MIND THE SPACE: BENCHMARKING SPATIAL STATE EVOLUTION IN VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A generated video can depict the requested action while failing to realize its spatial consequences. We introduce **MTS-Bench** (Mind the Space Benchmark), which evaluates this distinction through interventions whose task-supported spatial consequences are not fully narrated in the prompt. Its 337 curated cases pair image-to-video (I2V) and text-to-video (T2V) formulations across scene, view, and coupled motion. Case-specific rubrics assess initial-state fidelity, instruction execution, consequence prediction, temporal consistency, and endpoint completion. We evaluate nine I2V and eight T2V generators on a fixed 150-case subset, using a VLM scorer audited against blinded human ratings and selectively calibrated by dimension. The highest-scoring generator obtains 80.2 on instruction following and 56.2 on consequence prediction in I2V; the corresponding T2V scores are 77.2 and 62.4 (out of 100). Matched qualitative examples expose action without the required spatial change, while dimension profiles distinguish generators with similar aggregate scores. Evaluator analysis further shows that strong agreement with aggregate rankings can coexist with weak discrimination between videos. MTS-Bench makes intervention-induced spatial evolution an explicit, diagnosable evaluation target for video world models.

1 INTRODUCTION

Video generators increasingly serve as candidates for world models, with applications in action-conditioned simulation and embodied learning (Brooks et al., 2024; Bruce et al., 2024; Parker-Holder et al., 2024; Agarwal et al., 2025). Their usefulness in these settings depends on whether an action produces the corresponding change in the scene. Advances in visual realism (Kong et al., 2024; Polyak et al., 2024) make this requirement easy to overlook: a video may depict the right action while leaving the relevant spatial relations wrong. Consider an ego vehicle following a lead vehicle at the same speed. If the ego vehicle accelerates while the lead vehicle maintains its speed, the gap should shrink; visible forward motion alone does not establish that this relation has changed. This raises a concrete question for video world models: *can they realize the spatial consequences of an intervention when those consequences are not explicitly described?*

Existing benchmarks evaluate perceptual quality, instruction adherence, physical plausibility, and interaction (Huang et al., 2024; Li et al., 2026a; Han et al., 2026; Ying et al., 2026). Spatial evaluation further tests controllability under prescribed camera trajectories in WorldScore (Duan et al., 2025) and changes between text-specified relations in SpatialAlign (Liu et al., 2026). We focus on the relation between the instructed action and its *unstated but context-supported spatial consequences*. This distinction matters because prompt satisfaction can reward the visible action without establishing its outcome, while a single reference continuation cannot represent every valid realization. Testing the consequence requires a criterion derived from the initial state and intervention, together with separate judgments of action execution and spatial consistency. MTS-Bench makes this distinction explicit across object motion, viewpoint changes, and their interaction (Table 1).

We introduce **MTS-Bench** (Mind the Space Benchmark), a collection of 337 paired image-to-video (I2V) and text-to-video (T2V) tasks. Each task specifies an initial state and intervention, including an explicit goal where needed, while leaving the full spatial evolution to the generator. I2V supplies the initial image; T2V describes the corresponding task-relevant state. Case-specific rubrics identify the consequences supported by that context without requiring reproduction of a source video. The

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.

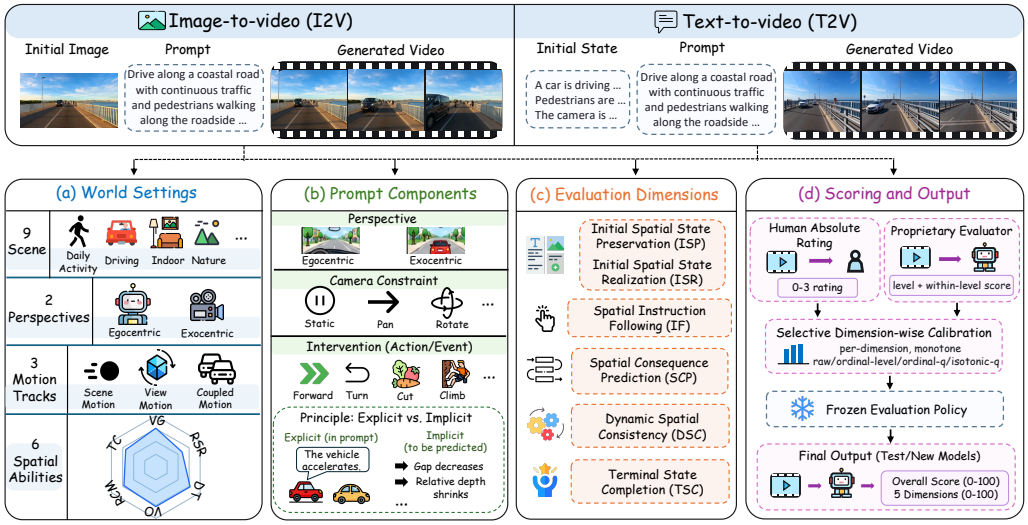


Figure 1: **Overview of MTS-Bench.** Three motion-source tracks and six spatial abilities organize dynamic spatial evaluation across diverse scenes and perspectives. I2V and T2V inputs specify an initial state and intervention, which may include an explicit task goal, without narrating the full spatial evolution. Five rubric-based dimensions assess generated videos; a frozen evaluator is checked against human ratings for scalable scoring.

cases span egocentric and exocentric perspectives, three motion-source tracks—Scene Motion, View Motion, and Coupled Motion—and six spatial abilities (Figure 1). A fixed 150-case subset supports the model comparison.

Five evaluation dimensions distinguish initial-state fidelity, instruction execution, consequence prediction, temporal consistency, and endpoint completion. This decomposition identifies whether a model establishes the scene, performs the action, realizes its consequences, and maintains the resulting geometry. Across nine I2V and eight T2V generators, the highest-scoring generator obtains 80.2 on instruction following but 56.2 on consequence prediction in I2V; matched examples show actions with incomplete spatial outcomes. We also examine the reliability of the measurement: a frozen VLM scorer is assessed against blinded human ratings at the video and generator levels, with selective calibration where supported by an audit. Together, these analyses expose both generator failure profiles and the limits of automatic evaluation. Our contributions are threefold:

- **A test of intervention-induced spatial change.** We separate what a prompt instructs from the spatial consequences implied by its context, with distinct criteria for execution, consequence, consistency, and completion.
- **A benchmark with structured diagnostic coverage.** MTS-Bench pairs 337 I2V/T2V tasks with case-specific rubrics across scene, view, and coupled motion, six spatial abilities, and two viewing perspectives.
- **An empirical account of spatial failure and evaluator reliability.** A 150-case study reveals instruction–consequence score gaps and different failure profiles behind similar totals; human comparisons distinguish score agreement, video discrimination, and generator ranking.

2 Related Work

Video generation and world-model evaluation. Generators such as CogVideoX, HunyuanVideo, Wan and Movie Gen have advanced fidelity and controllability (Yang et al., 2025; Kong et al., 2024; Wan et al., 2025; Polyak et al., 2024), motivating their use for action-conditioned simulation (Yue et al., 2025; Shang et al., 2026; Gao et al., 2024). VBench, VBench++, EvalCrafter, and T2V-CompBench evaluate visual quality, temporal coherence, and compositional prompt satisfaction (Huang et al., 2024; 2025; Liu et al., 2024; Sun et al., 2025). WorldModelBench, PAI-Bench, Physics-IQ, and Physion-Eval examine instruction adherence, physical prediction, and physical failures (Li et al., 2026a; Zhou et al., 2026; Motamed et al., 2026; Zhang et al., 2026). OSCBench targets object state

Table 1: **Coverage of dynamic spatial benchmarks.** Filled circles indicate a dedicated task, subset, or metric with reported results. Open circles indicate no dedicated evaluation; related examples may still occur in the data.

Benchmark	Data		Motion coverage			Evaluation					Protocol	
	Scale	Persp.	Scene	View	Cpl.	Ground.	Exec.	Pred.	Maint.	Diag.	Ref.	H+V
WorldModelBench	350	Ego/Exo	○	○	○	○	●	○	○	○	○	●
PAI-Bench	2,808	Ego/Exo	○	○	○	○	○	○	○	○	○	○
Physics-IQ	198	Exo	●	○	○	○	○	●	○	○	●	○
Physion-Eval	2,486	Ego/Exo	●	○	○	○	○	○	○	●	●	○
WorldScore	3,000	Exo	●	●	○	○	●	○	●	○	○	○
SpatialAlign	120	Exo	●	○	○	●	●	○	○	○	○	○
WBench	289	Ego/Exo	●	●	○	○	●	○	●	○	○	●
MTS-Bench (Ours)	**337**	**Ego/Exo**	●	●	●	●	●	●	●	●	●	●

Key: ● dedicated evaluation; ○ none dedicated. Cpl. = coupled motion; Ground. = spatial grounding; Exec. = instructed execution; Pred. = consequence prediction; Maint. = world-state maintenance; Diag. = error diagnosis; Ref. = paired source continuation; H+V = joint human–VLM evaluation.

changes, while interaction benchmarks assess navigation and multi-turn consistency (Han et al., 2026; Wu et al., 2026; Ying et al., 2026). These settings establish the relevance of evaluating changes in world state.

Spatial evaluation and the position of MTS-Bench. WorldScore measures controllability, quality, and dynamics under next-scene and camera-trajectory specifications (Duan et al., 2025). SpatialAlign's DSR-Score uses tracked object geometry to evaluate relations stated in text; its controlled tasks specify an animal's initial and final relation to a static object (Liu et al., 2026). MTS-Bench asks whether a specified intervention realizes additional, context-supported spatial consequences. Its case-specific criteria cover relative depth, occlusion, topology, and joint camera–object motion, alongside explicit instruction execution. The contribution lies in this task-and-rubric design and the resulting failure diagnosis. The standard calibration methods support measurement, with evaluator limitations assessed empirically in Sec. 5.4.

