WOVEN: WEAVING VISUAL WORLD MODELING INTO MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) exhibit scattered failures across spatial, embodied, physical, and temporal reasoning. We hypothesize that these failures reflect a shared deficit in **visual transition reasoning**, and test whether this capability can serve as a **shared training primitive**, one that different models can learn from different supervision sources and reuse across different tasks, further seeking a systematic training recipe that benefits diverse downstream tasks. Existing benchmarks document these deficits separately but do not support controlled comparisons across scenes, actions, and reasoning operations. To enable such comparisons, we introduce **WOVEN**, a training source and benchmark for visual transition reasoning that organizes transition supervision by scene, action, and reasoning type, using diverse, realistic rollouts from video-pretrained generative models. It comprises 36,076 examples across 20 scene, 5 action, and 8 reasoning types. WOVEN reveals a substantial gap to human performance, motivating a direct study of trainability and transfer. Post-training on WOVEN transfers broadly from little data: controlled training subsets of only about 2,000 items each improve 22 of 26 external benchmarks, with gains of up to 27.3 percentage points, and the transfer persists across model scales. Two properties confirm that visual transition reasoning serves as a shared training primitive: different supervision sources improve the same capability, and the downstream gains come from this capability rather than from separate task-specific effects. WOVEN data can even replace part of task-specific training data while retaining comparable accuracy. More importantly, controlled comparisons yield a **training recipe**, validated prospectively on held-out benchmarks. For the largest downstream benefit, transition supervision should be chosen by the reasoning operation it teaches rather than by the actions, scenes, or domains it shows. When robustness to state changes is needed, the supervision should involve larger changes to the visual state, such as object manipulation. Temporal and agent-driven tasks must be supervised directly, and tasks that involve no transition, such as static perception, do not benefit. Our work establishes visual transition reasoning as a reusable foundation for improving diverse downstream capabilities, paving the way for systematic visual world-model training in MLLMs.

1 INTRODUCTION

Multimodal large language models (MLLMs) perform well on visual recognition and language-conditioned reasoning (Bai et al., 2025b), yet remain unreliable on tasks involving visual change, including spatial, physical, embodied, and temporal reasoning (Yin et al., 2025; Chow et al., 2025; Ma et al., 2026; Huang et al., 2026). For example, they may misjudge camera rotation, predict outcomes that violate physical constraints, recommend actions that conflict with the task goal, or misorder video frames (Fig. 1(1)). Despite their different goals, many of these tasks share a common structure: an action a links an initial visual state s to a resulting state s' (Fig. 1(2)). Predicting outcomes, inferring actions from observed changes, and reasoning about alternative actions are different inferences over the same (s, a, s') triplet. We hypothesize that a shared deficit in *action-conditioned visual transition reasoning* underlies failures across these tasks.

This hypothesis motivates exploring a shared training primitive for otherwise distinct downstream capabilities: rather than improving each capability separately, we ask whether and how visual transition reasoning can serve as a common training target across models and benefit diverse downstream

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.

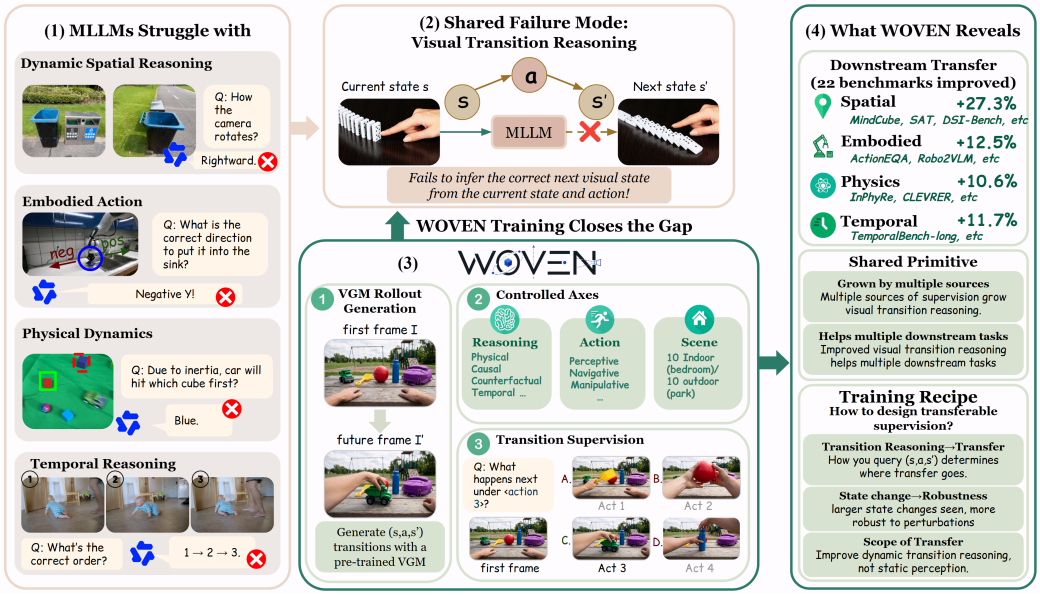

Figure 1: **Overview.** (1) MLLMs fail across spatial, embodied, physical, and temporal tasks. (2) These failures share a weakness: reasoning about how visual states, actions, and outcomes relate. (3) WOVEN generates (s, a, s') transitions with a video generation model, organizes them along controlled axes of reasoning, action, and scene, and turns them into error-typed multiple-choice supervision. (4) Training on WOVEN improves 22 of 26 external benchmarks; the gains come from a shared primitive that different sources build and different tasks use; and controlled comparisons yield a training recipe: the reasoning operation decides where transfer goes, larger state changes give robustness, and transfer covers transition tasks rather than static perception.

tasks. In this paper, we pursue this goal through four key questions. First, do failures in visual transition reasoning exhibit consistent patterns across model families, and how pronounced are these deficits? Second, can it be learned through transition supervision, and do the benefits extend beyond the training tasks? Third, do these gains arise from separate training components helping separate tasks, or from learning visual transition reasoning that different models can reuse across different tasks? Fourth, and more fundamentally, what governs the downstream transfer and should therefore guide a systematic training recipe: similarities in actions, scenes, or application domains, or the reasoning operation applied to the transition (s, a, s')?

Answering these questions calls for controlled comparisons that allow us to examine the roles of individual components of the (s, a, s') triplet. Existing task-specific benchmarks document important deficits, but do not readily support such comparisons (Yin et al., 2025; Chow et al., 2025; Foss et al., 2025; Cai et al., 2024; Ma et al., 2026; Huang et al., 2026). We therefore introduce **WOVEN** (**WO**rld modeling via controllable **V**ideo g**EN**eration), a diagnostic and training source that draws on the transition priors of video-pretrained generative models (Team Wan et al., 2025). Generated rollouts provide known actions and endpoints for forward and inverse supervision, outcomes under alternative actions for counterfactual supervision, and temporally coherent intermediate states for temporal supervision. Organizing this supervision along scene, action, and reasoning dimensions enables controlled studies of learning and transfer (Fig. 1(3)). The collection comprises 36,076 examples spanning 20 scene types (e.g., kitchen, park, and living room), five action types (e.g., passive physical events, camera motion, and object manipulation), and eight reasoning types across four families (e.g., causal dynamics, temporal coherence, and counterfactual reasoning). We also provide diagnostic splits which cover in-distribution evaluation, held-out scenes, and state perturbations, with error-typed distractors supporting fine-grained analysis of the effects of training (§3).

To examine visual transition reasoning as a common training target, we first ask whether its deficits recur across model families and whether learning it yields gains beyond the training tasks. Through fine-grained evaluation of 38 frontier MLLMs (e.g., GPT-5.4 (OpenAI, 2026), Qwen3-VL-235B-A22B (Bai et al., 2025a), and Llama-4-Maverick (Meta AI, 2025)), we identify recurring failure

Table 1: Eight reasoning operations considered in this work. The same operation can be applied to different actions and scenes (§3). Notation: a_{alt}, an alternative action from the same initial state; a_{exo}, a passive event; a_q, a_{exo} with a cue naming the physical principle; $\mathbf{s} = (s_0, \ldots, s_L)$, sampled states in time order with $s_0 = s$ and $s_L = s'$; $\widetilde{\mathbf{s}}$, the same states shuffled; s_-, s_+, the neighbors in time of a reference state s_{ref}. In the physical and temporal families the action description is optional.

Family	Reasoning operation	Distribution	Reasoning operation	Distribution
Causal	Forward dynamics	$P(s' \mid s, a)$	Inverse dynamics	$P(a \mid s, s')$
Counterfactual	Counterfactual removal	$P(s \mid a, s')$	Counterfactual substitution	$P(s'_{a_{\text{alt}}} \mid a, s', a_{\text{alt}})$
Physical	Outcome prediction	$P(s' \mid s, a_{\text{exo}})$	Cued prediction	$P(s' \mid s, a_q)$
Temporal	Temporal ordering	$P(\mathbf{s} \mid \widetilde{\mathbf{s}}, a)$	Temporal adjacency	$P(s_-, s_+ \mid s_{\text{ref}}, \widetilde{\mathbf{s}}, a)$

patterns across model families. These deficits are substantial: even the strongest model achieves only 65.8% accuracy against a 92.3% human baseline. These recurring deficits identify visual transition reasoning as a common target for improving current MLLMs, establishing a training need that extends beyond the weaknesses of any individual model (§3). We then ask whether this capability can be learned and whether the resulting gains extend beyond the training tasks. Crucially, our empirical results from post-training MLLMs at multiple scales (§4.2) demonstrate substantial gains: training on the full WOVEN training set raises the in-distribution accuracy of Qwen2.5-VL-3B (Bai et al., 2025b) from 26.4% to 89.3%, with 88.2% on held-out scenes (§4.2). The benefits extend to 12 of 26 external benchmarks across multiple domains. Controlled subsets containing approximately 2,000 examples each collectively expand this coverage to 22 of 26 benchmarks, with gains of up to 27.3 percentage points (Fig. 1(4)).

However, this is not enough to establish visual transition reasoning as a shared training primitive that different models can reuse from different sources in learning different tasks, as these downstream gains could just arise from separate parts of the training source, each helping only its corresponding tasks. To serve as a shared training primitive, visual transition reasoning must be a common capability that models acquire from any of the sources, and the gains on downstream tasks must also result from this common capability. Therefore, we first test whether different sources improve the same capability. Results show that each of the 11 controlled subsets, which represent different sources of transition supervision, improves WOVEN accuracy, including on the four action types that it does not train on (43 of 44 comparisons; §4), meeting the first requirement. Then, we examine whether the downstream gains result from this common capability. If they do, subsets that improve this capability more should transfer more, and the gains should not just come from an overall stronger checkpoint. Results show that the more a subset improves WOVEN accuracy, the larger its average gain on the external benchmarks (r=0.86), as expected. Moreover, all 22 transition-related benchmarks improve, whereas the four benchmarks of static perception and general video QA do not (§4.2), so the transfer is specific to tasks that involve visual transitions rather than an overall stronger checkpoint. We further test causality by two interventions on the training data: the same WOVEN format with permuted or mirrored transitions does not reproduce the gains, and replacing 30–50% of task-specific examples with WOVEN under a fixed training budget preserves comparable accuracy (Fig. 3), showing that transition supervision can take over part of the training normally supplied by task-specific data. Together with transfer across model scales (§4.2), these results establish visual transition reasoning as a shared training primitive.

More importantly, our controlled comparisons yield a training recipe for visual world modeling, validated prospectively on held-out benchmarks (§5). **(1) Select supervision by reasoning operation.** Surprisingly, supervision that teaches the same reasoning operation produces similar patterns of downstream gains even when the actions differ, whereas supervision that shares only an action type does not (Fig. 1(4)). The reasoning operation therefore provides a basis for selecting supervision with relevant downstream benefits. Once the operation is selected, **(2) choose supervision with larger changes to the visual state for robustness.** Supervision whose actions change more of the scene, such as object manipulation, yields larger robustness gains than supervision that only moves the camera. **(3) Supervise temporal and agent-driven tasks directly.** Transfer does not cross the boundary between the temporal family and the other three reasoning families, nor the boundary between passive physical events and agent-driven actions. **(4) Keep dedicated supervision for capabilities beyond state transitions.** Transition supervision improves diverse tasks that involve state transitions, but does not necessarily generalize to capabilities such as static perception, which therefore need their own supervision (§5). Taken together, our study establishes visual transition

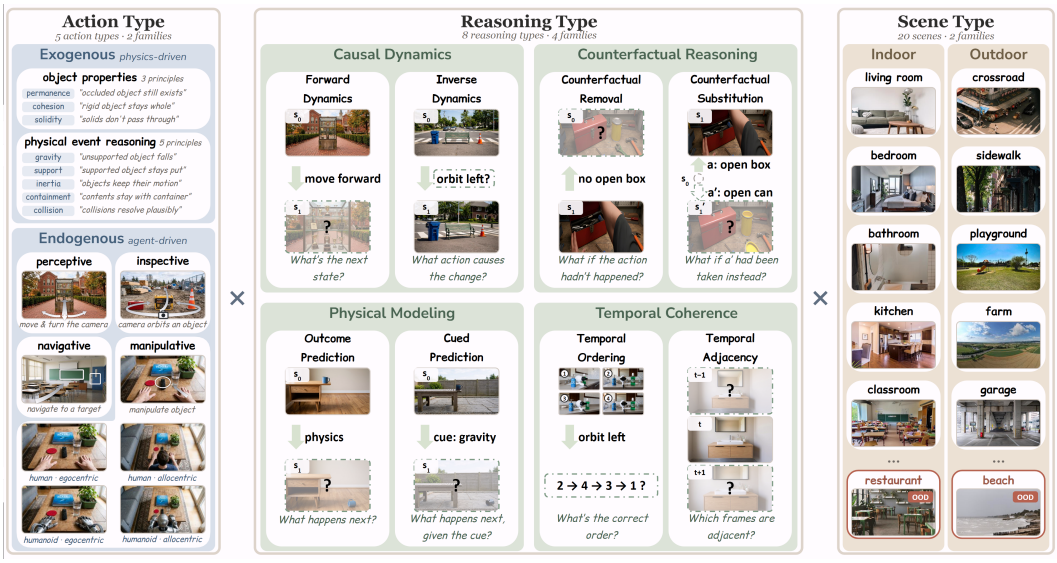

Figure 2: **The WOVEN taxonomy.** Every item is labeled along the three axes shown from left to right (§3.1). Left: the exogenous action type with its eight principles, and the four endogenous types, where manipulative actions combine subject (human, humanoid) and view (egocentric, allocentric). Middle: the eight reasoning types in four families; in each schematic, the dashed box marks the element of (s, a, s') that the question asks for. Right: example scenes; restaurant and beach are held out for the scene-OOD test.

reasoning as a shared, trainable primitive and provides a systematic recipe for training it, offering a controlled framework for advancing visual world modeling in MLLMs.

2 Visual Transition Reasoning

Given a transition (s, a, s'), where s and s' are the visual states of a scene before and after an action a, which may be agent-driven or passive, *visual transition reasoning* is the process of inferring unobserved component(s) of the transition from its known or observed parts. Different choices of which parts are given or to be inferred define different reasoning operations, each written as a conditional distribution such as $P(s' \mid s, a)$ (shorthand for $P(S' = s' \mid S = s, A = a)$). This work considers eight operations in four families (Table 1). Appendix A.1 gives the semantics of each operation.

3 WOVEN

Studying visual transition reasoning requires controlled comparisons of supervision: varying supervision along one axis such as scene, action, or reasoning operation, while holding the rest of them fixed. Real video offers neither controlled actions nor alternative outcomes from the same initial state, and simulators offer control but limited realism and limited scene and object diversity. Drawing on the rich rollout prior (Appendix P) of video generation models, we propose WOVEN, which produces controlled transitions from specified initial states and actions through action-conditioned video generation, making alternative outcomes and temporally coherent intermediate frames available for counterfactual and temporal supervision, so that every item is labeled along all three axes: scene, action, and reasoning type.

3.1 Controlled Axes

We define the **taxonomy** along three axes and categorize each item accordingly: (1) the *scene* in which its transition occurs, (2) the *action* producing the change, and (3) the *reasoning operation* applied to the transition (Fig. 2). *Scene type* has 20 categories, evenly divided between indoor and outdoor environments. *Action type* has 5 categories: one exogenous type (passive physical events),

divided into 8 principles from cognitive science (3 object properties and 5 physical events), and four endogenous types (agent-driven): perceptive, inspective, navigative, and manipulative (Fig. 2). *Reasoning type* instantiates the eight operations of Table 1, grouped in pairs into four families: causal dynamics, counterfactual reasoning, physical modeling, and temporal coherence. Within compatible combinations, each axis can be varied while the other two are held fixed. Category definitions and cognitive motivation are in Appendices B.2–B.5.

3.2 Rollout Generation

We use Wan2.2-I2V-A14B (Team Wan et al., 2025) to generate rollouts from initial frames and action descriptions, following two construction principles. **(P1) World-model isolation.** Resulting and intermediate states are produced by the rollout rather than prescribed in the conditioning prompt, so supervision draws on learned transition dynamics. **(P2) Counterfactual rollouts.** From the same initial state, we generate several independent rollouts under different actions, so that alternative outcomes of the same scene are available. Generation, frame extraction, and quality-control details are provided in Appendix B.1.

3.3 Transition Supervision

Question construction. We build multiple-choice questions from the rollouts, each referred to as an item and is defined on all three axes: its scene and action types are those of the transition it is built from, and its reasoning type (Table 1) determines which parts of the transition are given and which part is asked for. Each question is phrased with one of 30 templates drawn at random to ensure linguistic diversity. The three distractor options are usually drawn from other rollouts sharing the same initial state, each wrong in a specific way (e.g., the outcome of a different action) and labeled with its error type (e.g., "reversed direction" or "violating gravity"), enabling fine-grained error analysis and statistical attribution beyond overall accuracy. For physical-modeling items, the distractors are produced by an image editor, so the correct option is also passed through the same editor with an instruction to change nothing; a control experiment shows that editing artifacts do not reveal the answer. See Appendices B.8 and D for details.

Splits. WOVEN has one training set (22,728), one validation set (2,508), and three test sets, totaling 36,076 four-option multiple-choice items (Appendix B.6). Each test set differs from the training set in one respect. The *in-distribution* test set (6,228) contains new transitions from the same 18 scenes as the training set. The *scene-OOD* test set (3,496) contains transitions from two held-out scenes, restaurant and beach, following held-out-environment evaluation in embodied AI (Majumdar et al., 2024; Cheng et al., 2025). The *state-perturbation* test set (1,116) measures robustness to controlled changes in the initial state: following the logic of contrast sets (Gardner et al., 2020), each item perturbs one condition of an in-distribution item's initial state, either its spatial configuration or its appearance, while keeping the action and the multiple-choice question fixed, testing whether a model actually models the state rather than the typical effect of the action (Appendix B.9).

4 Visual Transition Reasoning as a Shared, Trainable Primitive

We extensively study visual transition reasoning as a core capability for visual world modeling in MLLMs. We start by asking a series of research questions: how far current models fall short of it, whether it is efficiently learnable and transferable beyond the training distribution, and finally whether it serves as a shared primitive when helping downstream tasks. With WOVEN as a controlled, error-typed, shortcut-resistant, and large diagnostic and training source, we are able to study them in depth by combining a tremendous range of techniques and analyses, including broad and detailed evaluation of frontier models, controlled training at multiple scales, fine-grained transfer analysis with error-type-level statistical attribution, and causal intervention on the training data. Throughout, training is carried out in two complementary ways: on the full WOVEN training set, and on controlled subsets, each containing only items of a specified action type and reasoning family. WOVEN's training set consists of 11 such subsets in total. We provide details in the following subsections.

Setup. Training uses supervised fine-tuning followed by GRPO (Shao et al., 2024) (Appendix G.1) on MLLMs of multiple scales (Appendix G.15); the full training set has 22,728 items and each of the 11 controlled subsets about 2,000. Evaluation uses the three WOVEN test sets and, zero-shot,

Table 2: **Zero-shot transfer after training on the complete WOVEN training set.** Accuracy (%) on the 12 improved benchmarks; Δ is the larger of the SFT and GRPO gains. References, results for all 26 benchmarks, and category breakdowns are in Appendix G.13.

	SAT	InPhyRe	Mind Cube	Action EQA	Temporal Bench-L	WM-ABench	World Prediction	TOMATO	Black Swan	ERQA	Spatial Viz	CV-Bench
Base	37.3	22.0	37.2	43.4	16.6	31.0	34.8	31.6	61.8	33.8	27.7	69.3
WOVEN SFT	55.3	24.6	37.8	49.5	11.5	37.1	39.6	31.5	55.9	**37.5**	28.8	72.0
WOVEN GRPO	**61.3**	**34.9**	**48.6**	**54.4**	**26.8**	**39.8**	**41.5**	**36.1**	**65.6**	34.5	**31.1**	**72.4**
Δ	+24.0	+13.0	+11.3	+11.0	+10.3	+8.9	+6.7	+4.5	+3.8	+3.8	+3.4	+3.0

Table 3: **Broad transfer from controlled training subsets.** Results on 8 of the 22 improved benchmarks; each column uses the best-performing subset for that benchmark. Results for all 22 benchmarks are in Appendix G.13 (Table 42).

	SAT	Action EQA	Temporal Bench-L	InPhyRe	Robo2VLM	World Prediction	DSI-Bench	Mind Cube
Base	37.3	43.4	16.6	22.0	33.5	34.8	34.9	37.2
Subset SFT	63.3	47.5	22.0	32.3	40.5	40.1	40.3	40.0
Subset GRPO	**64.7**	**55.9**	**28.2**	**32.6**	**42.4**	**42.9**	**40.9**	**43.1**
Δ	+27.3	+12.5	+11.7	+10.6	+8.9	+8.1	+6.0	+5.9
Subset	perc_causal	perc_causal	mani_cf	insp_cf	insp_cf	mani_causal	perc_causal	perc_temporal

26 external benchmarks: 22 that involve visual transitions, spanning spatial, embodied, physical, temporal reasoning, etc., and 4 benchmarks of static perception and general video QA; a benchmark counts as improved when the gain over the base model exceeds 2 percentage points.

4.1 How Far Do Current Models Fall Short?

To establish the need for training, we first quantify the size and distribution of the deficit in visual transition modeling by evaluating 38 frontier MLLMs on the in-distribution and state-perturbation test sets. The best model reaches 65.8% in-distribution accuracy against a 92.3% human baseline (the mean of three annotators), demonstrating a large gap that persists across model scales and cannot be closed by scaling alone. The gap also has a consistent structure across model families (e.g., temporal adjacency is harder than ordering, counterfactual removal is harder than substitution, and accuracy drops more under geometric than under appearance changes), suggesting a systematic deficit in visual transition modeling rather than an idiosyncrasy of any particular family or training recipe. These results also provide a fine-grained baseline for the training study that follows. The model inventory, human protocol, and detailed results are in Table 10 and Appendices C.2 and F; complementary evaluations of a VGM, V-JEPA 2 (Assran et al., 2025), and an IGM are in Appendix P.

4.2 Is It Efficiently Learnable and Transferable?

Supervised fine-tuning on the full training set raises in-distribution accuracy from 26.4% to 89.3%, with 88.2% on held-out scenes; the state-perturbation test set improves least, from 21.1% to 38.3% (Appendix G.2). By error type, fine-tuning reduces predictions that "nothing changed" and improves temporal localization, while errors in the direction of motion and under geometric changes remain (Appendix G.17). The deficit is therefore learnable and the learning generalizes across scenes, with robustness to state changes as the main remaining gap. Each controlled subset also raises WOVEN accuracy on its own, including on the four action types it does not train on (43 of 44 comparisons; Appendix G.17), so different kinds of supervision improve the same capability rather than only the action they were trained on. The learning is also data-efficient: with about 7 hours of generated video, WOVEN post-training yields a larger gain on a spatial-reasoning benchmark than Orca (Wang et al., 2026c), which mid-trains the same backbone on 125,000 hours of video (Appendix G.16).

We then examine whether the learned reasoning transfers to other tasks. The full training set improves as many as 12 of the 26 external benchmarks, with gains of up to 24.0 percentage points on SAT (Table 2). The controlled subsets transfer effectively as well, despite holding only about 2,000 items

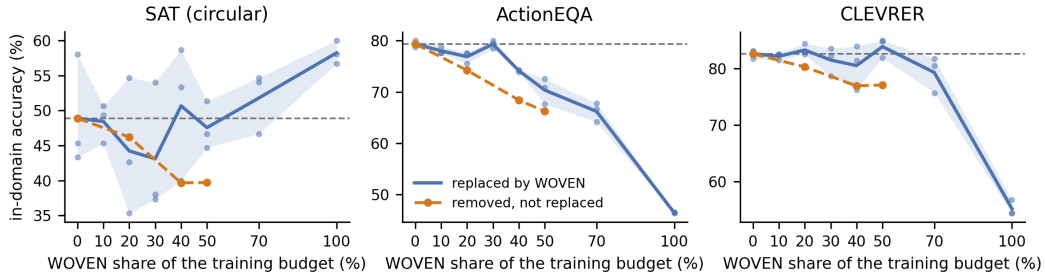

Figure 3: **WOVEN supervision substitutes for in-domain training data.** Each panel retrains one external benchmark on a fixed 2,000-item budget in which a share p of the benchmark's own training data is replaced by items from a selected WOVEN subset (blue dots: 3 seeds; blue line: seed mean; band: seed range; gray dashed: p=0 mean; orange dashed: deletion control). Replacing 30–50% of task-specific supervision preserves comparable accuracy; on SAT, training exclusively on WOVEN yields the highest accuracy in the sweep. Protocol and full tables in Appendix G.14.

each: with the best subset chosen for each benchmark, the coverage rises to a remarkable 22 of 26, and gains reach up to 27.3 percentage points (Table 3; selection and seed checks in Appendix M). The transfer also holds at scale (Appendix G.15) and across model families (Appendix J). The 4 benchmarks of static perception and general video QA do not improve under any training, so the gains are not due to an overall stronger model or to better multiple-choice answering (Appendix I). Full scores and evaluation procedures are in Appendices G.13 and D.

4.3 Does It Serve as a Shared Primitive?

The gains so far could still be a sum of separate effects, each subset helping the tasks closest to it. For visual transition reasoning to serve as a shared training primitive, two requirements must hold: different sources must improve the same capability, and the downstream gains must also result from that capability. The first requirement is met in §4.2, where every subset improves WOVEN accuracy beyond the action type it trains on; the capability also spans reasoning operations, as training on forward prediction alone raises counterfactual substitution and removal by 48 and 25 percentage points (Appendices G.8 and L), and it is not a shortcut (Table 86; Appendix K). For the second requirement, if the downstream gains result from the capability, all subsets should improve largely the same benchmarks, and subsets that improve the capability more should transfer more. Results show that the transfer profiles of the 11 subsets are all positively correlated (Fig. 4, left), so the subsets improve largely the same benchmarks (detailed analysis in Appendix G.13). Moreover, the more a subset improves WOVEN accuracy, the larger its average gain on the external benchmarks (r=0.86; Appendix G.13), as expected, and the same WOVEN format with permuted or mirrored transitions does not reproduce the downstream gains (Appendix H). Together with the specificity shown in §4.2, the downstream gains therefore come from the learned capability rather than from the individual subsets. We further test causality by intervention on the training data. With the training set fixed at 2,000 items, we replace a growing share of task-specific items with a WOVEN subset on three benchmarks (SAT (Ray et al., 2024), ActionEQA (Anonymous, 2025), and CLEVRER (Yi et al., 2020)) from three different domains, and average results over three seeds. Substituting up to 30–50% (depending on the benchmark) of the task-specific items leaves accuracy comparable to training on task-specific items alone, and above a control that simply deletes the replaced items (Fig. 3; Appendix G.14). Transition supervision thus supplies part of what these tasks would otherwise learn from their own data, supporting the conclusion that the downstream gains result from the learned capability, which satisfies the second requirement. Together, these results establish visual transition reasoning as a shared training primitive; how its transfer is organized beyond this common component is the subject of §5.

5 A Training Recipe for Visual World Modeling

Section 4 establishes visual transition reasoning as a shared, trainable primitive. From its fine-grained results, the transfer profiles, error-typed analyses, and perturbation robustness of the controlled

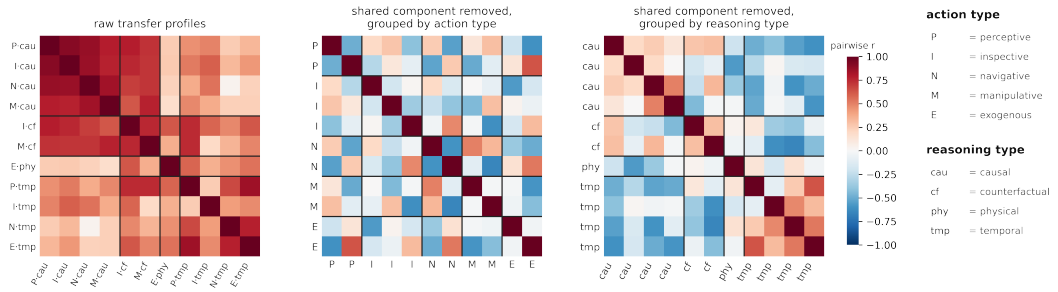

Figure 4: **Correlation of transfer profiles under the two candidate groupings.** Pairwise Pearson correlation of the 11 subsets' 26-benchmark transfer profiles. Left: raw profiles, all pairs positive (the common component of §4.3). Middle and right: the same residual matrix after the per-benchmark z-score, ordered by action type and by reasoning family; only the latter produces diagonal blocks.

subsets, we distill a training recipe for visual world modeling in MLLMs (Fig. 1(4)): (1) to improve a downstream task, select supervision that teaches the reasoning operation the task requires, not supervision that matches its actions, scenes, or domains; (2) for robustness, prefer supervision with large state changes, such as object manipulation over camera motion; (3) supervise temporal and agent-driven tasks directly: transfer does not cross the boundary between the temporal family and the other three, nor the boundary between passive physical events and agent-driven actions; and (4) for downstream tasks that involve no transition, such as static perception, transition supervision does not help. The following subsections give the evidence behind each item; Appendix O tests the recipe prospectively on held-out benchmarks.

5.1 Reasoning Operation Decides Where Transfer Goes

Similarity in scenes, actions, or application domains offers an intuitive basis for selecting transferable supervision. To test whether such similarity predicts downstream benefit, we compare subsets that share only an action type with those that share only a reasoning family. All transfer profiles contain the common component of §4.3, so we remove it by standardizing the gains within each benchmark, which leaves each subset's relative pattern of benefit, and correlate these residual profiles pairwise (Fig. 4).

Surprisingly, results show that subsets teaching the same reasoning family have similar residual profiles, with a mean pairwise correlation of $+0.32$, whereas subsets sharing only an action type are no more similar than subsets sharing neither (-0.22 versus -0.23), so the action shown in a transition does not predict where its supervision transfers. Moreover, the same ordering holds under three other normalizations of the gains (Fig. 4; Appendix G.13). Together with the transfer to held-out scenes and to external domains in §4.2, the reasoning operation a source of supervision teaches, not the actions or scenes it shows, decides which downstream tasks it helps. **Recipe (1).** To improve a downstream task, select supervision that teaches the reasoning operation the task requires, not supervision that matches its actions, scenes, or domains; Table 3 and Appendix G.13 list the best-transferring subset for each benchmark.

5.2 Extending Robustness Through State Interventions

Recipe (1) leaves the extent of state change open. We examine whether larger state changes improve robustness by comparing subsets that differ in the actions that transform the scene. Among the causal subsets, robustness gains increase from camera motion ($+1.7$) and object inspection ($+3.6$) to navigation ($+13.6$) and object manipulation ($+23.2$), and passive physical events give the largest gain of all subsets ($+29.9$; Appendix G.17). **Recipe (2).** For robustness, prefer supervision with large state changes, such as object manipulation over camera motion.

5.3 Reasoning Types That Need Their Own Supervision

Although supervision on one reasoning operation often teaches others, as forward-dynamics supervision alone raises counterfactual substitution and removal by 48 and 25 points (§4.3), two

kinds of transition reasoning are not learned this way. (a) Temporal localization is learned only from temporal supervision: the four temporal subsets raise temporal-adjacency accuracy from 25.7% to 52.7–56.7%, whereas the seven other subsets leave it near the baseline, and supervision on temporal adjacency and on temporal ordering do not transfer to each other (Appendices G.9 and G.17). (b) Endogenous (agent-driven) transitions are not learned from exogenous supervision: the subset of passive physical events contains no agent action to learn from, and it lowers accuracy on the two benchmarks built on an agent's action and its outcome, WorldPrediction (-10.2) and ActionEQA (-1.7; Appendix G.13). **Recipe (3).** Supervise temporal and agent-driven tasks directly: transfer does not cross the boundary between the temporal family and the other three reasoning families, nor the boundary between passive physical events and agent-driven actions.

5.4 Scope of Transfer

Finally, transfer covers all 22 transition-related benchmarks and none of the 4 benchmarks of static perception and general video QA, which do not improve under any training (§4.2). **Recipe (4).** For downstream tasks that involve no transition, transition supervision does not help. Together, the four items form our training recipe (Fig. 1(4)), and items (1), (3), and (4) hold prospectively (Appendix O).

6 Related Work

World models and video priors. World models predict how a scene evolves, either in a latent space (Hafner et al., 2023; Assran et al., 2025) or by generating video (OpenAI, 2024; Ball et al., 2025; Team Wan et al., 2025; NVIDIA Cosmos Team, 2025), and video models increasingly serve as simulators for planning (Du et al., 2023; Zhou et al., 2024) and as sources of training data for downstream learning (DreamGen Team, 2025). Transition-level supervision has also been used directly: state transition pretraining teaches GUI agents the forward and inverse dynamics of interface states (Wang et al., 2026a), and intervention-based world-state data support controlled generation, action prediction, and policy transfer (Cai et al., 2026). WOVEN uses video-pretrained transition priors for a different purpose, as a source of controlled supervision for studying which transition reasoning MLLMs learn and reuse, which complements evaluations of rollout quality (VBench Team, 2024; NVIDIA, 2025b) and of world models on planning and interaction tasks (WorldPrediction Team, 2025; Warrier et al., 2025).

Benchmarks and supervision for multimodal reasoning. A large body of benchmarks tests MLLMs on one facet of visual change at a time: spatial reasoning under viewpoint change (Yin et al., 2025; Yang et al., 2025b; Li et al., 2025a; Wang et al., 2026b), physical and intuitive-physics reasoning (Chow et al., 2025; Sreekumar & Boddeti, 2025; IntPhys 2 Team, 2025; Yi et al., 2020), counterfactual reasoning about videos (Foss et al., 2025; MuCR Team, 2024; CRIPP-VQA Team, 2022; Chinchure et al., 2025), temporal ordering and localization (Cai et al., 2024; Shangguan et al., 2025; Liu et al., 2024), and embodied and robotic question answering (Majumdar et al., 2024; Anonymous, 2025; Chen et al., 2025a), with recent suites targeting world modeling as a whole (Huang Team and Contributors, 2025; WorldPrediction Team, 2025). On the supervision side, synthetic data has been used to teach MLLMs spatial reasoning (Ray et al., 2024) and transferable reasoning through rule-based games (Xie et al., 2026), and fully controlled scenes have been used to study what fine-tuning transfers to real images (Rizzoli et al., 2026). WOVEN relates these facets through a common (s, a, s') representation and controlled training comparisons, separating the reasoning operation of a task from its action and scene content (Appendices A.6 and A.4).

