0.5%	0	0	0	1 / 0.01%
R08 Explain causes or physical effects	1 / 0.5%	0	0	0	15 / 0.2%
R10 Predict what happens next	0	1 / 0.5%	0	1 / 1.0%	13 / 0.2%*
R11 Reason under changed conditions	0	0	0	0	5 / 0.06%*
R12 Change the reference frame	1 / 0.5%*	0	0	0	19 / 0.2%*
R13 Assess evidence sufficiency	2 / 1.0%	0	0	0	0
S11 Material shape					
R01 Observe a specified moment	0	1 / 0.5%	1 / 1.0%	0	17 / 0.2%
R02 Compare change or persistence	5 / 2.5%*	5 / 2.5%	5 / 5.0%	0	63 / 0.8%*
R03 Infer from spatial cues	0	0	0	0	1 / 0.01%
R04 Determine order or timing	0	1 / 0.5%	1 / 1.0%	0	7 / 0.09%
R06 Infer action goals	0	0	0	0	1 / 0.01%
R08 Explain causes or physical effects	0	1 / 0.5%	1 / 1.0%	0	16 / 0.2%
R11 Reason under changed conditions	0	0	0	0	6 / 0.07%*
R13 Assess evidence sufficiency	0	0	0	0	1 / 0.01%
S12 Body pose and joint configuration					
R01 Observe a specified moment	0	0	0	0	20 / 0.2%*
R02 Compare change or persistence	5 / 2.5%	1 / 0.5%	0	1 / 1.0%	22 / 0.3%
R04 Determine order or timing	0	1 / 0.5%	0	1 / 1.0%	3 / 0.04%
R06 Infer action goals	0	0	0	0	1 / 0.01%
R08 Explain causes or physical effects	0	0	0	0	1 / 0.01%
R12 Change the reference frame	0	0	0	0	1 / 0.01%
S13 Part articulation and mechanical configuration					
R01 Observe a specified moment	0	0	0	0	19 / 0.2%*
R02 Compare change or persistence	2 / 1.0%*	2 / 1.0%	2 / 2.0%	0	20 / 0.2%
R04 Determine order or timing	0	0	0	0	5 / 0.06%
R07 Select a goal-directed action	0	0	0	0	1 / 0.01%
R08 Explain causes or physical effects	0	0	0	0	1 / 0.01%*
R10 Predict what happens next	0	0	0	0	1 / 0.01%
R12 Change the reference frame	0	0	0	0	1 / 0.01%

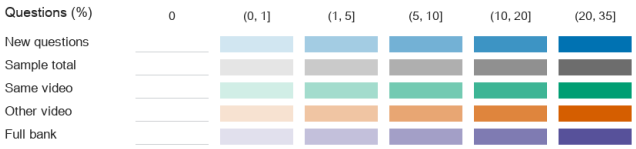
Each question is counted at most once per row. Denominators: 200, 200, 100, 100 and 8,000. Tentative labels do not contribute to cell shading.

First-pass model classification; some assignments remain unresolved. White cells do not establish absence. Observed labels do not exhaust all possible questions.

Supplementary Figure 6: Detailed task distributions (5/6). Cells show counts and within-group percentages for the corresponding rows of Supplementary Figure 1. White cells indicate no definite assignment; asterisks indicate additional assignments awaiting confirmation. Questions may receive multiple labels. Tentative assignments do not contribute to shading. Source data are provided as a Source Data file.

Fine-grained coverage: detailed values 6/6

Cells show question count / within-group percentage; data and colour bins match the overview