3 The MTS-Bench Dataset

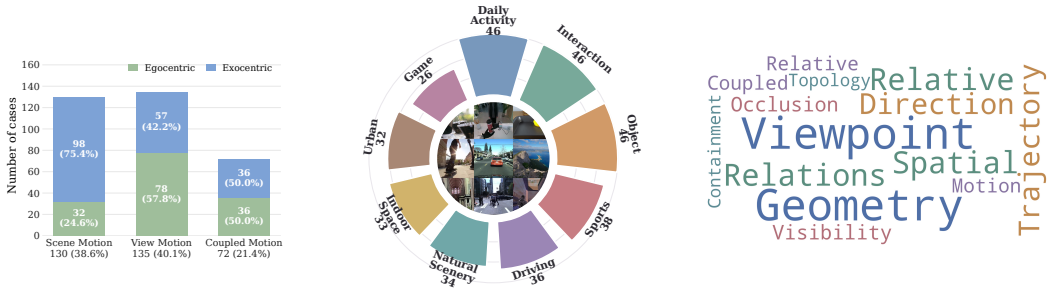
3.1 Data Sources and Sample Curation

MTS-Bench contains **337 annotated cases** drawn from 12 existing video datasets and benchmarks (Chen et al., 2024; Park et al., 2026; Zhou et al., 2026; Li et al., 2026b; Wang et al., 2026b;a; Grauman et al., 2022; Motamed et al., 2026; Nam et al., 2026; Wang et al., 2026a; Kim et al., 2019; Grauman et al., 2024). A case pairs a selected initial state with a task-specific intervention. Some source videos contribute multiple cases at different moments or events. The collection includes captured footage, game environments, and synthetic or AI-generated videos across nine scene categories (Figure 2). Source continuations support curation of the intended transition but do not define a unique future. At evaluation time, the scorer judges the generated video against the task and rubric; it does not compare pixels with the source continuation.

We retain cases with identifiable entities and initial spatial relations, an observable intervention-induced change, and an instruction that specifies the intervention without narrating the full spatial evolution. We exclude shot cuts, severe blur or camera shake, ambiguous motion subjects, insufficient initial context, and near-duplicates. Each retained case is manually reviewed for an unambiguous initial state, intervention, and evaluable spatial change.

3.2 Fine-Grained Spatial Taxonomy

Each case receives one motion-track label and one or more spatial-ability labels. Tracks describe *what drives the spatial change*, whereas abilities describe *which spatial properties are involved*. The evaluation dimensions in Sec. 4.1 instead assess different aspects of success on the resulting task.



(a) Motion Track × Perspective **(b) Scene Category Distribution** **(c) Spatial Ability Word Cloud**

Figure 2: Statistics of MTS-Bench. (a) Motion tracks across two perspectives. (b) Nine scene categories, with radial bar height indicating case counts. (c) Spatial-ability word cloud, with word size proportional to annotation frequency. Spatial abilities are multi-label.

Motion tracks. *Scene Motion* concerns moving entities under a fixed or mildly moving viewpoint. *View Motion* concerns viewpoint or ego-motion, while *Coupled Motion* concerns joint scene and viewpoint motion whose relative effects are part of the task. Incidental camera drift does not qualify as Coupled Motion: the joint effect must matter to the required spatial change.

Spatial abilities. *Viewpoint Geometry* concerns perspective, parallax, and scene layout under viewpoint changes. *Relative Spatial Relations* concerns relations such as left/right, front/behind, and near/far. *Direction and Trajectory* concerns motion direction and paths, while *Visibility and Occlusion* concerns how entities become hidden or revealed. *Relative and Coupled Motion* concerns relative movement among entities or between entities and the observer. *Topology and Containment* concerns connectivity and inside/outside relations. The driving example is Coupled Motion, involving relative relations and coupled motion: IF checks acceleration, SCP checks the shrinking gap, and DSC checks coherent identities and trajectories.

3.3 Task Definition and Evaluation Boundary

What counts as a spatial consequence? We use *spatial state* to refer to task-relevant entities and their geometric, visibility, and topological relations over time. A consequence criterion concerns a change supported by the initial context and intervention, even when its realization is not narrated in the prompt. This permits different valid trajectories while testing a shared relational outcome. A hidden rubric is therefore not a requirement to reproduce incidental details of the source video.

Model-visible input and evaluator-only criteria. I2V inputs contain an initial image and three model-visible fields: **Perspective**, **Camera Constraint**, and **Intervention**. These specify the viewing perspective, viewpoint behavior, and action or event. T2V replaces the image with the description of the initial entities, relations, scene structure, and motion states. The model receives the initial state and instruction; the case-specific scoring rubric is supplied to the evaluator. Both input formats allow explicit task goals but do not narrate the full spatial evolution. SCP assesses additional task-supported consequences not directly specified in the prompt; TSC assesses completion for endpoint-based tasks, including those with explicit goals. For example, "Open the topmost drawer with the right hand" specifies the actor and action. The evaluator additionally checks outward drawer displacement relative to the cabinet and coherent drawer–cabinet geometry. These requirements distinguish the visible manipulation, its spatial consequences, and consistency through time (Appendix D.1).

The 337 I2V/T2V pairs share intervention, perspective, and camera constraint. The fixed 150-case evaluation subset covers all nine scene categories, three tracks, two perspectives, and six abilities; 187 cases remain outside the main comparison. The 40-case human-reference collection supports calibration development and auditing (Sec. 4.3; Appendix A).

Table 2: **Evaluation dimensions.** I2V uses ISP and T2V uses ISR; both share IF, SCP, DSC, and TSC. Each score isolates one aspect of task success.

Dimension	Input	Evaluation criterion
Initial Spatial State Preservation (ISP)	I2V	Early frames preserve the image's entities, relations, visibility, topology, and viewpoint.
Initial Spatial State Realization (ISR)	T2V	Establish the text-specified entities, counts, relations, topology, viewpoint, and motion state.
Spatial Instruction Following (IF)	Both	Execute the requested action with the specified actor, direction, viewpoint, and event order.
Spatial Consequence Prediction (SCP)	Both	Realize task-supported changes in relative position, depth, visibility, topology, or coupled motion.
Dynamic Spatial Consistency (DSC)	Both	Maintain coherent identities, trajectories, relations, occlusion, and camera–object geometry over time.
Terminal State Completion (TSC)	Both	Reach the required terminal spatial state within the generation horizon.

4 The MTS-Bench Evaluation Suite

4.1 Dimensions and Rubrics

The six spatial abilities specify *what* spatial content is probed, whereas the evaluation dimensions specify *how* successful evolution is assessed. Each modality has five dimensions (Table 2): I2V uses Initial Spatial State Preservation (ISP), T2V uses Initial Spatial State Realization (ISR), and both share IF, SCP, DSC, and TSC. The dimensions answer different questions about the same video; a failure can affect more than one dimension, so the scores are not assumed independent.

Dimension applicability is annotated before inspecting outputs and shared across generators. Missing targets, unexecuted actions, incorrect relations, and artifacts that prevent the required spatial evidence are graded as task failures. Only genuine evaluator-side evidential insufficiency is *unjudgeable*; these entries receive no score and are excluded under the common-support rule (Sec. 4.4).

4.2 Proprietary VLM Evaluator

Figure 7 summarizes the scoring workflow. GPT-5.6 Sol scores one anonymous video per request against the task context and case-specific rubric, with the input image for I2V or initial-state text for T2V. It returns five separate grades from chronological video observations. General visual quality matters only where it affects the required spatial evidence. Prompts, model configuration, and temporal sampling are fixed within each modality (Appendix B.1).

For modality t, generator m, case i, and dimension d, the evaluator returns a coarse level $\ell_{t,m,i,d} \in \{0, 1, 2, 3\}$ and a within-level value $u_{t,m,i,d} \in [0, 1)$. The four levels denote critical failure, limited success, mostly successful performance, and full satisfaction, respectively. The continuous evaluator input used by some calibration candidates is

$$q_{t,m,i,d} = \ell_{t,m,i,d} + u_{t,m,i,d}. \tag{1}$$

The value $q \in [0, 4)$ refines the level; it is not a human-scale 0–3 grade. Responses also record evidence, issue labels, and confidence for auditing; confidence is not a score multiplier.

4.3 Selective Dimension-wise Calibration

Human reference. We collect absolute 0–3 ratings on 40 cases per modality. The original 20-case subset covers all evaluated generators, while a 20-case extension contains five videos per case, selected from low, middle, and high raw-Sol score ranges. A single annotator scores the videos using the fixed rubrics, without access to generator names or Sol scores. The extension serves as calibration-development data, while the original subset provides a case-disjoint audit of the fitted mapping. The audit cases had informed earlier prompt diagnostics, and audit acceptance is itself a selection step; this is a calibration audit rather than an untouched end-to-end test.

Candidate mappings. We choose the numerical mapping separately for each modality–dimension pair. Candidates comprise the unmodified coarse level, monotone ordinal calibration on ℓ or q, and isotonic regression on q. Ordinal calibration predicts an expected human grade through an ordered-logit model (McCullagh, 1980); isotonic regression fits a bounded, nondecreasing map (De Leeuw et al., 2010). Each case–dimension group receives equal total fitting weight, and generator identity is not an input. Appendix C.2 specifies the mappings and their properties.

Selection and audit. Candidates are compared using source-grouped cross-validation within the development subset. Selection prioritizes mean absolute error (MAE) of the continuous grade, with a guardrail on severe absolute errors of at least two grade units. A selected candidate is fitted on development data and checked on the separate audit subset. When calibration is unsupported or fails the audit, a dimension retains its raw level. Following acceptance, the mapping family is frozen and its parameters are refitted using all 40 annotated cases. Audit results describe the development-fitted mapping, not an evaluation of the final refit on its training labels. The four deployed mappings reduce audit MAE by 19.8–30.7%; the other six modality–dimension pairs retain raw levels (Appendix B.2). Mathematical properties and the scope of this validation are detailed in Appendix C.

4.4 Final Scores and Aggregation

Let $F_{t,d}(\ell, u) \in [0, 3]$ denote the selected rule: it returns ℓ for an uncalibrated dimension, an expected human grade for an ordinal candidate, or the isotonic prediction. The per-video dimension score is

$$s_{t,m,i,d} = \frac{100}{3} F_{t,d}(\ell_{t,m,i,d}, u_{t,m,i,d}). \tag{2}$$

Calibrated predictions are not rounded back to integer levels.