7 Conclusion

We used a factorized experimental design to study how the content and reasoning demands of supervision shape the acquisition and transfer of visual transition reasoning. The rollout priors and conditional generation of video generation models supplied the controlled transitions for these comparisons. To the best of our knowledge, we are the first to establish visual transition reasoning as a shared training primitive in MLLMs across diverse task domains and derive a systematic training recipe from controlled comparisons of its downstream transfer. More broadly, our work paves the way for developing visual world modeling in MLLMs through reusable reasoning capabilities whose training supervision is designed by controlled experiments.

ETHICS STATEMENT

The human baseline (Appendix C.2) was collected from three annotators who answered multiple-choice questions about WOVEN items; annotators were paid in accordance with local wage standards.

AI USE STATEMENT

We used AI tools to assist with writing and polishing, and used Wan2.2-I2V-A14B (Team Wan et al., 2025) and GPT-5.2 (OpenAI, 2025) to generate synthetic data.

REFERENCES

3D-R1 Team. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding, 2025. URL https://arxiv.org/abs/2507.23478.

3DSRBench Team. 3DSRBench: A comprehensive 3d spatial reasoning benchmark, 2025. URL https://arxiv.org/abs/2412.07825. ICCV 2025.

Anonymous. ActionEQA: Action interface for embodied question answering. In *OpenReview preprint*, 2025. https://openreview.net/forum?id=HY2ruqdMt4.

Mahmoud Assran et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. 2025b. URL https://arxiv.org/abs/2502.13923.

Renée Baillargeon. Object permanence in 3.5- and 4.5-month-old infants. *Developmental Psychology*, 23(5):655–664, 1987.

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, et al. Genie 3: A new frontier for world models. https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/, 2025. Google DeepMind blog post.

Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

Neil Burgess. Spatial memory: How egocentric and allocentric combine. *Trends in Cognitive Sciences*, 10(12):551–557, 2006.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. TemporalBench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024. URL https://arxiv.org/abs/2410.10818.

Ruojin Cai, Jason Y. Zhang, Philipp Henzler, Zhengqi Li, Noah Snavely, and Ricardo Martin-Brualla. Can generative video models help pose estimation? In *CVPR*, 2025. Highlight.

Yiming Cai, Fangjie Yu, Meiqing Yu, Ziyue Shi, Pengfei Yuan, and Yong Guo. CG-World: A large-scale world-state dataset and protocol for world models, 2026. URL https://arxiv.org/abs/2607.26452.

Kaiyuan Chen et al. Robo2VLM: Visual question answering from large-scale in-the-wild robot manipulation datasets. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025a. arXiv:2505.15517.

Shiqi Chen, Tongyao Zhu, Zian Wang, Jinghan Zhang, Kangrui Wang, Siyang Gao, Teng Xiao, Yee Whye Teh, Junxian He, and Manling Li. Internalizing world models via self-play finetuning for agentic rl, 2025b. URL https://arxiv.org/abs/2510.15047.

Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, Tianyu Yu, Weize Chen, et al. EmbodiedEval: Evaluate multimodal LLMs as embodied agents. *arXiv preprint arXiv:2501.11858*, 2025.

Aditya Chinchure et al. Black swan: Abductive and defeasible video reasoning in unpredictable events. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.05725.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025. ICLR 2025.

ChronoMagic-Bench Team. ChronoMagic-Bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation, 2024. URL https://arxiv.org/abs/2406.18522. NeurIPS 2024 Datasets & Benchmarks Spotlight.

CogVideoX Team. CogVideoX: Text-to-video diffusion models with an expert transformer, 2024. URL https://arxiv.org/abs/2408.06072. ICLR 2025.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in time: A new temporal benchmark for VideoLLMs. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2025. arXiv:2410.07752; introduces TVBench.

CRIPP-VQA Team. CRIPP-VQA: Counterfactual reasoning about implicit physical properties via video question answering, 2022. URL https://arxiv.org/abs/2211.03779.

DreamGen Team. DreamGen: Unlocking generalization in robot learning through video world models, 2025. URL https://arxiv.org/abs/2505.12705.

Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Kaelbling, Andy Zeng, and Jonathan Tompson. Video language planning, 2023. URL https://arxiv.org/abs/2310.10625. ICLR 2024.

ECBench Team. ECBench: Can multi-modal foundation models understand the egocentric world? a holistic embodied cognition benchmark, 2025. URL https://arxiv.org/abs/2501.05031. CVPR 2025.

EmbodiedBench Team. EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents, 2025. URL https://arxiv.org/abs/2502.09560. ICML 2025 Oral.

Russell A. Epstein, Eva Zita Patai, Joshua B. Julian, and Hugo J. Spiers. The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience*, 20(11):1504–1513, 2017.

Zheyu Fan, Jiateng Liu, Yuji Zhang, Zihan Wang, Yi R. Fung, Manling Li, and Heng Ji. EMCompress: Video-LLMs with endomorphic multimodal compression. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 137–162, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.8. URL https://aclanthology.org/2026.findings-acl.8/.

Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T. Kao. CausalVQA: A physically grounded causal reasoning benchmark for video models, 2025. URL https://arxiv.org/abs/2506.09943.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, 2010.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2404.12390.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Gemini Robotics Team. Gemini robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025. Introduces the ERQA benchmark.

Gemma Team, Google DeepMind. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.

Google DeepMind. Gemini 3.1 Flash Image. https://deepmind.google/models/gemini-image/flash/, 2026. Image generation and editing model.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2023. URL https://arxiv.org/abs/2301.04104. Nature 2025.

Patrick Haggard. Human volition: Towards a neuroscience of will. *Nature Reviews Neuroscience*, 9 (12):934–946, 2008.

Yuzhi Huang, Kairun Wen, Rongxin Gao, Dongxuan Liu, Yibin Lou, Jie Wu, Jing Xu, Jian Zhang, Zheng Yang, Yunlong Lin, Chenxin Li, Panwang Pan, Junbin Lu, Jingyan Jiang, Xinghao Ding, Yue Huang, and Zhi Wang. Thinking in dynamics: How multimodal large language models perceive, track, and reason dynamics in physical 4D world, 2026. URL https://arxiv.org/abs/2603.12746.

Huang Team and Contributors. Do vision-language models have internal world models? towards an atomic evaluation, 2025. URL https://arxiv.org/abs/2506.21876. Findings of ACL 2025; ICLR 2025 Workshop.

IntPhys 2 Team. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments, 2025. URL https://arxiv.org/abs/2506.09849.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. EgoTaskQA: Understanding human tasks in egocentric videos. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022. arXiv:2210.03929.

Klaus Kessler and Lindsey A. Thomson. The embodied nature of spatial perspective taking: Embodied transformation versus sensorimotor interference. *Cognition*, 114(1):72–88, 2010.

Nima Khalighinejad, Aaron Schurger, Andrea Desantis, Leor Zmigrod, and Patrick Haggard. Dissociating cognitive and motoric precursors of human self-initiated action. *bioRxiv*, 2018. doi: 10.1101/326686.

Benno Krojer et al. A shortcut-aware video-QA benchmark for physical understanding via minimal video pairs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.09987; introduces MVP.

David Lewis. *Counterfactuals*. Harvard University Press, 1973.

Dingming Li et al. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025a.

Eric Li et al. Critiques of world models, 2025b. URL https://arxiv.org/abs/2507.05169.

Yijiang Li et al. Core knowledge deficits in multi-modal language models. In *International Conference on Machine Learning (ICML)*, 2025c. arXiv:2410.10855; introduces CoreCognition.

Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, Rushikesh Zawar, Xue Bai, Yilun Du, Chuang Gan, and Deva Ramanan. Towards understanding camera motions in any video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xinran Ling, Chen Zhu, Meiqi Wu, Hangyu Li, Xiaokun Feng, Cundian Yang, Aiming Hao, Jiashu Zhu, Jiahong Wu, and Xiangxiang Chu. VMBench: A benchmark for perception-aligned video motion generation, 2025. URL https://arxiv.org/abs/2503.10076.

Yuanxin Liu et al. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2024. arXiv:2403.00476.

Yulin Luo, Chun-Kai Fan, Menghang Dong, Jiayu Shi, Mengdi Zhao, Bo-Wen Zhang, Cheng Chi, Jiaming Liu, Gaole Dai, Rongyu Zhang, Ruichuan An, Kun Wu, Zhengping Che, Shaoxuan Xie, Guocai Yao, Zhongxia Zhao, Pengwei Wang, Guang Liu, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. RoboBench: A comprehensive evaluation benchmark for multimodal large language models as embodied brain, 2025. URL https://arxiv.org/abs/2510.17801.

Wenxin Ma, Chenlong Wang, Ruisheng Yuan, Hao Chen, Nanru Dai, S. Kevin Zhou, Yijun Yang, Alan Yuille, and Jieneng Chen. CausalSpatial: A benchmark for object-centric causal spatial reasoning, 2026. URL https://arxiv.org/abs/2601.13304.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. OpenEQA: Embodied question answering in the era of foundation models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025. Blog post; April 2025.

A. David Milner and Melvyn A. Goodale. *The Visual Brain in Action*. Oxford University Press, 1995.

MuCR Team. MuCR: Multimodal causal reasoning benchmark: Challenging multimodal large language models to discern causal links across modalities, 2024. URL https://arxiv.org/abs/2408.08105. Findings of ACL 2025.

MVBench Team. MVBench: A comprehensive multi-modal video understanding benchmark, 2023. URL https://arxiv.org/abs/2311.17005.

NVIDIA. Cosmos-Reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025a.

NVIDIA. PBench: A physical ai benchmark for world models. https://research.nvidia.com/labs/dir/pbench/, 2025b. Repository.

NVIDIA Cosmos Team. Cosmos-Predict2: Video-to-world diffusion transformer. https://github.com/nvidia-cosmos/cosmos-predict2, 2025. GitHub repository.

Open3D-VQA Team. Open3D-VQA: A benchmark for embodied spatial reasoning with multimodal large language model in open space, 2025. URL https://arxiv.org/abs/2503.11094.

OpenAI. Video generation models as world simulators. OpenAI technical report, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

OpenAI. Update to GPT-5 system card: GPT-5.2. System card, 2025. URL https://openai.com/index/gpt-5-system-card-update-gpt-5-2/.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, March 2026. Model release; March 5, 2026.

Richard E. Passingham, Sara L. Bengtsson, and Hakwan C. Lau. The specialization of frontal cortex for self-initiated versus externally cued actions. *Trends in Cognitive Sciences*, 14(1):16–21, 2010.

Viorica Pătrăucean et al. Perception Test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. arXiv:2305.13786.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*, 2023. URL https://arxiv.org/abs/2308.11483.

Yu Qi, Haibo Zhao, Ziyu Guo, Siyuan Ma, Ziyan Chen, Yaokun Han, Renrui Zhang, Zitiantao Lin, Shiji Xin, Yijian Huang, et al. BEAR: Benchmarking and enhancing multimodal language models for atomic embodied capabilities. *arXiv preprint arXiv:2510.08759*, 2025.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. URL https://arxiv.org/abs/2412.07755.

Giacomo Rizzolatti, Leonardo Fogassi, and Vittorio Gallese. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2(9):661–670, 2001.

Massimo Rizzoli, Simone Alghisi, Seyed Mahed Mousavi, and Giuseppe Riccardi. Synthetic stimuli, real gains: Rethinking VLM fine-tuning through fully controlled data generation, 2026. URL https://arxiv.org/abs/2511.11440.

Neal J. Roese. Counterfactual thinking. *Psychological Bulletin*, 121(1):133–148, 1997.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision (ECCV)*, pp. 146–162, 2022.

Ziyao Shangguan et al. TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.23266.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024. URL https://arxiv.org/abs/2409.19256.

Elizabeth S. Spelke. *What Babies Know: Core Knowledge and Composition, Volume 1*. Oxford University Press, 2022.

Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

Gautam Sreekumar and Vishnu Naresh Boddeti. InPhyRe discovers: Large multimodal models struggle in inductive physical reasoning. *arXiv preprint arXiv:2509.12263*, 2025. URL https://arxiv.org/abs/2509.12263.

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

Tencent Hunyuan Team. HunyuanVideo: A systematic framework for large video generation model, 2024. URL https://arxiv.org/abs/2412.03603.

Shengbang Tong, Ellis Brown, Penghao Wu, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.16860; introduces CV-Bench.

Leslie G. Ungerleider and Mortimer Mishkin. Two cortical visual systems. In David J. Ingle, Melvyn A. Goodale, and Richard J. W. Mansfield (eds.), *Analysis of Visual Behavior*, pp. 549–586. MIT Press, Cambridge, MA, 1982.

VBench Team. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. CVPR 2024 Highlight; arXiv:2311.17982.

VLM4D Team. VLM4D: Towards spatiotemporal awareness in vision language models, 2025. URL https://arxiv.org/abs/2508.02095. ICCV 2025.

Haonan Wang, Meng Fang, Chao Du, Michael Qizhe Shieh, and Tianyu Pang. Scaling GUI agents with visual state transitions, 2026a. URL https://arxiv.org/abs/2607.24112.

Qi Wang, Zhipeng Zhang, Baao Xie, Xin Jin, Yunbo Wang, Shiyu Wang, Liaomo Zheng, Xiaokang Yang, and Wenjun Zeng. Disentangled world models: Learning to transfer semantic knowledge from distracting videos for reinforcement learning, 2025. URL https://arxiv.org/abs/2503.08751. ICCV 2025 Poster.

Siting Wang et al. SpatialViz-Bench: A cognitively-grounded benchmark for diagnosing spatial visualization in MLLMs. In *International Conference on Learning Representations (ICLR)*, 2026b. arXiv:2507.07610.

Yihao Wang et al. Orca: The world is in your mind. *arXiv preprint arXiv:2606.30534*, 2026c. URL https://arxiv.org/abs/2606.30534.

Archana Warrier, Dat Nguyen, Michelangelo Naim, Moksh Jain, Yichao Liang, Karen Schroeder, Cambridge Yang, Joshua B. Tenenbaum, Sebastian Vollmer, Kevin Ellis, and Zenna Tavares. Benchmarking world-model learning, 2025. URL https://arxiv.org/abs/2510.19788.

Daniel M. Wolpert, Zoubin Ghahramani, and Michael I. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995.

WorldPrediction Team. WorldPrediction: A benchmark for high-level world modeling and long-horizon procedural planning, 2025. URL https://arxiv.org/abs/2506.04363. ICML 2025 World Models Workshop.

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. ConceptMix: A compositional image generation benchmark with controllable difficulty. *arXiv preprint arXiv:2408.14339*, 2024. NeurIPS 2024 Datasets & Benchmarks Poster.

Shitao Xiao et al. OmniGen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. URL https://arxiv.org/abs/2409.11340.

Yunfei Xie, Yinsong Ma, Shiyi Lan, Alan Yuille, Junfei Xiao, and Chen Wei. Play to generalize: Learning to reason through game play. In *International Conference on Learning Representations (ICLR)*, 2026.

Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. Strengthened symbol binding makes large language models reliable multiple-choice selectors. *arXiv preprint arXiv:2406.01026*, 2024. URL https://arxiv.org/abs/2406.01026.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, June 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Position: Video as the new language for real-world decision making, 2024. URL https://arxiv.org/abs/2402.17139. ICML 2024.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence, 2025b. URL https://arxiv.org/abs/2505.23764.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. MindJourney: Test-time scaling with world models for spatial reasoning, 2025c. URL https://arxiv.org/abs/2507.12508.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1910.01442.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. Spatial mental modeling from limited views, 2025. URL https://arxiv.org/abs/2506.21458.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-world: World models in a closed-loop world, 2025a. URL https://arxiv.org/abs/2510.18135.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3D, 2025b. URL https://arxiv.org/abs/2503.22976.

Shiwei Zhang et al. I2VGen-XL: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. URL https://arxiv.org/abs/2311.04145.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities (COMFORT), 2024. URL https://arxiv.org/abs/2410.17385. ICLR 2025 Oral.

Ziang Zhang, Zehan Wang, Guanghao Zhang, et al. DSI-Bench: A benchmark for dynamic spatial intelligence. *arXiv preprint arXiv:2510.18873*, 2025c.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.03882. arXiv:2309.03882.

Fengzhe Zhou et al. PAI-Bench: A comprehensive benchmark for physical AI. *arXiv preprint arXiv:2512.01989*, 2025.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024. ICML 2024 Poster.

A Background and motivation

A.1 World modeling as a cognitive primitive

This appendix unfolds the developmental, neuroscientific, and AI-program evidence that world modeling is a cognitive primitive with an intrinsic structure, established independently of any benchmark, summarized in §1. The point of assembling this evidence is not decorative: it is what licenses treating the two axes of §3.1 as the primitive's own structure rather than a decomposition we impose, and it is why the scattered external failures noted in §1 can be read as manifestations of that structure.

Interpretation of the query formulation. The formulation in §2 connects forward and inverse internal models (Wolpert et al., 1995) with physical simulation (Battaglia et al., 2013) and counterfactual reasoning (Pearl, 2009). A shared trajectory mechanism makes the episode conditions explicit:

$$\tau_a = \mathcal{G}(s, a, u), \qquad \mathrm{start}(\tau_a) = s, \qquad \mathrm{end}(\tau_a) = s', \tag{1}$$

Here u collects the episode's remaining background conditions; evaluating an alternative action in $\mathcal{G}$ preserves the same initial state and background conditions. This notation describes the transition structure underlying the query distributions.

Forward, inverse, and removal recover the resulting state, action, and initial state, respectively; removal asks what the scene would look like had the action not occurred, so its target is the initial state. Substitution retains the factual action and observed endpoint as evidence for the alternative outcome; its input does not include the initial state. Outcome and cued prediction share a physical prediction target but differ in the supplied principle information, such as containment. Ordering and adjacency recover global and local temporal relations within the same transition, with adjacency defined among the sampled frames. In the physical and temporal families, the action description enters the condition only when it is given.

Developmental psychology. Predicting the visual consequence of an action and inferring the action behind a state change is documented in infants from the first months of life. Spelke's core-knowledge programme identifies object permanence, cohesion, and solidity as innate primitives (Spelke & Kinzler, 2007; Spelke, 2022), with violation-of-expectation paradigms (e.g., Baillargeon 1987) demonstrating sensitivity to physical events long before language acquisition. Counterfactual thought is a classical topic in cognitive science (Lewis, 1973; Roese, 1997) formalized in causal inference (Pearl, 2009). The temporal structure of a single action is represented within the forward model itself: forward models predict the time course of a movement, not only its endpoint, and this trajectory representation underlies online prediction and correction (Wolpert et al., 1995). Mental simulation of physical outcomes admits computational analogues in noisy-simulation accounts (Battaglia et al., 2013).

Cognitive neuroscience. Action-conditioned prediction maps onto the predictive-coding and forward-model accounts of motor and perception (Friston, 2010; Wolpert et al., 1995). Action-from-observation maps onto the mirror-neuron and action-understanding circuitry (Rizzolatti et al., 2001). The exo/endo dissociation distinguishes externally-triggered from self-initiated actions with distinct prefrontal signatures and behavioral dissociations (Haggard, 2008; Passingham et al., 2010; Khalighinejad et al., 2018). Within the visual system, the dorsal/ventral two-stream dissociation (Ungerleider & Mishkin, 1982; Goodale & Milner, 1992; Milner & Goodale, 1995) is the structural basis for the geometric/appearance partition of the perturbation probe (§3.3).

Recurring target across AI research programs. The primitive recurs under several labels: embodied planning requires forward rollout from current observation and a proposed action (Majumdar et al., 2024; Luo et al., 2025; EmbodiedBench Team, 2025; Du et al., 2023); counterfactual and

causal video reasoning requires reasoning over (s, a, s') triples under intervention on a (Foss et al., 2025; MuCR Team, 2024; CRIPP-VQA Team, 2022); mental simulation of "what-if" movement is the primary axis of recent spatial-reasoning benchmarks (Yin et al., 2025; VLM4D Team, 2025).

A.2 THE TWO AXES AS THE INTRINSIC STRUCTURE OF THE PRIMITIVE

The action and reasoning axes are not a decomposition we impose on the data; each enumerates distinctions that cognitive science documents as real and dissociable, so their cross-product is the primitive's own structure made measurable rather than a set of principal components fit to failures.

Reasoning axis. The operations one can perform over (s, a, s') are organized by three orthogonal structural properties of a predictive model. The first is causal depth: Pearl's hierarchy separates prediction from intervention ($\mathrm{do}(a)$) from counterfactual reasoning, a strictly nested ladder in which each rung expresses what the one below cannot (Pearl, 2009). Physical modeling occupies the predictive base, the passive dynamics of objects under intuitive physics (Spelke & Kinzler, 2007; Battaglia et al., 2013); causal dynamics occupies the intervention rung, the visual effect of a deliberate action; counterfactual reasoning occupies the top rung, the visual scene under an action that did not occur. The second property is direction of inference, the paired internal models of sensorimotor control: a forward model predicts the consequences of an action and an inverse model recovers the action from an observed change (Wolpert et al., 1995; Friston, 2010), and within the counterfactual rung the analogous split is removal of an action versus its substitution by an alternative. The third property is temporal granularity: because a forward model represents the time course of a single action rather than its endpoints alone (Wolpert et al., 1995), the temporal family probes that trajectory at two grains within one transition, coarse ordering of the intermediate states and fine adjacency of consecutive ones. The eight reasoning types are the cells of these three properties, and the causal-depth nesting further predicts a difficulty ordering, physical below causal below counterfactual, that the evaluation tests directly (Appendix F).

Action axis. The kinds of transition the primitive must represent track the core systems through which human cognition represents the physical world. The coarsest cut is agency: cognitive neuroscience separates externally triggered from self-initiated action with distinct prefrontal signatures (Haggard, 2008; Passingham et al., 2010; Khalighinejad et al., 2018), giving one exogenous type, the passive physics of objects, against agentic types. The exogenous type is organized by object core knowledge, with permanence, cohesion, and solidity as object primitives and gravity, support, inertia, containment, and collision as physical-event classes (Spelke & Kinzler, 2007; Battaglia et al., 2013). The agentic types are separated by the reference frame in which the action transforms the visual state: egocentric self-motion and object-centered transformation rely on different reference frames (Burgess, 2006; Epstein et al., 2017), and object-based mental rotation is behaviorally and neurally dissociable from egocentric perspective-taking, giving perceptive (egocentric viewpoint change) and inspective (object-centered examination) as distinct cells, with navigative (goal-directed locomotion) and manipulative (goal-directed action on an object) grounded in the agent core system. The five action types thus partition how a visual state can be caused to change, each a documented system rather than a label.

Why this makes the external failures legible. Because the two axes are the primitive's own degrees of freedom, any benchmark that exercises world modeling necessarily lands somewhere on their grid. Existing benchmarks each fix one axis and leave the other uncontrolled: intuitive-physics suites fix the exogenous, passive-prediction corner; spatial suites fix the perceptive and inspective actions while mixing operations; counterfactual video QA fixes the counterfactual operation while mixing actions; temporal suites fix the temporal operation. Their failures are therefore scattered marginals of one grid, which is why they look unrelated until the grid is drawn. WOVEN draws it by instantiating both axes jointly; the empirical correspondence between each external marginal and the WOVEN cell it falls in is quantified in §A.5.

A.3 FAILURE-MODE TRIANGULATION ACROSS EXTERNAL BENCHMARKS

This appendix unfolds the three-sub-domain triangulation summarized in §1.

Spatial sub-domain (dorsal-stream correspondence). Cognitive neuroscience distinguishes object identity, processed in the ventral "what" pathway, from spatial relations and viewpoint, processed in the dorsal "where" pathway (Ungerleider & Mishkin, 1982; Goodale & Milner, 1992; Milner & Goodale, 1995). The AI analogues of dorsal-stream competence are precisely the tasks on which MLLMs fail hardest: MindCube reports near-random MLLM performance on perspective-taking and "what-if" camera movement (Yin et al., 2025); VLM4D identifies systematic failure modes on translational and rotational motion (VLM4D Team, 2025); 3DSRBench documents degraded accuracy under uncommon viewpoints (3DSRBench Team, 2025); COMFORT finds VLMs fail to track reference frames across perspective change (Zhang et al., 2024). Each is, in world-modeling terms, a forward-prediction query under a self-initiated viewpoint or navigation action.

Physical sub-domain (Spelke "objects" core system). Developmental psychology documents infants' intuitive physics, object permanence, solidity, continuity, inertia, as core knowledge emerging in the first months of life (Spelke & Kinzler, 2007; Spelke, 2022; Baillargeon, 1987). AI counterparts consistently report MLLMs at or near chance: IntPhys 2 shows that state-of-the-art vision models approach 50% on violation-of-expectation items (IntPhys 2 Team, 2025); PhysBench reports persistent MLLM shortfalls against human baselines across four physical-reasoning domains (Chow et al., 2025); CRIPP-VQA probes counterfactual reasoning about implicit physical properties and finds comparable weakness (CRIPP-VQA Team, 2022).

Causal/counterfactual sub-domain. Direct video-reasoning analogues, CausalVQA for physically grounded counterfactual video QA (Foss et al., 2025) and MuCR for cross-modal causal-link identification (MuCR Team, 2024), together with the counterfactual-simulation dimension of WM-ABench (Huang Team and Contributors, 2025), report systematic MLLM weakness on intervention-on-action queries.

Parallel positive evidence: VGMs pass where MLLMs fail. Benchmarks evaluating VGMs directly on action-conditioned rollout report frontier models (Team Wan et al., 2025; CogVideoX Team, 2024; Tencent Hunyuan Team, 2024; NVIDIA Cosmos Team, 2025; Ball et al., 2025) passing well above chance: PBench evaluates world-model rollout across physics, robotics, and commonsense dimensions (NVIDIA, 2025b); VBench decomposes generation quality into 16 hierarchical dimensions with human-aligned metrics (VBench Team, 2024); ChronoMagic-Bench targets metamorphic temporal evolution (ChronoMagic-Bench Team, 2024). Downstream work assumes the world-modeling capability and exploits it as a black-box simulator: MindJourney couples a frozen VLM to a video-diffusion world model and reports an 8% boost on SAT (Yang et al., 2025c); DreamGen distills a video world model into robot policies that generalize to 22 new behaviors (DreamGen Team, 2025); InterPose uses VGM-hallucinated intermediate frames to outperform DUSt3R (Cai et al., 2025); RoboDreamer (Zhou et al., 2024) and VLP (Du et al., 2023) exploit the same rollout capability.

Existing world-model benchmarks do not close the gap. WM-ABench (Huang Team and Contributors, 2025), WorldPrediction (WorldPrediction Team, 2025), WorldTest (Warrier et al., 2025), and PBench (NVIDIA, 2025b) target world-modeling evaluation but leave the world modeling in MLLMs question open: they aggregate across capabilities without controlling interventions on a, treat VGMs rather than MLLMs as the subject, confine themselves to grid-world or text-only environments, or probe only a single sub-domain.

A.4 Cell-level prior-work coverage and downstream surfaces

This appendix collects the per-axis prior-work coverage discussion and the per-cell downstream-relevance argument deferred from §3. Together they substantiate the claim, made in §1, that prior benchmarks each touch a slice of world modeling but no single benchmark commits to the full action type $\times$ reasoning type cross-factoring.

Action axis: prior-work cell coverage. Existing benchmarks at the action-axis cell touch only subfragments of WOVEN's taxonomy. Embodied-AI evaluation suites are dominated by the *manipulative* and *navigative* cells (Cheng et al., 2025; EmbodiedBench Team, 2025; Majumdar et al., 2024; ECBench Team, 2025; Luo et al., 2025); spatial-reasoning benchmarks center on *perceptive* and

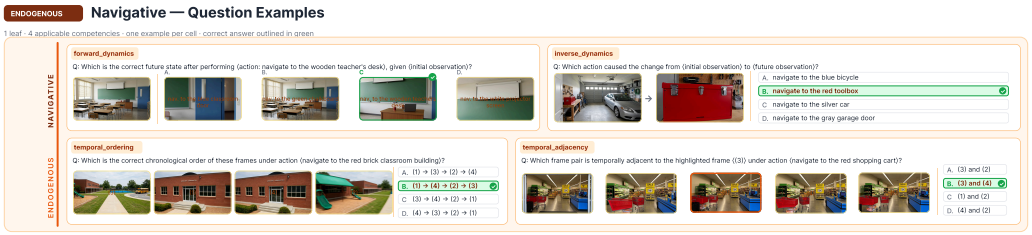

Figure 5: Navigative action type: representative QA items showing in-scene navigation targets across all four navigative reasoning types.

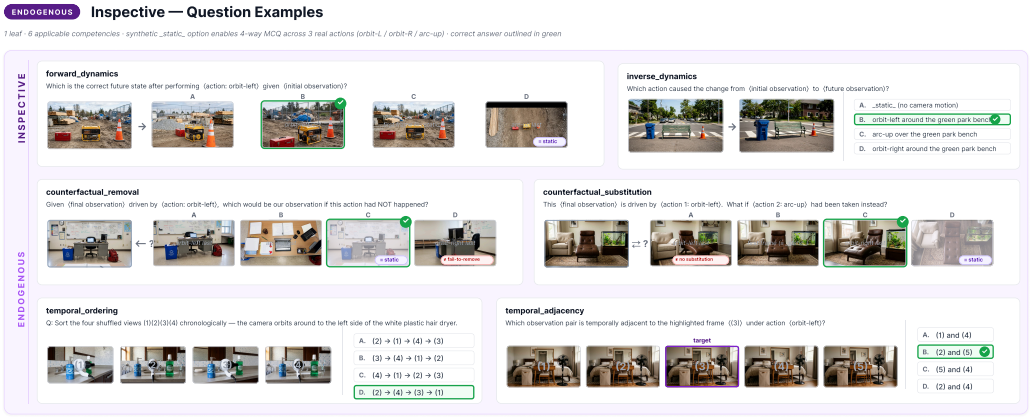

Figure 6: Inspective action type: representative QA items showing orbit-L / orbit-R / arc-up actions paired with all six inspective reasoning types.

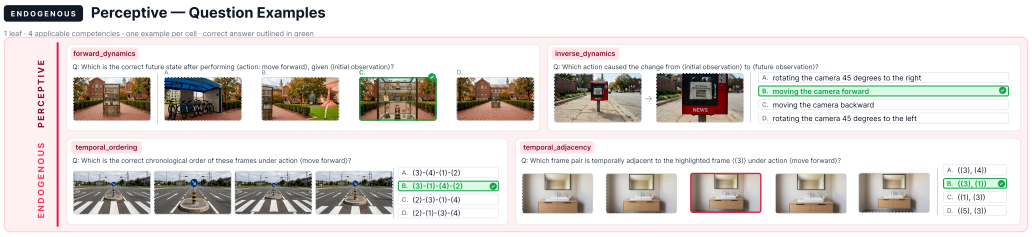

Figure 7: Perceptive action type: representative QA items across all four perceptive reasoning types (forward, inverse, t_ord, t_adj).

inspective (Yin et al., 2025; Yang et al., 2025c; Chen et al., 2025b; 3D-R1 Team, 2025; VLM4D Team, 2025; Open3D-VQA Team, 2025); physical-reasoning benchmarks operate primarily on *exogenous* (Chow et al., 2025; IntPhys 2 Team, 2025; Qi et al., 2025; NVIDIA, 2025b; Huang Team and Contributors, 2025). Each line implicitly probes a slice of world modeling but stops short of formalizing the action axis as such, and no single benchmark commits to a full cross-section. WOVEN's contribution at this axis is the explicit, exhaustive partition; failure attribution then becomes a question of which cell rather than whether the model is broken in some unspecified way.

Reasoning axis: prior-work cell coverage. Single reasoning types appear scattered across prior benchmarks: forward prediction is the standard probe in physical-modeling work (IntPhys 2 Team, 2025; Chow et al., 2025; NVIDIA, 2025b; Huang Team and Contributors, 2025); counterfactual reasoning is the dedicated focus of (Foss et al., 2025; MuCR Team, 2024; Zhang et al., 2024) and the concept-mixing variant of (Wu et al., 2024); temporal coherence is unpacked in video-generation evaluation (VBench Team, 2024; ChronoMagic-Bench Team, 2024; Ling et al., 2025) and in video-language understanding suites (MVBench Team, 2023; Yang et al., 2024). None co-instantiate all four families on a shared, controlled action axis, leaving the cross-family interaction structure unmeasured. WOVEN holds the four families jointly factored against the action axis of §3; the resulting (action type, reasoning type) cells are the elementary units of fine-grained diagnosis.

Action axis: downstream surfaces. Each action type tracks a distinct downstream surface: *exogenous* feeds physical-AI safety (Chow et al., 2025; NVIDIA, 2025b); *perceptive* and *inspective* feed visual spatial-mental-modeling (Yin et al., 2025; Yang et al., 2025c; Chen et al., 2025b; 3D-R1 Team, 2025); *navigative* and *manipulative* feed embodied agents and visual planners (Cheng et al., 2025; EmbodiedBench Team, 2025; Majumdar et al., 2024; Du et al., 2023; DreamGen Team, 2025). A failure localized to one action type identifies which downstream surface the model is most unfit for.

Reasoning axis: downstream surfaces. Forward dynamics is the central capability behind embodied planning and visual world-model rollouts (Du et al., 2023; DreamGen Team, 2025; Zhou et al., 2024; Hafner et al., 2023); inverse dynamics underlies action recognition and learning-from-observation (Rizzolatti et al., 2001); counterfactual templates underlie causal video reasoning (Foss et al., 2025; Zhang et al., 2024); physical-modeling templates underlie physical-AI safety (Chow et al., 2025; NVIDIA, 2025b); temporal-coherence templates underlie video-understanding pipelines (MVBench Team, 2023). As with the action type axis, failure localized to a reasoning type family identifies the downstream surface most affected.

Scene axis: scale and breadth. A central premise of the methodology is that scene diversity is a *controlled* dimension to be held fixed while the action and reasoning axes are perturbed. Real-world video corpora cannot guarantee both the action-axis controllability of §3 and full scene-axis breadth simultaneously. Simulators, prevalent in embodied-AI benchmarks (Cheng et al., 2025; EmbodiedBench Team, 2025; Majumdar et al., 2024; ECBench Team, 2025; Luo et al., 2025) and physical-AI benchmarks (Huang Team and Contributors, 2025; IntPhys 2 Team, 2025), deliver controllability but commit to a small number of pre-built environments with a documented visual-domain gap. A VGM stimulus engine delivers both: under fixed scene specification, the rollout produces visually realistic stimuli at every (action type, scene type) cell, and across scenes, lexical scene specification swaps cleanly without further pipeline change. This is the most direct realization of the VGM-as-stimulus-engine argument of §3.1.