White: no definite label observed

* Additional classifications await confirmation

Multiple labels per question; shared bins across columns

Sample total = 100 same-video + 100 other-video questions

f Content category / required operation

	Original sample: total and subsets				
	New questions n=200	Sample total n=200	Same video n=100	Other video n=100	Full bank n=8,000
S14 Forces and stability					
R01 Observe a specified moment	0	1 / 0.5%	1 / 1.0%	0	14 / 0.2%*
R02 Compare change or persistence	0	1 / 0.5%	1 / 1.0%	0	14 / 0.2%*
R03 Infer from spatial cues	0	0	0	0	5 / 0.06%
R04 Determine order or timing	0	0	0	0	1 / 0.01%
R06 Infer action goals	0	1 / 0.5%	1 / 1.0%	0	8 / 0.1%
R08 Explain causes or physical effects	0	1 / 0.5%	1 / 1.0%	0	26 / 0.3%*
R10 Predict what happens next	0	0	0	0	1 / 0.01%
R11 Reason under changed conditions	0	1 / 0.5%	0	1 / 1.0%	9 / 0.1%
R12 Change the reference frame	0	0	0	0	5 / 0.06%
S15 Non-shape states (e.g., on/off, displays)					
R01 Observe a specified moment	0	0	0	0	6 / 0.07%
R02 Compare change or persistence	1 / 0.5%	1 / 0.5%	1 / 1.0%	0	4 / 0.05%
R04 Determine order or timing	1 / 0.5%	1 / 0.5%	1 / 1.0%	0	3 / 0.04%
R08 Explain causes or physical effects	0	0	0	0	1 / 0.01%
R11 Reason under changed conditions	0	0	0	0	1 / 0.01%
R13 Assess evidence sufficiency	0	0	0	0	1 / 0.01%
S16 Actions and events					
R01 Observe a specified moment	0	1 / 0.5%	1 / 1.0%	0	58 / 0.7%
R02 Compare change or persistence	0	0	0	0	4 / 0.05%*
R03 Infer from spatial cues	0	0	0	0	1 / 0.01%
R04 Determine order or timing	0	1 / 0.5%	0	1 / 1.0%	99 / 1.2%
R06 Infer action goals	0	4 / 2.0%	2 / 2.0%	2 / 2.0%	89 / 1.1%
R07 Select a goal-directed action	0	0	0	0	3 / 0.04%
R08 Explain causes or physical effects	0	1 / 0.5%	1 / 1.0%	0	30 / 0.4%
R09 Infer unobserved past events	0	0	0	0	2 / 0.03%*
R10 Predict what happens next	0	1 / 0.5%	0	1 / 1.0%	24 / 0.3%*
R11 Reason under changed conditions	0	1 / 0.5%	0	1 / 1.0%	24 / 0.3%
R12 Change the reference frame	0	0	0	0	2 / 0.03%
S17 Quantities and event counts					
R02 Compare change or persistence [object count]	0	0	0	0	1 / 0.01%
R05 Count or identify occurrence order [event count]	0	1 / 0.5%	0	1 / 1.0%	9 / 0.1%
R05 Count or identify occurrence order [occurrence order]	0	2 / 1.0%	0	2 / 2.0%	11 / 0.1%
R05 Count or identify occurrence order [object count]	0	2 / 1.0%	1 / 1.0%	1 / 1.0%	47 / 0.6%
R10 Predict what happens next [object count]	0	0	0	0	1 / 0.01%

Each question is counted at most once per row. Denominators: 200, 200, 100, 100 and 8,000. Tentative labels do not contribute to cell shading.

First-pass model classification; some assignments remain unresolved. White cells do not establish absence. Observed labels do not exhaust all possible questions.

Supplementary Figure 7: Detailed task distributions (6/6). Cells show counts and within-group percentages for the corresponding rows of Supplementary Figure 1. White cells indicate no definite assignment; asterisks indicate additional assignments awaiting confirmation. Questions may receive multiple labels. Tentative assignments do not contribute to shading. Source data are provided as a Source Data file.

Supplementary Note 9. Source videos for the illustrated cases

Supplementary Table 17 provides links to the original videos underlying the text cases in Figures 3a and 7c–e. Time ranges refer to the original videos; timestamps within questions refer to the corresponding extracted clips. Figure 3d uses schematic video symbols to illustrate the construction pipeline. The text presentation does not change the video inputs used for evaluation or the reported responses and results.

Supplementary Table 17: Original video links and clip time ranges for the cases in Figures 3a and 7c–e. Each linked identifier opens the corresponding original video.

Figure / case	Scene	Video identifier	Time range (min:s)
3a, case 01	Ball catching and walking	`10CKm0h0ZvE`	05:29.329–05:42.242
3a, case 02	City view and viewing platform	`KRVIcZifq0M`	07:23.360–07:46.279
3a, case 03	Bamboo pole and roof	`JoNF5uRkoXQ`	01:45.038–02:04.090
3a, case 04	Repeated jumps across a line	`K23Kb4BvwyE`	00:16.399–00:38.166
3a, case 05	Fishing rod and reel	`ZOYahjTf2M4`	00:47.799–01:03.640
3a, case 06	Puppy and picnic table	`N_qdIrFeKSE`	00:39.072–01:25.985
7c	Bridge and castle gate	`btZR6ryrsZM`	24:55.920–25:09.039
7d	Golf green and flagstick	`gabXO4pMBQY`	16:10.666–16:25.233
7e	Surfing maneuver	`g5nVWDNaz-0`	01:33.793–01:54.414