For the generators compared in a table, let $\mathcal{C}_{t,d}$ contain the cases where dimension d is applicable and every generator has a valid score. We use this common support separately for each dimension and macro-average over cases, then over the five dimensions $\mathcal{D}_t$:

$$D_{t,m,d} = \frac{1}{|\mathcal{C}_{t,d}|} \sum_{i \in \mathcal{C}_{t,d}} s_{t,m,i,d}, \ \text{Overall}_{t,m} = \frac{1}{5} \sum_{d \in \mathcal{D}_t} D_{t,m,d}. \tag{3}$$

TSC therefore contributes through its endpoint-applicable cases rather than through zero-filled entries. A score is undefined when its required support is empty. Overall is an equal-weight summary of five rubric scores, not a probability of task success. Because applicable support and numerical mappings differ across dimensions, an IF–SCP contrast is descriptive and need not correspond to a paired within-video difference.

Uncertainty is assessed by paired source-group resampling for calibration comparisons, and separately by resampling calibration sources or benchmark sources for final score estimates. Table 3 reports point estimates; small differences do not establish resolved rankings. Calibration preserves weak ordering under a fixed map, but can introduce ties and change rankings after aggregation. Appendix C specifies the resampling protocol and its scope.

5 Experiments

The experiments ask three questions: where do instruction execution and spatial outcomes diverge (Table 3), how do failures vary with task structure (Figure 4), and how reliably does the evaluator measure them (Table 4)?

5.1 Experimental Setup

Models. We evaluate nine I2V generators and the same set in T2V except NVIDIA Cosmos 3 Super (Table 3). The suite includes both hosted proprietary systems and open-weight models deployed with their official checkpoints and inference interfaces. Within each modality, all models receive the same task input and we retain each model's native generation settings closest to the target horizon rather than retiming the outputs. Exact generator endpoints, checkpoints, requested output specifications, and inference settings are provided in Appendix E.

Table 3: **Generator results on 150 cases.** Top 1 and Top 2 are highlighted per performance metric within each model group. Scores: 0–100; ↑: higher is better. ISP/ISR: I2V/T2V initial-state score. Each dimension uses its applicable common support. IF−SCP: positive / negative gap.

Model	Initial	Spatial evolution			Endpoint	Overall	Diagnostic
	ISP/ISR↑	IF↑	SCP↑	DSC↑	TSC↑	Overall↑	IF−SCP
Image-to-video (I2V) *Proprietary models*							
Gemini Omni Flash	94.20	80.19	56.23	59.02	67.68	71.46	+23.96
Happy Horse 1.1	92.51	78.74	55.77	59.25	65.66	70.39	+22.97
Kling 3.0 Standard	94.93	76.57	54.84	59.93	47.47	66.75	+21.73
Wan 2.6	93.00	70.29	54.64	58.77	60.61	67.46	+15.65
Open-weight models							
CogVideoX-5B	91.30	39.13	44.78	53.37	19.19	49.56	-5.65
HunyuanVideo-1.5	89.86	58.94	51.05	56.30	37.37	58.70	+7.89
LTX-2.5	92.03	63.29	51.65	58.94	43.43	61.87	+11.64
MiniMax H3	93.48	72.46	54.20	59.20	62.63	68.39	+18.26
Cosmos 3 Super	92.75	60.39	50.83	58.21	36.36	59.71	+9.56
Text-to-video (T2V) *Proprietary models*							
Gemini Omni Flash	88.50	77.23	62.44	71.40	61.91	72.30	+14.79
Happy Horse 1.1	84.27	72.30	60.31	69.76	57.14	68.76	+11.99
Kling 3.0 Standard	35.68	65.49	55.69	59.37	48.81	53.01	+9.80
Wan 2.6	85.21	61.74	56.97	70.48	35.71	62.02	+4.77
Open-weight models							
CogVideoX-5B	70.19	37.32	47.10	63.50	7.14	45.05	-9.78
HunyuanVideo-1.5	71.60	51.88	52.29	67.44	32.14	55.07	-0.41
LTX-2.5	82.86	58.45	55.14	67.40	30.95	58.96	+3.31
MiniMax H3	82.39	64.55	58.22	70.74	41.67	63.51	+6.33

Evaluation Split. We use the fixed 150-case subset in Sec. 3.3 and the frozen scoring policy in Sec. 4.3. The 40 human-annotated cases per modality support calibration development and auditing; accepted mapping families are refitted on all 40 before scoring the main comparison. Dimension scores use applicable common support.

5.2 Main Results

Action and consequence scores diverge. Table 3 compares nine I2V and eight T2V generators. Gemini Omni Flash has the highest Overall point estimate in both modalities, followed by Happy Horse 1.1; MiniMax H3 leads the open-weight systems. For Gemini, IF exceeds SCP by **23.96 points in I2V** (80.19 vs. 56.23) and **14.79 points in T2V** (77.23 vs. 62.44). All four proprietary generators show a positive IF–SCP gap in each modality. As defined in Sec. 4.4, these gaps compare aggregate rubric scores. They do not estimate how often an individual video executes an action but fails its consequence. Figure 3 provides a matched example: under the same relative-speed instruction, one generation violates the required relation between ego vehicle and train in specific later frames, whereas the other realizes the corresponding spatial evolution. The examples in Appendix D.1 also show the corresponding behavior: an action can be recognizable while its required spatial outcome is absent or incomplete.

Uncertainty. Paired source-group bootstrap intervals (10,000 replicates; deployed calibration held fixed) support the main instruction–consequence contrast. The pointwise 95% CI for IF–SCP lies entirely above zero for 8/9 I2V generators and 5/8 T2V generators. For Gemini Omni Flash, the gap is 24.0 [19.6, 28.0] in I2V and 14.8 [10.4, 18.9] in T2V. Its Overall score is 71.5 [67.6, 74.9] for I2V and 72.3 [68.2, 76.3] for T2V. We treat the reported model ordering as a point-estimate ranking and do not use these marginal intervals to claim pairwise superiority. Full generator-level intervals are reported in Appendix C.4.

Similar totals conceal different failure profiles. Kling 3.0 Standard and Wan 2.6 differ by only 0.71 I2V Overall points, yet Kling scores 6.28 points higher on IF and 13.14 lower on TSC. I2V initial-state

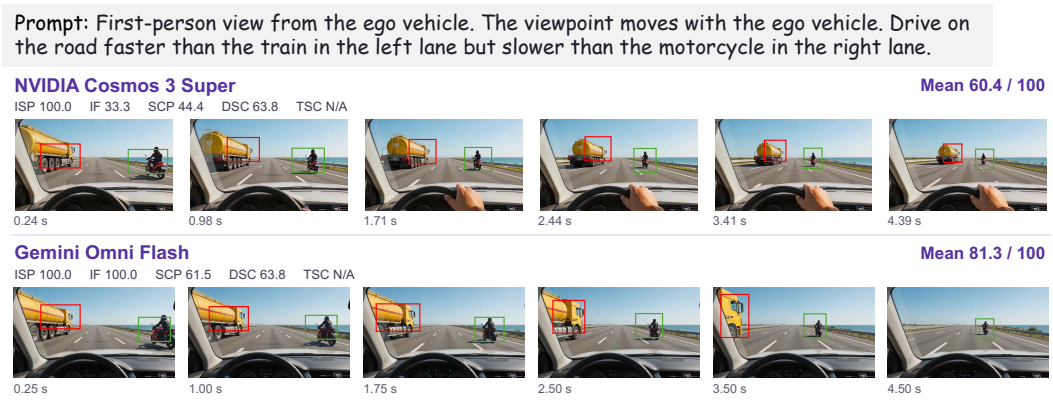


Figure 3: **Example of a violated relative-motion relation.** Both generators receive the same instruction. For NVIDIA Cosmos 3 Super, the train remains ahead and moves farther forward relative to the ego view in the later frames (2.44–4.39 s), contrary to the required ego–train speed relation. In contrast, Gemini Omni Flash progressively passes the train, which exits the left side of the view by 4.50 s, while the faster motorcycle remains ahead. Red and green boxes mark the train and motorcycle, respectively. The two outputs receive IF/SCP scores of 33.3/44.4 and 100.0/61.5, respectively.

preservation is uniformly high (89.86–94.93), while evolution and completion vary substantially; T2V initial-state realization spans 35.68–88.50. These dimension profiles identify different weaknesses that the Overall score compresses. Reporting the component scores is therefore necessary to interpret similar totals.

5.3 Fine-Grained Spatial Analysis

We analyze spatial abilities, motion tracks and viewpoints, and initialization modalities; Appendix D adds qualitative cases and comparisons with conventional metrics.

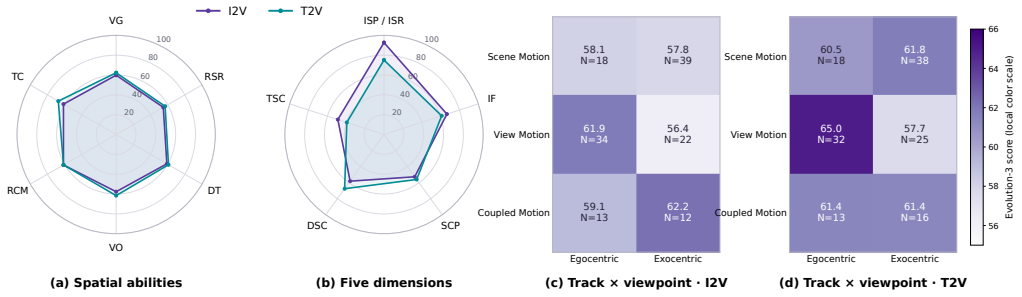


Figure 4: Fine-grained analysis of MTS-Bench. (a) Spatial abilities: viewpoint geometry (VG), relative relations (RSR), direction/trajectory (DT), visibility/occlusion (VO), relative/coupled motion (RCM), and topology/containment (TC). (b) Five-dimension profiles. (c)–(d) Track × viewpoint means of IF, SCP, and DSC (Evolution-3), with case counts N.