A.5 Empirical correspondence between external failures and WOVEN cells

This appendix makes the claim of §A.2 quantitative: the failures external benchmarks report in isolation are the marginals of WOVEN's grid, so the cell where each external failure falls should be hard in WOVEN exactly when the external task is hard. We establish the correspondence in two layers, a conceptual mapping of each benchmark to its WOVEN cell and three statistical tests of that mapping, all computed from published baselines and our existing 38-model evaluation without additional experiments.

Conceptual mapping. Table 4 assigns each external benchmark, at the sub-task level where one is provided, to the WOVEN (action type, reasoning type) cell it instantiates. The mapping is many-to-

one by construction: several external benchmarks fall in the same WOVEN cell because each isolates one axis while WOVEN crosses both, which is the precise sense in which WOVEN unifies rather than duplicates them.

Table 4: Conceptual mapping of external benchmarks to the WOVEN cells they instantiate. Several benchmarks fall in one cell (many-to-one), the sense in which WOVEN unifies scattered marginals rather than duplicating any one of them. **Stat** marks which correspondence test the pair primarily supports: **M** model-level rank correlation, **D** cell-level difficulty correlation.

External benchmark (sub-task)	WOVEN cell / axis	Stat
VSI-Bench (Yang et al., 2025a) (relative direction)	perceptive/inspective × forward	M, D
VSI-Bench (appearance order)	perceptive × temporal	M, D
VSI-Bench (route planning)	navigative × forward	M, D
ViewSpatial-Bench (camera vs. human)	manipulative view-axis (ego/allo)	M, D
3DSRBench (orientation, uncommon view)	inspective × forward; orientation perturbation	M, D
VLM4D (rotational/directional motion)	inspective/perceptive rotation	D
MindCube (perspective-taking, what-if)	inspective × {forward, counterfactual}	M, D
IntPhys 2 (object principles)	exogenous × physical	D
PhysBench (physical reasoning)	exogenous × {physical, forward}	M, D
CLEVRER (predictive, counterfactual)	exogenous × {physical, counterfactual}	D
CausalVQA (counterfactual, hypothetical)	manipulative × counterfactual	D
TemporalBench (event order/adjacency)	all actions × temporal	M, D

Statistical tests. Three complementary analyses, all from published numbers and our evaluation, test whether the mapping holds empirically. (i) *Model-level rank correlation.* Our cohort and the external benchmarks share a backbone of commonly evaluated models, principally the Qwen2.5-VL ladder ($3/7/32/72$B) and Gemma-3 ($4/12/27$B), which appear in most recent spatial and physical benchmarks, often with the sub-task breakdowns that map to WOVEN cells. For each mapped pair we compute the Spearman correlation between the shared models' WOVEN-cell accuracy and their external sub-task accuracy; the Qwen2.5-VL ladder is the strongest anchor because its four scale points trace a controlled difficulty gradient, so a monotone co-trend is itself evidence. As the shared set per benchmark is small (about 4 to 8 models), this layer is corroborating. (ii) *Cell-level difficulty correlation.* For each mapped pair we take the WOVEN-cell difficulty (frontier accuracy or human-frontier gap, from our evaluation) and the external sub-task difficulty (published frontier-vs-human gap), and correlate the difficulty profile across all pairs. This needs no model overlap and tests the structural claim directly, whether the WOVEN cells that are hard are the external tasks that are hard, and is the statistical backbone of the correspondence. (iii) *Common-reference normalization.* For benchmarks reporting few of our models, we normalize both sides to a shared anchor, performance relative to Qwen2.5-VL-72B or to the human ceiling, so that thin-overlap benchmarks still enter the cell-level correlation on a common scale.

Reading the result. A positive correspondence supports the central claim that WOVEN's taxonomy is the primitive's intrinsic structure and reflects the same failure modes external work reports, while the many-to-one mapping preserves WOVEN's distinct contribution: the correspondence is convergence onto one grid rather than redundancy with any benchmark on it.

A.6 Extended related work

WOVEN sits at the intersection of three programs; cell-level prior-work coverage and downstream-surface mapping are in Appendix A.4.

World modeling. The term *world model* spans heterogeneous formal commitments: latent predictive architectures (DreamerV3 (Hafner et al., 2023), V-JEPA 2 (Assran et al., 2025)); generative pixel-rollout simulators (Sora (OpenAI, 2024), Genie 3 (Ball et al., 2025), Cosmos-Predict2 (NVIDIA Cosmos Team, 2025), HunyuanVideo (Tencent Hunyuan Team, 2024); quality benchmarks (VBench Team, 2024; ChronoMagic-Bench Team, 2024; Ling et al., 2025)); symbolic-environment learners (Warrier et al., 2025; Chen et al., 2025b); and procedural-planning formulations (WorldPrediction Team, 2025). Recent surveys note that no single operationalization dominates (Li et al., 2025b), and

existing world-model evaluation benchmarks inherit this fragmentation (Huang Team and Contributors, 2025; NVIDIA, 2025b; Zhang et al., 2025a; Wang et al., 2025). We focus on action-conditioned visual transition reasoning over (s, a, s'), using a common transition representation to study the shared and distinct supervision requirements of visual reasoning tasks.

VGMs as world simulators. The Sora technical report (OpenAI, 2024) introduced the "video generation as world simulator" framing, since adopted by CogVideoX (CogVideoX Team, 2024), HunyuanVideo (Tencent Hunyuan Team, 2024), the Wan family (Team Wan et al., 2025), Cosmos-Predict2 (NVIDIA Cosmos Team, 2025), and Genie 3 (Ball et al., 2025), with Yang et al. (2024) articulating the strongest version of the claim. Empirical support comes from direct rollout measurement (NVIDIA, 2025b; VBench Team, 2024; ChronoMagic-Bench Team, 2024; Ling et al., 2025) and downstream exploitation as black-box simulators (Yang et al., 2025c; DreamGen Team, 2025; Cai et al., 2025; Zhou et al., 2024; Du et al., 2023; Wang et al., 2025); closed-loop evaluation (Zhang et al., 2025a) additionally validates that VGM-rollout *controllability* drives embodied utility. These findings motivate our use of Wan2.2-I2V-A14B (Team Wan et al., 2025) to generate diverse, controllable transitions for structured supervision.

Multimodal reasoning benchmarks. Independent literatures document MLLM failures along axes that each touch a slice of action-conditioned visual world modeling without naming it: *counterfactual video reasoning* (CausalVQA (Foss et al., 2025), MuCR (MuCR Team, 2024), CRIPP-VQA (CRIPP-VQA Team, 2022)); *post-event physical prediction* (IntPhys 2 (IntPhys 2 Team, 2025), PhysBench (Chow et al., 2025), BEAR (Qi et al., 2025)); *viewpoint-conditioned spatial inference* (MindCube (Yin et al., 2025), VLM4D (VLM4D Team, 2025), 3DSRBench (3DSRBench Team, 2025), COMFORT (Zhang et al., 2024), Open3D-VQA (Open3D-VQA Team, 2025), 3D-R1 (3D-R1 Team, 2025)); and *video understanding plus embodied evaluation* (MVBench (MVBench Team, 2023), ECBench (ECBench Team, 2025), OpenEQA (Majumdar et al., 2024), EmbodiedBench (EmbodiedBench Team, 2025), EmbodiedEval (Cheng et al., 2025), RoboBench (Luo et al., 2025); on the training side, EMCompress (Fan et al., 2026)). These four families have been pursued in isolation with disjoint distractor designs, yet their failure modes factor through a shared structural deficit (Appendix A.3 makes the triangulation explicit). WOVEN builds on these diagnostic insights by expressing reasoning tasks over a common (s, a, s') representation and varying the semantic components of supervision, enabling controlled study of shared training requirements and the organization of transfer.

B Benchmark Construction Details

B.1 Construction Pipeline Details

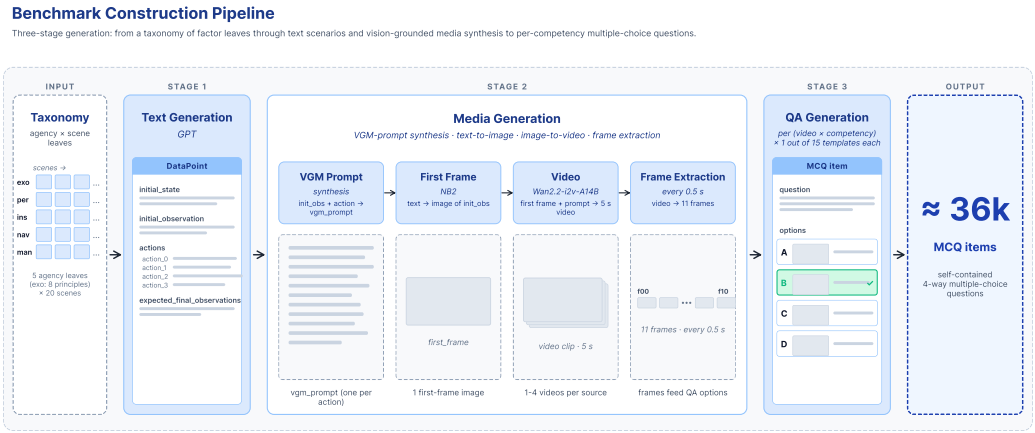

Figure 8: Construction pipeline: stage-1 textual specification, stage-2 VGM rollout (Wan2.2-I2V-A14B), stage-3 typed-distractor MCQ assembly, with quality-control filters at each stage.

This appendix documents the technical realization of the three-stage construction pipeline summarized in §3.3.

Stage 1 (text specification). For each (action type, scene) combination we generate a textual specification consisting of: (i) the initial environment state and the first-person observation of s_t; (ii) one or more action descriptions with deterministic identifiers; and (iii) the expected post-action state. Action vocabularies are action-type-specific: *exogenous* draws one principle per item from the closed inventory of 8 principles; *perceptive* draws all 4 fixed camera-translation/rotation actions; *inspective* draws all 3 fixed camera-orbit actions; *navigative* draws 4 in-scene targets; *manipulative* draws 4 object interactions consistent with the scenario-group definition (§3).

Stage 2 (media synthesis). The first frame is synthesized from the initial-observation text by an image generative model. The post-action 5-second video at 24 fps is then synthesized by Wan2.2-I2V-A14B (Team Wan et al., 2025) conditioned on (s_t, a). We extract 11 uniformly spaced frames per video, indexed $f_{00}, \ldots, f_{10}$, with $f_{00} = s_t$ and $f_{10} = s_{t+1}$. Intermediate frames are reserved for temporal-coherence reasoning types.

Stage 3 (MCQ assembly). Four-way MCQs are assembled from the resulting frames per the reasoning-type-specific distractor strategies of §3. Each item additionally exposes structured metadata, taxonomy path, action-type-specific auxiliary fields (e.g., manipulated object, paired-variant indices, exogenous principle), per-frame paths, and the deterministic distractor type label where applicable, to support fine-grained downstream analysis.

Quality control. Media-level rule-based and model-based filters are applied to first frames and videos to reject outputs that fail per-action-type physical plausibility checks; layout and orientation perturbations carry an additional automated edit-faithfulness, scene-consistency, and artifact-absence check. Quality control operates only on media; it does not gate question generation.

B.2 Why action type over motor command

Existing world-model action taxonomies organize a by *embodiment*: discrete vs. continuous in latent-policy world models (Hafner et al., 2023), joint/wheel/keyboard by application domain in driving and robotics WMs, or DOF-counted action vectors in egocentric WMs. These conventions are appropriate for embodiment-bound WM training, where the action space is fixed by the agent's effectors and the state space by the simulator. WOVEN instead evaluates a cognitive primitive across heterogeneous embodiments, bodiless cameras, simulated humans, simulated humanoids, externally-caused physical events. An embodiment-centric taxonomy would force inter-action-type alignment via low-level motor parameters that have no shared semantics across these settings (a wheel-radian and a translation-meter are not commensurable diagnostics for world modeling). Organizing by *action type*, "who initiates the change and how it changes the visual state", operationalizing the cognitive-science notion of *agency* (Haggard, 2008; Passingham et al., 2010; Khalighinejad et al., 2018), preserves cross-embodiment comparability while remaining cognitively grounded. We retain the cognitive-science terms *agency* (for the action axis) and *competency* (for the reasoning axis) only here, as the interface that motivates the taxonomy; the benchmark itself is named by action type, reasoning type, and scene type.

B.3 Scene set

The indoor set is {*living_room, bedroom, bathroom, restaurant, kitchen, classroom, office, supermarket, gym, hospital*}; the outdoor set is {*crossroad, sidewalk, playground, farm, garage, park, construction_site, yard, school_campus, beach*}. The held-out OOD pair is (*restaurant*, *beach*).

B.4 Cognitive Science support of Exogenous principles

The eight exogenous principles in WOVEN are organized in two cog-sci-grounded groups, each instantiating a separate strand of intuitive-physics evidence.



Figure 9: Exogenous *physical events* (gravity, support, inertia, containment, collision): per-principle illustration showing the action-conditioned violation-of-expectation distractor design.

Group 1: object properties (*permanence*, *cohesion*, *solidity*). These are the three foundational primitives of Spelke's core-knowledge inventory (Spelke & Kinzler, 2007; Spelke, 2022). *Permanence*: an object continues to exist when occluded; the diagnostic violation is dis-appearance during occlusion (Baillargeon, 1987). *Cohesion*: an object's parts move together as a connected whole; the diagnostic violation is mid-event part-splitting. *Solidity*: distinct objects do not interpenetrate; the diagnostic violation is one object passing through another. Each principle has its own Stage 1 prompt template in the construction pipeline, ensuring coverage of the principle-specific scene type and event type.

Group 2: physical events (*gravity*, *support*, *inertia*, *containment*, *collision*). These are five principal physical-event classes routinely catalogued in computational intuitive-physics work (Battaglia et al., 2013) and in violation-of-expectation studies (Baillargeon, 1987). *Gravity*: unsupported objects fall in expected directions and trajectories. *Support*: an object resting on a surface remains supported absent additional cause. *Inertia*: motion persists or stops in cause-consistent ways. *Containment*: contained objects move with their containers and constrain motion. *Collision*: contact events transfer motion in expected ways.

The two groups jointly span the principles of object-state-as-property (Group 1) and object-state-as-event (Group 2), giving the exogenous action type a clean cross-section. Each principle is tied to its own IEM violation-of-expectation distractor inventory in `outcome_prediction` and `cued_prediction` (§3).

B.5 Manipulative cross-factoring

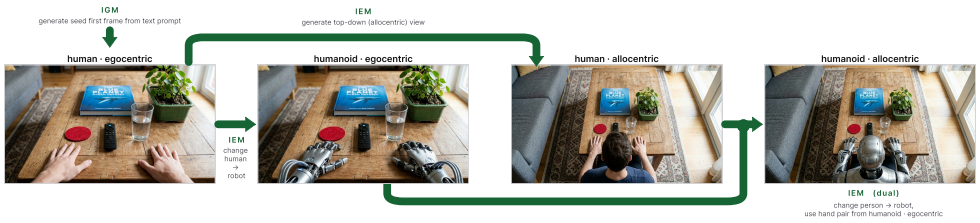

Figure 11: Manipulative cross-factoring: {human, humanoid} × {egocentric, allocentric} derivation showing the four pair-mate cells used by F-H pair-mate independence and ENHANCE F5 cross-action-type analyses.

Why 2×2. The cross-factoring along **subject** $\in \{\text{human}, \text{humanoid}\} \times$ **view** $\in \{\text{egocentric}, \text{allocentric}\}$ is motivated by two complementary cog-sci traditions. The subject axis distinguishes biologically natural actors from kinematically constrained robotic actors, isolating an MLLM's robustness to embodiment-shape variation in action-recognition pathways (Rizzolatti et al., 2001; Gibson, 1979). The view axis instantiates the egocentric/allocentric reference-frame distinction, a foundational dichotomy in spatial cognition (Burgess, 2006; Kessler & Thomson, 2010; Epstein et al., 2017): egocentric views ground the action in the actor's effector frame, while allocentric views ground it in a third-person scene frame. Cross-factoring isolates whether world-modeling competence is invariant to actor identity at fixed view, invariant to view at fixed actor, or jointly entangled across the two.

Generation pipeline. Each manipulative scenario specifies five non-rotationally-symmetric, color-distinct everyday objects within a fixed scene. Four of them are designated as targets of four *independent* ("parallel-universe") actions; the fifth, the bystander, is held visible and unmoved across all four actions, anchoring the geometric/appearance perturbations of §3.3. The four paired (subject, view) variants share the same scenario specification, the same five-object configuration, and the same four action verbs, differing only in actor identity and camera frame. Per scenario, this

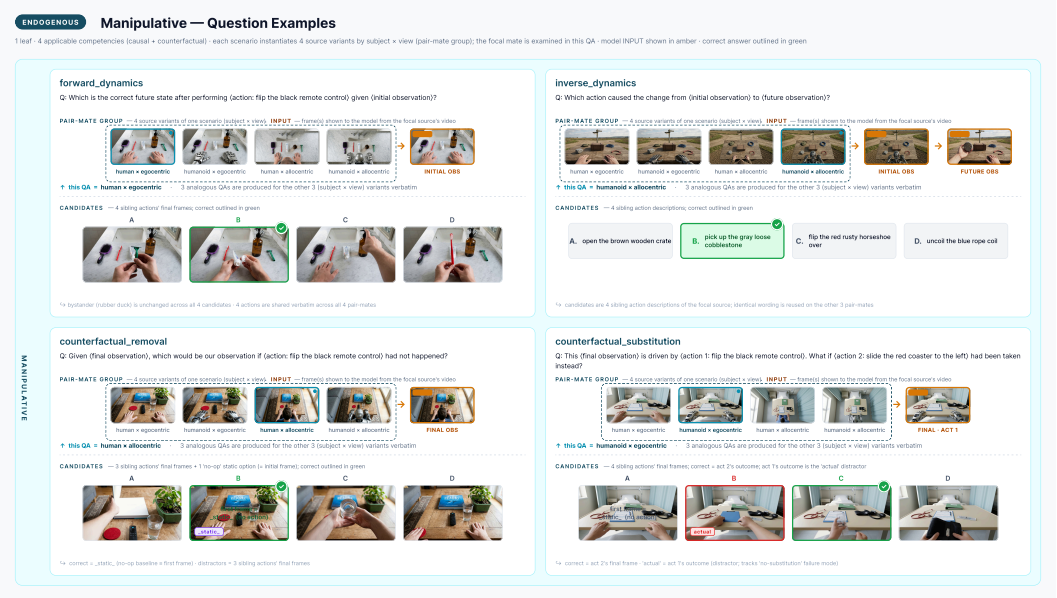

Figure 12: Manipulative action type: representative QA items showing object-interaction A/B/C/D action structure with all four manipulative reasoning types (forward, inverse, cf_sub, cf_rem).

yields four sources that share underlying content but differ in embodiment and frame-of-reference; within-scenario consistency tests are the basis for the embodiment-vs-frame-of-reference dissociation analysis in Appendix F.8. The construction pipeline is multi-stage and inter-stage-dependent, subject-and-view selection → five-object selection (with diversity and rotational-asymmetry constraints) → action-verb selection (four mutually-independent verbs targeting four different objects) → bystander designation → per-(subject, view)-variant Stage 2 media synthesis with view-dependent prompt instructions.

Split-time invariant. The four-way scenario-group is the atomic unit of partitioning at split time, ensuring all four paired variants land in the same split. Validation is carved at the scenario-group level (10% of groups, seed 42); training/test partition is at the source level within remaining groups. This prevents the most direct visual-leakage failure mode, a humanoid-allocentric variant in train and the matching human-egocentric variant in val/test sharing the underlying scene, object configuration, and action verbs.

B.6 Splits: stratification and pool sizes

Partitioning is performed at the pre-MCQ-assembly stage. For each action type × scene stratum, the underlying generation pool is partitioned 80%/20% into a train candidate pool and a test candidate pool. Manipulative is partitioned at the four-way scenario-group level (Appendix B.5). The validation set is then carved as ∼ 10% of the train pool, stratified by action type and exogenous principle (random seed 42). The scene-OOD test set is generated by an identical Stage 1–3 pipeline restricted to the two held-out scenes (*restaurant*, *beach*). The perturbation test is generated by the same pipeline from ID-test sources whose initial frames are edited with a geometric or appearance change before rollout (§3.3, Appendix B.9). The resulting per-split item counts are reported in Table B.6 (Appendix).

B.7 Action type–reasoning type adjacency

Not every (action type, reasoning type) cell is instantiated in the benchmark: counterfactual reasoning applies where the action meaningfully alters scene state, and physical modeling is exogenous-specific. The full instantiation matrix is given in Table 5.



Figure 13: WOVEN splits: `train_pool` (used only for ENHANCE), `test_id` (n=6,228), `test_perturbation_id` (n=1,116), and held-out OOD scenes (`beach`, `restaurant`). The OOD set is reserved for future fine-tuning evaluations.

Table 5: Action type–reasoning type adjacency: which reasoning types instantiate items for each action type. ✓ = instantiated;, = not instantiated.

Action type	outcome_pred.	cued_pred.	temporal_ord.	temporal_adj.	forward_dyn.	inverse_dyn.	cf_removal	cf_substitution
Exogenous	✓	✓	✓	✓	—	—	—	—
Perceptive	—	—	✓	✓	✓	✓	—	—
Inspective	—	—	✓	✓	✓	✓	✓	✓
Navigative	—	—	✓	✓	✓	✓	—	—
Manipulative	—	—	—	—	✓	✓	✓	✓

Why each cell is or is not instantiated. *Physical Modeling* (`outcome_prediction`, `cued_prediction`) is exogenous-specific: only externally-driven physical events admit a violation-of-expectation outcome judgement that does not depend on agent choice. *Causal Dynamics* (`forward_dynamics`, `inverse_dynamics`) is endogenous-specific: the source datapoint must contain multiple sibling actions for the cross-action MCQ to admit non-trivial distractors, which exogenous's single-action structure does not afford. *Temporal Coherence* (`temporal_ordering`, `temporal_adjacency`) applies wherever visual change unfolds over time within a single rollout, which excludes manipulative whose inter-frame dynamics are dominated by hand-occlusion artifacts unrelated to scene-state evolution. *Counterfactual Reasoning* (`counterfactual_substitution`, `counterfactual_removal`) requires the action to meaningfully alter scene state, producing an s_{t+1} informatively different from s_t, which holds for inspective (camera viewpoint changes the visible object configuration) and manipulative (object configuration itself changes), but not perceptive (camera-only, scene unchanged) or navigative (scene-positional change without state change).

B.8 Distractor design: scope of typed labeling

The scope of the typed-label property of distractors (§3) varies by reasoning type family.

Fully-typed distractors. For reasoning types that draw distractors from sibling actions in action types with a small fixed action vocabulary, *perceptive* (4 fixed actions: translation and rotation) and *inspective* (3 fixed actions: orbit-L, orbit-R, arc-up; plus a static-camera padding option to bring the option set to four), each distractor carries a deterministic, globally trackable label such as `rotate_left`. Item-level error attribution to specific failure modes is therefore directly cross-comparable across items within an action type.

Family-level-typed distractors. For reasoning types whose action vocabulary varies per item, *navigative* (the four targets are scene-specific) and *manipulative* (the four object interactions are scenario-specific), distractors are still drawn from the closed sibling-action set, but their labels are not globally trackable across items. Failure-mode attribution is informative at the family level (e.g., "selected a sibling action that shares object identity but differs in motion direction") but does not align item-by-item.

Closed-class distractors. For temporal reasoning types (temporal ordering and temporal adjacency), distractors are drawn from a closed inventory of permutation classes (swap, off-by-one, endpoint-swap, adjacent-swap variants). Each class corresponds to a specific temporal-cognition error mode and is family-level trackable.

IEM-edit distractors. For physical-modeling reasoning types (outcome prediction and cued prediction), distractors are constructed by editing s_{t+1} via Gemini 3.1 Flash Image (Google DeepMind, 2026) from a closed inventory of physically-violating edit types (object-property and physical-event violations). Each edit type is globally trackable and aligned with a specific violation-of-expectation hypothesis. The correct outcome frame is also passed once through the same editor under a null instruction that asks for the same image with nothing changed, so every image option of a physical-modeling item has been through the editor exactly once (all $1{,}488$ exogenous sources). All four options of an item are released at the same size and in the same format.

Editing-artifact control. The null pass gives the correct option and the distractors the same editing history only if the editor re-renders the whole image rather than inpainting the edited region. We test this on all 864 edited distractor frames of the test set. Each edited frame is resized to the size of its source frame and compared with the source in the half of the image farthest from the edit; the control is the source frame resized to the editor's output size and back, which is what an untouched region looks like after resampling alone. Table 6 reports the result. The far-region residual is about $20\times$ the resampling floor, present in every frame and every violation type, spatially structured (lag-1 autocorrelation 0.58), and not reduced by any global shift, so it is neither a resampling nor an alignment effect. The null-edited correct options show the same footprint (far-region residual median 2.0, whole-image median 3.4). The editor therefore re-renders the full image regardless of

the instruction, and the correct option and the distractors share the same editing process everywhere in the image.

We further test whether any residual difference between a null pass and a violation edit is detectable. The test is designed to see editing artifacts only and no task-relevant content. We randomly sample 166 test sources and use their 278 distractors that are edits of the outcome frame, since only these share unedited content with the correct option. A probe sees only 28×28 patches from the half of each image farthest from the edited region, where the content is identical across options, taken at the same positions for the correct option and its distractors. The probe is a six-layer convolutional network (3×3 convolutions with 32–128 channels, batch normalization, ReLU, global average pooling, and a linear output), trained with AdamW (learning rate 2×10^{-3}, weight decay 10^{-4}, one-cycle schedule) for 1,500 steps on batches of 1,024 patches balanced across labels, with random horizontal flips. Evaluation uses five-fold cross-validation over sources with three training seeds, so no source appears in both training and test folds. We report patch-level balanced accuracy (chance 50%) and image-level AUC from each image's mean patch logit (chance 0.50), against a permutation null that relabels one randomly chosen option per source as correct and reruns the same pipeline (10 seeds). Table 7 shows no detectable signal above the null. As a positive control, we run the same probe, cross-validation, and permutation null on the same 166 option sets with the correct option replaced by the original outcome frame that has not been through the editor (Table 8). There the correct option is clearly separable (55.9% balanced accuracy, AUC 0.718; z=11.3 and 14.2), so the probe has the power to detect editing artifacts when they exist. Without the null edit the correct option is separable from the distractors; with it, it is not. The null pass therefore removes the editing signature that would otherwise mark the correct option, and the only systematic difference between the correct option and a distractor is the edited content itself.

Table 6: The editor re-renders the whole image. Far-region residual between edited distractor frames and their source frames (test set, 864 edited frames; gray levels on a 0–255 scale).

Measurement	Value
Mean absolute residual, far half of the image, edited vs. source	3.8 (median 1.9, p90 8.0)
Same measurement for the resize round-trip of the source frame (control)	0.17
Ratio edited / control, per frame	median 20.6×; 0 of 864 below 1.5×
Lag-1 spatial autocorrelation of the far-region residual	0.58
Reduction of the far-region residual by the best global shift (±3 px, 0.5 px steps, 80 frames)	0%
Ratio edited / control by violation type (23 types)	10×–36×, every type

Table 7: Editing-artifact probe on the released data (test set, 166 sources, 166 correct options, 278 distractors).

Metric	Observed	Permutation null (10 seeds)	z
Patch-level balanced accuracy (%)	51.1	50.9 ± 0.4	0.5
Image-level AUC	0.532	0.520 ± 0.008	1.4

Table 8: Positive control: the same probe on a constructed subset without the null edit, with the correct option replaced by the original, unedited outcome frame (test set, 166 sources, 166 correct options, 278 distractors).

Metric	Observed	Permutation null (10 seeds)	z
Patch-level balanced accuracy (%)	55.9	50.5 ± 0.5	11.3
Image-level AUC	0.718	0.508 ± 0.015	14.2

Per-cell traceability coverage. Table 9 summarizes which (action type, reasoning type) cells receive fully-typed (item-level globally trackable) versus family-level (closed-class but item-dependent) distractors.

Table 9: Distractor-type coverage by (action type, reasoning type) cell. **F** = fully-typed sibling-action distractors, globally trackable item-by-item. **C** = closed-class sibling-action distractors, family-level trackable. **I** = IEM-edit distractors, globally trackable. **S** = sequence-permutation distractors, family-level trackable.: = cell not instantiated.

	outcome_pred.	cued_pred.	temporal_ord.	temporal_adj.	forward_dyn.	inverse_dyn.	cf_removal	cf_substitution
Exogenous	I	I	S	S	—	—	—	—
Perceptive	—	—	S	S	F	F	—	—
Inspective	—	—	S	S	F	F	F	F
Navigative	—	—	S	S	C	C	—	—
Manipulative	—	—	—	—	C	C	C	C

A cell labeled **F** or **I** contributes per-item failure-mode counts that license the per-cell binomial, χ^2, and two-proportion-z tests of §3.3. A cell labeled **C** or **S** contributes to family-level distributional tests but not to per-item attribution. The fully-typed cells cover all of perceptive's and inspective's causal-dynamics + counterfactual structure, i.e., the action types on which the perturbation probe (§3.3) also concentrates, ensuring that the most diagnostically loaded subset of the benchmark is also the subset with the strongest statistical instrumentation.

B.9 Perturbation probe: construction and scope

Construction. The perturbation test follows the same Stage 1–3 pipeline as the other splits. For each source, the designated object in the initial frame is edited with a geometric (layout or orientation) or an appearance change, the edited frame passes the edit-faithfulness, scene-consistency, and artifact-absence checks of Appendix B.1, and the VGM is rolled out again from the edited frame under the same action. Questions are assembled from the new rollout as for the in-distribution test, so each perturbed item shares its action and query with an unperturbed item and differs only in the initial state.

Scope: inspective + manipulative only. Geometric and appearance perturbations are well-defined when the action operates on a discrete object whose perturbation can be cleanly localized. *Inspective* actions orbit/arc around a designated main object, and the perturbation targets that main object. *Manipulative* actions act on one of four target objects per scenario; the perturbation targets the bystander object that is held visible and unmoved across all four actions (§3, Appendix B.5), so the perturbation does not disturb the action's primary target. For *perceptive* and *navigative*, no such designated object exists and the action changes the whole frame, so a localized edit of the initial frame has no stable target whose change the prediction should track. *Exogenous* similarly lacks a clean target: the relevant "state" is the multi-object configuration jointly, and the perturbation would re-define the principle being probed. Inspective and manipulative thus offer the cleanest diagnostic signal for the dorsal/ventral state-axis dualization the probe targets.

C Evaluation cohort and human baseline

C.1 Cohort: model inventory

We evaluate 38 frontier MLLMs spanning closed-source frontier, open-source dense ladders, and MoE families (Table 10). Where vendors do not publish parameter counts (e.g., GPT-5.x, Qwen-VL-Max, Intern-S1, MiniMax, Seed, GLM-4.6V), the column is marked N/A; where they publish active vs. total separately (e.g., 30B-A3B = 30B total / 3B active), we list both.

Table 10: T01 Model inventory.

Model	Family	Params
GPT-5.4	GPT	N/A
GPT-5.2	GPT	N/A
Qwen-VL-Max	Qwen-max	N/A
Qwen2.5-VL-3B	Qwen2.5	3B
Qwen2.5-VL-7B	Qwen2.5	7B
Qwen2.5-VL-32B	Qwen2.5	32B
Qwen2.5-VL-72B	Qwen2.5	72B
Qwen3-VL-8B-Instruct	Qwen3	8B
Qwen3-VL-30B-A3B	Qwen3	30B-A3B (MoE)
Qwen3-VL-32B-Instruct	Qwen3	32B
Qwen3-VL-235B-A22B	Qwen3	235B-A22B (MoE)
Intern-S1	Intern	N/A
Intern-S1-Mini	Intern	N/A
Intern-S1-Pro	Intern	N/A
InternVL3.5-241B-A28B	Intern	241B-A28B (MoE)
InternVL3.5-Latest	Intern	N/A
Gemma-3-4B	Gemma3	4B
Gemma-3-12B	Gemma3	12B
Gemma-3-27B	Gemma3	27B
Gemma-4-26B-A4B	Gemma4	26B-A4B (MoE)
Gemma-4-31B	Gemma4	31B
Llama-4-Scout	Llama4	Scout (109B-MoE)
Llama-4-Maverick	Llama4	Maverick (400B-MoE)
Mistral-Small-3.2	Mistral	24B
Mistral-Large-2512	Mistral	N/A
Pixtral-Large-2411	Mistral	123B
MiniMax-01	MiniMax	N/A
Qwen3.5-9B	Qwen3.5	9B
Qwen3.5-27B	Qwen3.5	27B
Qwen3.5-35B-A3B	Qwen3.5	35B-A3B (MoE)
Qwen3.5-122B-A10B	Qwen3.5	122B-A10B (MoE)
Qwen3.5-397B-A17B	Qwen3.5	397B-A17B (MoE)
GLM-4.6V	GLM	N/A
Seed-1.6-Flash	Seed	N/A
Seed-2.0-Mini	Seed	N/A
Nemotron-Nano-12B-V2-VL	Nemotron	12B
ERNIE-4.5-VL-28B-A3B	ERNIE	28B-A3B (MoE)
ERNIE-4.5-VL-424B-A47B	ERNIE	424B-A47B (MoE)

C.2 HUMAN BASELINE

Three annotators, none of whom are authors, answered 990 in-distribution items, sampled uniformly over the combinations of action type and reasoning type, and 1,000 perturbation items (500 appearance, 500 geometric) through the interface in Fig. 14, which presents the same frames, question, and four options as the model input. Each answer is saved when chosen and cannot be changed. We compute each annotator's accuracy per category and report the mean over the three annotators: 92.3% on the in-distribution items (Table 11) and 98.0% on the perturbation items (Table 12). Annotators were paid in accordance with local wage standards.

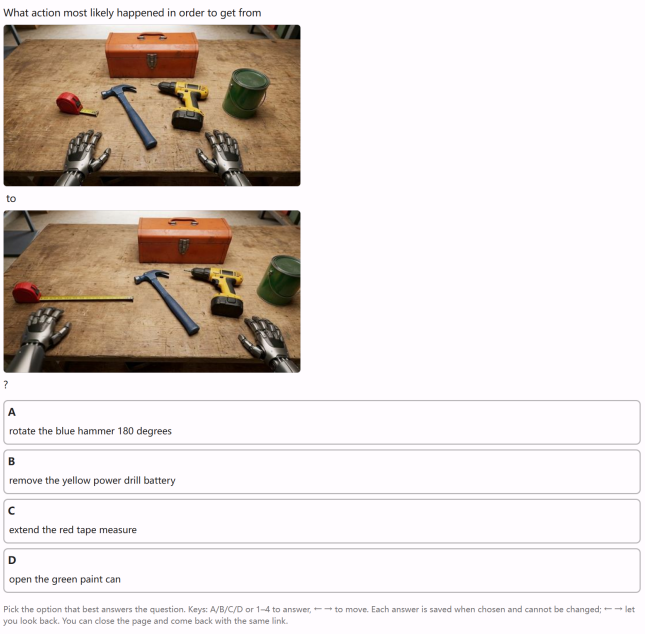

Figure 14: **Human study interface.** Annotators see the same frames, question, and four options as the models. Each answer is saved when chosen and cannot be changed.