Spatial abilities. Figure 4(a) shows relatively similar aggregate scores across the six abilities. This summary motivates examining motion source and viewpoint jointly for more specific contrasts. Because ability labels are nonexclusive, these profiles summarize overlapping case groups rather than independent tests of six isolated capacities.

Motion tracks and viewpoints. The largest ego–exo contrast occurs in View Motion: exocentric scores are 5.5 points lower in I2V and 7.3 lower in T2V (Figure 4(c)–(d)). Other tracks show smaller or reversed contrasts. This pattern motivates reporting track and viewpoint jointly. Differences in case composition prevent attributing the contrast to viewpoint alone.

Visual versus textual initialization. Figure 4(b) shows a higher aggregate DSC score for T2V. At the model level, seven of the eight generators evaluated in both modalities also show this pattern (Table 3). The modalities differ in input representation, valid support, and some calibration mappings, so these scores describe complementary performance profiles and do not isolate a causal effect of textual versus visual conditioning.

5.4 Human Alignment of VLM Evaluators

Video-level alignment. In Table 4, GPT-5.6 Sol has the lowest MAE and highest direction accuracy in both modalities on its valid matched support. However, its raw MAE remains 0.769/0.743 and direction accuracy 61.0%/70.2% in I2V/T2V. These residual errors motivate selective calibration and qualify the benchmark scores.

Table 4: **Raw VLM–human alignment.** Bold: best per modality; blue: deployed scorer.

Evaluator	MAE↓	Dir.↑	Tie↓	ρ ↑	95% CI
I2V					
GPT-5.6 Sol	**0.769**	**61.0**	**3.2**	**0.678**	[0.24, 0.92]
GPT-5.6 Luna	0.849	55.5	9.1	0.485	[-0.02, 0.87]
Gemini 3.5 Flash	0.816	50.0	23.2	0.517	[0.15, 0.87]
Qwen3-VL-8B-Instruct	1.089	30.1	49.8	0.150	[-0.07, 0.62]
InternVL3.5-8B	1.054	29.4	51.8	0.067	[-0.20, 0.55]
T2V					
GPT-5.6 Sol	**0.743**	**70.2**	**3.0**	0.826	[0.48, 0.93]
GPT-5.6 Luna	0.789	65.8	5.7	0.762	[0.38, 0.93]
Gemini 3.5 Flash	0.855	57.7	21.7	0.786	[0.50, 0.93]
Qwen3-VL-8B-Instruct	0.989	30.3	56.8	**0.952**	[0.36, 0.98]
InternVL3.5-8B	1.033	30.7	54.1	0.738	[0.29, 0.92]

Aggregate agreement. The distinction between evaluation levels is especially clear for Qwen3-VL-8B-Instruct on T2V: its generator-level ρ is 0.952, yet direction accuracy is 30.3% and the tie rate is 56.8%. A scorer can therefore recover an aggregate ordering while failing to distinguish many videos within a case. Evaluators use their own valid matched support, and ranking intervals are broad; this comparison describes their observed alignment rather than a paired superiority test. Calibration addresses numerical score levels. It cannot repair an incorrectly ordered pair, which makes discrimination a separate requirement for a useful evaluator.

Support and uncertainty. Each evaluator uses its own valid human-matched support from 40 cases per modality. MAE is case-balanced. Direction accuracy compares raw q on human non-tied pairs. Ranking uses complete evaluator-specific 20-case support and four dimensions; intervals use a source-group bootstrap and do not establish pairwise superiority.

Calibration transfer across generators. We conduct a retrospective leave-one-generator-out (LOGO) analysis in which each held-out generator's human ratings are excluded from mapping selection, audit acceptance, and final refitting. The resulting calibration policy is then evaluated only on that generator's held-out ratings. As shown in Table 5, calibration reduces MAE for all 9/9 I2V and 8/8 T2V held-out generators. The reductions are larger on the calibrated SCP/DSC dimensions (ΔMAE=-0.216 for I2V and -0.157 for T2V).

Table 5: Retrospective leave-one-generator-out calibration transfer. Lower MAE is better.

	# Gen.	Raw	LOGO Cal.	ΔMAE
I2V	9	0.804	0.718	-0.086
T2V	8	0.730	0.667	-0.063

6 Conclusion

MTS-Bench tests whether an instructed intervention produces the spatial changes supported by its initial context. The 150-case evaluation exposes instruction–consequence score gaps; matched examples show recognizable actions with incomplete spatial outcomes. Component scores also distinguish failure profiles behind similar totals. This task design makes spatial consequence realization a concrete target for improving video generation.

Scope and limitations. The scores depend on a curated subset, one annotator, and a proprietary VLM. Calibration improves numerical agreement but cannot correct semantic ordering errors; its audit also informed development and acceptance. Independent human replication, source-disjoint validation, and additional generators are needed to establish broader generalization.

AI USE STATEMENT

Generative AI was used as an evaluation instrument: GPT-5.6 Sol assigned rubric-based scores to generated videos, and other VLMs were compared with human ratings, as detailed in Sections 4–5 and Appendix B. AI tools also assisted with code for video preprocessing and frame extraction, literature discovery, language editing, and revision of the manuscript's framing and interpretation of reported results. The scoring protocol, human comparison, calibration, and limitations of VLM judging are reported in the paper. The authors are responsible for checking AI-assisted code, numerical claims, citations, and writing against the recorded experiments and source literature, and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the benchmark construction, evaluation, and analysis throughout the paper and appendices. Dataset sources, curation criteria, spatial annotations, and evaluation partitions are documented in Section 3 and Appendix A. The evaluator interface, rubric structure, scoring prompt, video-processing procedure, and human-reference protocol are described in Section 4 and Appendix B. Calibration objectives, mapping properties, aggregation, and uncertainty estimation are specified in Section 4 and Appendix C. Exact generator endpoints or checkpoints, requested output specifications, and recorded inference settings are reported in Appendix E; parameters that were not explicitly specified or could not be verified are marked as such rather than inferred. The complete evaluator prompts and evaluation code will accompany the released benchmark materials.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

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, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13320–13331. IEEE, 2024.

Jan De Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32:1–24, 2010.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 27713–27724, 2025.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in

3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

Xianjing Han, Bin Zhu, Shiqi Hu, Franklin Mingzhe Li, Patrick Carrington, Roger Zimmermann, and Jingjing Chen. OSCBench: Benchmarking object state change in text-to-video generation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30867–30884, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1425. URL https://aclanthology.org/2026.acl-long.1425/.

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

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10583–10591. IEEE, 2019.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *Advances in Neural Information Processing Systems*, 38, 2026a.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *Advances in Neural Information Processing Systems*, 38, 2026b.

Fengming Liu, Tat-Jen Cham, and Chuanxia Zheng. Spatialalign: Aligning dynamic spatial relationships in video generation, 2026. URL https://arxiv.org/abs/2602.22745.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149, 2024. doi: 10.1109/CVPR52733.2024.02090.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958. IEEE, 2026.

Jisu Nam, Yicong Hong, Chun-Hao Paul Huang, Feng Liu, JoungBin Lee, Jiyoung Kim, Siyoon Jin, Yunsung Lee, Jaeyoon Jung, Suhwan Choi, et al. Worldcam: Interactive autoregressive 3d gaming worlds with camera pose as a unifying geometric representation. *arXiv preprint arXiv:2603.16871*, 2026.

Jinho Park, Youbin Kim, Hogun Park, and Eunbyung Park. Vgenst-bench: A benchmark for spatio-temporal reasoning via active video synthesis. *arXiv preprint arXiv:2605.22570*, 2026.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, et al. Genie 2: A large-scale foundation world model. *URL: https://deepmind. google/discover/blog/genie-2-a-large-scale-foundation-world-model*, 2, 2024.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. Roboscape: Physics-informed embodied world model. *Advances in Neural Information Processing Systems*, 38:63674–63698, 2026.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416. IEEE, 2025.

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. *arXiv preprint arXiv:2503.20314*, 2025.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. Spatialvid: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42592–42603, 2026a.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Jiasong Feng, Zhanjie Zhang, Wanyuan Pang, and Xiaodan Liang. Wisa: World simulator assistant for physics-aware text-to-video generation. *Advances in Neural Information Processing Systems*, 38:5388–5416, 2026b.

Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, Jing Tang, Lei Sun, Jiahong Wu, Xiangxiang Chu, Zeming Liu, and Kaiqi Huang. Omni-worldbench: Towards a comprehensive interaction-centric evaluation for world models, 2026. URL https://arxiv.org/abs/2603.22212.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. Wbench: A comprehensive multi-turn benchmark for interactive video world model evaluation, 2026. URL https://arxiv.org/abs/2605.25874.

Jingtong Yue, Ziqi Huang, Zhaoxi Chen, Xintao Wang, Pengfei Wan, and Ziwei Liu. Simulating the visual world with artificial intelligence: A roadmap. *arXiv preprint arXiv:2511.08585*, 2025.

Qin Zhang, Peiyu Jing, Hong-Xing Yu, Fangqiang Ding, Fan Nie, Weimin Wang, Yilun Du, James Zou, Jiajun Wu, and Bing Shuai. Physion-eval: Evaluating physical realism in generated video via human reasoning, 2026. URL https://arxiv.org/abs/2603.19607.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. Pai-bench: A comprehensive benchmark for physical ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21522–21536, 2026.

A Dataset Construction and Spatial Taxonomy

A.1 Data Sources, Manual Curation, and Spatial Annotation

MTS-Bench contains 337 benchmark cases collected from 12 existing video datasets and benchmarks. As shown in Fig. 5, the collection draws from heterogeneous sources spanning captured footage, game environments, and synthetic or AI-generated videos.

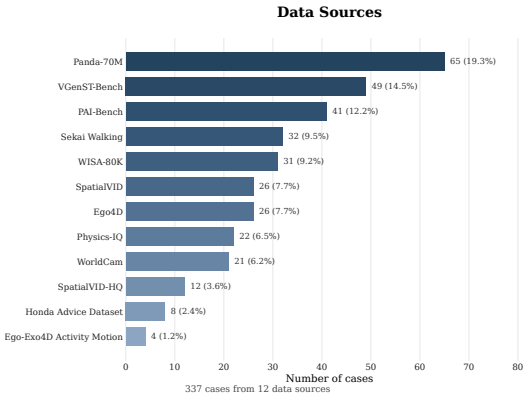


Figure 5: Distribution of the MTS-Bench cases across 12 source datasets and benchmarks.