Table 11: T37 Human accuracy (%) on the ID test, averaged over three annotators.

	Per-action type					Per-reasoning type								
	exo	per	insp	nav	man	cued	out	t-adj	t-ord	fwd	inv	cf-rem	cf-sub	overall
human	92.8	92.8	92.6	90.6	92.8	91.1	93.3	91.7	92.8	92.8	92.8	91.1	92.2	92.3

Table 12: T38 Human accuracy (%) on the perturbation probe, averaged over three annotators.

	appearance	geometric	overall
human	98.7	97.3	98.0

D STATISTICAL METHODOLOGY

This appendix has two parts. **O.1** enumerates the non-default methodological choices made across the paper, each with a one-line rationale. **O.2** gives per-finding statistical details (test, n, statistic, raw and Bonferroni-adjusted p, effect size) for every quantitative claim that enters the main text.

D.1 METHODOLOGICAL CHOICES

Above-chance threshold. A model is treated as "above chance" on a cell when accuracy $> 25\%$ *and* one-sided binomial $p<10^{-4}$. Twenty-five percent is the four-way-MCQ chance floor; the 10^{-4} threshold gives Bonferroni-style headroom for the cohort-level screen of 38 models $\times$ up to 8 reasoning types (≈ 300 simultaneous binomials), preserving family-wise $\alpha < 0.05$ before any downstream test is applied.

OLS on $\log_{10} N$. Within-family scaling fits use $y = a + b \log_{10} N$ with b in pts/decade. The log axis is the standard scaling-law parameterisation; pts/decade is directly comparable across axes (geometric vs. appearance, ID vs. perturbation).

Cramér's V. Effect-size companion to per-model χ^2 goodness-of-fit on distractor-class concentrations. Bounded in $[0, 1]$ and normalized by sample size, so it is comparable across models with different total error counts; conventional thresholds (small 0.10, medium 0.30, large 0.50) apply at $\mathrm{df}=2$.

Centred cosine similarity. Cosine similarity on the 8-dimensional per-principle accuracy vector after subtracting each model's own mean. The centring step isolates representational *shape* (which principles a model is relatively strong on) from *level* (overall accuracy), so two strong models with the same competence profile look similar even if one is uniformly higher.

Two-proportion z for within-model gap. Used for the within-model appearance-vs-geometric perturbation gap (e.g., Qwen2.5-VL-7B at 50.5% vs. 23.5%). Standard form, no continuity correction at the n values reported here (n_{geo}=n_{app}=558).

Paired Wilcoxon signed-rank. Per-model paired test for the counterfactual-removal cost asymmetry (§O.2, F-G), pairing $\mathrm{cost}(\texttt{fwd}\rightarrow\texttt{cf_rem})$ vs. $\mathrm{cost}(\texttt{fwd}\rightarrow\texttt{cf_sub})$ within each model. Two-sided; ties handled by mid-rank (Pratt method); restricted to the 24 above-chance models so that "cost" is measured against a non-degenerate `forward` baseline.

Bootstrap, $1000\times$ on residuals. OLS scaling-fit confidence intervals come from parametric bootstrap on regression residuals, 1000 resamples, slope CIs as 2.5/97.5 quantiles. The slope estimates and CIs are stable to two decimal places at this resample count given the four-rung ladders in question.

Bonferroni for family-wise control. The seven main-text findings (F-A through F-H, see O.2) form the relevant family. We apply Bonferroni at $k=7$ on top of each finding's raw p value. Bonferroni rather than Benjamini–Hochberg is chosen for conservativeness: at our cohort size, controlling family-wise error is the stronger guarantee a benchmark paper can offer reviewers. All seven findings remain significant ($p_{\text{adj}} < 0.001$) under this correction.

Scene-CV per model. Per-scene accuracy variability is summarized by $\mathrm{CV}=\sigma/\mu$ within each model across the 15 ID scenes used for the noise-floor estimate. The dimensionless form makes the noise floor (median CV 5.6%) interpretable across models at different overall accuracy levels.

Pair-mate Pearson aggregation. Pair-mate independence (F-H) reports Pearson r between the four (subject $\times$ view) cells of each manipulative scenario, averaged over the ~ 152 available (model $\times$ reasoning type) combinations. Aggregation is by Fisher-z transform, mean over combinations, inverse-transform back; reported r is the inverse-transformed mean.

D.2 PER-FINDING STATISTICAL DETAILS

Every main-text finding is summarized in one row. Raw and Bonferroni-adjusted p are reported with $k=7$ across the family of main-text findings. Effect sizes use Cramér's V for χ^2, Cohen's d for t-tests, slope-R^2 for OLS, r for Spearman/Pearson, and the matched within-model gap for paired Wilcoxon.

Notes on aggregation. For F-A's mirror-confusion test, we report the worst-of-seven raw p; the seven families are individually corrected (i.e. effective $k=49$ if treated as separate hypotheses), and all seven remain $p<10^{-9}$ adjusted. For F-D and F-G, paired Wilcoxon is computed on the per-model paired difference Δ_i; ties handled by Pratt's method; only above-chance models contribute to F-G to keep "cost" well-defined. For F-H, the claim is the absence of a strong correlation, so we report the magnitude bound rather than a p value: all six pair-mate correlations satisfy $|\bar{r}| < 0.20$, well below conventional moderate-correlation thresholds. For F-B's slope CIs, parametric bootstrap on

Table 13: Per-finding statistical details for main-text claims. Family-wise correction applies Bonferroni at k=7. All seven findings remain significant under correction.

ID	Claim	n	Test	Statistic (df)	Raw p	Bonf. adj. p	Effect size
F-A	Inspective below manipulative, frontier	11 models	sign test	11/11 positive	9.8×10^{-4}	6.9×10^{-3}	gap range 16.2–35.6pp
	Mirror confusion, frontier	7 families	1-sided binomial vs. 1/3	$\hat{p}_{\text{mir}}$=0.55	$\ll10^{-10}$ each	$\ll10^{-9}$ each	$\hat{p}-1/3\in[0.22,0.56]$
	Orientation concentration, GPT-5.4	231 err	χ^2 vs. uniform 1/3	65.22 (df=2)	6.9×10^{-15}	4.8×10^{-14}	V=0.376 (med–large)
	Orientation concentration, Qwen2.5-VL-7B	477 err	χ^2 vs. uniform 1/3	31.74 (df=2)	1.3×10^{-7}	9.1×10^{-7}	V=0.182 (small)
F-B	Appearance vs. geometric scaling slope	4 dense ladders, 11 models	OLS log-linear, per-ladder intercepts, slope b	app b=39.8, geo b=11.1 pts/dec (t=4.43, 4.83; df=6)	4.4×10^{-3}, 2.9×10^{-3}	3.1×10^{-2}, 2.0×10^{-2}	slope difference 28.7 pts/dec, 95% CI [8.1, 49.4]
	Within-model app–geo gap, Qwen2.5-VL-7B	558+558	two-proportion z	z=9.34	$<10^{-20}$	$<10^{-19}$	Δ=27pp
F-C	ID vs. Appearance correlation	38	Spearman	ρ=+0.90 (df=36)	1.5×10^{-14}	1.1×10^{-13}	R^2=0.86
	ID vs. Geometric correlation	38	Spearman	ρ=+0.76 (df=36)	3.1×10^{-8}	2.2×10^{-7}	R^2=0.52
F-D	t-ord−t-adj within-model gap	11 frontier	paired Wilcoxon, two-sided	W=66 (n_{nonzero}=11)	9.8×10^{-4}	6.9×10^{-3}	median gap +20.7pp
F-E	Dense premium at matched total params	3 within-family pairs	one-sample t vs. 0	t=14.3 (df=2)	4.8×10^{-3}	3.4×10^{-2}	d=8.3, mean +12.4pp
F-F	Cross-ladder emergence ordering replication	3 ladders	Spearman rank concordance	Kendall's W=0.93	$<10^{-3}$	$<7\times10^{-3}$	W>0.9 (strong)
	Gravity `rise` suppressed below 1/3	12 models	one-sample t vs. 1/3	t=−3.85 (df=11)	2.7×10^{-3}	1.9×10^{-2}	d=−1.11, mean 0.205
F-G	cf-rem vs. cf-sub cost asymmetry	24 above-chance	paired Wilcoxon, two-sided	W=293 (n=24, Pratt zero handling)	1.4×10^{-6}	9.8×10^{-6}	median gap +11.0pp
F-H	Pair-mate independence	$\sim$152 groups	Pearson r, Fisher-z aggregated	$\bar{r}\in[0.02,0.16]$	per-pair >0.05 (n.s.)	—	all $\lvert\bar{r}\rvert<0.20$

residuals (1000 resamples) gives appearance slope 95% CI $[24.0, 38.4]$ and geometric slope 95% CI $[6.5, 11.3]$; the two CIs do not overlap, reproducing the $3.5\times$ ratio finding under resampling.

E PER-COHORT EVALUATION TABLES

Table 14: T02 Benchmark scale.

Set	QA count	Sources
ID reasoning type items	6,228	594
ID perturbation items	1,116	320
Scene-OOD reasoning type	3,496	332

Table 15: T03 Action type × reasoning type adjacency.

Action type	#Comp	Reasoning type list
exogenous	4	out, cued, t-ord, t-adj
perceptive	4	fwd, inv, t-ord, t-adj
inspective	6	fwd, inv, cf-rem, cf-sub, t-ord, t-adj
navigative	4	fwd, inv, t-ord, t-adj
manipulative	4	fwd, inv, cf-rem, cf-sub

F DIAGNOSTIC RESULTS

This appendix details the diagnostic baseline summarized in §4.1. Figure 15 compares the action, reasoning, and perturbation profiles of eight frontier models; Tables 17, 18, and 21 report the full cohort. The following analyses examine geometric reasoning, scaling, temporal and counterfactual asymmetries, architecture, and scenario factors.

F.1 F-A. ROTATION DEFICIT DETAILS

The geometric reasoning deficit is supported by four converging observations. (A.1) Inspective accuracy falls below manipulative accuracy for $11/11$ frontier models, with manipulative−inspective gaps of $+16.2$ to $+35.6$pp (Table 22). (A.2) Within perceptive, translation $\gg$ rotation accuracy by $+8$ to $+55$pp ($28/29$ above-chance models positive; Table 23). (A.3) Errors on rotation cells concentrate on the mirror-direction distractor at frontier scale. (A.4) $\chi^2(\text{df}=2) > 100$ for $9/11$ frontier models on perturbation geometric distractors when restricted to orientation-flip items, Cramér's $V > 0.30$.

F.2 F-B. GEOMETRIC VS. APPEARANCE SCALING DISSOCIATION

Within-family scaling fit on Qwen2.5-VL dense ladder is OLS log-linear $y = a + b\log_{10} N$ with $b_{\text{app}} = 31.2$ pts/decade (R^2=0.70), $b_{\text{geo}} = 8.9$ pts/decade (R^2=0.73), ratio $3.5\times$ (Figure 16). Pooled over the four dense ladders (Qwen2.5-VL $\{3, 7, 32, 72\}$B, Qwen3-VL $\{8, 32\}$B, Gemma-3

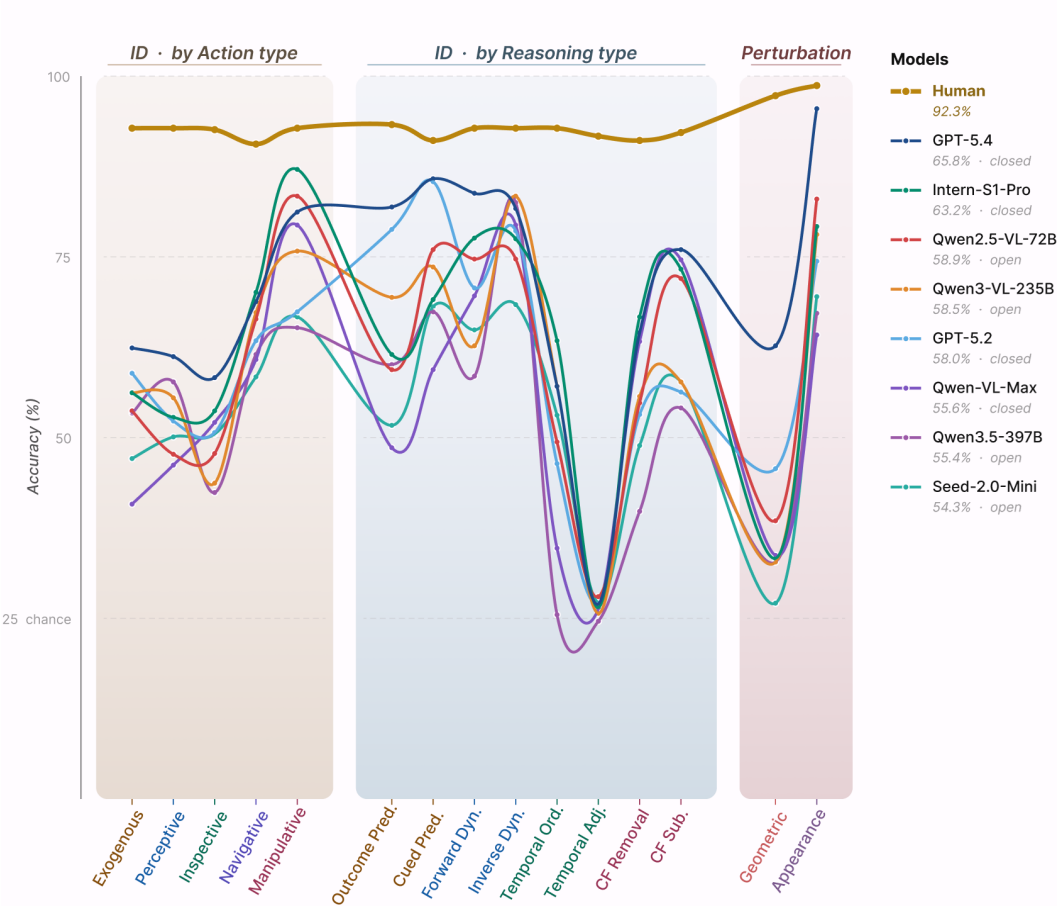

Figure 15: **Diagnostic profiles of eight frontier MLLMs against the human baseline.** Performance is grouped by action type (left), reasoning type (middle), and state-axis perturbation (right). Full 38-model results are reported in Appendix E.

Table 16: T04 ID overall accuracy (n=6,228, sorted).

Model	Family	Params	ID acc
GPT-5.4	GPT	N/A	65.8%
Intern-S1-Pro	Intern	N/A	63.2%
Qwen2.5-VL-72B	Qwen2.5	72B	58.9%
Qwen3-VL-235B-A22B	Qwen3	235B-A22B (MoE)	58.5%
GPT-5.2	GPT	N/A	58.0%
Gemma-4-31B	Gemma4	31B	57.8%
Qwen-VL-Max	Qwen-max	N/A	55.6%
Qwen3.5-397B-A17B	Qwen3.5	397B-A17B (MoE)	55.4%
Qwen3-VL-32B-Instruct	Qwen3	32B	55.2%
Seed-2.0-Mini	Seed	N/A	54.3%
ERNIE-4.5-VL-424B-A47B	ERNIE	424B-A47B (MoE)	51.7%
Pixtral-Large-2411	Mistral	123B	48.4%
Qwen2.5-VL-32B	Qwen2.5	32B	47.7%
Gemma-3-27B	Gemma3	27B	47.0%
Seed-1.6-Flash	Seed	N/A	45.3%
ERNIE-4.5-VL-28B-A3B	ERNIE	28B-A3B (MoE)	44.9%
Gemma-4-26B-A4B	Gemma4	26B-A4B (MoE)	43.9%
Qwen3-VL-30B-A3B	Qwen3	30B-A3B (MoE)	42.8%
Mistral-Small-3.2	Mistral	24B	42.5%
Qwen2.5-VL-7B	Qwen2.5	7B	41.6%
Gemma-3-12B	Gemma3	12B	41.5%
Mistral-Large-2512	Mistral	N/A	40.4%
Qwen3.5-27B	Qwen3.5	27B	33.3%
Qwen3-VL-8B-Instruct	Qwen3	8B	33.1%
Llama-4-Scout	Llama4	Scout (109B-MoE)	32.6%
MiniMax-01	MiniMax	N/A	31.9%
Nemotron-Nano-12B-V2-VL	Nemotron	12B	31.5%
Qwen3.5-9B	Qwen3.5	9B	30.1%
Gemma-3-4B	Gemma3	4B	28.9%
Llama-4-Maverick	Llama4	Maverick (400B-MoE)	28.7%
Qwen3.5-122B-A10B	Qwen3.5	122B-A10B (MoE)	26.9%
Qwen2.5-VL-3B	Qwen2.5	3B	26.3%
InternVL3.5-Latest	Intern	N/A	25.1%
InternVL3.5-241B-A28B	Intern	241B-A28B (MoE)	25.1%
Intern-S1-Mini	Intern	N/A	24.4%
Intern-S1	Intern	N/A	24.2%
GLM-4.6V	GLM	N/A	22.6%
Qwen3.5-35B-A3B	Qwen3.5	35B-A3B (MoE)	22.5%

$\{4, 12, 27\}$B, Qwen3.5 $\{9, 27\}$B; 11 models) with a separate intercept per ladder, both slopes are significant (appearance 39.8 pts/decade, p=0.004; geometric 11.1 pts/decade, p=0.003; df=6), and the appearance slope exceeds the geometric slope by 28.7 pts/decade (95% CI $[8.1, 49.4]$, p=0.015). The lower appearance R^2 reflects a non-monotonicity at 7B; we leave the point in the fit. First-order extrapolation places 50% geometric at $N \approx 2.9\times10^{12}$ and 60% at $N \approx 3.8\times10^{13}$ parameters; 'we report this to communicate the gap's severity, with the log-linear fit's R^2=0.73 holding within the 3B–72B range only and no claim made beyond it. Within-model two-proportion z-test on Qwen2.5-VL-7B: $p_{\text{app}} = 0.505$, $p_{\text{geo}} = 0.235$, $n_{\text{app}} = n_{\text{geo}} = 558$, z=9.34, $p<10^{-20}$.

F.3 F-C. Capability–representation decoupling

ID-overall vs. perturbation Spearman correlations across n=38 models: ρ_{app}=+0.90 (R^2=0.86), ρ_{geo}=+0.76 (R^2=0.52). Roughly 48% of geometric-robustness variance is independent of overall capability. The rank-shift table makes the decoupling visible: ERNIE-4.5-VL-424B-A47B drops from ID rank 11 to geometric rank 30 (Δ=+19); Seed-1.6-Flash rises from ID rank 15 to geometric rank 5 (Δ=−10).

Table 17: T05 ID accuracy by reasoning type (%, all 38 models).

Model	fwd	inv	cf-rem	cf-sub	out	cued	t-ord	t-adj
GPT-5.4	83.8	81.7	64.5	76.0	81.9	85.8	57.1	27.0
GPT-5.2	70.7	78.6	53.2	56.3	78.8	85.4	46.4	27.1
Qwen-VL-Max	69.6	79.4	63.3	74.6	48.6	59.4	34.7	26.3
Qwen2.5-VL-3B	19.7	42.2	21.0	17.6	22.6	23.3	26.9	25.1
Qwen2.5-VL-7B	53.4	53.8	37.1	54.8	36.5	43.8	31.2	24.3
Qwen2.5-VL-32B	49.4	70.5	40.3	46.1	36.1	49.0	49.3	28.6
Qwen2.5-VL-72B	74.7	74.7	54.8	72.0	59.4	76.0	49.4	28.0
Qwen3-VL-8B-Instruct	24.8	69.4	24.9	19.2	24.0	25.3	28.8	24.8
Qwen3-VL-30B-A3B	40.2	76.2	36.7	36.2	32.3	41.0	38.5	24.6
Qwen3-VL-32B-Instruct	71.4	81.2	45.9	71.9	51.4	62.8	37.1	26.3
Qwen3-VL-235B-A22B	62.7	83.4	55.7	57.7	69.4	73.6	57.2	25.7
Intern-S1	24.8	24.1	23.3	25.8	23.6	22.2	23.8	24.6
Intern-S1-Mini	25.7	23.9	23.1	27.6	22.6	23.3	23.0	24.5
Intern-S1-Pro	77.6	77.5	66.7	73.3	61.5	69.1	63.4	26.5
InternVL3.5-241B-A28B	23.5	26.5	25.4	24.9	25.3	28.5	24.7	25.0
InternVL3.5-Latest	26.3	24.8	25.3	24.2	24.3	24.7	25.6	24.7
Gemma-3-4B	24.5	53.8	21.5	18.5	21.5	24.0	25.6	23.7
Gemma-3-12B	49.5	56.5	41.6	46.2	40.6	49.7	29.5	26.1
Gemma-3-27B	62.2	62.6	47.8	62.7	39.2	47.9	30.6	26.2
Gemma-4-26B-A4B	51.1	59.4	39.2	55.6	49.7	56.2	29.2	26.1
Gemma-4-31B	76.3	76.6	63.6	64.5	64.9	70.5	30.8	25.3
Llama-4-Scout	19.8	67.7	18.8	18.1	18.4	15.3	32.6	31.9
Llama-4-Maverick	14.9	59.2	12.5	13.6	15.6	17.4	49.1	23.9
Mistral-Small-3.2	49.8	53.0	39.4	44.8	38.9	45.1	27.1	24.6
Mistral-Large-2512	52.0	54.7	29.2	50.9	40.3	54.5	26.9	22.7
Pixtral-Large-2411	56.4	59.9	45.2	50.5	35.8	47.9	21.8	27.1
MiniMax-01	23.6	59.8	22.6	26.5	29.5	24.3	28.0	26.0
Qwen3.5-9B	16.5	67.5	19.5	15.2	22.6	13.5	29.8	24.5
Qwen3.5-27B	28.2	67.8	25.4	21.7	25.3	32.2	35.6	24.6
Qwen3.5-35B-A3B	12.2	67.0	14.6	12.9	4.2	5.9	26.2	25.4
Qwen3.5-122B-A10B	19.6	18.6	26.5	19.5	50.0	56.9	17.6	23.9
Qwen3.5-397B-A17B	75.0	67.7	47.7	53.4	53.5	68.4	25.5	24.6
GLM-4.6V	10.5	25.3	9.0	18.0	39.6	60.4	42.2	26.1
Seed-1.6-Flash	50.7	65.0	27.9	60.4	44.4	56.2	25.7	25.9
Seed-2.0-Mini	76.0	67.4	50.5	50.9	43.4	53.5	53.1	27.2
Nemotron-Nano-12B-V2-VL	35.0	49.0	27.6	37.5	25.2	27.2	27.5	25.9
ERNIE-4.5-VL-28B-A3B	53.7	67.7	37.7	56.3	41.5	53.8	43.9	23.0
ERNIE-4.5-VL-424B-A47B	64.6	64.4	33.5	62.4	47.9	58.4	47.3	27.9

F.4 F-D. Temporal ordering–adjacency dissociation

The frontier median $\texttt{t_ord} - \texttt{t_adj}$ gap is $+20.7$pp (mean $+20.9$pp, range $+5.5$ to $+36.9$, $11/11$ frontier models positive; Table 29). For GPT-5.4, `t_adj` is 27.0%, within 2pp of chance, consistent with temporal adjacency being the most isolated reasoning type family in WOVEN. Both queries operate within a single action's frame sequence, so the dissociation indicates that models recover the coarse order of the transition's intermediate states but not their fine temporal resolution, the two grains of the trajectory representation of §A.2.

F.5 F-E. MoE active-parameter tax

Three matched-total within-family pairs yield dense premia of $+12.4$, $+13.9$, $+10.9$pp respectively (Table 30). Visual world-modeling capability tracks active parameters rather than total. Scaling plots that compare MoE to dense must use the active-parameter axis under penalty of $\sim$12pp overstatement.

F.6 F-F. Emergence ordering and gravity violation asymmetry

Emergence ordering. Eight exogenous principles cross the 40% ID threshold in a cross-ladder order on Qwen2.5-VL: *gravity, inertia* at 7B; *cohesion, permanence, containment* at 32B; *support, solidity* at 72B; *collision* never crosses (frontier mean 31.4%). The ordering replicates on Qwen3-VL ladder (8B$\rightarrow$ 235B-A22B) and Gemma-3 ladder (4B$\rightarrow$ 27B), aligning with the developmental sequence in infant cognition (Spelke & Kinzler, 2007; Baillargeon, 1987).

Gravity violation asymmetry. The violation-of-expectation distractor breakdown on `cued_prediction` for the gravity principle shows directional asymmetry: across 12 above-chance models from 7 architecturally distinct families, $P(\texttt{rise} \mid \text{err})$=$20.5\%$ vs. $P(\texttt{float} \mid \text{err})$=$37.0\%$

Table 18: T06 ID accuracy by action type (%, all 38 models).

Model	exo	per	insp	nav	man
GPT-5.4	62.4	61.2	58.3	68.8	81.2
GPT-5.2	58.9	52.3	50.7	63.4	67.4
Qwen-VL-Max	40.8	46.2	52.1	60.8	79.4
Qwen2.5-VL-3B	24.2	24.9	22.8	32.2	28.9
Qwen2.5-VL-7B	33.5	32.2	34.4	46.4	64.3
Qwen2.5-VL-32B	39.6	43.2	39.6	61.0	58.3
Qwen2.5-VL-72B	53.7	47.7	47.8	66.4	83.4
Qwen3-VL-8B-Instruct	28.1	29.0	29.5	42.3	38.2
Qwen3-VL-30B-A3B	33.0	42.1	38.0	51.2	51.4
Qwen3-VL-32B-Instruct	43.7	52.3	47.8	61.9	73.2
Qwen3-VL-235B-A22B	56.2	55.5	43.7	67.3	75.8
Intern-S1	24.5	22.9	26.2	23.9	22.9
Intern-S1-Mini	25.2	23.5	24.9	23.0	24.9
Intern-S1-Pro	56.2	52.8	53.7	70.1	87.1
InternVL3.5-241B-A28B	26.6	24.6	26.7	22.7	24.3
InternVL3.5-Latest	25.5	25.3	25.8	24.0	24.8
Gemma-3-4B	22.7	26.9	27.3	39.1	29.3
Gemma-3-12B	37.3	28.0	39.8	46.4	56.4
Gemma-3-27B	36.6	31.1	44.6	54.0	69.6
Gemma-4-26B-A4B	40.5	32.1	42.9	47.9	56.0
Gemma-4-31B	50.8	49.4	53.1	61.2	74.3
Llama-4-Scout	23.6	34.0	32.2	42.4	30.5
Llama-4-Maverick	20.9	27.7	26.8	42.5	26.3
Mistral-Small-3.2	34.1	27.5	44.4	64.6	44.1
Mistral-Large-2512	36.5	31.8	40.8	51.2	41.3
Pixtral-Large-2411	33.6	35.5	55.6	80.6	48.6
MiniMax-01	27.7	27.5	29.8	37.6	37.9
Qwen3.5-9B	22.5	28.3	27.0	40.1	33.9
Qwen3.5-27B	30.3	29.5	21.5	44.2	43.1
Qwen3.5-35B-A3B	10.3	29.3	15.3	27.6	30.0
Qwen3.5-122B-A10B	41.5	31.4	22.7	23.3	21.6
Qwen3.5-397B-A17B	53.4	57.7	42.4	61.5	65.2
GLM-4.6V	32.2	16.0	20.3	30.2	17.2
Seed-1.6-Flash	35.5	31.2	42.4	49.5	68.8
Seed-2.0-Mini	47.1	50.1	50.5	58.4	66.7
Nemotron-Nano-12B-V2-VL	26.0	24.4	31.9	34.6	40.7
ERNIE-4.5-VL-28B-A3B	36.6	34.2	40.8	51.4	63.0
ERNIE-4.5-VL-424B-A47B	40.1	40.5	48.7	66.6	63.4

Table 19: T07 Exogenous sub-principle accuracy (%, ID; capable models shown, full 38 in repo).

Model	permanence	cohesion	solidity	gravity	support	inertia	collision	containment
GPT-5.4	56.2	61.8	69.4	75.0	62.5	60.4	43.8	70.1
GPT-5.2	57.6	60.4	62.5	75.0	50.7	55.6	43.8	66.0
Qwen-VL-Max	29.9	44.4	38.9	56.9	41.7	47.2	18.8	48.6
Qwen2.5-VL-7B	37.5	37.5	28.5	47.2	34.0	41.0	21.5	20.8
Qwen2.5-VL-32B	41.0	42.3	36.4	44.2	37.6	41.6	32.4	41.8
Qwen2.5-VL-72B	54.2	51.4	45.8	76.4	58.3	59.0	32.6	52.1
Qwen3-VL-32B-Instruct	34.7	51.4	41.7	56.9	43.8	44.4	27.1	49.3
Qwen3-VL-235B-A22B	52.8	61.1	49.3	73.6	61.8	61.1	33.3	56.2
Intern-S1-Pro	45.8	41.7	61.8	77.8	68.8	61.1	36.8	56.2
Gemma-3-12B	37.5	44.4	35.4	61.8	38.2	33.3	25.7	22.2
Gemma-3-27B	34.7	38.9	31.9	61.8	34.0	47.2	27.8	16.7
Gemma-4-26B-A4B	34.0	47.1	46.5	56.4	42.6	44.2	22.9	30.4
Gemma-4-31B	55.5	54.2	48.5	49.6	58.6	58.1	23.7	58.9
Pixtral-Large-2411	39.3	31.2	31.2	50.7	26.3	26.8	25.4	37.3
Qwen3.5-397B-A17B	51.0	58.6	46.5	60.0	54.1	62.5	30.8	66.3
Seed-2.0-Mini	51.5	50.7	42.0	66.2	46.1	62.9	29.9	28.2
ERNIE-4.5-VL-424B-A47B	41.1	37.0	37.7	60.4	37.0	36.8	25.2	45.3

vs. $P(\texttt{lateral} \mid \text{err})$=38.6% (one-sample t=−3.85, p=0.003). The pattern replicates on `outcome_prediction`.

F.7 F-G. Counterfactual cost asymmetry

Define $\text{cost}(\texttt{fwd} \to \texttt{cf}_X) := \text{acc}(\texttt{fwd}) - \text{acc}(\texttt{cf}_X)$. Across 24 above-chance models, 23 satisfy $\text{cost}(\texttt{cf_rem}) > \text{cost}(\texttt{cf_sub})$ with mean gap +11.0pp (paired Wilcoxon $p<10^{-5}$). Mechanism is scale-dependent: frontier models distribute `cf_rem` errors away from the actually-occurring action's

Table 20: T07b Manipulative subject × view accuracy (%, ID; capable models).

Model	h/ego	h/allo	hh/ego	hh/allo
GPT-5.4	87.2	80.9	82.6	74.3
GPT-5.2	69.8	67.7	64.9	67.4
Qwen-VL-Max	85.1	79.2	81.2	72.2
Qwen2.5-VL-7B	68.8	65.6	66.7	56.2
Qwen2.5-VL-32B	66.7	59.0	60.8	46.9
Qwen2.5-VL-72B	91.3	84.4	85.8	72.2
Qwen3-VL-32B-Instruct	83.0	70.5	75.0	64.2
Qwen3-VL-235B-A22B	83.0	77.1	79.2	63.9
Intern-S1-Pro	89.6	87.5	89.6	81.6
Gemma-3-27B	76.7	67.4	75.0	59.4
Gemma-4-31B	78.1	74.7	77.8	66.7
Pixtral-Large-2411	60.8	40.6	56.6	36.5
Qwen3.5-397B-A17B	68.1	61.8	65.6	65.3
Seed-2.0-Mini	70.1	66.7	70.1	59.7
ERNIE-4.5-VL-424B-A47B	75.7	63.5	69.8	44.4

Table 21: T08 Perturbation accuracy (overall / geo / app, n=1,116, all 38).

Model	overall	geo	app
GPT-5.4	79.1	62.7	95.5
GPT-5.2	60.0	45.7	74.4
Qwen-VL-Max	48.9	33.7	64.2
Qwen2.5-VL-3B	21.1	22.6	19.7
Qwen2.5-VL-7B	37.0	23.5	50.5
Qwen2.5-VL-32B	37.4	30.1	44.6
Qwen2.5-VL-72B	60.8	38.5	83.0
Qwen3-VL-8B-Instruct	23.0	25.1	21.0
Qwen3-VL-30B-A3B	34.1	25.6	42.5
Qwen3-VL-32B-Instruct	50.4	32.4	68.5
Qwen3-VL-235B-A22B	55.5	32.8	78.1
Intern-S1	26.7	26.5	27.0
Intern-S1-Mini	26.4	26.4	26.4
Intern-S1-Pro	56.3	33.3	79.2
InternVL3.5-241B-A28B	26.7	26.3	27.1
InternVL3.5-Latest	26.8	26.0	27.6
Gemma-3-4B	24.4	24.6	24.2
Gemma-3-12B	40.3	32.6	48.0
Gemma-3-27B	49.5	36.4	62.5
Gemma-4-26B-A4B	34.1	25.4	42.8
Gemma-4-31B	77.2	59.0	95.5
Llama-4-Scout	15.2	17.0	13.4
Llama-4-Maverick	11.3	9.9	12.7
Mistral-Small-3.2	34.3	27.3	41.3
Mistral-Large-2512	31.7	27.1	36.4
Pixtral-Large-2411	50.1	35.1	65.1
MiniMax-01	25.0	26.3	23.7
Qwen3.5-9B	20.6	20.8	20.4
Qwen3.5-27B	36.8	31.2	42.5
Qwen3.5-35B-A3B	12.6	10.6	14.7
Qwen3.5-122B-A10B	25.8	24.4	27.2
Qwen3.5-397B-A17B	50.0	32.8	67.2
GLM-4.6V	7.2	5.4	9.0
Seed-1.6-Flash	54.9	38.4	71.5
Seed-2.0-Mini	48.3	27.1	69.5
Nemotron-Nano-12B-V2-VL	25.5	27.1	24.0
ERNIE-4.5-VL-28B-A3B	42.5	32.5	52.5
ERNIE-4.5-VL-424B-A47B	35.3	25.1	45.5

outcome (mean $P_{\text{actual|err}}$=20.4%, all below $1/3$ uniform; GPT-5.4 24.0%); mid-scale models revert to it (Qwen2.5-VL-7B 63.6%). Pearson r(ID overall, $P_{\text{actual|err}}$)=−0.06 across 38 models. ENHANCE F6 provides converse evidence (Appendix G.8).