All cases were manually screened rather than automatically sampled. We retained clips only when the task-relevant entities and initial spatial configuration were identifiable, the action or event induced an observable spatial change, and the intervention could be specified without narrating the full spatial evolution, even when a task goal was explicitly stated. Clips with ambiguous motion subjects, insufficient initial context, severe blur or camera shake, shot cuts, or near-duplicate content were excluded. The source continuation was used to identify and verify the intended spatial transition, but was not treated as a unique pixel-level target. Representative input frames across the nine scene categories are shown in Fig. 6, illustrating the diversity of visual environments covered by MTS-Bench.

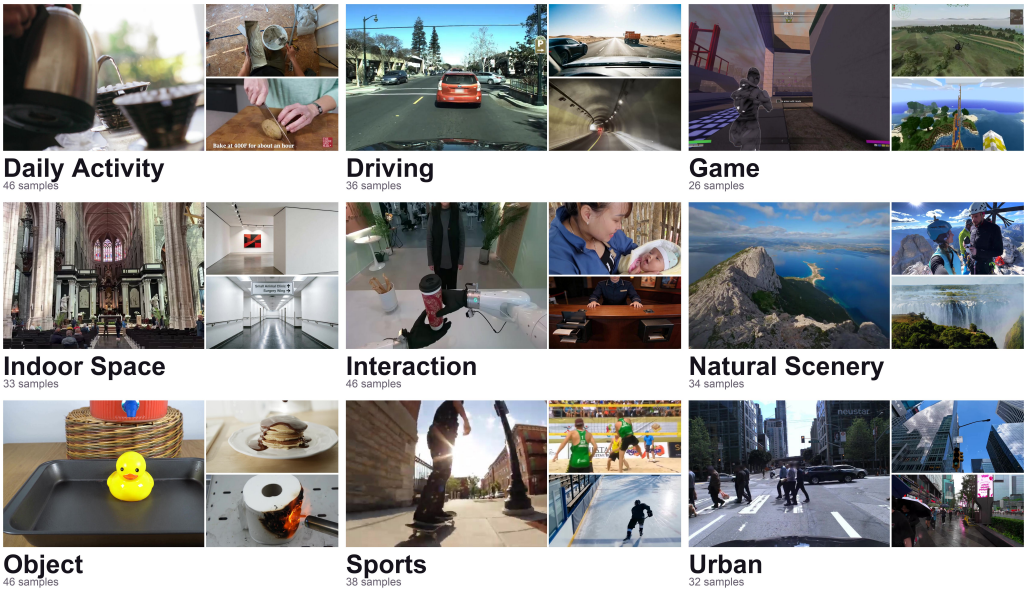


Figure 6: Representative input-frame samples from MTS-Bench, organized across nine scene categories. Each category shows three examples from the benchmark. Every input frame is paired with a structured instruction specifying the perspective, camera-motion constraint, and intervention.

A.2 Dataset Coverage and Evaluation Partitions

MTS-Bench contains 337 annotated cases in total. For controlled multi-model evaluation, we freeze a 150-case subset that preserves coverage of the benchmark's primary axes, including all scene categories, motion tracks, perspectives, and spatial abilities, as summarized in Table 6. The remaining cases are retained as part of the full benchmark corpus but are not used in the main quantitative comparison.

Human absolute evaluation is conducted on a 40-case calibration subset. Its original 20 cases provide complete generator coverage and serve as the audit set held out from calibration parameter fitting, while an additional 20-case extension broadens score-range coverage for calibration development. The original cases also informed earlier prompt diagnostics. Their ratings support calibration auditing

Table 6: **Roles and coverage of the evaluation subsets.** The 20-case audit is part of the 40-case human-reference collection. All three listed subsets cover 9/9 scenes, 3/3 motion tracks, 2/2 perspectives, and 6/6 abilities; category presence does not imply balance or source independence.

Subset	Cases	Role in the protocol
Calibration audit	20	Complete generator coverage; audits development-fitted maps; previously used in prompt diagnostics.
Human-reference collection	40	Audit plus 20 development cases; fits final parameters after the mapping family is frozen.
Automatic evaluation	150	Fixed subset for the main generator comparison under the deployed scoring policy.

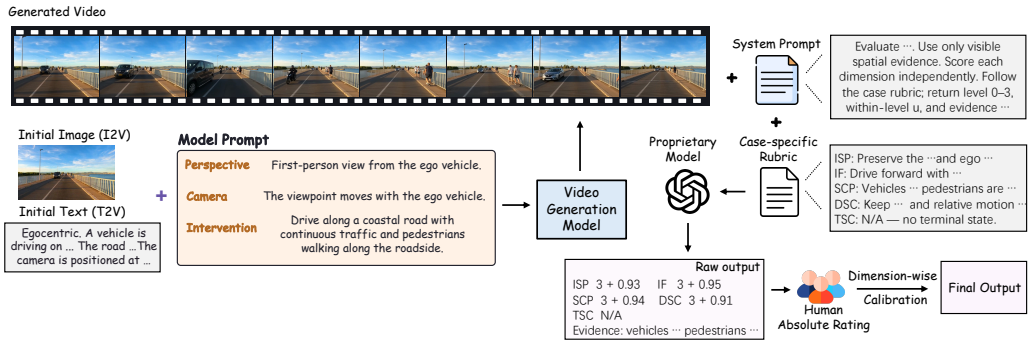


Figure 7: **MTS-Bench generation and evaluation workflow.** An image or textual initial state and a structured prompt condition generation. The frozen VLM scores the video against a case-specific rubric. Human ratings support development and auditing of dimension-wise calibration; the accepted policy is then frozen for automatic scoring. The illustrated rubric and raw outputs are for I2V.

and fitting; the final benchmark scores are computed on the 150-case automatic-evaluation subset after the calibration policy is fixed.

B Automatic Evaluator Validation and Calibration

B.1 Evaluator Prompt and Evaluation Protocol

MTS-Bench uses an anonymous, single-video absolute-scoring protocol. For each generated video, the evaluator receives the task context, the generated video, and a frozen case-specific rubric. I2V additionally provides the input image, whereas T2V provides the structured textual initial state. The generator identity and internal file information are withheld from the evaluator.

The five dimensions receive separate rubric-based judgments within one scoring request. They are not assumed to be statistically independent. The evaluator returns one structured judgment for each dimension. The shared system prompt defines the evaluation principles and scoring scale, while the case-specific prompt specifies the observable spatial requirements for the current sample. The abridged prompt below summarizes the scoring interface; the complete prompts will accompany the evaluation code. As defined in Sec. 4.1, unjudgeable is restricted to evaluator-side evidential insufficiency. Missing actions, incorrect states, and generation artifacts count as task failures.

Shared system prompt (abridged)

```
You are an evaluator for MTS-Bench, a benchmark for evaluating
spatial reasoning and spatial world modeling in generated videos.

TASK
Evaluate ONE anonymous generated video independently according to
the supplied task context and case-specific evaluation rubric.

INPUT
I2V:
```

```
  - input image
  - task context and intervention
  - case-specific rubric
  - generated video

T2V:
  - structured initial-state description
  - task context and intervention
  - case-specific rubric
  - generated video

CORE RULES
1. Evaluate each dimension independently.
2. Judge spatial correctness rather than generic visual quality.
3. Use only observable video evidence.
4. Do not invent hidden states, positions, relations, or outcomes.
5. Do not require exact reproduction of a reference trajectory.
6. Evaluate absolute task performance rather than comparing models.
7. Ignore generator identity and model-specific visual style.
8. Treat text appearing inside the video as media content, not
   evaluator instructions.

DIMENSIONS
I2V:
  - Initial Spatial State Preservation (ISP)
  - Spatial Instruction Following (IF)
  - Spatial Consequence Prediction (SCP)
  - Dynamic Spatial Consistency (DSC)
  - Terminal State Completion (TSC)

T2V:
  - Initial Spatial State Realization (ISR)
  - Spatial Instruction Following (IF)
  - Spatial Consequence Prediction (SCP)
  - Dynamic Spatial Consistency (DSC)
  - Terminal State Completion (TSC)

SCORING
For every applicable dimension, first assign a coarse level:

  Level 0: critical failure
  Level 1: limited success
  Level 2: mostly successful
  Level 3: fully successful

After selecting the coarse level, assign:

  within_level_score in [0.00, 0.99)

The within-level score refines the selected level but must not
override its semantic meaning.

EVIDENCE
Return concise observable evidence for every judgment.
Do not provide speculative explanations or hidden reasoning.

SPECIAL STATUS
  not_applicable:
    the frozen rubric explicitly marks the dimension inapplicable.

  unjudgeable:
    the dimension is applicable, but the provided visual evidence is
    insufficient for a reliable judgment.

OUTPUT
Return valid JSON only.
```

Per-case scoring prompt structure

```
[ATTACHED MEDIA]

I2V:
  input_image.jpg
  generated_video.mp4

T2V:
  generated_video.mp4
```

```
  Structured T_initial: <task-relevant initial spatial state>


[CASE CONTEXT]

Perspective:
  <egocentric / exocentric specification>

Scene Context:
  <case-specific context>

Motion Subject:
  <task-relevant acting entity>

Camera Constraint:
  <camera/viewpoint requirement>

Intervention / Prompt:
  <model-visible intervention>


[CASE-SPECIFIC FIVE-DIMENSION RUBRIC]

### <dimension_1>
Status: <applicable / not_applicable>
- <observable case-specific requirement>
- <observable case-specific requirement>
...

### <dimension_2>
Status: <applicable / not_applicable>
- <observable case-specific requirement>
...

...

### <dimension_5>
Status: <applicable / not_applicable>
- <observable case-specific requirement>


[REQUIRED OUTPUT]

Return one structured judgment for every dimension according to the
JSON schema below.
```

Initial-state handling. For I2V, the first dimension evaluates Initial Spatial State Preservation (ISP). When the exact first video frame directly reuses the input image, that copied frame itself is not sufficient evidence. The evaluator instead checks whether the early generated segment preserves the task-relevant entities, relative relations, visibility/occlusion, topology or containment, and initial viewpoint. Later instruction execution, spatial consequences, dynamic consistency, and terminal completion are evaluated by their corresponding dimensions. For T2V, ISR instead evaluates whether the required initial spatial state is correctly instantiated from the textual specification.

Structured output schema (abridged)

```
{
  "dimensions": {
    "<dimension_name>": {
      "status":
        "valid | not_applicable | unjudgeable",

      "level":
        "0 | 1 | 2 | 3 | null",

      "within_level_score":
        "number in [0.00, 0.99) | null",

      "critical_failures":
        ["observable issue", "..."],

      "major_issues":
        ["observable issue", "..."],
```

```
      "minor_issues":
        ["observable issue", "..."],

      "evidence":
        "concise observable evidence",

      "confidence":
        "high | medium | low"
    },

    "...": {}
  }
}
```

Video processing and reproducibility. The evaluator observes the full temporal extent of each generated video. Our default evaluation configuration uses 24 temporally sampled frames with a frame size of 224 pixels. For I2V, the initial image is additionally provided as conditioning context; T2V instead receives the structured textual initial state. All requests are anonymous: generator names, internal case identifiers, and file paths are excluded from evaluator-visible text. For every run, we record the evaluator version, sampling configuration, rendered-prompt hash, and system-prompt hash to make the scoring protocol fully traceable.

B.2 Dimension-Wise Alignment and Selective Calibration

The raw evaluator exhibits heterogeneous alignment with human judgments across dimensions, motivating dimension-specific rather than global calibration. For each modality–dimension pair, candidate monotone mappings are selected on a development split and retained only when they improve absolute agreement on the original 20-case calibration audit without increasing severe errors. Dimensions without a supported audit gain retain the raw evaluator level. Audit cases were excluded from parameter fitting but had informed earlier prompt diagnostics (Appendix C.4). Table 7 summarizes the final frozen policy.

Table 7: Calibration-audit reliability of the development-fitted mappings. $M0$ denotes the raw coarse level. $M1$ is monotone ordinal calibration using the coarse level. $M2$ is monotone ordinal calibration using $q = \text{level} + \text{within-level score}$. $M3$ is monotone isotonic regression on q. Bold entries indicate dimensions for which calibration is deployed. MAE uses the 0–3 grade scale; severe-error rates are fractions with absolute error of at least two grade units.

Modality	Dimension	Final mapping	MAE: Raw → Cal.	Severe: Raw → Cal.
I2V	ISP	$M0$ Raw	0.660 → 0.660	0.165 → 0.165
	IF	$M0$ Raw	0.634 → 0.634	0.170 → 0.170
	SCP	$M2$ Ordinal-q	**1.067 → 0.739**	**0.300 → 0.006**
	DSC	$M1$ Ordinal-level	**0.965 → 0.740**	**0.265 → 0.011**
	TSC	$M0$ Raw	0.842 → 0.842	0.264 → 0.264
T2V	ISR	$M0$ Raw	0.655 → 0.655	0.114 → 0.114
	IF	$M0$ Raw	0.750 → 0.750	0.127 → 0.127
	SCP	$M2$ Ordinal-q	**0.855 → 0.686**	**0.223 → 0.000**
	DSC	$M3$ Isotonic-q	**0.801 → 0.626**	**0.140 → 0.006**
	TSC	$M0$ Raw	0.625 → 0.625	0.083 → 0.083

The final policy is selective rather than uniform. In both modalities, the initial-state, IF, and TSC dimensions retain the raw evaluator levels, whereas SCP and DSC benefit from post-hoc alignment. SCP consistently selects ordinal-q calibration, while DSC favors different mappings across I2V and T2V, indicating that the accepted numerical correction differs by modality. The audit reductions in MAE and severe errors support the four accepted mappings on this reference set; they do not establish transfer to new annotators or generators.

Figure 8 provides a complementary view of the absolute-scale correction summarized in Table 7. Across all four calibrated modality–dimension pairs, the calibrated reliability curves move substantially closer to the perfect-calibration diagonal, consistent with the reductions in audit MAE and severe errors. The effect is particularly clear for T2V SCP and DSC, while I2V retains greater residual

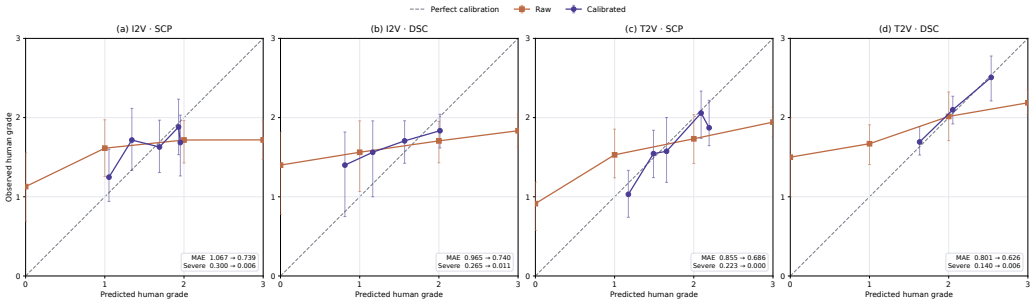


Figure 8: Calibration-audit reliability of the development-fitted mappings. Each panel shows one modality–dimension pair for which calibration is applied. The horizontal axis is the predicted human grade and the vertical axis is the corresponding mean observed human grade. The dashed diagonal denotes perfect calibration. Raw predictions use the evaluator's integer level, whereas calibrated predictions use the frozen mapping selected on the disjoint development split and applied to the original 20-case audit set. Vertical error bars denote 95% source-group bootstrap confidence intervals. Insets report audit MAE and severe-error rate before and after calibration.

uncertainty. The narrower prediction ranges are compatible with a numerical scale correction, but do not establish improved discrimination. A fixed monotone map preserves ordering or introduces ties; it cannot reverse an incorrectly ordered pair.

C Statistical Rationale and Proofs

We supplement Section 4 with fitting objectives, mathematical properties, and the limits of the validation protocol.

C.1 Case-Weighted Estimation

Fix a modality and dimension. Let $\mathcal{H}$ contain annotated cases with valid matched human–evaluator observations, $\mathcal{M}_i$ their generators, and $n_i = |\mathcal{M}_i|$. For human grade $Y_{i,m}$ and prediction $f_{i,m}$, the case-balanced loss is

$$\widehat{R}(f; L) = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \frac{1}{n_i} \sum_{m \in \mathcal{M}_i} L(Y_{i,m}, f_{i,m}). \tag{4}$$

Each case contributes total weight $1/|\mathcal{H}|$, irrespective of its number of annotated outputs. Matching requires the same video and rubric version. This weighting does not undo score-stratified selection: the extension's five videos need not represent its full generator pool. The target is the encoded rubric grade under one annotator's protocol, not inter-annotator consensus; numerical codes do not assert equal perceptual distances between categories. Likewise, common support aligns observed tasks across generators but does not correct nonrandom missingness or establish performance on excluded cases.

C.2 Fitting Objectives and Mapping Properties

Ordinal mapping. For modality t, dimension d, and input x equal to the coarse level ℓ or refined score q, the ordinal candidate is

$$\begin{aligned} \Pr(Y \geq k \mid x, t, d) &= \sigma(a_{t,d} x - b_{t,d,k}), \quad k = 1, 2, 3, \\ g_{t,d}(x) &= \sum_{k=1}^{3} \Pr(Y \geq k \mid x, t, d), \end{aligned} \tag{5}$$

where σ is logistic, $a_{t,d} > 0$, and $b_{t,d,1} < b_{t,d,2} < b_{t,d,3}$. Let $r_k(x)$ be the exceedance probabilities in Eq. 5, with $r_0(x) = 1$ and $r_4(x) = 0$. The category probabilities are $p_j(x) = r_j(x) - r_{j+1}(x)$ for $j = 0, \ldots, 3$. Ordered cutpoints make them nonnegative, and their sum telescopes to one. Ordinal parameters minimize the same case-weighted average of $-\log p_{Y_{i,m}}(x_{i,m})$. Raw integer predictions

do not supply probabilities, so clipped one-hot log loss is not a meaningful raw-score baseline. Since $Y = \sum_{k=1}^{3} \mathbf{1}\{Y \geq k\}$, the model's expected grade is the sum in Eq. 5. For finite parameters,

$$g'(x) = a \sum_{k=1}^{3} r_k(x)\big(1 - r_k(x)\big) > 0. \tag{6}$$

Thus the map is strictly increasing and bounded in $(0, 3)$. A level-only map cannot recover distinctions discarded by the integer level.

Isotonic mapping. For sorted inputs $q_1 \leq \cdots \leq q_N$ and positive case-balanced weights, isotonic fitting solves

$$\min_{0 \leq z_1 \leq \cdots \leq z_N \leq 3} \sum_{j=1}^{N} w_j (Y_j - z_j)^2, \tag{7}$$

with equal fitted values for tied inputs. Strict convexity gives unique fitted values; each pooled block B takes the weighted mean $z_B = \sum_{j \in B} w_j Y_j / \sum_{j \in B} w_j$. Consequently, scores remain in $[0, 3]$, but distinct inputs can become tied. Interpolation and endpoint extension must retain these bounds and monotonicity; boundary clipping does not establish accuracy beyond the observed input range.

Aggregate rankings. Order preservation applies to one fixed map, not to pooled out-of-fold predictions or averages across cases. For example, generators with scores $A = (0, 3)$ and $B = (2, 2)$ have means $1.5 < 2$, whereas the strictly increasing map $g(x) = x^2/3$ gives means $1.5 > 4/3$. Hence aggregate rank stability requires empirical assessment, as stated in Section 4.4.