Table 22: T21 Manipulative−inspective accuracy gap (frontier).

Model	manipulative	inspective	gap
GPT-5.4	81.2	58.3	+23.0
GPT-5.2	67.4	50.7	+16.7
Qwen-VL-Max	79.4	52.1	+27.3
Qwen2.5-VL-7B	64.3	34.4	+29.9
Qwen2.5-VL-32B	58.3	39.6	+18.8
Qwen2.5-VL-72B	83.4	47.8	+35.6
Qwen3-VL-32B	73.2	47.8	+25.4
Qwen3-VL-235B-A22B	75.8	43.7	+32.1
Intern-S1-Pro	87.1	53.7	+33.4
Gemma-4-31B	74.3	53.1	+21.2
Seed-2.0-Mini	66.7	50.5	+16.2

Table 23: T24 Perceptive translation vs. rotation (frontier).

Model	translation	rotation	gap
GPT-5.4	92.4	74.0	+18.4
GPT-5.2	88.2	56.6	+31.6
Qwen-VL-Max	82.6	42.7	+39.9
Qwen2.5-VL-7B	56.6	17.0	+39.6
Qwen2.5-VL-32B	58.0	41.3	+16.7
Qwen2.5-VL-72B	77.1	42.7	+34.4
Qwen3-VL-32B	86.1	61.5	+24.7
Qwen3-VL-235B-A22B	88.2	60.1	+28.1
Intern-S1-Pro	76.7	57.6	+19.1
Gemma-4-31B	91.0	47.9	+43.1
Pixtral-Large-2411	67.4	12.2	+55.2
Seed-2.0-Mini	79.9	34.4	+45.5

Table 24: T25 Qwen2.5-VL dense scaling (%).

Model	Params	ID	overall	geo	app
Qwen2.5-VL-3B	3B	26.4	21.1	22.6	19.7
Qwen2.5-VL-7B	7B	41.6	37.0	23.5	50.5
Qwen2.5-VL-32B	32B	47.7	37.4	30.1	44.6
Qwen2.5-VL-72B	72B	58.9	60.8	38.5	83.0

Table 25: T26 Qwen3-VL family.

Model	Params	ID	overall	geo	app
Qwen3-VL-8B-Instruct	8B	33.1	23.0	25.1	21.0
Qwen3-VL-30B-A3B	30B-A3B (MoE)	42.8	34.1	25.6	42.5
Qwen3-VL-32B-Instruct	32B	55.2	50.4	32.4	68.5
Qwen3-VL-235B-A22B	235B-A22B (MoE)	58.5	55.5	32.8	78.1

Table 26: T27 Gemma-3/Gemma-4.

Model	Params	ID	overall	geo	app
Gemma-3-4B	4B	28.9	24.4	24.6	24.2
Gemma-3-12B	12B	41.5	40.3	32.6	48.0
Gemma-3-27B	27B	47.0	49.5	36.4	62.5
Gemma-4-26B-A4B	26B-A4B (MoE)	43.9	34.1	25.4	42.8
Gemma-4-31B	31B	57.8	77.2	59.0	95.5

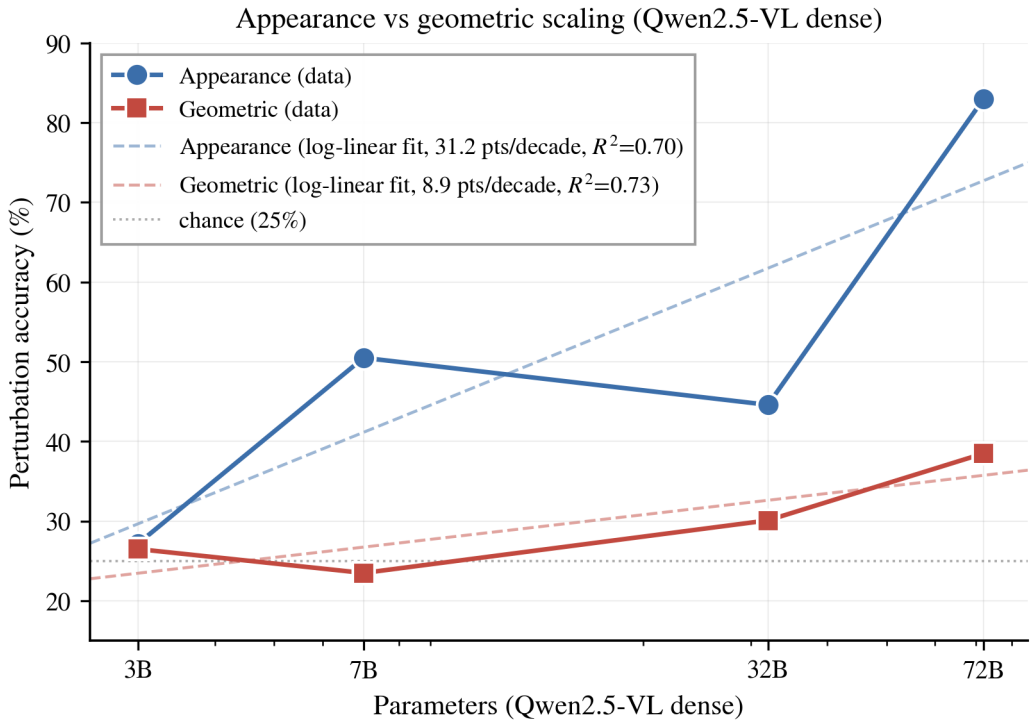

Figure 16: **Appearance and geometric robustness across the Qwen2.5-VL dense model ladder ($\{3, 7, 32, 72\}$B).** Appearance robustness scales 3.5× faster than geometric robustness, with fitted slopes of 31.2 and 8.9 percentage points per decade of parameters, respectively.

Table 27: T28 Qwen3.5 family (reasoning disabled).

Model	Params	ID	overall	geo	app
Qwen3.5-9B	9B	30.1	20.6	20.8	20.4
Qwen3.5-27B	27B	33.3	36.8	31.2	42.5
Qwen3.5-35B-A3B	35B-A3B (MoE)	22.5	12.6	10.6	14.7
Qwen3.5-122B-A10B	122B-A10B (MoE)	26.9	25.8	24.4	27.2
Qwen3.5-397B-A17B	397B-A17B (MoE)	55.4	50.0	32.8	67.2

Table 28: T31 ID vs. perturbation-geometric rank shift (subset; positive Δ = relatively worse on geometric).

Model	ID acc	ID rank	Geo acc	Geo rank	Δ
ERNIE-4.5-VL-424B-A47B	51.7	11	25.1	30	+19
Qwen2.5-VL-7B	41.6	20	23.5	33	+13
Gemma-4-26B-A4B	43.9	17	25.4	28	+11
Llama-4-Scout	32.6	25	17.0	35	+10
Qwen3-VL-30B-A3B	42.8	18	25.6	27	+9
Seed-2.0-Mini	54.3	10	27.1	19	+9
Intern-S1-Pro	63.2	2	33.3	9	+7
GPT-5.4	65.8	1	62.7	1	0
GPT-5.2	58.0	5	45.7	3	−2
Gemma-4-31B	57.8	6	59.0	2	−4
Seed-1.6-Flash	45.3	15	38.4	5	−10
Intern-S1	24.2	36	26.5	22	−14

Table 29: T22 Temporal `t_ord` − `t_adj` gap (frontier).

Model	t-ord	t-adj	gap
GPT-5.4	57.1	27.0	+30.1
GPT-5.2	46.4	27.1	+19.3
Qwen-VL-Max	34.7	26.3	+8.4
Qwen2.5-VL-32B	49.3	28.6	+20.7
Qwen2.5-VL-72B	49.4	28.0	+21.4
Qwen3-VL-32B	37.1	26.3	+10.8
Qwen3-VL-235B-A22B	57.2	25.7	+31.5
Intern-S1-Pro	63.4	26.5	+36.9
Gemma-4-31B	30.8	25.3	+5.5
Seed-2.0-Mini	53.1	27.2	+25.9
ERNIE-4.5-VL-424B-A47B	47.3	27.9	+19.4

Table 30: T29 Dense vs. MoE comparisons (matched total).

Comparison	Dense acc	MoE acc	gap
Qwen3-VL-32B vs. 30B-A3B	55.2	42.8	+12.4
Gemma-4-31B vs. 26B-A4B	57.8	43.9	+13.9
Qwen3.5-27B vs. 35B-A3B	33.3	22.5	+10.9

F.8 F-H. Manipulative pair-mate independence + view-axis dominance

Cell-hardness ladder: `h_ego` > `hh_ego` > `h_allo` > `hh_allo` (cohort-mean source-acc $54.2 > 52.6 > 49.5 > 44.6$). View axis (ego vs. allo) shifts $\sim$5pp; combined with subject shift adds another $\sim$5pp.

Pair-mate independence. Pearson r between cell pairs across 18 scenario-groups, averaged over $\sim$152 (model $\times$ reasoning type) combinations: all six pairwise r in $[+0.02, +0.16]$. Diagonal cells (`h_ego` $\leftrightarrow$ `hh_allo`) essentially uncorrelated; same-axis (`allo` $\leftrightarrow$ `allo`) highest. Models do not have a unified scenario representation.

View-axis dominates subject-axis in single-axis failures. Of 14 (model, reasoning type) combinations with $\geq 4/18$ scenarios exhibiting single-axis failure, 11 are view-axis (ego/allo) and only 3 are subject-axis (human/humanoid). Within view-axis failures, ego-only-correct dominates.

F.9 F-I. Per-scene invariance within ID

Across 12 above-chance models on 15 ID scenes, per-model scene CV ranges 4.04 to 8.58%, median 5.6%. Per-model max−min scene gap 8 to 16pp. Most uniform: GPT-5.4 (CV 4.04%, gap 8.4pp); least uniform: ERNIE-4.5-VL-424B (CV 8.58%, gap 16.1pp). This sets a measurement-validity floor for OOD-gap interpretation: `beach` + `restaurant` OOD vs. ID gaps larger than $\sim$5 to 9pp can be attributed to scene novelty.

F.10 F-Minor. Small-scale no-op prior

Qwen2.5-VL-3B picks the `static` ("no action") option on 34.0% of inspective `forward_dynamics` items, well above the 25% uniform-chance baseline (one-sided binomial, k=102, N=300, p=3.1×10^{-4}). When the smallest model is uncertain about camera motion, it defaults to "nothing happened." This signature disappears at $\geq$ 7B and is absent from frontier models.

F.11 F-Fingerprint. Family fingerprinting via principle vectors

Each model's 8-dim per-principle accuracy vector serves as a representational fingerprint. Centred cosine similarity (after subtracting each model's own mean, measures *shape* rather than capability level) shows within-family similarity is high but not uniform.

Table 31: F-Fingerprint: centered cosine similarity on 8-dim per-principle accuracy vectors.

Pair	centered cos
Qwen2.5-VL-72B ↔ Qwen3-VL-235B-A22B	$+0.951$
GPT-5.4 ↔ GPT-5.2	$+0.895$
Qwen3-VL-235B ↔ Qwen3-VL-32B	$+0.895$
ERNIE-424B ↔ GPT-5.4	$+0.825$
Qwen3-VL-235B ↔ GPT-5.4	$+0.761$
Qwen2.5-VL-72B ↔ GPT-5.4	$+0.712$
Seed-2.0-Mini ↔ GPT-5.4	$+0.341$
Gemma-4-31B ↔ Gemma-3-27B (same provider, gen step)	$\mathbf{+0.104}$

Gemma's 3→4 generation transition produces an *orthogonal* per-principle reasoning type profile ($+0.104$), unlike Qwen2.5 → Qwen3 which preserves shape ($+0.951$). Within-provider scaling does not preserve representational shape.

G Enhance recipes and finding details

G.1 Shared supervised fine-tuning protocol

Base models and reporting. The main teachability and transfer comparisons use Qwen2.5-VL-3B-Instruct (Bai et al., 2025b). Model-scale experiments use Qwen2.5-VL-7B-Instruct and Qwen2.5-VL-32B-Instruct (Appendix G.15); the matched-backbone video-supervision comparison uses Qwen3.5-4B (Appendix G.16). SFT and subsequent GRPO (Shao et al., 2024) results are reported separately where both stages are evaluated.

Shared SFT optimization. All SFT experiments, including full-mix, single-cell, and temporal supervision, use LoRA with r=16, α=4, dropout 0.05, and target modules $\{q, k, v, o\}_{\text{proj}}$. Training uses 3 epochs, batch size 32, AdamW with learning rate 1×10^{-4}, cosine scheduling, 5% warmup, and fp16 precision.

Option-permutation augmentation. For WOVEN-only training, each item is presented under 4 permutations of its MCQ options, with the answer updated consistently, to reduce dependence on answer position. Fixed-budget substitution experiments use the answer-balanced data construction described in Appendix G.14.

G.2 E-F1. Full-mix SFT calibration

Full-mix SFT uses the 22,728-item WOVEN training set and the shared optimization protocol in Appendix G.1. On Qwen2.5-VL-3B-Instruct, it reaches 89.3% on the in-distribution test and 88.2% on held-out scenes (§4.2).

G.3 E-F2. Cross-benchmark transfer

Full-mix cross-benchmark transfer under the final locked protocol is reported in the main text (Tables 2 and 3); per-benchmark, per-category breakdowns for all training configurations are in Appendix G.13. Within WM-ABench, the gains concentrate on the transition dimensions (Prediction/Transitivity $+21.7$, Prediction/Mechanistic $+18.6$) while the static Visual dimension is flat (±1); see the WM-ABench table in Appendix G.13.

G.4 Targeted training and diagnostic correspondence

Inverse-text is the strongest narrow world-modeling signal. Narrow SFT on inverse dynamics with text-described action options across the 4 endogenous action types (n=4,132, shared SFT protocol; Appendix G.5) yields $+54$pp own-cell average (inspective $+77$, perceptive $+70$, navigative $+39$, manipulative $+32$). Holding option modality constant and switching only direction from forward to inverse gives $+51$pp: the gain comes from the inverse direction rather than the text-options modality.

Table 32: **SFT accuracy on WOVEN's own test sets (ID, scene-OOD, state-axis perturbation probe) (%).** Full-mix SFT is on Qwen2.5-VL-3B; inverse-text is narrow SFT on inverse dynamics with text options (4 endogenous action types); mv-balanced is balanced `mv_fwd`+`mv_bwd` training on Qwen2.5-VL-7B. Downstream transfer is in Tables 2 and 3. Per-cell own deltas, the remaining narrow-run families, and full recipe details in the appendix.

			Perturbation probe		
Variant	ID	OOD	all	geo	app
Qwen2.5-VL-3B (base)	26.4	—	21.1	22.6	19.7
+full-mix SFT	**89.3** (↑62.9)	**88.2**	**38.3** (↑17.1)	**35.0** (↑12.4)	**41.6** (↑21.9)
+inverse-text SFT	95.5† (↑54)	—	—	—	—
Qwen2.5-VL-7B (base)	41.6	—	37.0	23.5	50.5
+mv-balanced SFT	50.1 (↑8.5)	57.5	52.4 (↑15.4)	38.4 (↑14.9)	66.5 (↑16.0)

† inverse-text own-cell mean over 4 endogenous action types: base 41%, post-SFT 95.5% (↑54).

Cross-action-type forward transfer is asymmetric. Single-action-type forward-dynamics training across the four endogenous action types yields three asymmetries (Appendix G.7): (i) navigative↔manipulative double-strong ($+61$pp both directions); (ii) perceptive↔inspective weakly symmetric ($+13$/$+14$pp); (iii) P/I→N/M strong but N/M→P/I weak ($+47$ to $+57$ vs. $+1$ to $+11$). One hypothesis is that P/I training teaches both visual grounding and spatial reasoning, while N/M training teaches only grounding.

Imagining absence stays hard even when its neighbors are trained. Within an action type (Appendix G.8), inspective forward dynamics teaches own $+65$pp, transfers to `cf_substitution` $+48$ and `cf_removal` $+25$, but not to temporal templates. `cf_substitution` training transfers to `forward` $+66$ and `inverse` $+34$ but to `cf_rem` ≈ 0pp. The `cf_sub` $\not\to$ `cf_rem` non-transfer triangulates the cohort-level cost asymmetry: "imagining absence" is intrinsic to the reasoning type rather than a coverage shortage.

Temporal reasoning is the most isolated family in WOVEN. `temporal_adjacency` is teachable in narrow SFT (own $+30$ to $+35$pp across 4 action types; Appendix G.9); `temporal_ordering` is not under our setup; the two sister reasoning types show ≈ 0pp transfer in both directions. Combined with the cohort-level dissociation of §4.1, this means temporal skill neither arrives with the other families nor spreads to them: it must be trained directly.

Chirality is direction-asymmetric and capacity-dependent. Balanced rotation training teaches both directions (Appendix G.10, G.11) ($+55$ to $+61$pp own); single-direction teaches own-direction but cross-direction $\Delta \approx 0$ at 3B. At 7B, perceptive cross-direction becomes *negative* (model specialises harder than at 3B), and inspective single-direction breaks symmetry: `orb_l` → `orb_r` jumps $+27$pp but reverse stays at -4. Balanced `mv_fwd`+`mv_bwd` on 7B saturates own directions to 100% while cross-onto-rotation drops -25 to -27pp: move and rotation are treated as separate action classes.

The developmental emergence ordering works as a training curriculum. The exogenous emergence ordering of Appendix F.6 (Appendix G.12) (gravity, inertia → cohesion, permanence, containment → support, solidity) operationalizes as a curriculum: training the same data in emergence order vs. random shuffle gives $+4.8$pp own-bench at 1 epoch. Developmental ordering can be a useful prior on data scheduling.

Five of seven Assess findings replicate train-side. In detail: (i) the appearance/geometric scaling dissociation (Appendix F.2) replicates as $+21.9$ vs. $+12.4$pp on the perturbation probe after full-mix SFT, preserving the stronger improvement under appearance perturbations; (ii) temporal isolation appears in both the cohort and the training transfers (F7); (iii) the `cf_rem` cost asymmetry is intrinsic rather than a coverage shortage (F6); (iv) chirality blindness manifests train-side as direction-asymmetric, capacity-dependent transfer (F8/F9); (v) the exogenous developmental ordering operationalizes as a curriculum prior with $+4.8$pp gain (F10). The two non-replicated findings, the MoE active-parameter dependence (F-E) and scenario pair-mate independence (F-H), target base-model architecture and item-factor structure, axes outside the post-training intervention surface. The convergence of observation and intervention supports the cohort-side structural claims as representational rather than artifactual.

G.5 E-F3. INVERSE-TEXT PER-CELL OWN RESULTS

Table 33: Inverse-text SFT per-cell own (Δ vs. base 3B). Columns are inverse-dynamics cells by action type; the last column is the mean over the four own cells.

	inspective	perceptive	navigative	manipulative	mean
base 3B	16	27	59	62	41
inverse-text SFT	**93**	**97**	**98**	**94**	**95.5**
Δ	+77	+70	+39	+32	+54

The matched forward-text run (text options, the same SFT protocol and n=4,132) has own-cell mean 44.3%; switching only direction from forward to inverse yields +51pp, driven by inverse direction.

G.6 E-F4. NARROW SFT DATA-SCALE MARGIN

A 1×-data run (n=4132) vs. a 2×-data run (n=8264), all other config identical. Own-bench overall: 1× 50.7%, 2× 56.4%, Δ=+5.7pp on 2× data. Sample efficiency exhibits diminishing marginal returns under narrow SFT.

G.7 E-F5. CROSS-ACTION-TYPE FORWARD TRANSFER MATRIX

Table 34: Cross-action-type forward transfer (4×4). $n\approx$1,020 each, shared SFT protocol.

Source ↓ / Target →	per fwd	insp fwd	nav fwd	man fwd
Base 3B (acc, %)	28.1	28.1	30.9	21.9
perceptive (n=1036) Δ	+**68.8** own	+13.8	+55.6	+50.0
inspective (n=1020) Δ	+12.9	+**64.9** own	+57.3	+46.9
navigative (n=1020) Δ	+11.1	+10.0	+**67.0** own	+61.4
manipulative (n=1020) Δ	+7.0	+1.5	+61.5	+**74.6** own

G.8 E-F6. WITHIN-ACTION-TYPE CROSS-REASONING-TYPE

Table 35: (a) Within-action-type cross-comp transfer (inspective forward run).

	fwd	cf_sub	cf_rem	inv	t_adj	t_ord
Base 3B (acc, %)	28.1	26.7	25.6	15.2	26.1	27.0
SFT Δ	+**65** own	+48	+25	+19	0	−1

Table 36: (b) `CF_SUB` within-action-type cross-comp, per-action-type split.

	fwd	cf_sub	cf_rem	inv
Base 3B insp (acc, %)	28.1	26.7	25.6	15.2
CF_SUB Δ insp	+60	+**66** own	−1	+48
Base 3B man (acc, %)	21.9	20.8	24.0	62.5
CF_SUB Δ man	+71	+**74** own	0	+20

G.9 E-F7. TEMPORAL ORDERING AND ADJACENCY

One run trains `temporal_ordering` cross 4 action types (n=4,132); a second trains `temporal_adjacency` cross 4 action types (n=4,132). Both use the shared SFT optimization protocol (Appendix G.1). Across 4 action types (averaged): the adjacency run own `t_adj` +30 to +35pp, cross `t_ord` ≈ 0pp. The ordering run own `t_ord` ≈ 0pp, cross `t_adj` ≈ 0pp. Conversely, the four temporal subsets leave every other reasoning type within 3.3 points of the base model under SFT (Table 76).

G.10 E-F8 (7B SCALING). ROTATION CHIRALITY AT 7B

Table 37: Balanced rotation: 3B vs. 7B (`arc_up` interference reverses). Columns are inspective × forward cells by camera move.

	orb_l (own)	orb_r (own)	arc_up (orth)
base 3B (%)	30.0	20.0	31.1
3B Δ	+40.0	+45.6	**−24.4**
7B Δ	**+50.0**	**+75.6**	**+43.3**

Table 38: Perceptive single-direction runs: 3B vs. 7B. Columns are perceptive cells by camera move.

	rot_l	rot_r
base 3B (%)	27.8	33.3
rot_l only (7B) Δ	**+65.3** own	−15.2 cross
rot_r only (7B) Δ	−13.9 cross	**+63.9** own

Table 39: Inspective single-direction runs: 3B vs. 7B, asymmetric chirality sharing. Columns are inspective cells by camera move.

	orb_l	orb_r	arc_up
base 3B (%)	30.0	20.0	31.1
orb_l only (7B) Δ	**+58.9** own	**+26.7** cross	+36.7 cross
orb_r only (7B) Δ	−4.4 cross	**+72.2** own	+37.8 cross

G.11 E-F9. MOVE CHIRALITY AT 7B

Table 40: mv-balanced (7B) move chirality (perceptive forward). Columns are perceptive cells by camera move.

	mv_fwd (own)	mv_bwd (own)	rot_l (cross)	rot_r (cross)
base 3B (%)	20.8	19.4	27.8	33.3
mv-balanced 7B (%)	**100.0**	**100.0**	2.8	5.6
Δ	+79.2	+80.6	**−25.0**	**−27.7**

G.12 E-F10. ASSESS-EMERGENCE ORDERING AS CURRICULUM

Train data: 7 exogenous principles (cohesion, solidity, gravity, support, inertia, containment, permanence; collision excluded), 1,820 raw items ×4× within-item permutation = 7,280 items. Both curriculum conditions use the shared SFT optimization protocol (Appendix G.1); Table 41 reports accuracy after the first epoch.

Table 41: Curriculum vs. shuffled at 1 epoch (own_exo subset).

Variant	own_exo subset acc
V_1 (shuffled)	33.2%
V_2 (emergence-ordered)	**38.0%**
Δ ($V_2 - V_1$)	**+4.8**pp

Order: gravity → inertia → cohesion → containment → permanence → support → solidity.

G.13 E-F11. DOWNSTREAM TRANSFER: PROTOCOL, FULL RESULTS, AND ANALYSES

Benchmark suite. The downstream suite comprises 26 external benchmarks: 22 transition-relevant targets, SAT (Ray et al., 2024), MindCube (Yin et al., 2025), ViewSpatial (Li et al., 2025a), SpatialViz (Wang et al., 2026b), DSI-Bench (Zhang et al., 2025c), WM-ABench (Huang Team and Contributors, 2025), WorldPrediction (WorldPrediction Team, 2025), TemporalBench (Cai et al., 2024), TOMATO (Shangguan et al., 2025), TempCompass (Liu et al., 2024), TVBench (Cores et al., 2025), BlackSwan (Chinchure et al., 2025), CLEVRER (Yi et al., 2020), InPhyRe (Sreekumar & Boddeti, 2025), MVP (Krojer et al., 2025), ActionEQA (Anonymous, 2025), Robo2VLM (Chen et al., 2025a), ERQA (Gemini Robotics Team, 2025), Cosmos (NVIDIA, 2025a), PaiBench (Zhou et al., 2025), CV-Bench (Tong et al., 2024), and BLINK (Fu et al., 2024), plus 4 adjacent-capability probes (static perceptual grounding: 3DSRBench (3DSRBench Team, 2025), CoreCognition (Li et al., 2025c); holistic video QA: PerceptionTest (Pătrăucean et al., 2023), EgoTaskQA (Jia et al., 2022)). Every benchmark is evaluated under its official full protocol and split: SAT uses circular evaluation (150 questions $\times$ 2 answer orders, both orders must be correct); TemporalBench uses dual-order binary accuracy (BA); MVP uses official paired accuracy; TVBench uses the official per-task macro average; DSI-Bench reports the official sample-wise aggregate over its four spatio-temporal flip variants; CLEVRER and ENACT use stratified samples (12,000 of 70,862 per-option rows; not used in the main tables). Answers are generated greedily with `max_new_tokens`$= 8$; unparseable outputs count as incorrect. A gain is called *clear* above $+2$ points, roughly $5\times$ the observed run-to-run standard deviation of the full-mix RL configuration.

RL optimization. The RL stage uses Group Relative Policy Optimization (Shao et al., 2024), which estimates advantages by comparing each sampled response to a group of responses under the same prompt, with no separate value network. We run GRPO via verl (Sheng et al., 2024) + vLLM on Qwen2.5-VL-3B initialized from a world modeling SFT checkpoint. Shared config: KL-to-reference `kl_loss_coef` 0.005 (low-variance estimator), lr 1×10^{-6}, entropy coefficient 0, train batch 16 prompts $\times$ 8 rollouts (mini-batch 8), max prompt 1,536 tokens, max response 8 tokens (single-letter answers), 300 optimization steps with checkpoints every 50; the reported checkpoint is step 150 (validation accuracy saturates over steps 50–150).

RL reward. To mitigate answer-symbol bias in multiple-choice prediction (Zheng et al., 2024; Pezeshkpour & Hruschka, 2023; Xue et al., 2024), we combine answer correctness with an anti-bias penalty:

$$\text{reward}(\text{rollout}_i) = \text{letter_match}_i \; - \; \lambda \cdot \text{KL}\Big(\hat{P}_i \,\|\, \text{uniform}\Big)\,,$$

where letter_match_i is 1 if the predicted answer letter matches the gold letter and 0 otherwise. The empirical distribution $\hat{P}_i$ is computed over the $W{=}64$ most recent predicted letters at the time rollout i is scored, including rollout i itself; uniform assigns probability 0.25 to each of A/B/C/D, and $\lambda{=}0.1$. The window is updated for each scored rollout, allowing the penalty to vary within a GRPO group. The uniform target reflects the answer-letter balance of the training data. All reported RL experiments use this reward.

Transfer-fingerprint normalizations. Table 43 reports the fingerprint analysis of §5 under four treatments of benchmark difficulty. The ordering (same reasoning family $>$ same action type $\approx$ unrelated) is invariant.

Figure 4 visualizes the analysis as the full 11×11 correlation matrices. The left panel shows the raw profiles: every cell pair is positively correlated, including pairs sharing neither axis, which is the shared primitive in correlational form. The middle and right panels display the identical residual matrix (after the per-benchmark z-score) under the two candidate orderings. Ordered by action type, no coherent structure appears along the diagonal; reordered by reasoning family, the remaining positive correlation concentrates into the diagonal family blocks (causal, counterfactual, temporal). The structure is revealed purely by reordering, so the family organization resides in the data rather than in the display. The exogenous-physical cell is visible as its own singleton: within the causal-adjacent region it correlates weakly with the agentic causal cells, consistent with its negative transfer to endogenous transitions discussed below.

Common component and dose–response across subsets. Two analyses connect the transfer of the 11 controlled subsets to the capability they learn on WOVEN (§4.3). Both use the 22 benchmarks with per-configuration tables in this appendix (the 18 improved benchmarks reported there and the

Table 42: **Best controlled-subset training for all 22 improved benchmarks.** Each subset is defined by action type and reasoning family and contains approximately 2,000 unique examples. Each column reports the subset with the largest gain on that benchmark (named below the benchmark); best SFT/GRPO accuracy is bold. Benchmarks: SAT (Ray et al., 2024), ActionEQA (Anonymous, 2025), TemporalBench-long (Cai et al., 2024), InPhyRe (Sreekumar & Boddeti, 2025), Robo2VLM (Chen et al., 2025a), WorldPrediction (WorldPrediction Team, 2025), DSI-Bench (Zhang et al., 2025c), MindCube (Yin et al., 2025), CLEVRER (Yi et al., 2020), WM-ABench (Huang Team and Contributors, 2025), MVP (Krojer et al., 2025), ERQA (Gemini Robotics Team, 2025), SpatialViz (Wang et al., 2026b), ViewSpatial (Li et al., 2025a), PaiBench (Zhou et al., 2025), CV-Bench (Tong et al., 2024), Cosmos (NVIDIA, 2025a), BLINK (Fu et al., 2024), TOMATO (Shangguan et al., 2025), TVBench (Cores et al., 2025), BlackSwan (Chinchure et al., 2025), TempCompass (Liu et al., 2024). The four remaining benchmarks and all subset results are reported below.

	SAT perc_causal	ActionEQA perc_causal	TemporalBench-L mani_cf	InPhyRe insp_cf	Robo2VLM insp_cf	WorldPrediction mani_causal	DSI-Bench perc_causal	MindCube perc_temporal	CLEVRER insp_cf	WM-ABench mani_cf	MVP perc_causal	ERQA mani_causal	SpatialViz navi_temporal	ViewSpatial mani_cf	PaiBench insp_causal	CV-Bench insp_causal	Cosmos insp_causal	BLINK exo_physical	TOMATO insp_temporal	TVBench perc_causal	BlackSwan mani_causal	TempCompass mani_causal
Base	37.3	43.4	16.6	22.0	33.5	34.8	34.9	37.2	58.2	31.0	24.4	33.8	27.7	35.3	48.3	69.3	76.7	46.2	31.6	36.5	61.8	67.7
Subset SFT	63.3	47.5	22.0	32.3	40.5	40.1	40.3	40.0	62.7	**35.6**	28.1	37.5	**32.2**	37.9	51.5	72.4	**80.0**	48.9	34.2	38.6	63.8	69.9
Subset GRPO	**64.7**	**55.9**	**28.2**	**32.6**	**42.4**	**42.9**	**40.9**	**43.1**	**63.3**	35.4	**29.0**	**38.3**	31.8	**39.6**	**51.8**	**72.9**	79.5	**49.5**	**34.7**	**39.4**	**64.3**	**70.2**
Δ	+27.3	+12.5	+11.7	+10.6	+8.9	+8.1	+6.0	+5.9	+5.1	+4.6	+4.5	+4.5	+4.5	+4.3	+3.5	+3.5	+3.3	+3.3	+3.1	+2.8	+2.5	+2.4

Table 43: Mean pairwise Pearson correlation of the 11 cells' 26-dimensional transfer profiles, grouped by what the cell pair shares, under four normalizations of the per-benchmark gains.

Normalization	same reasoning family	same action type	neither
raw Δ (points)	$+0.71$	$+0.53$	$+0.42$
relative Δ/base	$+0.65$	$+0.49$	$+0.42$
headroom $\Delta/(100 - \text{base})$	$+0.70$	$+0.46$	$+0.36$
per-benchmark z-score	$+0.32$	-0.22	-0.23

4 static-perception and general video QA benchmarks) and the WOVEN in-distribution gains of Table 75. *Common component.* A subset's transfer profile is its vector of gains over the base model across the benchmarks. All 55 pairwise Pearson correlations between the 11 profiles are positive (mean 0.56, range 0.11–0.92; Fig. 4, left). For an entrywise-positive correlation matrix, the Perron–Frobenius theorem guarantees a leading eigenvector whose components all share one sign, that is, a single factor on which every subset loads positively; here its eigenvalue is 6.7 of 11. Equivalently, the first principal component of the profile matrix explains 75% of the variance of the raw profiles (54% after centering each benchmark, 42% after standardizing each benchmark), and its scores correlate with the subsets' WOVEN gains at $|r|$=0.92–0.93 under all three normalizations. The subsets therefore improve largely the same benchmarks, and the shared part of their transfer tracks how much each improves WOVEN. *Dose–response.* Across the 11 subsets, the WOVEN in-distribution gain after SFT predicts the mean external gain: Pearson r=0.86 (p=0.0007), Spearman ρ=0.86; r=0.82 with the post-GRPO WOVEN gain, 0.82 against the number of benchmarks improved by more than 2 points, 0.84–0.92 under leave-one-subset-out, and positive within the temporal (0.68, n=4) and non-temporal (0.79, n=7) subsets separately, so the relation is not driven by the difference between these two groups. Specificity complements the dose–response: the benchmarks that improve are exactly the transition-related ones, and none of the 4 static-perception and general video QA benchmarks gains more than 2 points under any subset or under the full training set.

Complementary reasoning operations. Two within-WOVEN cases show that supervision of one reasoning operation does not necessarily teach another. (i) Temporal adjacency training does not produce temporal ordering nor vice versa (Appendix G.9). (ii) The exogenous-physical cell, whose queries never condition on an agent's action, attains the largest state-perturbation robustness gain of any cell ($+29.9$) yet transfers negatively to the two benchmarks built on endogenous, agent-driven transitions (WorldPrediction -10.2, ActionEQA -1.7): exogenous supervision does not teach endogenous transitions.

InPhyRe subset anomaly. A subset of InPhyRe (the "irregular" scenes) deliberately constructs fictional, real-world-violating physics (e.g. objects swap colors after collision) and marks the real-world-correct answer wrong; base models therefore score far below chance on these subsets (0.0% on `color_constancy`). Post-training moves several subsets violently in both directions (`color_constancy` 0.0 $\rightarrow$ 84.2; `off_the_wall_regular` 91.0 $\rightarrow$ 0.3), so the aggregate InPhyRe gain of $+13.0$ nets over large opposing subset swings; the per-category table below gives the full picture, and we flag this benchmark's aggregate as protocol-sensitive.

Per-benchmark, per-category results. The tables below give, for each suite benchmark, accuracy by official category for the base model and all 12 training configurations (columns: `base`; `full`=full-mix; then the 11 cells, abbreviated by action type `p`erceptive/`i`nspective/`n`avigative/`m`anipulative/`e`xogenous $\times$ reasoning family `ca`usal/`cf` counterfactual/`te`mporal/`ph`ysical). Each configuration column reports its selected checkpoint (§4); category rows come from the same checkpoint as that configuration's overall score.

G.14 E-F12. In-domain substitution sweep: protocol and full tables

Protocol. Three external benchmarks with public training splits are paired with their best-matching WOVEN cell: SAT with `insp_causal`, ActionEQA with `mani_causal`, and CLEVRER with `insp_cf`. For each share $p \in \{0, 10, 20, 30, 40, 50, 70, 100\}\%$, the training set holds a fixed 2,000-item budget: $2{,}000 \cdot (1 - p)$ items drawn from the benchmark's own training split in its native format, and $2{,}000 \cdot p$ items from the WOVEN cell. Items are packed into 16-row blocks that are answer-letter balanced within each source type, and the answer-letter sequence is identical across

Table 44: SAT (circular): accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	action_conseq	ego_movement	goal_aim	obj_movement	perspective
base	**37.3**	62.2	34.8	35.3	8.7	33.3
full-mix	**61.3**	73.0	87.0	79.4	26.1	36.4
perc_causal	**64.7**	73.0	100.0	73.5	30.4	45.5
perc_temporal	**46.0**	64.9	43.5	50.0	17.4	42.4
insp_causal	**63.3**	81.1	78.3	64.7	39.1	48.5
insp_cf	**56.7**	73.0	52.2	67.6	34.8	45.5
insp_temporal	**42.7**	59.5	60.9	44.1	8.7	33.3
navi_causal	**54.0**	54.1	82.6	55.9	21.7	54.5
navi_temporal	**40.7**	64.9	52.2	38.2	4.3	33.3
mani_causal	**51.3**	54.1	78.3	61.8	13.0	45.5
mani_cf	**53.3**	67.6	73.9	61.8	17.4	39.4
exo_physical	**45.3**	59.5	73.9	41.2	13.0	36.4
exo_temporal	**42.7**	56.8	30.4	64.7	4.3	39.4

Table 45: ActionEQA: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	bridge_high_forward	bridge_high_inverse	droid_high_forward	droid_high_inverse	rt1_high_forward	rt1_high_inverse
base	**43.4**	25.0	37.0	36.8	64.9	62.0	59.3
full-mix	**54.4**	49.4	48.7	47.0	68.0	59.1	64.7
perc_causal	**55.9**	56.4	45.9	40.2	61.2	76.4	58.9
perc_temporal	**51.4**	49.6	46.2	40.8	66.9	57.5	54.4
insp_causal	**54.3**	52.3	39.7	53.5	62.9	74.4	58.6
insp_cf	**55.9**	52.5	45.0	52.4	66.0	67.4	65.4
insp_temporal	**45.2**	37.1	44.4	34.0	62.0	53.5	48.8
navi_causal	**47.8**	33.3	40.2	43.6	62.9	64.5	63.1
navi_temporal	**51.4**	50.5	50.3	36.5	62.3	55.1	53.9
mani_causal	**55.0**	40.5	57.8	36.5	64.0	73.9	62.7
mani_cf	**53.2**	40.4	50.5	43.1	65.7	67.2	65.0
exo_physical	**41.7**	31.2	41.5	26.1	61.8	42.9	56.6
exo_temporal	**47.4**	40.1	42.2	41.9	60.3	57.5	54.4

Table 46: TemporalBench-long (BA): accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	ActivityNet	COIN	Charades	EgoExo4D	FineGym
base	**16.6**	17.5	16.4	15.8	16.7	14.9
full-mix	**26.8**	34.1	26.5	15.8	28.2	19.3
perc_causal	**17.7**	18.9	18.2	19.0	16.6	14.9
perc_temporal	**25.7**	30.4	25.3	20.2	27.4	14.6
insp_causal	**15.5**	17.5	16.6	13.7	12.8	19.3
insp_cf	**20.3**	21.0	21.2	16.0	20.3	21.8
insp_temporal	**11.8**	12.6	10.9	10.3	13.7	8.0
navi_causal	**14.9**	15.8	13.8	13.7	14.3	21.8
navi_temporal	**24.9**	27.3	23.7	21.6	27.0	19.3
mani_causal	**19.5**	22.2	20.4	19.9	16.0	21.8
mani_cf	**28.2**	29.6	29.0	26.2	28.8	21.8
exo_physical	**22.2**	24.3	23.3	19.9	21.3	17.9
exo_temporal	**22.3**	24.2	21.5	20.5	23.3	19.3

Table 47: InPhyRe: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	ang_mom_cons	color_constancy	color_lin_mom_cons	color_passthrough	lin_mom_cons	lin_mom_cons_regular	off_the_wall	off_the_wall_regular	size_bias	size_bias_regular
base	**22.0**	29.2	0.0	2.9	2.9	0.1	5.7	8.9	91.0	0.3	79.2
full-mix	**34.9**	73.2	84.2	0.6	0.8	0.0	99.8	23.2	0.2	14.6	52.8
perc_causal	**29.1**	37.4	3.7	5.1	4.8	2.4	64.7	1.6	96.8	0.2	75.3
perc_temporal	**29.7**	45.3	5.3	3.0	2.9	0.5	69.7	34.7	63.0	2.2	71.1
insp_causal	**29.9**	51.7	30.1	3.3	3.2	0.1	79.8	45.9	25.1	0.2	59.5
insp_cf	**32.6**	57.8	3.0	8.4	9.7	9.9	69.7	41.6	38.0	0.8	87.4
insp_temporal	**26.8**	39.2	5.5	4.5	4.0	1.1	66.8	45.5	45.0	4.6	52.2
navi_causal	**25.9**	32.0	3.9	4.3	4.0	1.0	52.6	20.2	68.7	5.2	67.3
navi_temporal	**29.3**	46.5	18.4	4.0	4.2	0.8	78.2	41.2	42.1	0.5	57.4
mani_causal	**28.3**	15.2	0.2	14.6	14.6	14.6	46.1	2.4	97.5	12.8	65.8
mani_cf	**29.9**	40.5	10.7	3.4	3.5	0.3	71.0	26.7	63.9	0.8	79.1
exo_physical	**30.6**	62.1	16.1	5.0	5.3	3.3	25.4	25.9	72.8	0.0	90.5
exo_temporal	**31.6**	49.3	20.2	4.8	5.0	2.5	63.8	52.2	43.9	2.6	72.0

Table 48: Robo2VLM: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	robo2vlm
base	**33.5**	33.5
full-mix	**31.8**	31.8
perc_causal	**39.1**	39.1
perc_temporal	**34.2**	34.2
insp_causal	**39.1**	39.1
insp_cf	**42.4**	42.4
insp_temporal	**37.7**	37.7
navi_causal	**38.4**	38.4
navi_temporal	**34.2**	34.2
mani_causal	**37.2**	37.2
mani_cf	**41.6**	41.6
exo_physical	**35.0**	35.0
exo_temporal	**33.9**	33.9

Table 49: WorldPrediction-WM: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	COIN	CrossTask	EPIC-KITCHENS-100	EgoExo4D	IKEAASM
base	**34.8**	29.7	30.3	48.2	36.7	27.7
full-mix	**41.5**	35.2	32.1	57.0	43.8	36.5
perc_causal	**37.2**	30.5	28.4	54.4	32.0	36.5
perc_temporal	**35.4**	33.1	27.5	49.7	34.4	27.7
insp_causal	**39.9**	34.3	33.0	55.4	39.8	34.0
insp_cf	**29.7**	28.0	24.8	37.3	30.5	25.8
insp_temporal	**31.4**	31.4	25.7	39.9	32.0	24.5
navi_causal	**38.9**	29.2	35.8	55.4	40.6	34.0
navi_temporal	**29.7**	28.4	25.7	37.8	31.2	23.3
mani_causal	**42.9**	31.8	33.0	59.6	52.3	38.4
mani_cf	**42.5**	29.7	34.9	65.3	40.6	40.9
exo_physical	**24.6**	16.9	21.1	33.2	23.4	28.9
exo_temporal	**32.7**	28.8	24.8	44.6	33.6	28.9

Table 50: DSI-Bench: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	cate0	cate1	cate2	cate3	cate4	cate5
base	**34.9**	34.7	39.3	26.3	32.4	48.0	22.9
full-mix	**35.2**	35.5	31.7	40.9	29.4	65.9	31.2
perc_causal	**40.9**	47.3	44.3	33.5	36.6	53.8	30.0
perc_temporal	**36.7**	40.0	43.3	23.8	32.4	51.6	23.8
insp_causal	**34.0**	36.9	33.0	32.2	28.1	62.1	32.4
insp_cf	**38.9**	45.9	46.3	30.1	35.7	34.8	25.6
insp_temporal	**36.5**	41.4	43.3	20.2	33.3	47.5	26.5
navi_causal	**39.5**	48.8	46.0	25.2	34.8	50.7	25.9
navi_temporal	**32.2**	35.7	35.5	25.3	26.7	47.5	30.9
mani_causal	**38.4**	45.9	44.7	25.4	32.8	52.2	27.9
mani_cf	**39.0**	47.0	45.9	27.1	33.8	48.2	25.0
exo_physical	**37.8**	44.1	42.1	24.9	34.1	53.6	28.2
exo_temporal	**33.9**	39.1	37.4	26.4	28.6	50.9	25.9

Table 51: DSI-Bench: accuracy (%) by spatio-temporal variant.

Config	Overall	std	hflip	reverse	reverse_hflip
base	**34.9**	42.3	41.8	28.0	27.4
full-mix	**35.2**	31.1	30.9	39.2	39.6
perc_causal	**40.9**	42.0	41.6	39.5	40.4
perc_temporal	**36.7**	48.0	46.9	25.6	26.5
insp_causal	**34.0**	42.6	42.3	25.7	25.5
insp_cf	**38.9**	49.6	49.9	27.9	28.4
insp_temporal	**36.5**	42.4	41.5	31.0	31.1
navi_causal	**39.5**	46.5	46.5	32.3	32.9
navi_temporal	**32.2**	39.1	38.3	25.6	25.8
mani_causal	**38.4**	48.3	48.3	28.3	28.7
mani_cf	**39.0**	49.1	49.0	28.7	29.2
exo_physical	**37.8**	50.0	48.7	26.5	26.2
exo_temporal	**33.9**	41.0	40.3	27.4	27.0

Table 52: MindCube: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	linear/O-O/mw/self	linear/P-O/mw/level2	linear/P-P/mw/self	perp/O-O/mw/level2	perp/O-O/mw/self	perp/P-O/mw/level1	perp/P-O/mw/level2	perp/P-O/mw/self	perp/P-O/rot/self	perp/P-O/scq/self	perp/P-P/mw/self
base	**37.2**	51.5	55.6	41.0	33.3	21.7	50.0	25.0	18.2	33.0	50.3	40.8
full-mix	**48.6**	75.7	44.4	58.1	33.3	27.4	75.0	75.0	18.2	29.5	87.6	44.4
perc_causal	**31.4**	40.4	55.6	33.3	33.3	21.7	75.0	25.0	19.5	32.5	41.8	25.4
perc_temporal	**43.1**	65.4	55.6	41.9	27.8	21.7	75.0	75.0	18.2	34.0	71.2	44.4
insp_causal	**38.3**	42.6	55.6	49.5	38.9	29.7	50.0	25.0	19.5	34.0	48.4	40.2
insp_cf	**41.4**	53.7	66.7	45.7	27.8	25.1	75.0	50.0	22.1	30.5	58.8	50.9
insp_temporal	**35.7**	52.9	44.4	33.3	38.9	22.3	50.0	25.0	26.0	33.5	50.3	30.2
navi_causal	**32.2**	41.2	11.1	32.4	27.8	24.6	50.0	75.0	16.9	35.0	32.7	36.1
navi_temporal	**40.4**	55.1	77.8	49.5	33.3	20.0	50.0	50.0	20.8	34.5	55.6	44.4
mani_causal	**33.8**	50.0	11.1	33.3	33.3	20.6	50.0	50.0	15.6	32.5	48.4	32.0
mani_cf	**39.0**	47.1	55.6	48.6	38.9	21.7	100.0	75.0	23.4	29.5	54.9	45.6
exo_physical	**41.4**	60.3	55.6	46.7	27.8	21.7	50.0	50.0	24.7	35.5	55.6	45.6
exo_temporal	**41.0**	52.9	55.6	48.6	44.4	24.0	50.0	25.0	27.3	35.5	50.3	47.3

Table 53: CLEVRER: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	counterfactual	explanatory	predictive
base	**58.2**	52.3	49.5	72.8
full-mix	**56.8**	51.1	51.2	68.2
perc_causal	**62.0**	53.6	50.6	81.7
perc_temporal	**61.7**	52.8	54.7	77.5
insp_causal	**58.9**	52.3	51.5	73.0
insp_cf	**63.3**	56.0	56.1	77.7
insp_temporal	**58.8**	52.4	52.3	71.6
navi_causal	**57.8**	51.0	47.9	74.5
navi_temporal	**58.7**	51.4	51.4	73.3
mani_causal	**55.2**	50.1	45.7	69.8
mani_cf	**58.7**	54.9	51.8	69.3
exo_physical	**56.7**	50.6	49.8	69.7
exo_temporal	**62.6**	54.1	56.8	76.8

arms, so arms differ only in the substituted content. Training uses the shared SFT optimization protocol of Appendix G.1 for 3 epochs, with seeds $42/43/44$, followed by 300 GRPO steps with the anti-bias reward. The fixed-budget datasets use the answer-balanced blocks described above without additional option-permutation augmentation. Evaluation uses each benchmark's own protocol (SAT scored circularly).

Trend statistics. An OLS fit of in-domain accuracy on the share over $p \leq 50\%$ (n=18 runs per benchmark) gives slopes, per $+10\%$ share, of -0.03 ± 1.00 pp for SAT ($t(16)$=-0.03, p=0.98), -1.54 ± 0.30 pp for ActionEQA ($t(16)$=-5.08, p=1.1×10^{-4}), and -0.01 ± 0.30 pp for CLEVRER ($t(16)$=-0.04, p=0.97). Per-arm seed ranges are visible in Table 71 (SAT carries the widest), so the sweep is read at the level of the fitted trend rather than of individual arms.

Deletion control. To test whether preserved accuracy under substitution merely reflects saturation of the task-specific training budget, we ablate the WOVEN contribution at $p \in \{20, 40, 50\}\%$. Each deletion arm retains exactly the same task-specific subset as its substitution counterpart but omits the WOVEN examples, leaving 1,600, 1,200, or 1,000 training items under the same optimization recipe. At 40% and 50%, substitution outperforms deletion on all three benchmarks by 3.60–11.00 percentage points (Table 72); at 20%, it improves ActionEQA and CLEVRER but is 1.99 points lower on SAT. This pattern supports an additional contribution from WOVEN supervision beyond simply discarding redundant task-specific data. In Fig. 3, the deletion curves connect these three ablation settings to the shared p=0 reference.

G.15 E-F13. Transfer at scale: 7B and 32B

The transfer of §4 persists when the same full-mix recipe is applied at larger scale. Table 73 reports zero-shot transfer gains over the same-scale base for Qwen2.5-VL-7B (SFT and GRPO stages, following the 3B recipe) and Qwen2.5-VL-32B (SFT stage). Gains of the same character as at 3B appear at both scales: egocentric spatial reasoning (SAT $+20.0$ at 7B, $+22.7$ at 32B), paired video perception (MVP), world-model probes (WM-ABench), and dynamic-scene inference (DSI), alongside single-scale gains on temporal, embodied, and intuitive-physics surfaces.

G.16 E-F14. Same-backbone comparison with video-SSL mid-training

Orca-4B (Wang et al., 2026c) mid-trains the Qwen3.5-4B backbone on latent next-state prediction over 125K hours of video with 160M event annotations. We post-train the *same* Qwen3.5-4B backbone on WOVEN supervision alone: SFT on ≈107K QA pairs drawn from ≈7 hours of generated video, four orders of magnitude less video and a post-training-only budget. The comparison uses 3DSRBench, a benchmark on which Orca-4B reports results. On this backbone, WOVEN post-training with about four orders of magnitude less video yields a larger gain over the shared base than Orca-4B mid-training ($+6.7$ vs. $+4.0$ points; Table 74).

Table 54: WM-ABench: accuracy (%) by official category. Rows: training configurations; columns: categories.

| Config | Overall | Perc/Motion/MotDir_maniskill_subtask | Perc/Motion/MotDir_maniskill_towards | Perc/Motion/MotDir_tdw | Perc/Motion/MotDir_tdw_subtask | Perc/Motion/MotSpd_maniskill | Perc/Motion/MotSpd_tdw | Perc/Motion/MotTraj_maniskill_path_choices | Perc/Motion/MotTraj_path_choices | Perc/Motion/speed_tdw | Perc/Quantity/DiscreteCounting_tdw_top | Perc/Quantity/QuantityContinuous_tdw_continuous | Perc/Quantity/RelativeCounting_maniskill | Perc/Quantity/RelativeCounting_tdw_color_letter | Perc/Spatial/Multiview_maniskill | Perc/Spatial/Multiview_tdw | Perc/Spatial/RelPos_front_top | Perc/Spatial/RelPos_front_top_tdw | Perc/Spatial/SpatialOccupancy_maniskill | Perc/Spatial/SpatialOccupancy_occupancy_fitting_tdw_top | Perc/Spatial/SpatialOccupancy_occupancy_size_tdw_front_large | Perc/Spatial/SpatialOccupancy_volume_object_only_unfloat_top | Perc/Temporal/TemporalExtension_temporal_tdw_yjx_01_05 | Perc/Temporal/TemporalPositioning_maniskill | Perc/Visual/VisualAttribute_color | Perc/Visual/VisualAttribute_color_tdw | Perc/Visual/VisualAttribute_material_diff_tdw | Perc/Visual/VisualAttribute_shape_tdw | Pred/Compositionality/Compositionality_maniskill_lift | Pred/Compositionality/Compositionality_maniskill_push | Pred/Compositionality/Compositionality_maniskill_rotate | Pred/Compositionality/Compositionality_tdw_two_sphere_2_14 | Pred/Mechanistic/MechanisticKnowledge_maniskill_drop | Pred/Mechanistic/MechanisticKnowledge_maniskill_lift | Pred/Mechanistic/MechanisticKnowledge_maniskill_push | Pred/Mechanistic/MechanisticKnowledge_maniskill_rotate | Pred/Mechanistic/MechanisticKnowledge_phyision_type_collision_v2 | Pred/Mechanistic/MechanisticKnowledge_phyision_type_drop | Pred/Mechanistic/MechanisticKnowledge_phyision_type_slide | Pred/Mechanistic/PredNextState_carla_next_state_prediction_02_08 | Pred/Mechanistic/Transitivity_habitat_lab_trans_ver_2_noback | Pred/Transitivity/Transitivity_carla_transitivity_yjx_0218 | Pred/Transitivity/Transitivity_habitat_lab_seq_ver_3_noback | Pred/Transitivity/Transitivity_maniskill_pick_rotate | Pred/Transitivity/Transitivity_maniskill_push_pick |
|---|---|---

Table 55: MVP: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	human_object_interactions	intuitive_physics	robot_object_interactions	temporal_reasoning
base	**24.4**	22.6	17.0	25.8	32.2
full-mix	**26.2**	25.6	14.7	29.6	35.1
perc_causal	**28.9**	25.7	18.4	32.4	39.3
perc_temporal	**21.6**	19.1	16.4	19.8	30.9
insp_causal	**27.1**	24.5	16.3	33.3	34.2
insp_cf	**24.8**	22.2	18.1	28.7	30.2
insp_temporal	**22.8**	21.2	16.5	22.2	31.4
navi_causal	**25.7**	24.1	17.5	28.2	33.0
navi_temporal	**22.1**	19.9	14.3	23.4	30.8
mani_causal	**28.4**	26.2	16.8	34.4	36.1
mani_cf	**26.0**	25.3	16.4	27.3	35.1
exo_physical	**25.7**	23.3	15.1	28.4	35.9
exo_temporal	**20.4**	20.9	16.5	13.7	30.7

Table 56: ERQA: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	Action Reasoning	Multi-view Reasoning	Other	Pointing	Spatial Reasoning	State Estimation	Task Reasoning	Trajectory Reasoning
base	**33.8**	40.3	27.0	14.3	35.3	29.8	40.0	42.1	28.8
full-mix	**37.5**	38.9	13.5	35.7	38.2	36.9	50.9	47.4	33.3
perc_causal	**35.8**	37.5	32.4	28.6	35.3	33.3	36.4	50.0	31.8
perc_temporal	**35.2**	41.7	29.7	14.3	38.2	38.1	40.0	26.3	31.8
insp_causal	**37.0**	37.5	24.3	35.7	38.2	44.0	40.0	42.1	28.8
insp_cf	**34.8**	34.7	21.6	28.6	29.4	35.7	40.0	47.4	33.3
insp_temporal	**35.5**	40.3	18.9	28.6	41.2	32.1	43.6	42.1	31.8
navi_causal	**34.5**	34.7	24.3	28.6	47.1	28.6	45.5	44.7	27.3
navi_temporal	**37.5**	34.7	29.7	28.6	44.1	34.5	49.1	42.1	34.8
mani_causal	**38.2**	45.8	18.9	28.6	44.1	35.7	34.5	52.6	37.9
mani_cf	**34.8**	40.3	18.9	35.7	29.4	25.0	43.6	52.6	34.8
exo_physical	**36.2**	40.3	35.1	14.3	35.3	34.5	47.3	39.5	28.8
exo_temporal	**32.8**	34.7	16.2	7.1	38.2	32.1	49.1	31.6	30.3

Table 57: SpatialViz: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	MentalAnimation/ArrowMoving	MentalAnimation/BlockMoving	MentalAnimation/MechanicalSystem	MentalFolding/CubeReconstruction	MentalFolding/CubeUnfolding	MentalFolding/PaperFolding	MentalRotation/2DRotation	MentalRotation/3DRotation	MentalRotation/3ViewProjection	VisualPenetration/CrossSection	VisualPenetration/CubeAssembly	VisualPenetration/CubeCounting
base	**27.7**	21.2	43.8	52.5	32.5	28.3	20.8	20.0	8.8	22.0	7.5	50.0	34.2
full-mix	**31.1**	33.8	31.2	32.5	30.8	25.0	29.2	25.0	33.8	29.0	16.7	46.2	45.0
perc_causal	**29.0**	30.0	45.0	61.2	23.3	24.2	22.5	27.5	7.5	30.0	6.7	50.0	35.8
perc_temporal	**30.0**	23.8	40.0	45.0	21.7	31.7	27.5	28.8	33.8	30.0	11.7	46.2	32.5
insp_causal	**30.3**	21.2	38.8	50.0	32.5	25.0	22.5	33.8	35.0	26.0	6.7	57.5	31.7
insp_cf	**28.6**	23.8	30.0	53.8	24.2	25.0	26.7	32.5	8.8	32.0	8.3	45.0	41.7
insp_temporal	**30.8**	36.2	38.8	46.2	31.7	27.5	19.2	31.2	37.5	34.0	10.0	41.2	32.5
navi_causal	**30.7**	27.5	41.2	56.2	25.0	30.0	36.7	25.0	13.8	32.0	5.8	51.2	34.2
navi_temporal	**32.2**	30.0	40.0	51.2	30.8	30.8	30.0	25.0	35.0	35.0	7.5	50.0	34.2
mani_causal	**31.6**	33.8	38.8	48.8	30.0	29.2	35.8	26.2	17.5	39.0	7.5	42.5	37.5
mani_cf	**28.6**	33.8	32.5	38.8	25.8	28.3	30.0	30.0	1.2	30.0	8.3	51.2	38.3
exo_physical	**30.8**	22.5	36.2	50.0	26.7	30.8	30.8	25.0	36.2	28.0	9.2	48.8	36.7
exo_temporal	**31.6**	28.8	38.8	53.8	23.3	32.5	23.3	31.2	42.5	33.0	6.7	47.5	35.8

Table 58: ViewSpatial: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	Camera perspective - Object View Orientation	Camera perspective - Relative Direction	Person perspective - Object View Orientation	Person perspective - Relative Direction	Person perspective - Scene Simulation Relative Direction
base	**35.3**	29.7	42.9	42.4	28.6	26.8
full-mix	**34.1**	28.1	43.3	35.4	28.3	28.1
perc_causal	**38.1**	34.6	43.8	50.6	30.3	26.7
perc_temporal	**38.3**	30.5	43.2	49.2	34.2	30.5
insp_causal	**37.2**	29.4	44.1	47.5	28.7	30.1
insp_cf	**37.2**	27.9	41.4	54.1	34.0	26.1
insp_temporal	**32.7**	28.2	40.7	36.8	27.2	24.4
navi_causal	**36.6**	28.0	44.2	48.6	28.3	27.7
navi_temporal	**31.6**	30.2	41.3	29.8	25.4	23.6
mani_causal	**37.0**	28.0	43.1	50.9	32.1	26.6
mani_cf	**39.6**	26.2	44.9	59.1	34.6	29.4
exo_physical	**37.4**	31.2	44.6	45.8	30.3	29.5
exo_temporal	**35.8**	25.8	40.7	47.0	33.8	28.5

Table 59: PaiBench: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	Common Sense/Physical world	Common Sense/Space: Affordance	Common Sense/Space: Environment	Common Sense/Space: Plausibility	Common Sense/Space: Relationship	Common Sense/Time: Actions	Common Sense/Time: Camera	Common Sense/Time: Causality	Common Sense/Time: Order	Common Sense/Time: Planning	Embodied Reasoning/AV	Embodied Reasoning/Agibot	Embodied Reasoning/BridgeData V2	Embodied Reasoning/HoloAssist	Embodied Reasoning/RoboFail	Embodied Reasoning/RoboVQA
base	**48.3**	43.4	18.8	68.2	23.5	72.0	51.0	46.1	25.0	40.0	66.7	27.0	36.0	31.0	52.0	67.0	85.5
full-mix	**50.2**	47.3	43.8	59.1	41.2	68.0	47.6	60.7	50.0	43.6	66.7	38.0	34.0	31.0	48.0	62.0	85.5
perc_causal	**50.3**	42.0	37.5	77.3	17.6	60.0	52.4	43.8	25.0	47.3	33.3	49.0	36.0	24.0	54.0	72.0	87.3
perc_temporal	**50.2**	46.0	31.2	72.7	70.6	60.0	50.3	44.9	0.0	50.9	33.3	30.0	37.0	33.0	54.0	65.0	87.3
insp_causal	**51.8**	46.0	25.0	63.6	17.6	72.0	55.8	55.1	75.0	47.3	66.7	40.0	34.0	33.0	52.0	73.0	83.6
insp_cf	**49.6**	41.6	25.0	59.1	29.4	56.0	51.0	47.2	50.0	50.9	66.7	35.0	35.0	29.0	57.0	72.0	86.4
insp_temporal	**50.2**	49.6	31.2	72.7	52.9	76.0	49.7	39.3	25.0	36.4	66.7	41.0	37.0	27.0	53.0	63.0	87.3
navi_causal	**48.7**	43.4	25.0	77.3	23.5	60.0	47.6	43.8	25.0	50.9	66.7	27.0	36.0	29.0	58.0	68.0	86.4
navi_temporal	**49.8**	48.2	18.8	81.8	41.2	72.0	50.3	41.6	50.0	43.6	66.7	33.0	35.0	30.0	52.0	67.0	85.5
mani_causal	**49.5**	43.4	31.2	68.2	29.4	68.0	50.3	42.7	25.0	38.2	66.7	37.0	38.0	34.0	53.0	69.0	85.5
mani_cf	**48.5**	39.4	25.0	72.7	17.6	60.0	46.9	46.1	0.0	45.5	33.3	29.0	38.0	36.0	57.0	69.0	88.2
exo_physical	**48.0**	42.5	31.2	77.3	23.5	68.0	44.2	50.6	0.0	38.2	66.7	35.0	36.0	30.0	46.0	72.0	83.6
exo_temporal	**50.1**	50.0	25.0	77.3	47.1	68.0	51.0	42.7	50.0	40.0	66.7	36.0	33.0	30.0	52.0	64.0	86.4

Table 60: CV-Bench: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	2D/Count/ADE20K	2D/Count/COCO	2D/Relation/ADE20K	2D/Relation/COCO	3D/Depth/Omni3D	3D/Distance/Omni3D
base	**69.3**	50.9	71.7	69.4	76.0	75.5	67.8
full-mix	**72.4**	55.8	76.0	72.9	77.4	77.5	70.7
perc_causal	**72.3**	54.4	71.3	75.9	79.4	78.2	71.3
perc_temporal	**70.4**	56.4	72.0	71.1	77.2	78.3	64.7
insp_causal	**72.9**	53.5	72.4	79.7	83.0	78.0	69.7
insp_cf	**70.5**	53.2	72.2	72.5	76.3	75.7	69.7
insp_temporal	**68.0**	52.6	71.1	70.8	74.7	75.5	61.8
navi_causal	**72.1**	50.3	70.6	75.6	80.5	80.3	70.5
navi_temporal	**68.2**	52.0	70.6	71.8	78.0	76.5	59.5
mani_causal	**69.5**	51.5	70.6	72.2	74.9	78.2	65.7
mani_cf	**71.4**	51.5	71.3	77.0	80.5	80.2	66.0
exo_physical	**68.7**	49.7	69.7	68.4	75.8	72.8	70.7
exo_temporal	**69.1**	54.1	70.4	72.9	77.7	79.2	59.7

Table 61: Cosmos: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	robofail	robovqa
base	**76.7**	67.0	85.5
full-mix	**75.2**	63.0	86.4
perc_causal	**80.0**	72.0	87.3
perc_temporal	**75.2**	63.0	86.4
insp_causal	**80.0**	71.0	88.2
insp_cf	**79.0**	69.0	88.2
insp_temporal	**75.7**	65.0	85.5
navi_causal	**79.0**	72.0	85.5
navi_temporal	**76.7**	67.0	85.5
mani_causal	**78.6**	70.0	86.4
mani_cf	**79.5**	71.0	87.3
exo_physical	**76.7**	70.0	82.7
exo_temporal	**75.7**	64.0	86.4

Table 62: BLINK: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	Art_Style	Counting	Forensic_Detection	Functional_Correspondence	IQ_Test	Jigsaw	Multi-view_Reasoning	Object_Localization	Relative_Depth	Relative_Reflectance	Semantic_Correspondence	Spatial_Relation	Visual_Correspondence	Visual_Similarity
base	**46.2**	51.3	64.2	26.5	24.6	28.0	50.7	44.4	55.7	65.3	47.8	27.3	81.8	31.4	55.6
full-mix	**45.4**	64.1	69.2	25.8	23.1	29.3	47.3	44.4	49.2	62.9	26.9	27.3	76.9	34.3	64.4
perc_causal	**49.0**	53.0	65.8	45.5	29.2	32.0	46.7	36.1	57.4	71.8	32.8	27.3	85.3	40.7	68.9
perc_temporal	**46.4**	52.1	59.2	39.4	24.6	30.0	47.3	51.9	55.7	58.1	43.3	26.6	79.7	28.5	62.2
insp_causal	**47.1**	65.8	65.8	25.8	26.9	28.7	46.0	49.6	55.7	73.4	38.1	23.0	80.4	24.4	69.6
insp_cf	**48.4**	55.6	65.8	33.3	26.9	28.0	42.0	50.4	55.7	66.1	45.5	27.3	82.5	39.0	67.4
insp_temporal	**46.5**	54.7	67.5	29.5	31.5	29.3	46.0	44.4	55.7	60.5	48.5	29.5	81.8	29.1	52.6
navi_causal	**46.8**	59.0	67.5	18.2	29.2	27.3	42.0	40.6	60.7	73.4	48.5	22.3	81.8	32.0	63.7
navi_temporal	**45.9**	59.0	61.7	25.8	25.4	28.0	48.7	49.6	58.2	62.9	43.3	23.0	80.4	26.2	60.7
mani_causal	**49.0**	61.5	65.0	40.2	27.7	33.3	45.3	48.9	58.2	69.4	41.8	25.2	84.6	26.7	69.6
mani_cf	**47.6**	53.0	65.0	29.5	27.7	25.3	50.0	42.9	55.7	69.4	47.0	33.8	84.6	29.7	61.5
exo_physical	**49.4**	58.1	66.7	32.6	26.2	29.3	46.7	45.1	56.6	66.9	50.7	30.9	83.2	40.1	66.7
exo_temporal	**45.0**	56.4	66.7	28.0	23.8	27.3	51.3	44.4	57.4	59.7	38.1	28.8	82.5	25.0	51.1

Table 63: TOMATO: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	count/human	count/object	count/simulated	direction/human	direction/object	direction/simulated	rot/human	rot/object	rot/simulated	shape&trend/human	shape&trend/simulated	velocity&frequency/human	velocity&frequency/object	velocity&frequency/simulated	visual cues/object
base	**31.6**	19.3	47.6	63.4	21.8	42.4	28.6	31.6	32.3	28.0	34.5	34.7	9.4	20.5	14.3	32.9
full-mix	**36.1**	36.8	43.9	59.2	26.3	40.2	24.5	23.7	28.8	28.0	41.4	46.9	42.2	37.9	50.0	32.9
perc_causal	**29.2**	10.5	46.3	57.7	20.2	52.2	26.5	10.5	29.3	32.0	29.3	32.7	9.4	15.2	14.3	34.3
perc_temporal	**33.2**	15.8	48.2	56.3	24.8	45.7	36.7	34.2	32.3	26.0	39.1	38.8	9.4	22.7	21.4	34.3
insp_causal	**34.3**	17.5	45.7	62.0	24.0	51.1	28.6	18.4	33.3	22.0	40.2	46.9	14.1	34.1	14.3	32.9
insp_cf	**30.4**	10.5	40.2	67.6	23.7	45.7	38.8	21.1	29.3	24.0	36.2	38.8	6.2	13.6	7.1	35.7
insp_temporal	**34.7**	17.5	47.0	62.0	26.7	37.0	28.6	31.6	31.3	28.0	43.7	49.0	15.6	33.3	21.4	30.0
navi_causal	**30.7**	21.1	46.3	62.0	24.4	42.4	24.5	23.7	31.3	26.0	28.7	42.9	4.7	18.2	14.3	34.3
navi_temporal	**34.7**	17.5	53.7	60.6	26.7	37.0	32.7	26.3	32.8	16.0	42.0	49.0	10.9	30.3	21.4	34.3
mani_causal	**30.1**	19.3	42.7	49.3	22.5	40.2	24.5	31.6	36.4	24.0	30.5	46.9	6.2	15.9	14.3	32.9
mani_cf	**31.7**	15.8	45.1	57.7	27.9	41.3	30.6	36.8	30.3	28.0	31.6	42.9	6.2	18.2	21.4	35.7
exo_physical	**32.3**	21.1	45.7	56.3	23.3	45.7	32.7	39.5	31.8	22.0	37.9	42.9	6.2	19.7	21.4	35.7
exo_temporal	**34.7**	21.1	51.8	53.5	26.0	39.1	24.5	34.2	31.3	26.0	43.7	49.0	12.5	31.1	21.4	34.3