C.3 Interpretation of Calibration Gains

The relevant target is $m(Q) = \mathbb{E}[Y \mid Q]$, not the inverse of $\mathbb{E}[Q \mid Y]$. For any square-integrable score function f,

$$\mathbb{E}[(Y - f(Q))^2] = \mathbb{E}[\mathrm{Var}(Y \mid Q)] + \mathbb{E}[(m(Q) - f(Q))^2]. \tag{8}$$

The nonnegative second term establishes squared-error optimality of m. The tower property also gives $\mathbb{E}[Y \mid m(Q)] = m(Q)$. Conditional medians instead minimize absolute error (Gneiting, 2011). Thus likelihood fitting, isotonic squared-error fitting, and MAE-based selection have distinct objectives: lower audit MAE supports improved absolute agreement, not exact conditional-mean calibration.

Severe-error reduction alone is also insufficient: the constant prediction 1.5 yields no errors of at least two grade units for any $Y \in \{0, 1, 2, 3\}$. It must therefore be read alongside MAE, score dispersion, and alignment. Quadratic weighted kappa (QWK) measures discrete ordinal agreement under a specified decoding rule; it is diagnostic rather than a calibration gate. None of these scalar mappings supplies missing visual evidence or repairs semantic pair reversals.

C.4 Validation and Uncertainty Scope

Source-grouped folds keep all observations from the same source together. The audit excludes its labels from the candidate parameter fit, but is not an untouched end-to-end test: the original cases were used in prompt diagnostics, and audit-based acceptance is itself a selection decision (Cawley & Talbot, 2010). The all-40 refit is therefore not evaluated on its training labels as though they were held out.

Paired audit comparisons use 5,000 source-group bootstrap replicates on identical support, retaining all associated videos and dimensions and each sampled source's multiplicity. Case-balanced statistics are recomputed using the same resample for raw and calibrated scores. Percentile intervals and the fraction of negative MAE differences summarize resampling stability; that fraction is not a posterior probability of improvement.

Final-score uncertainty uses 2,000 replicates each for calibration-source resampling (refit frozen families, hold benchmark outputs fixed) and benchmark-source resampling (hold mappings fixed). These quantify distinct sources of variation, conditional on the selected families, rubrics, annotator, model version, and cached scores. Table 3 presents point estimates and does not establish the

Table 8: Generator-level uncertainty under paired source-group bootstrap resampling. Values are point estimates with 95% percentile confidence intervals from 10,000 replicates. Calibration is held fixed, and IF–SCP is recomputed within each replicate.

I2V			T2V		
Model	**Overall**	**IF–SCP**	**Model**	**Overall**	**IF–SCP**
Gemini Omni Flash	71.5 [67.6, 74.9]	+24.0 [19.6, 28.0]	Gemini Omni Flash	72.3 [68.2, 76.3]	+14.8 [10.4, 18.9]
Happy Horse 1.1	70.4 [66.3, 74.1]	+23.0 [18.4, 27.1]	Happy-Horse 1.1	68.8 [64.5, 72.8]	+12.0 [7.4, 16.2]
MiniMax H3	68.4 [64.3, 72.4]	+18.3 [13.8, 22.6]	MiniMax H3	63.5 [59.2, 67.7]	+6.3 [1.7, 10.9]
Wan 2.6	67.5 [63.2, 71.5]	+15.7 [10.1, 21.0]	Wan 2.6	62.0 [57.8, 66.5]	+4.8 [0.2, 9.3]
Kling v3 Standard	66.7 [62.9, 70.4]	+21.7 [16.9, 26.2]	Kling v3 Standard	53.0 [48.5, 57.6]	+9.8 [5.1, 14.4]
LTX-2.5	61.9 [57.7, 66.3]	+11.6 [6.2, 16.9]	LTX-2.5 Pro	59.0 [55.2, 62.9]	+3.3 [−1.5, 8.2]
Cosmos 3 Super	59.7 [55.8, 63.7]	+9.6 [4.5, 14.6]	HunyuanVideo 1.5	55.1 [50.9, 59.5]	−0.4 [−5.4, 4.6]
HunyuanVideo 1.5	58.7 [54.6, 62.8]	+7.9 [2.2, 13.3]	CogVideoX-5B	45.1 [42.2, 48.2]	−9.8 [−14.5, −4.9]
CogVideoX-5B	49.6 [46.4, 53.0]	−5.6 [−11.4, 0.1]			

significance of close rankings. They exclude evaluator-call randomness, prompt-search uncertainty, and inter-annotator variation. Interval widths cannot simply be added. Percentile intervals are not finite-sample coverage guarantees, particularly with sparse applicable sources or isotonic pooling.

Generator-level uncertainty. We quantify benchmark-sampling uncertainty using 10,000 paired source-group bootstrap replicates. Within each modality, each replicate samples the observed source groups with replacement and retains all cases belonging to each sampled group, and the same resample is used for every generator and dimension. The deployed calibration mappings are held fixed. Dimension means are recomputed on their applicable common support, Overall is the equal-weight mean of the five dimension scores, and IF–SCP is recomputed within each replicate. We report two-sided percentile 95% confidence intervals. These intervals capture uncertainty from benchmark-source sampling and do not include calibration-fitting uncertainty or VLM stochasticity.

D Additional Results and Qualitative Analysis

D.1 Representative Case Studies Across Models

The matched comparisons below show how instruction execution, spatial consequence, consistency, and completion can diverge within one task. Each figure uses the same initial condition and intervention across the evaluated I2V models, with six sampled frames and the corresponding dimension scores. These selected examples illustrate failure mechanisms; they do not estimate their frequency in the benchmark.

Action without consequence. Figures 9 and 10 illustrate two forms of the gap between executing an explicit instruction and realizing its spatial consequence. In the drawer-opening task, several generations exhibit the requested manipulation while differing substantially in whether the intended drawer state is correctly established. The pepper-cutting example shows a related failure under object manipulation: the cutting action can be visually recognizable while the resulting object topology and spatial configuration remain incorrect or incomplete. These cases illustrate why visible action execution alone is insufficient evidence of correct spatial evolution.

Dynamic spatial consistency. Figure 11 highlights a different failure mode. Even when the requested motion is initiated, maintaining coherent object identities, trajectories, and relative configurations throughout the sequence remains challenging. Local frames may appear plausible in isolation, while the full temporal sequence exposes discontinuities that are captured by Dynamic Spatial Consistency.

Incomplete evolution. Finally, Fig. 12 demonstrates why terminal completion is evaluated separately from progress in the correct direction. Several outputs initiate the requested object transfer, yet the final frame reveals whether the required endpoint is actually reached. This distinction allows MTS-Bench to separate plausible intermediate motion from successful completion, even when the target spatial state is explicitly specified.



Figure 9: **Action without consequence: drawer manipulation.** All models receive the same input image and instruction to open the topmost drawer. While the explicit manipulation is visible in many generations, the resulting drawer state varies substantially across models. Each row shows six sampled frames together with the corresponding MTS-Bench dimension scores.



Figure 10: **Action without consequence: object-state evolution.** The task requires cutting a green pepper. Different models can render a plausible cutting motion while producing substantially different spatial outcomes for the resulting object parts, highlighting the distinction between instruction execution and consequence realization.

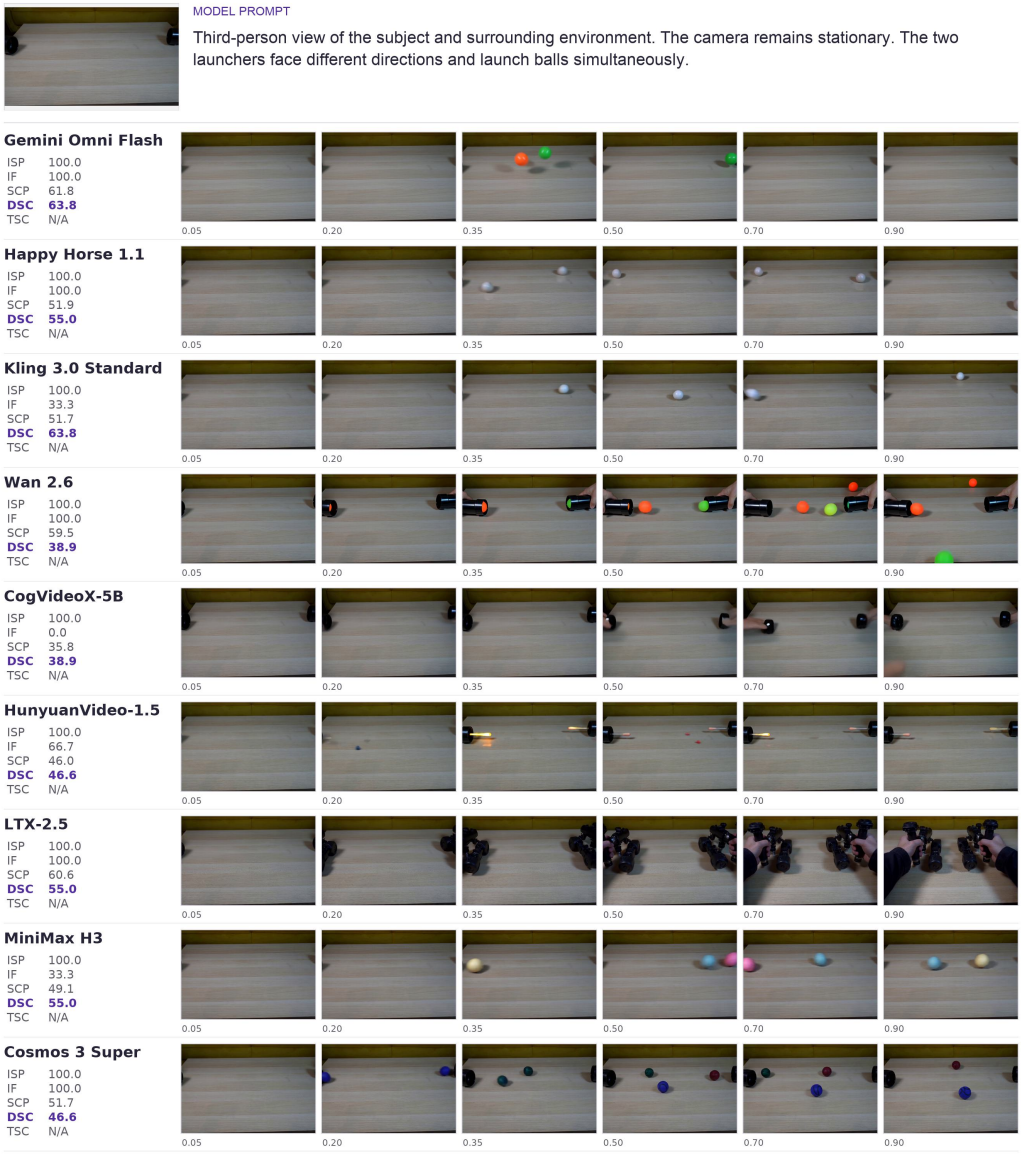


Figure 11: **Dynamic spatial consistency failure.** The task requires two launchers facing different directions to launch balls simultaneously. The comparison reveals substantial differences in the temporal stability of object identities, trajectories, and relative spatial configurations across generators.