Table 64: TVBench: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	tvbench_action_count	tvbench_action_localization	tvbench_action_seq	tvbench_egocentric_seq	tvbench_moving_direction	tvbench_object_count	tvbench_object_shuffle	tvbench_scene_transition	tvbench_unexpected_action
base	**36.5**	0.0	36.2	62.0	31.5	24.6	24.3	39.6	78.9	31.7
full-mix	**37.4**	0.6	38.1	62.9	32.5	34.1	21.6	37.8	80.5	28.0
perc_causal	**39.4**	1.9	35.6	59.0	41.0	36.6	23.6	37.3	80.0	39.0
perc_temporal	**35.8**	1.7	36.2	63.4	37.5	22.8	25.0	36.0	73.5	25.6
insp_causal	**36.4**	0.2	35.0	62.7	33.5	29.7	25.7	36.9	73.5	30.5
insp_cf	**38.4**	1.5	35.6	63.6	37.5	31.0	25.7	38.7	78.9	32.9
insp_temporal	**36.7**	20.7	34.4	60.4	34.0	22.4	23.6	40.0	71.9	23.2
navi_causal	**37.1**	1.3	34.4	59.3	35.5	26.3	28.4	41.8	78.9	28.0
navi_temporal	**35.0**	0.6	35.6	61.6	27.0	29.3	25.0	35.1	69.2	31.7
mani_causal	**36.3**	0.2	32.5	60.9	34.0	25.4	23.0	42.2	79.5	29.3
mani_cf	**38.7**	2.4	35.0	62.2	39.5	25.9	24.3	43.6	78.9	36.6
exo_physical	**35.7**	0.7	31.9	61.3	25.0	28.4	23.6	39.1	81.1	30.5
exo_temporal	**35.4**	7.6	35.6	62.2	30.0	25.9	26.4	34.2	71.4	25.6

Table 65: BlackSwan: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	Detective/easy	Detective/hard	Detective/medium	Reporter/easy	Reporter/hard	Reporter/medium
base	**61.8**	64.0	45.6	55.6	77.9	62.4	64.9
full-mix	**65.6**	66.9	50.4	59.5	80.5	66.7	70.6
perc_causal	**60.6**	63.8	51.2	56.6	68.7	59.8	61.3
perc_temporal	**60.5**	63.4	44.0	54.1	75.4	64.6	61.3
insp_causal	**62.7**	66.3	47.2	58.9	77.4	55.0	64.9
insp_cf	**60.8**	63.0	43.2	57.0	75.9	57.7	63.7
insp_temporal	**62.7**	65.3	47.2	55.3	79.5	64.6	66.1
navi_causal	**62.5**	67.6	46.4	53.9	77.9	61.9	65.7
navi_temporal	**62.4**	64.8	44.0	55.3	80.5	64.6	65.3
mani_causal	**64.2**	65.1	44.8	59.1	79.5	67.2	68.5
mani_cf	**59.6**	63.4	44.8	54.9	71.3	56.1	62.5
exo_physical	**59.5**	62.8	44.8	54.5	70.8	61.4	60.1
exo_temporal	**62.0**	59.4	47.2	56.4	81.0	69.3	66.1

Table 66: TempCompass: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	capmatch/action	capmatch/attribute_change	capmatch/direction	capmatch/order	capmatch/speed	mc/action	mc/attribute_change	mc/direction	mc/order	mc/speed	yesno/action	yesno/attribute_change	yesno/direction	yesno/order	yesno/speed
base	**67.7**	96.6	74.0	60.9	82.7	62.2	95.9	68.1	37.9	62.3	47.9	87.7	65.6	56.1	61.3	57.9
full-mix	**65.9**	96.6	79.9	62.7	81.0	61.9	95.3	74.7	44.5	59.9	45.1	78.0	63.4	53.2	55.0	51.2
perc_causal	**70.0**	97.0	78.1	56.9	86.3	68.7	95.6	73.6	43.9	65.6	47.9	89.2	72.1	55.3	63.9	58.8
perc_temporal	**68.2**	96.6	75.3	60.6	82.7	61.5	95.6	68.8	41.5	59.9	48.3	85.8	68.1	55.3	62.6	60.9
insp_causal	**68.7**	96.6	80.6	63.9	84.7	59.1	93.8	77.8	44.2	68.2	47.6	87.3	69.0	54.0	61.8	51.0
insp_cf	**68.7**	97.3	76.7	59.0	84.3	67.0	96.2	69.8	36.1	66.2	49.5	87.5	69.0	52.4	62.3	60.1
insp_temporal	**67.8**	97.3	76.7	61.5	79.0	61.5	95.3	66.3	40.9	59.9	51.4	87.7	66.5	55.5	59.4	58.3
navi_causal	**69.5**	97.6	81.6	60.9	85.0	64.6	94.7	78.1	42.4	66.2	48.9	82.1	76.3	53.6	58.9	60.1
navi_temporal	**67.8**	97.3	75.3	61.2	77.7	62.9	95.0	67.4	42.4	60.3	51.1	86.5	67.9	55.7	58.1	59.6
mani_causal	**70.2**	98.0	82.6	63.0	86.0	68.0	95.3	77.1	43.0	65.6	51.7	86.0	71.2	54.0	63.1	57.2
mani_cf	**68.8**	97.3	77.4	57.8	83.7	67.4	97.0	69.1	41.8	61.9	47.3	89.4	68.8	51.4	61.5	61.8
exo_physical	**69.0**	97.3	81.6	59.3	86.3	63.2	97.9	76.7	40.0	67.9	50.2	87.0	69.9	54.3	55.2	56.2
exo_temporal	**67.4**	97.6	73.3	62.7	75.0	57.7	95.6	67.7	41.2	60.3	50.8	87.7	68.1	54.3	61.5	56.6

Table 67: PerceptionTest (val): accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	abstraction	memory	physics	semantics
base	**62.9**	54.6	47.6	58.4	78.0
full-mix	**58.1**	54.6	46.8	50.7	71.8
perc_causal	**63.7**	55.1	46.4	58.9	79.7
perc_temporal	**61.8**	53.7	46.2	57.8	76.2
insp_causal	**60.6**	52.5	48.8	53.7	78.0
insp_cf	**63.2**	54.5	44.9	58.8	79.0
insp_temporal	**60.4**	53.4	44.4	55.8	74.7
navi_causal	**62.8**	54.1	46.0	58.0	79.1
navi_temporal	**60.4**	53.6	44.6	56.4	73.7
mani_causal	**62.6**	53.7	45.3	58.6	78.1
mani_cf	**62.4**	53.6	45.8	56.6	79.8
exo_physical	**63.7**	55.0	46.4	59.5	79.2
exo_temporal	**60.5**	53.7	46.0	55.6	74.5

Table 68: EgoTaskQA: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	egotaskqa
base	**81.5**	81.5
full-mix	**74.1**	74.1
perc_causal	**77.8**	77.8
perc_temporal	**81.3**	81.3
insp_causal	**79.8**	79.8
insp_cf	**81.6**	81.6
insp_temporal	**80.2**	80.2
navi_causal	**78.3**	78.3
navi_temporal	**80.7**	80.7
mani_causal	**74.2**	74.2
mani_cf	**81.6**	81.6
exo_physical	**82.6**	82.6
exo_temporal	**80.9**	80.9

Table 69: 3DSRBench (circular): accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	height_higher	location_above	location_closer_to_camera	location_next_to	multi_object_closer_to	multi_object_facing	multi_object_parallel	multi_object_same_direction	multi_object_viewpoint_towards_object	orientation_in_front_of	orientation_on_the_left	orientation_viewpoint
base	**51.2**	53.6	61.8	71.8	51.3	49.9	59.4	52.1	53.1	28.1	56.5	49.9	33.8
full-mix	**50.6**	50.7	60.8	71.5	56.6	47.3	59.7	51.2	49.9	29.7	50.9	50.1	34.8
perc_causal	**51.1**	53.6	59.4	73.5	53.7	49.6	56.5	52.2	53.6	27.4	54.7	49.9	35.8
perc_temporal	**50.7**	54.7	59.2	72.9	48.4	49.1	59.1	52.2	51.7	26.7	58.6	50.1	33.0
insp_causal	**51.6**	52.2	64.0	74.8	54.0	48.1	63.6	50.6	55.1	26.2	50.0	50.0	35.9
insp_cf	**51.1**	53.3	61.6	70.0	51.3	51.3	60.0	53.1	52.3	28.9	57.4	50.0	32.8
insp_temporal	**51.2**	52.6	61.1	71.3	55.5	53.6	60.0	52.5	52.6	27.5	53.8	49.9	34.1
navi_causal	**52.6**	49.4	67.3	71.6	62.5	52.7	63.4	53.7	53.5	27.0	53.3	50.0	37.5
navi_temporal	**50.5**	50.9	59.9	74.0	49.7	51.6	60.5	51.5	51.7	26.5	56.4	50.3	31.9
mani_causal	**52.3**	51.7	64.8	70.6	57.2	53.0	70.1	48.8	50.9	27.1	56.4	50.1	37.1
mani_cf	**52.6**	54.6	64.2	70.9	57.5	50.7	67.6	52.7	51.3	28.1	58.6	50.0	34.8
exo_physical	**50.7**	55.0	62.9	70.4	51.3	47.7	58.1	52.2	48.3	27.3	61.0	50.1	31.0
exo_temporal	**50.4**	50.7	59.1	72.6	52.4	49.1	58.5	50.6	52.0	27.6	56.5	49.9	34.3

Table 70: CoreCognition: accuracy (%) by official category. Rows: training configurations; columns: categories.

Config	Overall	boundary/Stage Sensorimotor	conservation/Stage Concrete Operational	continuity/Stage Sensorimotor	hierarchy/Stage Concrete Operational	intentionality/Stage Formal Operational	intuitivephysics/Stage Concrete Operational	mechanicalreasoning/Stage Formal Operational	perceptualconstancy/Stage Sensorimotor	permanence/Stage Sensorimotor	perspective/Stage Concrete Operational	spatiality/Stage Sensorimotor	toolusing/Stage Formal Operational
base	**59.8**	76.4	69.5	67.2	67.0	76.1	58.6	65.9	61.5	43.8	33.5	36.6	74.8
full-mix	**53.2**	74.2	63.8	60.9	57.3	60.6	50.0	61.4	50.3	25.0	24.3	37.4	77.3
perc_causal	**57.4**	76.4	58.1	64.1	60.2	76.1	53.4	65.9	64.3	37.5	33.5	35.0	69.7
perc_temporal	**57.8**	76.4	59.0	67.2	62.1	75.2	56.9	65.9	60.1	37.5	33.0	36.6	73.1
insp_causal	**52.6**	79.8	21.9	67.2	66.0	71.6	56.9	67.0	48.3	43.8	30.8	39.0	63.9
insp_cf	**57.7**	78.7	65.7	64.1	67.0	72.5	53.4	63.6	65.7	50.0	30.8	33.3	66.4
insp_temporal	**57.5**	76.4	65.7	60.9	59.2	77.1	51.7	60.2	53.1	31.2	42.2	39.8	66.4
navi_causal	**56.0**	78.7	49.5	62.5	61.2	74.3	63.8	60.2	55.2	43.8	33.0	32.5	75.6
navi_temporal	**58.3**	73.0	65.7	62.5	63.1	76.1	56.9	62.5	51.0	43.8	41.1	39.0	73.1
mani_causal	**57.9**	73.0	61.0	64.1	57.3	74.3	58.6	65.9	62.2	43.8	32.4	39.8	74.8
mani_cf	**56.5**	71.9	46.7	64.1	65.0	70.6	58.6	69.3	64.3	37.5	29.7	35.8	74.8
exo_physical	**58.8**	78.7	68.6	67.2	66.0	70.6	58.6	67.0	62.9	37.5	33.0	31.7	73.9
exo_temporal	**57.4**	71.9	65.7	62.5	65.0	76.1	53.4	64.8	51.7	43.8	39.5	33.3	70.6

Table 71: **In-domain substitution sweep, full per-seed results** (in-domain accuracy, %). p is the WOVEN share of the fixed 2,000-item training budget; the p=0 row is the pure in-domain baseline.

	SAT (circular)				ActionEQA				CLEVRER			
p (%)	s42	s43	s44	mean	s42	s43	s44	mean	s42	s43	s44	mean
0	45.33	43.33	58.00	**48.89**	79.30	80.01	78.69	**79.33**	83.20	81.73	83.00	**82.64**
10	45.33	50.67	49.33	**48.44**	77.60	77.55	78.82	**77.99**	82.47	81.53	82.60	**82.20**
20	54.67	42.67	35.33	**44.22**	77.50	77.61	75.46	**76.86**	82.53	84.47	82.93	**83.31**
30	37.33	54.00	38.00	**43.11**	79.51	79.98	78.42	**79.30**	83.60	78.67	82.33	**81.53**
40	58.67	40.00	53.33	**50.67**	74.30	74.19	73.88	**74.12**	76.27	81.47	83.93	**80.56**
50	44.67	46.67	51.33	**47.56**	67.66	70.84	72.55	**70.35**	85.00	84.80	81.93	**83.91**
70	54.67	54.00	46.67	**51.78**	64.15	66.68	67.77	**66.20**	75.73	81.80	80.47	**79.33**
100	58.00	60.00	56.67	**58.22**	46.54	46.51	46.38	**46.47**	56.80	54.40	54.40	**55.20**

Table 72: **Deletion ablation for in-domain substitution.** Accuracy (%) after deleting the WOVEN portion (Del.) or retaining it (Sub.). Δ is substitution minus deletion in percentage points. Both arms retain the same task-specific subset.

	SAT (circular)			ActionEQA			CLEVRER		
p (%)	Del.	Sub.	Δ	Del.	Sub.	Δ	Del.	Sub.	Δ
20	46.21	44.22	-1.99	74.21	76.86	$+2.65$	80.31	83.31	$+3.00$
40	39.67	50.67	$+11.00$	68.41	74.12	$+5.71$	76.96	80.56	$+3.60$
50	39.72	47.56	$+7.84$	66.20	70.35	$+4.15$	77.12	83.91	$+6.79$

Table 73: **Transfer at scale.** Zero-shot transfer gains (pp over the same-scale base) after WOVEN full-mix post-training of Qwen2.5-VL-7B and Qwen2.5-VL-32B. The recipe follows the 3B protocol; the 32B run is SFT-only. '—': not reported.

	SAT (circular)	DSI	MVP	WM-ABench	TemporalBench-L	ViewSpatial	Robo2VLM	InPhyRe	ERQA	BlackSwan	CLEVRER	TVBench
7B SFT	+20.00	+6.32	+5.08	+4.39	+4.30	—	+3.24	—	—	—	—	—
7B RL	+20.00	+5.96	+4.49	+4.83	+5.34	+2.10	+3.54	—	+2.75	—	+2.19	+2.04
32B SFT	+22.67	—	+6.28	+5.91	—	+4.94	—	+3.54	—	+2.29	—	—

Table 74: **Same-backbone comparison with video-SSL mid-training.** Both models start from Qwen3.5-4B; gains in parentheses are over the shared base. Orca-4B: latent next-state mid-training (125K hours of video, 160M event annotations). Ours: WOVEN post-training (SFT, ≈107K QA pairs, ≈7 hours of generated video).

Model	3DSRBench
Qwen3.5-4B base	48.1
Orca-4B	52.1 (+4.0)
+WOVEN (ours)	**54.8** (+6.7)

G.17 E-F15. WOVEN RESULTS BY SUPERVISION SOURCE AND TRAINING STAGE

We report the complete WOVEN test results for the base Qwen2.5-VL-3B model, full-mix training, and all 11 individual supervision subsets. Accuracies and error breakdowns were recomputed from model generations. The three test sets are in-distribution (ID; $n = 6{,}228$), held-out scenes (OOD; $n = 3{,}496$), and state perturbations (Pert; $n = 1{,}116$); chance accuracy is 25%. Tables 75–91 have separate panels for supervised fine-tuning (SFT) and reinforcement learning (RL). Full-mix RL results are not available in this set of evaluations and are marked with dashes; the base model is repeated in both panels for comparison.

Configuration prefixes denote camera motion (`perc`), object inspection (`insp`), navigation (`navi`), object manipulation (`mani`), and passive physical events (`exo`). Suffixes identify causal, counterfactual (`cf`), temporal, or physical supervision, following the training configurations in Appendix G.13. Unless a caption specifies otherwise, accuracies are percentages over all items in the named category, and overall accuracy is computed over the complete named test split. Error shares describe the distractors selected on incorrect answers; captions specify the relevant subset and any categories omitted from the display.

Overall and cross-action gains. All 11 individual subsets improve overall ID accuracy over the 26.38% base score: SFT accuracies range from 31.87% to 50.75%, and RL accuracies from 32.45% to 51.65% (Table 75). To distinguish these gains from improvement restricted to the trained action category, we also compare each source on the other four action categories. At both stages, 43 of these 44 comparisons improve over the base model: ten sources improve all four other action categories, and the remaining source improves three. Table 76 further identifies which reasoning skills improve under each source.

G.17.1 ACCURACY BY ACTION, REASONING TYPE, AND TEST SPLIT

Table 75: ID accuracy (%) by action type ($n = 6{,}228$). Overall covers the complete ID split. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT						
Config.	Perceptive	Inspective	Navigative	Manipulative	Exogenous	Overall
base	27.00	23.46	30.03	28.82	23.78	26.38
full-mix	90.71	90.49	90.02	95.31	79.43	89.29
perc_causal	60.76	34.75	56.94	64.67	29.77	48.28
perc_temporal	38.98	27.28	36.81	29.25	31.16	32.29
insp_causal	37.07	62.28	56.34	58.85	30.30	49.97
insp_cf	33.42	58.89	53.21	71.96	32.99	50.75
insp_temporal	33.94	27.96	37.15	30.47	32.29	32.03
navi_causal	34.11	34.14	60.33	78.04	31.25	46.56
navi_temporal	35.42	27.78	37.59	29.08	31.16	31.87
mani_causal	31.42	25.86	55.56	90.36	31.77	45.41
mani_cf	30.99	26.54	56.68	92.53	34.55	46.63
exo_physical	26.74	28.83	35.16	33.51	51.30	34.63
exo_temporal	35.16	28.15	37.59	29.43	32.29	32.19
RL						
Config.	Perceptive	Inspective	Navigative	Manipulative	Exogenous	Overall
base	27.00	23.46	30.03	28.82	23.78	26.38
full-mix	—	—	—	—	—	—
perc_causal	59.98	36.91	56.94	68.58	28.91	49.26
perc_temporal	37.93	28.40	38.02	30.12	32.20	32.96
insp_causal	35.94	63.33	57.20	66.49	30.56	51.65
insp_cf	34.29	61.98	54.86	71.53	30.38	51.46
insp_temporal	33.42	26.85	41.23	30.03	32.99	32.45
navi_causal	37.07	36.17	60.07	79.77	31.42	47.94
navi_temporal	39.24	27.47	38.63	31.68	31.86	33.30
mani_causal	30.90	29.20	56.25	90.19	32.03	46.32
mani_cf	32.20	28.95	57.20	93.06	34.38	47.64
exo_physical	25.78	29.20	36.72	38.45	51.22	35.74
exo_temporal	34.11	28.95	40.10	31.16	32.20	32.98

Table 76: ID accuracy (%) by reasoning type ($n = 6{,}228$). Overall covers all eight reasoning types. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT

Config.	Forward dynamics	Inverse dynamics	CF substitution	CF removal	Outcome prediction	Cued prediction	Temporal ordering	Temporal adjacency	Overall
base	21.34	40.39	19.35	18.64	21.18	23.96	27.34	25.66	26.38
full-mix	97.09	94.80	94.80	96.95	80.90	83.33	92.77	69.66	89.29
perc_causal	75.31	76.63	62.19	28.85	33.33	33.68	25.13	26.28	48.28
perc_temporal	20.81	41.71	18.82	20.61	23.26	25.35	30.34	52.73	32.29
insp_causal	73.72	75.84	74.19	41.22	32.64	34.03	25.40	25.75	49.97
insp_cf	73.72	50.53	83.15	86.92	39.93	40.62	26.63	23.72	50.75
insp_temporal	20.72	41.36	19.00	21.86	22.57	25.00	24.96	56.70	32.03
navi_causal	64.90	64.90	59.68	49.64	29.86	39.93	26.46	27.95	46.56
navi_temporal	21.25	42.50	17.20	20.43	22.57	20.83	25.93	55.82	31.87
mani_causal	63.32	62.70	60.93	48.39	29.86	38.89	25.84	26.28	45.41
mani_cf	63.23	54.76	68.10	57.35	38.54	42.71	27.25	28.48	46.63
exo_physical	25.93	47.62	15.23	38.17	76.74	77.43	28.13	23.10	34.63
exo_temporal	20.55	42.77	18.10	21.51	22.92	24.65	25.31	56.61	32.19

RL

Config.	Forward dynamics	Inverse dynamics	CF substitution	CF removal	Outcome prediction	Cued prediction	Temporal ordering	Temporal adjacency	Overall
base	21.34	40.39	19.35	18.64	21.18	23.96	27.34	25.66	26.38
full-mix	—	—	—	—	—	—	—	—	—
perc_causal	76.01	79.28	67.56	31.00	29.86	34.72	24.51	25.84	49.26
perc_temporal	19.93	43.92	17.74	20.97	21.88	25.35	30.60	55.56	32.96
insp_causal	75.22	76.10	83.51	44.80	33.33	32.64	25.93	26.54	51.65
insp_cf	72.13	54.94	84.23	87.28	33.33	36.11	26.54	26.98	51.46
insp_temporal	20.55	43.12	14.87	23.12	23.26	24.65	24.96	58.73	32.45
navi_causal	66.31	70.55	60.39	47.49	29.51	39.24	25.40	30.51	47.94
navi_temporal	20.99	48.85	17.38	22.94	23.61	22.57	23.81	57.67	33.30
mani_causal	64.90	62.87	64.70	48.39	32.29	39.93	24.69	27.95	46.32
mani_cf	66.14	55.03	70.25	60.57	38.89	43.40	26.63	28.57	47.64
exo_physical	27.60	50.26	15.95	42.65	73.96	77.43	26.98	24.16	35.74
exo_temporal	21.34	44.36	17.56	23.66	25.35	23.26	26.54	56.26	32.98

Table 77: Held-out-scene (OOD) accuracy (%) by action type ($n = 3{,}496$). Overall covers the complete OOD split. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT						
Config.	Perceptive	Inspective	Navigative	Manipulative	Exogenous	Overall
base	27.19	24.04	33.75	25.16	30.47	27.77
full-mix	84.69	87.18	90.78	95.47	83.12	88.16
perc_causal	59.69	37.93	59.84	58.12	25.94	47.43
perc_temporal	33.28	23.50	33.28	31.41	28.91	29.52
insp_causal	31.88	59.62	59.22	53.44	32.81	48.43
insp_cf	36.25	62.61	59.69	63.75	35.31	52.46
insp_temporal	29.38	25.11	32.50	25.47	30.94	28.38
navi_causal	35.16	34.08	66.88	70.47	34.38	47.00
navi_temporal	30.16	25.64	35.94	28.12	35.00	30.52
mani_causal	33.75	29.70	66.72	87.50	44.06	50.43
mani_cf	31.72	29.81	61.56	87.03	45.62	49.34
exo_physical	26.09	29.49	38.12	32.03	62.19	36.90
exo_temporal	32.66	29.49	34.38	25.94	36.56	31.61
RL						
Config.	Perceptive	Inspective	Navigative	Manipulative	Exogenous	Overall
base	27.19	24.04	33.75	25.16	30.47	27.77
full-mix	—	—	—	—	—	—
perc_causal	58.59	39.74	59.22	60.47	24.69	47.80
perc_temporal	33.44	25.75	37.66	28.75	30.31	30.72
insp_causal	32.19	63.68	58.91	54.69	26.41	48.57
insp_cf	35.31	62.71	58.75	65.94	34.84	52.46
insp_temporal	25.16	22.44	30.00	24.38	28.59	25.80
navi_causal	37.66	36.86	66.72	72.81	33.75	48.48
navi_temporal	34.69	30.24	40.78	27.19	35.78	33.44
mani_causal	36.72	34.19	67.81	87.66	47.03	52.95
mani_cf	36.09	33.87	62.19	88.75	43.91	51.34
exo_physical	29.53	30.56	41.56	35.62	61.09	38.90
exo_temporal	33.44	32.37	37.03	25.94	37.97	33.27

Table 78: Held-out-scene (OOD) accuracy (%) by reasoning type ($n = 3{,}496$). Overall covers all eight reasoning types. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT									
Config.	Forward dynamics	Inverse dynamics	CF substitution	CF removal	Outcome prediction	Cued prediction	Temporal ordering	Temporal adjacency	Overall
base	19.18	35.69	14.56	24.37	24.38	31.88	37.26	27.04	27.77
full-mix	97.33	95.44	94.94	97.78	92.50	92.50	85.22	64.31	88.16
perc_causal	75.47	74.06	63.61	30.70	30.00	25.62	28.93	21.38	47.43
perc_temporal	20.44	37.11	23.42	24.05	27.50	30.62	25.31	41.19	29.52
insp_causal	70.60	76.10	74.05	38.29	41.25	42.50	22.01	20.60	48.43
insp_cf	72.96	47.48	79.75	85.13	45.62	42.50	33.18	30.66	52.46
insp_temporal	21.38	36.32	19.62	23.10	29.38	31.25	17.61	44.18	28.38
navi_causal	60.69	64.15	55.38	47.15	35.00	36.88	32.86	31.60	47.00
navi_temporal	21.70	37.89	21.52	22.15	31.25	35.62	26.10	43.55	30.52
mani_causal	61.95	62.89	61.08	45.25	53.75	53.75	40.09	32.39	50.43
mani_cf	63.52	50.47	65.82	56.96	53.75	57.50	31.13	37.11	49.34
exo_physical	24.21	44.50	15.82	42.09	94.38	93.75	30.82	27.20	36.90
exo_temporal	21.07	36.48	19.62	25.32	28.75	36.25	24.53	52.99	31.61
RL									
Config.	Forward dynamics	Inverse dynamics	CF substitution	CF removal	Outcome prediction	Cued prediction	Temporal ordering	Temporal adjacency	Overall
base	19.18	35.69	14.56	24.37	24.38	31.88	37.26	27.04	27.77
full-mix	—	—	—	—	—	—	—	—	—
perc_causal	77.83	78.30	67.09	31.01	30.00	22.50	25.16	19.50	47.80
perc_temporal	22.17	38.84	22.15	23.10	30.00	30.00	25.00	45.28	30.72
insp_causal	71.23	78.93	74.68	42.41	30.00	30.62	23.58	19.81	48.57
insp_cf	72.48	52.20	80.70	87.34	44.38	40.62	31.76	27.04	52.46
insp_temporal	20.28	38.36	16.77	20.57	29.38	31.88	12.74	36.48	25.80
navi_causal	63.05	66.98	55.70	47.78	28.75	31.25	33.33	36.64	48.48
navi_temporal	21.23	49.37	21.20	21.52	31.25	30.62	30.82	45.60	33.44
mani_causal	62.74	64.31	64.24	45.89	55.62	57.50	41.19	39.62	52.95
mani_cf	66.35	52.67	68.35	60.13	48.12	56.88	33.18	39.78	51.34
exo_physical	25.79	48.74	17.09	45.25	91.88	92.50	33.02	28.93	38.90
exo_temporal	21.70	38.36	19.62	23.73	28.75	33.75	28.93	56.60	33.27

Table 79: State-perturbation (Pert) accuracy (%) by perturbation group (n = 1,116). Geometric changes affect size, orientation, or layout; appearance changes affect color, lighting, or background. Overall covers the complete Pert split. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT			
Config.	Geometric	Appearance	Overall
base	22.58	19.71	21.15
full-mix	34.95	41.58	38.26
perc_causal	23.12	19.53	21.33
perc_temporal	24.19	20.43	22.31
insp_causal	25.09	24.37	24.73
insp_cf	26.34	24.37	25.36
insp_temporal	23.66	23.66	23.66
navi_causal	37.63	30.29	33.96
navi_temporal	25.27	24.19	24.73
mani_causal	34.95	50.00	42.47
mani_cf	35.13	44.62	39.87
exo_physical	37.10	62.37	49.73
exo_temporal	24.91	26.16	25.54
RL			
Config.	Geometric	Appearance	Overall
base	22.58	19.71	21.15
full-mix	—	—	—
perc_causal	23.48	22.22	22.85
perc_temporal	28.14	22.40	25.27
insp_causal	23.66	24.55	24.10
insp_cf	25.81	22.94	24.37
insp_temporal	25.09	27.60	26.34
navi_causal	38.53	31.00	34.77
navi_temporal	29.39	26.52	27.96
mani_causal	35.66	53.05	44.35
mani_cf	36.56	46.06	41.31
exo_physical	37.99	64.16	51.08
exo_temporal	28.14	27.96	28.05

Table 80: ID accuracy (%) by camera or viewpoint action. Direction-specific columns include only perceptive and inspective items. Overall ID is accuracy on the complete ID split ($n = 6{,}228$), not the camera/viewpoint subset. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT								
Config.	Rotate left	Rotate right	Move forward	Move backward	Orbit left	Orbit right	Arc up	Overall ID
base	19.79	19.79	39.58	28.82	22.04	19.63	28.70	26.38
full-mix	90.97	87.50	93.40	90.97	89.44	89.44	92.59	89.29
perc_causal	58.33	60.07	60.76	63.89	29.81	39.44	35.00	48.28
perc_temporal	34.38	32.64	42.71	46.18	28.89	21.48	31.48	32.29
insp_causal	34.38	30.21	39.58	44.10	62.41	64.81	59.63	49.97
insp_cf	26.74	27.08	42.01	37.85	53.70	56.48	66.48	50.75
insp_temporal	29.17	27.43	41.67	37.50	28.15	23.89	31.85	32.03
navi_causal	22.57	17.71	53.47	42.71	29.63	34.07	38.70	46.56
navi_temporal	34.38	28.82	40.62	37.85	28.70	22.41	32.22	31.87
mani_causal	16.67	16.32	37.15	55.56	24.81	21.48	31.30	45.41
mani_cf	14.58	14.93	48.26	46.18	25.19	26.67	27.78	46.63
exo_physical	11.46	14.24	38.89	42.36	28.15	30.37	27.96	34.63
exo_temporal	30.90	32.99	39.58	37.15	28.52	22.96	32.96	32.19
RL								
Config.	Rotate left	Rotate right	Move forward	Move backward	Orbit left	Orbit right	Arc up	Overall ID
base	19.79	19.79	39.58	28.82	22.04	19.63	28.70	26.38
full-mix	—	—	—	—	—	—	—	—
perc_causal	58.33	59.38	61.11	61.11	34.63	37.04	39.07	49.26
perc_temporal	35.07	32.64	39.93	44.10	27.04	24.44	33.70	32.96
insp_causal	33.33	32.64	41.67	36.11	65.00	66.30	58.70	51.65
insp_cf	28.12	26.39	47.92	34.72	55.56	60.00	70.37	51.46
insp_temporal	29.86	26.04	42.36	35.42	27.78	20.93	31.85	32.45
navi_causal	23.96	20.83	55.21	48.26	32.59	35.19	40.74	47.94
navi_temporal	40.28	34.72	43.06	38.89	27.04	20.93	34.44	33.30
mani_causal	17.71	17.71	33.33	54.86	27.41	25.37	34.81	46.32
mani_cf	18.06	17.36	50.35	43.06	26.11	29.81	30.93	47.64
exo_physical	8.68	14.93	39.24	40.28	29.63	29.63	28.33	35.74
exo_temporal	27.43	29.17	43.06	36.81	28.52	22.59	35.74	32.98

Table 81: ID accuracy (%) by answer-option modality (n = 6,228). Image and Text refer to the options presented to the model; Overall covers the complete ID split. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT			
Config.	Image	Text	Overall
base	20.67	31.13	26.38
full-mix	93.56	85.74	89.29
perc_causal	55.02	42.68	48.28
perc_temporal	21.09	41.59	32.29
insp_causal	59.16	42.33	49.97
insp_cf	71.37	33.63	50.75
insp_temporal	21.23	41.01	32.03
navi_causal	54.74	39.77	46.56
navi_temporal	20.38	41.42	31.87
mani_causal	54.00	38.27	45.41
mani_cf	58.42	36.83	46.63
exo_physical	36.66	32.95	34.63
exo_temporal	20.91	41.56	32.19
RL			
Config.	Image	Text	Overall
base	20.67	31.13	26.38
full-mix	—	—	—
perc_causal	56.55	43.21	49.26
perc_temporal	20.45	43.36	32.96
insp_causal	62.24	42.86	51.65
insp_cf	69.89	36.16	51.46
insp_temporal	20.63	42.27	32.45
navi_causal	54.92	42.15	47.94
navi_temporal	21.09	43.45	33.30
mani_causal	55.73	38.51	46.32
mani_cf	60.76	36.74	47.64
exo_physical	38.08	33.80	35.74
exo_temporal	21.66	42.39	32.98

G.17.2 ERRORS BY QUESTION TYPE

Table 82: Temporal-adjacency errors on ID ($n = 1{,}134$). Accuracy is in percent; Errors is the number of incorrect answers. Error shares (%) distinguish gross localization errors, off-by-one errors, and reversed before/after direction. Shares sum to approximately 100 within each run, allowing for rounding. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT

Config.	Accuracy (%)	Errors	Gross	Off-by-one	Direction
base	25.66	843	17.2	34.4	48.4
full-mix	69.66	344	1.7	38.4	59.9
perc_causal	26.28	836	16.0	35.9	48.1
perc_temporal	52.73	536	1.5	25.6	72.9
insp_causal	25.75	842	17.6	31.4	51.1
insp_cf	23.72	865	16.5	35.1	48.3
insp_temporal	56.70	491	1.8	26.7	71.5
navi_causal	27.95	817	14.7	33.3	52.0
navi_temporal	55.82	501	3.0	27.5	69.5
mani_causal	26.28	836	15.3	34.0	50.7
mani_cf	28.48	811	14.7	35.0	50.3
exo_physical	23.10	872	15.9	34.6	49.4
exo_temporal	56.61	492	2.4	29.3	68.3

RL

Config.	Accuracy (%)	Errors	Gross	Off-by-one	Direction
base	25.66	843	17.2	34.4	48.4
full-mix	—	—	—	—	—
perc_causal	25.84	841	18.7	34.4	47.0
perc_temporal	55.56	504	3.6	28.0	68.5
insp_causal	26.54	833	18.0	28.2	53.8
insp_cf	26.98	828	15.6	33.0	51.4
insp_temporal	58.73	468	2.6	26.3	71.2
navi_causal	30.51	788	12.4	33.1	54.4
navi_temporal	57.67	480	3.3	26.0	70.6
mani_causal	27.95	817	15.1	34.5	50.4
mani_cf	28.57	810	15.1	33.5	51.5
exo_physical	24.16	860	17.4	34.0	48.6
exo_temporal	56.26	496	2.2	31.5	66.3