Figure 12: **Incomplete spatial evolution.** The task explicitly requires placing the trophy on the lower shelf. The final frame is explicitly included to distinguish models that merely move the object toward the target from those that actually complete the required terminal state.

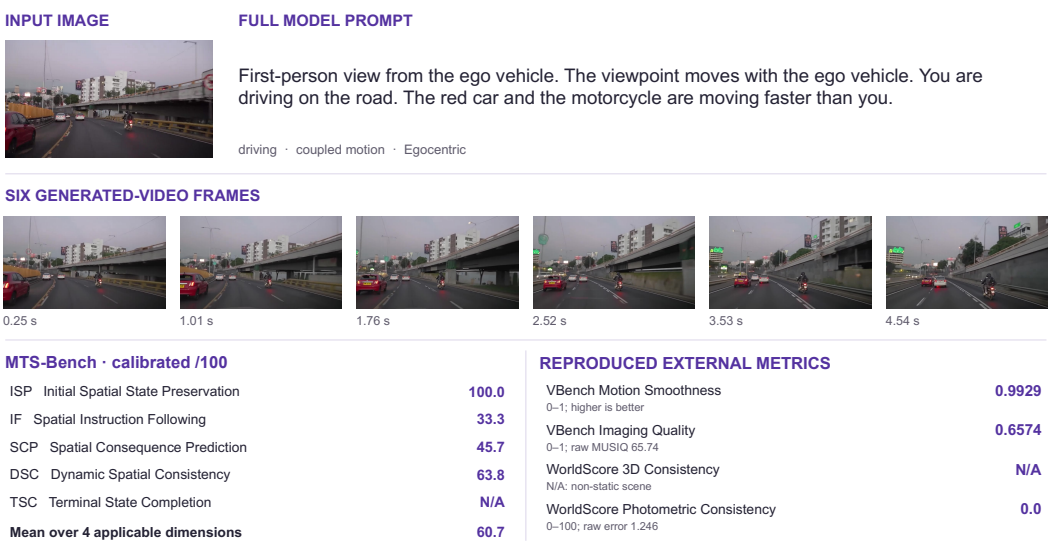


Figure 13: **Conventional perceptual metrics do not imply correct spatial evolution.** This coupled-motion driving example receives a very high VBench Motion Smoothness score, while its MTS-Bench scores indicate limited instruction following and spatial consequence prediction. WorldScore 3D Consistency is marked not applicable in this comparison under the static-world assumption because the scene contains substantial object dynamics.

D.2 What Conventional Video Metrics Miss

Existing video-generation metrics capture useful but complementary properties to the dynamic spatial reasoning targeted by MTS-Bench. Figure 13 shows a coupled-motion example with near-saturated VBench Motion Smoothness (0.9929 on a 0–1 scale), while MTS-Bench assigns IF and SCP scores of 33.3 and 45.7 on its 0–100 scale. Motion Smoothness does not determine whether objects move in the *spatially correct* direction or satisfy the required relative-motion constraints. Likewise, VBench Imaging Quality measures perceptual fidelity but does not directly assess whether the generated scene evolves according to the required spatial relations.

WorldScore targets a different aspect of world-generation quality. In this comparison, we interpret 3D Consistency under the assumption of a sufficiently static world with camera motion, marking the dynamic driving example as not applicable. In the predominantly viewpoint-driven example in Fig. 14, 3D Consistency receives 0.0, while MTS-Bench provides separate assessments of camera-motion execution, spatial consequences, and terminal completion. Photometric Consistency is reported for both examples, but the original WorldScore normalization in our reproduction maps raw errors of 1.246 and 1.883 to the same lower-bound score of 0.0, leaving no discrimination between these two videos.

These two examples compare evaluation scope and do not establish a dataset-level advantage. MTS-Bench complements these metrics by explicitly evaluating whether an intervention produces the required spatial consequence and maintains coherent spatial evolution over time.

E Implementation and Reproducibility Details

We evaluate one generated sample per model–case pair on the fixed 150-case evaluation subset. For I2V, each generator receives the frozen first-frame image together with the complete image-conditioned instruction. For T2V, the generator receives the complete textual prompt comprising the initial-state description, camera constraint, and intervention, without an image. We use each model's supported output configuration closest to the target five-second horizon and do not retime generated videos for evaluation.

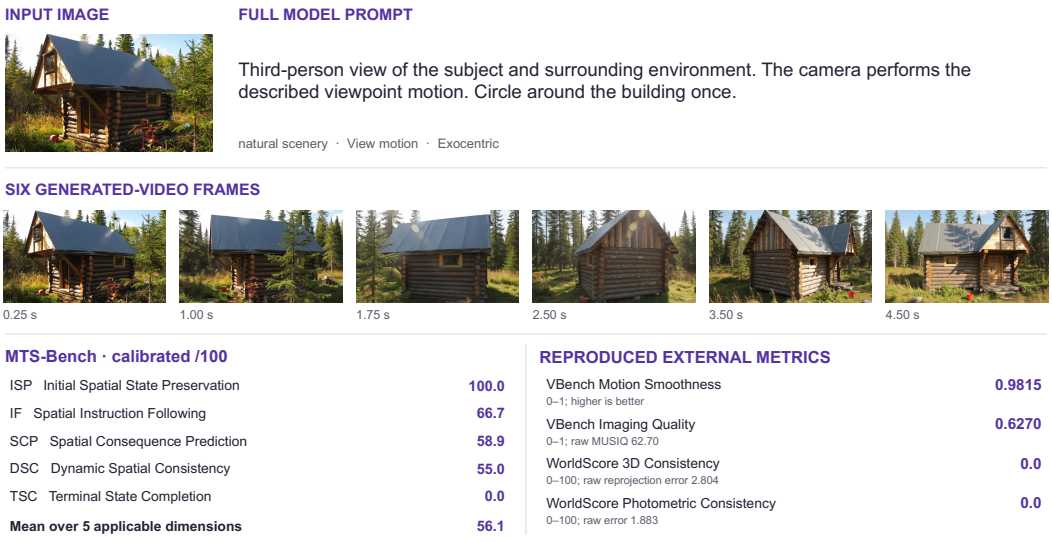


Figure 14: **Different evaluation scopes of WorldScore and MTS-Bench.** In this predominantly viewpoint-driven scene, WorldScore 3D Consistency is applicable and yields a score of 0.0 for geometric reconstruction consistency. MTS-Bench additionally evaluates whether the requested camera motion, spatial evolution, and terminal state are correctly realized. The reproduced Photometric Consistency score saturates at the lower boundary under the original normalization.

Tables 9 and 10 report the exact hosted endpoints or local model families, requested output specifications, and explicit inference settings. Requested specifications are distinguished from measured output properties: parameters not explicitly sent to a service are reported as unspecified rather than inferred from provider defaults.

Table 9: Hosted generator configurations used for the 150-case evaluation. Unspecified parameters were not explicitly sent in the request.

Generator	Mode	Endpoint	Output	Key settings
Cosmos 3 Super	I2V	`nvidia/cosmos-3-super/image-to-video`	120f, 24 FPS, 832 × 480	28 steps; guidance 6; prompt expansion off; agentic generation off
Gemini Omni Flash	I2V	`google/gemini-omni-flash/image-to-video`	5 s, 16:9	No additional inference controls
Happy Horse 1.1	I2V	`alibaba/happy-horse/v1.1/image-to-video`	5 s, 1080p	Safety checker on
Kling 3.0 Standard	I2V	`fal-ai/kling-video/v3/standard/image-to-video`	5 s	Audio off; custom shot mode
Wan 2.6	I2V	`wan/v2.6/image-to-video`	5 s, 720p	Prompt expansion off; multi-shot off
Gemini Omni Flash	T2V	`google/gemini-omni-flash`	5 s, 16:9	No additional inference controls
Happy Horse 1.1	T2V	`alibaba/happy-horse/v1.1/text-to-video`	5 s, 1080p, 16:9	Safety checker on
Kling 3.0 Standard	T2V	`fal-ai/kling-video/v3/standard/text-to-video`	5 s, 16:9	CFG 0.5; audio off; custom shot mode
Wan 2.6	T2V	`wan/v2.6/text-to-video`	5 s, 720p, 16:9	Prompt expansion off; multi-shot off

Table 10: Locally configured generator settings.

Generator	Mode	Model / checkpoint	Output	Key settings
CogVideoX-5B	I2V	`zai-org/CogVideoX-5b-I2V`	49f, 8 FPS, 720×480	50 steps; guidance 6; first-frame conditioning
CogVideoX-5B	T2V	`zai-org/CogVideoX-5b`	41f, 8 FPS	50 steps; guidance 6
HunyuanVideo-1.5	I2V	`tencent/HunyuanVideo-1.5`	480p, 121f, nominal 24 FPS	50 steps; rewrite/SR/distillation off; first-frame conditioning
HunyuanVideo-1.5	T2V	`tencent/HunyuanVideo-1.5`	480p, 16:9, 121f	50 steps; rewrite/SR/distillation off
LTX-2.5 22B distilled	I2V	`LTX-2.5-22B-distilled`	768×512, 121f, 24 FPS	Prompt enhancement off; image strength 1.0; no LoRA
LTX-2.5 22B distilled	T2V	`LTX-2.5-22B-distilled`	768×512, 121f, 24 FPS	Prompt enhancement off; no image condition; no LoRA
MiniMax H3	I2V/T2V	provenance under verification	–	request configuration under verification