Table 83: Temporal-ordering errors on ID (n = 1,134). Accuracy is in percent; Errors counts incorrect answers. Random 0–2 are the three independently sampled permutation distractors. Their shares (%) describe the distribution of errors and sum to approximately 100, allowing for rounding. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT

Config.	Accuracy (%)	Errors	Random 0	Random 1	Random 2
base	27.34	824	32.8	37.1	30.1
full-mix	92.77	82	36.6	35.4	28.0
perc_causal	25.13	849	33.8	34.5	31.7
perc_temporal	30.34	790	33.4	31.8	34.8
insp_causal	25.40	846	32.6	35.5	31.9
insp_cf	26.63	832	34.3	35.0	30.8
insp_temporal	24.96	851	33.0	34.0	33.0
navi_causal	26.46	834	32.0	35.4	32.6
navi_temporal	25.93	840	32.6	34.9	32.5
mani_causal	25.84	841	31.6	33.8	34.6
mani_cf	27.25	825	33.6	35.5	30.9
exo_physical	28.13	815	32.8	36.8	30.4
exo_temporal	25.31	847	35.3	33.5	31.2

RL

Config.	Accuracy (%)	Errors	Random 0	Random 1	Random 2
base	27.34	824	32.8	37.1	30.1
full-mix	—	—	—	—	—
perc_causal	24.51	856	31.9	36.0	32.1
perc_temporal	30.60	787	32.0	34.8	33.2
insp_causal	25.93	840	35.1	32.5	32.4
insp_cf	26.54	833	32.7	34.6	32.8
insp_temporal	24.96	851	35.0	33.4	31.6
navi_causal	25.40	846	32.9	35.1	32.0
navi_temporal	23.81	864	34.3	31.6	34.1
mani_causal	24.69	854	31.7	35.1	33.1
mani_cf	26.63	832	33.5	32.9	33.5
exo_physical	26.98	828	36.0	34.3	29.7
exo_temporal	26.54	833	32.7	34.8	32.5

Table 84: Physical-modeling errors on ID: exogenous outcome and cued prediction ($n = 576$). Accuracy is in percent; Errors counts all incorrect answers. The ten displayed violation types are the most frequent in the base model's errors. Each share (%) uses all errors in that run as its denominator, including unlisted types; displayed shares are not renormalized and need not sum to 100. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT

Config.	Accuracy (%)	Errors	Pass through	Transform	No effect	Split	Lateral	Float	Instant stop	Embed surface	Escape	Separate midair
base	22.57	446	8.1	5.6	5.6	5.2	5.2	5.2	5.2	4.7	4.7	4.5
full-mix	82.12	103	12.6	0.0	4.9	1.9	2.9	0.0	0.0	4.9	5.8	3.9
perc_causal	33.51	383	9.4	2.1	2.3	10.7	2.1	0.5	4.2	1.0	6.0	6.8
perc_temporal	24.31	436	7.3	5.3	4.8	4.6	5.3	6.2	6.0	5.0	3.0	3.9
insp_causal	33.33	384	14.1	1.0	1.0	1.3	0.5	1.6	0.0	1.3	7.3	4.2
insp_cf	40.28	344	12.8	0.9	2.9	2.6	2.3	1.2	0.0	0.6	4.9	4.1
insp_temporal	23.78	439	7.7	5.5	4.6	3.9	5.5	5.7	6.2	4.6	3.2	4.6
navi_causal	34.90	375	7.7	1.6	4.0	0.8	2.9	2.1	4.0	2.7	2.7	4.8
navi_temporal	21.70	451	7.8	4.9	3.8	4.0	5.3	5.8	6.7	4.4	3.8	4.9
mani_causal	34.38	378	8.2	0.0	4.2	0.0	4.0	2.9	5.6	1.3	4.2	2.9
mani_cf	40.62	342	7.6	0.0	6.7	2.0	5.3	1.8	1.5	1.5	5.8	2.9
exo_physical	77.08	132	4.5	0.0	16.7	0.0	0.8	3.8	0.0	10.6	0.0	3.0
exo_temporal	23.78	439	8.9	5.5	4.6	3.2	4.8	6.2	5.7	4.6	3.6	3.6

RL

Config.	Accuracy (%)	Errors	Pass through	Transform	No effect	Split	Lateral	Float	Instant stop	Embed surface	Escape	Separate midair
base	22.57	446	8.1	5.6	5.6	5.2	5.2	5.2	5.2	4.7	4.7	4.5
full-mix	—	—	—	—	—	—	—	—	—	—	—	—
perc_causal	32.29	390	5.1	0.8	1.5	8.2	1.5	1.3	2.8	1.3	6.4	8.2
perc_temporal	23.61	440	8.2	5.2	5.2	5.7	4.1	6.4	4.3	4.5	3.2	4.8
insp_causal	32.99	386	12.4	1.0	1.6	1.8	0.0	0.0	0.0	2.1	7.0	2.3
insp_cf	34.72	376	13.8	0.5	4.3	1.6	4.3	3.5	0.0	3.2	4.3	4.3
insp_temporal	23.96	438	7.1	5.0	4.3	4.3	5.5	5.0	7.1	4.3	2.5	4.1
navi_causal	34.38	378	6.3	0.8	5.3	0.5	2.9	2.1	3.4	2.4	2.4	4.8
navi_temporal	23.09	443	7.2	5.2	5.0	6.8	3.8	6.3	5.0	5.0	3.8	4.1
mani_causal	36.11	368	9.5	0.0	3.5	0.3	4.1	1.9	5.2	0.8	4.6	1.9
mani_cf	41.15	339	8.0	0.0	6.5	3.2	2.4	1.5	0.9	0.9	5.0	1.8
exo_physical	75.69	140	3.6	0.0	19.3	0.0	1.4	5.0	0.0	8.6	0.0	6.4
exo_temporal	24.31	436	8.7	3.4	5.0	5.0	4.4	4.6	3.9	4.8	4.6	5.5

Table 85: Incorrect option choices on the state-perturbation probe (Pert, $n = 1{,}116$). Accuracy is in percent; Errors counts incorrect answers. Shares (%) identify the perturbation dimension of the selected incorrect option; None denotes an unperturbed option. Shares sum to approximately 100, allowing for rounding. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT

Config.	Accuracy (%)	Errors	None	Size	Orientation	Layout	Color	Lighting	Background
base	21.15	880	0.0	16.2	17.6	15.2	17.0	16.2	17.6
full-mix	38.26	689	0.0	20.8	22.6	9.3	14.8	20.2	12.3
perc_causal	21.33	878	0.0	12.2	19.2	17.4	20.0	16.1	15.0
perc_temporal	22.31	867	0.0	15.5	18.2	15.1	16.3	15.6	19.4
insp_causal	24.73	840	0.0	16.8	19.0	13.9	13.8	17.6	18.8
insp_cf	25.36	833	0.0	19.3	16.8	13.2	20.0	17.3	13.3
insp_temporal	23.66	852	0.0	18.0	18.1	14.0	17.4	15.5	17.1
navi_causal	33.96	737	0.0	11.1	27.8	8.3	17.9	22.8	12.1
navi_temporal	24.73	840	0.0	16.7	19.0	13.9	16.2	17.1	17.0
mani_causal	42.47	642	0.0	22.9	25.4	8.3	17.4	15.9	10.1
mani_cf	39.87	671	0.0	21.5	23.8	8.6	17.1	16.4	12.5
exo_physical	49.73	561	0.0	26.4	28.2	8.0	13.2	20.0	4.3
exo_temporal	25.54	831	0.0	17.7	18.9	13.8	17.4	16.6	15.5

RL

Config.	Accuracy (%)	Errors	None	Size	Orientation	Layout	Color	Lighting	Background
base	21.15	880	0.0	16.2	17.6	15.2	17.0	16.2	17.6
full-mix	—	—	—	—	—	—	—	—	—
perc_causal	22.85	861	0.0	12.1	18.4	19.2	21.6	15.7	13.1
perc_temporal	25.27	834	0.0	17.0	16.8	14.3	16.4	15.6	19.9
insp_causal	24.10	847	0.0	18.1	19.2	13.0	13.5	15.9	20.3
insp_cf	24.37	844	0.0	18.6	17.9	12.6	15.4	19.0	16.6
insp_temporal	26.34	822	0.0	16.8	19.2	14.8	17.8	16.1	15.3
navi_causal	34.77	728	0.0	11.4	27.2	8.5	18.4	23.8	10.7
navi_temporal	27.96	804	0.0	19.7	18.3	11.1	17.8	18.9	14.3
mani_causal	44.35	621	0.0	20.8	28.0	9.0	15.5	17.9	8.9
mani_cf	41.31	655	0.0	19.8	25.2	9.0	13.6	15.9	16.5
exo_physical	51.08	546	0.0	30.8	28.2	4.4	11.2	23.1	2.4
exo_temporal	28.05	803	0.0	18.8	16.3	14.8	20.0	15.4	14.6

Table 86: Default-state choices on ID. Forward/inverse dynamics uses the subset with a no-change option ($n = 540$); counterfactual substitution and removal each use $n = 558$. Accuracy and $P(\text{static} \mid \text{error})$ are percentages. The latter is reported separately for forward/inverse dynamics and substitution, where the default-state choice is wrong. In removal, the unchanged state is correct, so accuracy measures selecting it appropriately. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT

	Forward/inverse dynamics		CF substitution		CF removal
Config.	Accuracy (%)	$P(\text{static} \mid \text{error})$ (%)	Accuracy (%)	$P(\text{static} \mid \text{error})$ (%)	Accuracy (%)
base	22.22	63.8	19.35	27.8	18.64
full-mix	93.52	0.0	94.80	51.7	96.95
perc_causal	43.52	7.2	62.19	36.5	28.85
perc_temporal	23.70	56.3	18.82	29.1	20.61
insp_causal	93.70	2.9	74.19	51.4	41.22
insp_cf	55.19	89.3	83.15	19.1	86.92
insp_temporal	20.93	57.6	19.00	26.8	21.86
navi_causal	39.07	35.9	59.68	47.6	49.64
navi_temporal	22.22	57.4	17.20	28.8	20.43
mani_causal	30.00	58.5	60.93	33.5	48.39
mani_cf	22.78	75.1	68.10	54.5	57.35
exo_physical	29.44	37.5	15.23	48.6	38.17
exo_temporal	24.81	43.1	18.10	28.0	21.51

RL

	Forward/inverse dynamics		CF substitution		CF removal
Config.	Accuracy (%)	$P(\text{static} \mid \text{error})$ (%)	Accuracy (%)	$P(\text{static} \mid \text{error})$ (%)	Accuracy (%)
base	22.22	63.8	19.35	27.8	18.64
full-mix	—	—	—	—	—
perc_causal	42.96	8.1	67.56	32.0	31.00
perc_temporal	24.44	56.9	17.74	31.4	20.97
insp_causal	89.63	0.0	83.51	51.1	44.80
insp_cf	60.00	90.3	84.23	37.5	87.28
insp_temporal	22.59	57.7	14.87	28.4	23.12
navi_causal	46.11	27.1	60.39	46.2	47.49
navi_temporal	28.52	40.4	17.38	32.1	22.94
mani_causal	34.81	51.1	64.70	32.5	48.39
mani_cf	26.48	72.5	70.25	45.2	60.57
exo_physical	31.30	38.0	15.95	53.5	42.65
exo_temporal	23.52	51.3	17.56	27.4	23.66

Table 87: Temporal-ordering errors on held-out scenes (OOD, $n = 636$). Accuracy is in percent; Errors counts incorrect answers. Error shares (%) distinguish adjacent swaps in the middle, adjacent swaps at an end, and exchanges of the first and last states. Shares sum to approximately 100, allowing for rounding. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT					
Config.	Accuracy (%)	Errors	Center swap	Edge swap	Endpoint swap
base	37.26	399	44.9	39.8	15.3
full-mix	85.22	94	26.6	71.3	2.1
perc_causal	28.93	452	31.0	41.2	27.9
perc_temporal	25.31	475	36.4	27.2	36.4
insp_causal	22.01	496	35.3	36.3	28.4
insp_cf	33.18	425	39.8	34.1	26.1
insp_temporal	17.61	524	27.3	32.1	40.6
navi_causal	32.86	427	40.3	37.7	22.0
navi_temporal	26.10	470	40.4	41.5	18.1
mani_causal	40.09	381	41.7	34.6	23.6
mani_cf	31.13	438	38.8	34.5	26.7
exo_physical	30.82	440	38.0	35.9	26.1
exo_temporal	24.53	480	39.6	42.7	17.7
RL					
Config.	Accuracy (%)	Errors	Center swap	Edge swap	Endpoint swap
base	37.26	399	44.9	39.8	15.3
full-mix	—	—	—	—	—
perc_causal	25.16	476	31.5	39.1	29.4
perc_temporal	25.00	477	38.6	27.3	34.2
insp_causal	23.58	486	32.3	39.5	28.2
insp_cf	31.76	434	38.7	37.1	24.2
insp_temporal	12.74	555	27.4	29.5	43.1
navi_causal	33.33	424	43.6	37.0	19.3
navi_temporal	30.82	440	32.5	42.5	25.0
mani_causal	41.19	374	38.2	38.2	23.5
mani_cf	33.18	425	40.2	34.6	25.2
exo_physical	33.02	426	37.3	37.6	25.1
exo_temporal	28.93	452	36.5	51.3	12.2

Table 88: Temporal-adjacency errors on held-out scenes (OOD, $n = 636$). Accuracy is in percent; Errors counts incorrect answers. Successor and predecessor off-by-one errors extend one step beyond the correct neighbor; Swapped reverses before and after. Error shares (%) sum to approximately 100, allowing for rounding. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT

Config.	Accuracy (%)	Errors	Successor off-by-one	Predecessor off-by-one	Swapped
base	27.04	464	29.3	30.8	39.9
full-mix	64.31	227	0.0	4.8	95.2
perc_causal	21.38	500	35.4	32.4	32.2
perc_temporal	41.19	374	1.6	7.2	91.2
insp_causal	20.60	505	23.8	39.4	36.8
insp_cf	30.66	441	24.9	32.9	42.2
insp_temporal	44.18	355	1.1	3.9	94.9
navi_causal	31.60	435	23.4	25.7	50.8
navi_temporal	43.55	359	0.0	1.4	98.6
mani_causal	32.39	430	22.3	23.0	54.7
mani_cf	37.11	400	23.2	23.2	53.5
exo_physical	27.20	463	37.4	28.9	33.7
exo_temporal	52.99	299	0.0	2.3	97.7

RL

Config.	Accuracy (%)	Errors	Successor off-by-one	Predecessor off-by-one	Swapped
base	27.04	464	29.3	30.8	39.9
full-mix	—	—	—	—	—
perc_causal	19.50	512	36.9	31.1	32.0
perc_temporal	45.28	348	1.7	4.0	94.3
insp_causal	19.81	510	25.9	36.5	37.6
insp_cf	27.04	464	24.1	31.7	44.2
insp_temporal	36.48	404	0.7	3.7	95.5
navi_causal	36.64	403	19.9	18.4	61.8
navi_temporal	45.60	346	0.6	0.0	99.4
mani_causal	39.62	384	20.6	18.2	61.2
mani_cf	39.78	383	19.1	17.5	63.4
exo_physical	28.93	452	36.1	29.4	34.5
exo_temporal	56.60	276	0.4	0.4	99.3

Table 89: Action-confusion errors in forward dynamics on perceptive and inspective ID items ($n = 558$). Options are outcome images. Mirror and Reverse select outcomes of the left/right or forward/backward counterpart; Other selects another action's outcome; Static selects the unchanged scene. Accuracy and error shares are percentages; Errors counts incorrect answers. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT

Config.	Accuracy (%)	Errors	Mirror (L/R)	Reverse (F/B)	Other action	Static option
base	24.37	422	15.4	10.7	58.3	15.6
full-mix	97.31	15	73.3	0.0	26.7	0.0
perc_causal	71.33	160	40.6	0.0	51.9	7.5
perc_temporal	24.19	423	15.1	8.5	64.3	12.1
insp_causal	70.43	165	22.4	7.3	69.7	0.6
insp_cf	71.51	159	40.9	11.3	40.9	6.9
insp_temporal	24.73	420	16.0	8.6	60.0	15.5
navi_causal	37.63	348	13.5	8.3	64.7	13.5
navi_temporal	25.09	418	16.3	12.2	55.7	15.8
mani_causal	30.11	390	9.2	11.5	48.7	30.5
mani_cf	33.87	369	14.9	12.7	50.9	21.4
exo_physical	29.75	392	20.4	18.9	59.7	1.0
exo_temporal	24.19	423	16.8	11.8	53.9	17.5

RL

Config.	Accuracy (%)	Errors	Mirror (L/R)	Reverse (F/B)	Other action	Static option
base	24.37	422	15.4	10.7	58.3	15.6
full-mix	—	—	—	—	—	—
perc_causal	70.61	164	52.4	0.0	39.0	8.5
perc_temporal	22.04	435	14.9	7.8	60.9	16.3
insp_causal	68.82	174	18.4	5.7	75.9	0.0
insp_cf	71.68	158	31.6	20.9	38.0	9.5
insp_temporal	24.01	424	17.2	10.6	55.9	16.3
navi_causal	39.25	339	13.3	6.8	64.9	15.0
navi_temporal	23.12	429	17.5	9.3	55.0	18.2
mani_causal	33.15	373	10.7	11.0	51.7	26.5
mani_cf	38.35	344	16.6	12.2	53.5	17.7
exo_physical	30.65	387	20.7	18.1	60.2	1.0
exo_temporal	24.19	423	16.1	10.2	55.8	18.0

Table 90: Action-confusion errors in inverse dynamics on perceptive and inspective ID items ($n = 558$). Options are action descriptions, not outcome images. Mirror and Reverse select the left/right or forward/backward counterpart; Other selects another action; Static denotes the no-action choice. Accuracy and error shares are percentages; Errors counts incorrect answers. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT						
Config.	Accuracy (%)	Errors	Mirror (L/R)	Reverse (F/B)	Other action	Static option
base	23.30	428	1.6	15.9	35.3	47.2
full-mix	94.62	30	73.3	0.0	26.7	0.0
perc_causal	72.04	156	44.2	1.3	48.1	6.4
perc_temporal	23.48	427	3.0	18.0	36.5	42.4
insp_causal	68.46	176	23.3	14.2	62.5	0.0
insp_cf	24.01	424	0.0	17.9	33.7	48.3
insp_temporal	23.12	429	6.8	14.0	37.1	42.2
navi_causal	39.43	338	12.1	13.6	53.3	21.0
navi_temporal	24.91	419	9.1	14.3	34.8	41.8
mani_causal	35.13	362	10.8	7.5	53.6	28.2
mani_cf	22.76	431	0.5	12.5	32.7	54.3
exo_physical	27.96	402	2.5	13.7	49.3	34.6
exo_temporal	29.03	396	18.7	15.7	40.2	25.5
RL						
Config.	Accuracy (%)	Errors	Mirror (L/R)	Reverse (F/B)	Other action	Static option
base	23.30	428	1.6	15.9	35.3	47.2
full-mix	—	—	—	—	—	—
perc_causal	71.51	159	50.9	3.1	39.0	6.9
perc_temporal	25.27	417	4.6	18.2	38.6	38.6
insp_causal	65.23	194	20.1	17.5	62.4	0.0
insp_cf	29.39	394	0.3	18.0	36.0	45.7
insp_temporal	25.27	417	7.9	16.1	34.8	41.2
navi_causal	49.64	281	18.1	12.8	59.1	10.0
navi_temporal	36.92	352	23.0	13.4	41.5	22.2
mani_causal	36.02	357	14.3	10.1	52.9	22.7
mani_cf	21.86	436	0.2	14.4	33.3	52.1
exo_physical	27.96	402	1.2	15.9	48.8	34.1
exo_temporal	26.88	408	4.9	13.5	48.3	33.3

Table 91: Action-confusion errors in counterfactual substitution on inspective ID items only ($n = 270$). Options are outcome images. Mirror selects the opposite left/right orbit, Other another action's outcome, and Static the unchanged scene. The forward/backward Reverse column is retained as supplied. Accuracy and error shares are percentages; Errors counts incorrect answers. Panels report SFT and RL separately; — denotes unavailable full-mix RL results.

SFT						
Config.	Accuracy (%)	Errors	Mirror (L/R)	Reverse (F/B)	Other action	Static option
base	18.52	220	8.2	0.0	61.4	30.5
full-mix	94.81	14	14.3	0.0	35.7	50.0
perc_causal	47.78	141	7.1	0.0	53.9	39.0
perc_temporal	20.74	214	8.9	0.0	59.3	31.8
insp_causal	80.00	54	0.0	0.0	13.0	87.0
insp_cf	93.70	17	0.0	0.0	82.4	17.6
insp_temporal	19.26	218	8.3	0.0	62.8	28.9
navi_causal	34.81	176	4.0	0.0	48.9	47.2
navi_temporal	18.52	220	8.6	0.0	59.5	31.8
mani_causal	28.52	193	3.6	0.0	70.5	25.9
mani_cf	38.89	165	4.2	0.0	40.0	55.8
exo_physical	11.85	238	11.8	0.0	38.2	50.0
exo_temporal	18.52	220	9.1	0.0	59.5	31.4
RL						
Config.	Accuracy (%)	Errors	Mirror (L/R)	Reverse (F/B)	Other action	Static option
base	18.52	220	8.2	0.0	61.4	30.5
full-mix	—	—	—	—	—	—
perc_causal	54.44	123	7.3	0.0	59.3	33.3
perc_temporal	17.04	224	8.5	0.0	58.9	32.6
insp_causal	86.30	37	0.0	0.0	16.2	83.8
insp_cf	95.19	13	0.0	0.0	53.8	46.2
insp_temporal	14.81	230	13.0	0.0	56.1	30.9
navi_causal	36.30	172	4.1	0.0	49.4	46.5
navi_temporal	16.30	226	8.0	0.0	59.3	32.7
mani_causal	35.19	175	3.4	0.0	71.4	25.1
mani_cf	43.33	153	5.9	0.0	45.8	48.4
exo_physical	12.59	236	11.0	0.0	37.3	51.7
exo_temporal	15.19	229	14.4	0.0	58.1	27.5

G.17.3 QUESTION-DEPENDENT STATE JUDGMENTS AND THE SCENE GENERALIZATION GAP

Shortcut check for the default-state result. A deflationary reading of Table 86 is that causal supervision, whose correct options are always changed frames, merely teaches a surface rule, *prefer the changed frame*, which would lower the static pull as a side effect. The counterfactual-removal probes separate the two readings: their correct option is exactly the unchanged scene, so a blanket changed-frame rule would suppress it and push accuracy *below* base. At the SFT stage, the opposite happens: every causally-trained cell improves counterfactual-removal accuracy over base ($18.6\% \rightarrow 28.9/41.2/49.6/48.4$ for `perc/insp/navi/mani_causal`) while simultaneously cutting the static pull on forward probes. The same weights move the unchanged-scene option in opposite directions depending on what the question asks, the signature of a removed prior with question-conditioned state selection.

Table 86 reports both stages and includes counterfactual substitution and removal alongside the forward/inverse probes. Table 82 gives the corresponding temporal-adjacency error analysis. Table 79 reports absolute accuracies under geometric and appearance perturbations, with the base row providing the reference for gains.

Competency-matched scene gap. Table 92 reports the SFT comparison that averages OOD-minus-ID accuracy differences equally across reasoning types. This statistic differs from subtracting the overall accuracies in Tables 75 and 77, which weight individual test items equally. Temporal-ordering distractors also differ between the two tests: ID uses random permutations, whereas OOD uses structured swaps. Their error distributions are therefore reported separately in Tables 83 and 87.

Table 92: **Scene-OOD minus ID accuracy gap by supervision** (competency-matched mean, pp).

Run	OOD − ID (pp)
base	$+2.06$
full-mix	$+1.21$
perc_causal	-1.45
perc_temporal	-0.50
insp_causal	$+0.33$
insp_cf	$+1.51$
insp_temporal	-1.17
navi_causal	$+0.05$
navi_temporal	$+1.66$
mani_causal	$+6.87$
mani_cf	$+4.48$
exo_physical	$+5.05$
exo_temporal	$+1.57$

H CONTENT CONTROL: TRANSITION CONTENT VERSUS TRAINING FORMAT

The downstream gains of WOVEN training come from learned transition reasoning: they appear only when the model acquires the correct action–outcome relations, not from the question format, the answer-letter balance, or the amount of training. Training on WOVEN teaches more than transitions: answering four-option questions, spreading answers over the letters, reading several images with images as options, and simply more training. We therefore train on four data sets that match a WOVEN subset in everything except that content and compare them with the real data (Table 93). The real data are the perceptive causal-dynamics subset. Two controls keep its items, images, options, and answer-letter balance and change only the action–outcome mapping: a per-source permutation maps the four camera actions to the outcomes by a random permutation without fixed points within each source, so that no consistent rule can be learned (the answer changes in 95.1% of the examples), and a global mirror reverses the direction of every action under one mapping for all sources (the answer changes in 95.0%), a consistent rule with the wrong semantics. A third control trains on A-OKVQA multiple-choice questions (Schwenk et al., 2022) with a balanced answer-letter distribution and no WOVEN content. If the gains came from the format or from the amount of training, the controls

would gain as much as the real data; if they require the transition content, only the real data gains, which is what we observe.

Table 93: Content control: the four training sets (9,750 training examples each, same number of optimizer steps).

Arm	Training data	Relation to the real data
Real	WOVEN perceptive × causal dynamics (forward and inverse dynamics questions, four option orderings).	—
Per-source permutation	Same items, images, options, and answer-letter balance. Within each source, the four camera actions are mapped to the outcomes by a random permutation with no fixed point: the forward question for action a_i has the outcome of $a_{\pi(i)}$ as its answer, and the inverse question showing outcome s'_k has $a_{\pi^{-1}(k)}$ as its answer. The answer changes in 95.1% of the examples.	Identical format; the action–outcome mapping differs across sources, so no consistent rule can be learned.
Global mirror	Same items, images, options, and answer-letter balance. One mapping for all sources: move forward ↔ move backward, rotate left 45° ↔ rotate right 45°. The answer changes in 95.0% of the examples.	Identical format; a consistent, learnable rule with the direction of every action reversed.
A-OKVQA	2,072 four-option questions from the A-OKVQA training set (Schwenk et al., 2022) (COCO images), each in its four option orderings, plus 1,462 additional random orderings.	Multiple-choice training with a balanced answer-letter distribution, without WOVEN content.

All four training sets have 9,750 examples and the same number of optimizer steps. Each is trained on Qwen2.5-VL-3B-Instruct with LoRA (rank 16, $\alpha = 32$, dropout 0.05) on the language model's q/k/v/o projections, vision encoder frozen, vision–language merger trained, learning rate 10^{-4}, 5% warmup, cosine schedule, effective batch size 64, 3 epochs, final checkpoint, with 3 seeds (12 models). Evaluation uses greedy decoding (vLLM) with the model asked to output only the option letter, on SAT (Ray et al., 2024) (150 two-option questions, scored circularly: each question is asked once with each option order and counts as correct only if both answers are right), SPAR-Bench view change (Zhang et al., 2025b) on ScanNet and ScanNet++ (261 questions with two real images and ground-truth camera poses, asking how the camera moved between them, the format of WOVEN's perceptive inverse-dynamics questions on real images; the numeric questions are converted to four options as in Appendix O), ActionEQA (Anonymous, 2025) (800 questions) and WorldPrediction (WorldPrediction Team, 2025) (825 questions), both scored circularly (each question is asked four times with the options rotated so that the correct option appears at every letter once, and counts as correct only if all four answers are right, so a preference for any answer letter earns no credit), and WOVEN OOD (the 3,496 held-out-scene questions). Circular scoring and the two-option SAT subset are stricter than the single-pass accuracy of the main text, so the base scores here differ from those in Table 2. For each question, the score of an arm is averaged over its 3 seeds; two arms are compared by a paired sign-flip permutation test over questions, with a bootstrap 95% confidence interval of the difference.

Table 94: Content control: accuracy (%), mean over 3 seeds.

Benchmark	Base	Real	Permutation	Mirror	A-OKVQA
SAT (circular)	39.3	**63.3**	35.1	37.8	44.0
SPAR-Bench view change (real images)	26.4	**54.9**	27.5	35.8	21.8
ActionEQA (circular)	25.6	**31.8**	22.2	26.5	22.6
WorldPrediction (circular)	14.7	**16.6**	12.4	11.3	10.3
WOVEN OOD	36.2	**45.9**	36.8	32.9	32.6

Only the real data gains (Tables 94–96). The two matched controls fall 4.2 to 28.2 points below the real arm on every benchmark (all $p < 0.001$; Table 95), so the multiple-choice format, the answer-letter balance, multi-image input with image options, and the amount of training do not by themselves produce the gains; under circular scoring, where a letter preference earns no credit, the

Table 95: Content control: Real minus control, percentage points with bootstrap 95% confidence intervals; every difference has $p < 0.001$ under the paired permutation test.

Benchmark	Real − Permutation	Real − Mirror	Real − A-OKVQA
SAT (circular)	$+28.2$ $[+21.1, +35.8]$	$+25.6$ $[+18.2, +33.3]$	$+19.3$ $[+11.8, +26.9]$
SPAR-Bench view change	$+27.5$ $[+20.7, +34.5]$	$+19.2$ $[+12.4, +25.5]$	$+33.1$ $[+26.4, +39.3]$
ActionEQA (circular)	$+9.6$ $[+7.0, +12.3]$	$+5.3$ $[+3.5, +7.2]$	$+9.2$ $[+6.7, +11.8]$
WorldPrediction (circular)	$+4.2$ $[+1.9, +6.3]$	$+5.3$ $[+3.6, +7.1]$	$+6.3$ $[+4.0, +8.6]$
WOVEN OOD	$+9.1$ $[+7.2, +11.0]$	$+13.0$ $[+11.4, +14.5]$	$+13.3$ $[+11.4, +15.3]$

Table 96: Content control: per-seed accuracy (%).

Benchmark	Arm	s1	s2	s3	Mean
SAT (circular)	Real	65.3	63.3	61.3	63.3
	Per-source permutation	35.3	36.7	33.3	35.1
	Global mirror	37.3	40.0	36.0	37.8
	A-OKVQA	42.7	44.7	44.7	44.0
SPAR-Bench view change	Real	56.3	55.2	53.3	54.9
	Per-source permutation	29.9	25.7	26.8	27.5
	Global mirror	42.9	39.5	24.9	35.8
	A-OKVQA	25.3	19.2	21.1	21.8
ActionEQA (circular)	Real	32.2	31.8	31.5	31.8
	Per-source permutation	22.2	22.2	22.1	22.2
	Global mirror	27.8	26.6	25.2	26.5
	A-OKVQA	22.1	22.6	23.0	22.6
WorldPrediction (circular)	Real	15.4	19.5	14.9	16.6
	Per-source permutation	12.2	12.1	13.0	12.4
	Global mirror	9.0	8.1	16.7	11.3
	A-OKVQA	9.0	11.2	10.7	10.3
WOVEN OOD	Real	45.5	46.3	45.9	45.9
	Per-source permutation	37.3	37.4	35.8	36.8
	Global mirror	34.8	35.2	28.7	32.9
	A-OKVQA	32.6	33.3	31.8	32.6

real arm stays above all controls. The A-OKVQA arm stays below the real arm on every benchmark and below the base model on three, so general multiple-choice training does not produce the gains either. The mirror arm learns its reversed rule: on the perceptive forward and inverse dynamics questions of the WOVEN ID test set (288 each, from sources not seen in training), it answers 85–92% of the inverse questions by the reversed rule, whereas the permutation arm stays near chance because its mapping does not carry over to a new source (Table 97). The mirror arm nevertheless obtains at most a fraction of the gains, 9.4 points above the base model on SPAR-Bench and 0.9 on ActionEQA and below it on the other three benchmarks, so the gains require the correct action semantics and not only a consistent rule. On SPAR-Bench, real images in WOVEN's inverse-dynamics format, the real arm doubles the base accuracy ($26.4 \rightarrow 54.9$) while the permutation arm (27.5) and the A-OKVQA arm (21.8) stay at or below the base model and the mirror arm reaches 35.8, so what is learned from the generated rollouts carries over to real images. Several controls fall below the base model (all three on WorldPrediction, the mirror and A-OKVQA arms on WOVEN OOD, the A-OKVQA arm on SPAR-Bench), so part of each difference in Table 95 comes from the control falling below the base model rather than from the real arm rising above it. For the perceptive causal-dynamics subset, the external gains therefore come from learned transition reasoning: they require the correct action–outcome relations and are not produced by the format or the amount of training.

Table 97: Content control, manipulation check: answers on WOVEN ID perceptive dynamics questions (%), per seed.

Arm	Forward: real answer	Forward: mirrored answer	Inverse: real answer	Inverse: mirrored answer
Base	24.7	27.1	25.3	28.8
Real	97.2 / 98.6 / 95.8	1.0 / 0.3 / 1.7	89.2 / 93.1 / 85.8	6.6 / 3.8 / 6.2
Per-source permutation	21.5 / 24.3 / 21.9	18.4 / 20.5 / 21.9	22.9 / 24.3 / 22.2	31.2 / 29.2 / 32.3
Global mirror	37.2 / 42.0 / 20.5	26.7 / 32.3 / 71.2	5.9 / 8.3 / 5.6	87.2 / 84.7 / 92.4
A-OKVQA	24.7 / 25.0 / 22.6	25.0 / 25.7 / 26.7	31.9 / 33.3 / 25.3	29.5 / 30.2 / 31.9

I LETTER-BALANCE CONTROL: CIRCULAR SCORING

The external gains of WOVEN training remain when a letter preference earns no credit. SFT uses all four option orderings of every question and the GRPO reward penalizes an uneven answer-letter distribution (Appendix G.1), so WOVEN-trained models spread their answers over the letters more evenly than the base model, which prefers some letters (Table 98); on a benchmark whose base score is lowered by a letter preference, this alone raises accuracy. We therefore re-score with circular evaluation: each question is asked once per option rotation, so that the correct option appears at every letter once, and it counts as correct only if every rotation is answered correctly, so a model that answers without knowing the answer scores at most $(1/n)^n$ per question (0.4% for four options, 25% for two) whatever its letter distribution. A gain that remains under circular scoring comes from the content of the options.

We re-score the base model and four WOVEN-trained models from the main experiments, the full mixture after SFT and after SFT + GRPO and the perceptive causal-dynamics subset after SFT and after SFT + GRPO, on one benchmark per domain: SAT (Ray et al., 2024) (a 150-question two-option subset), ActionEQA (Anonymous, 2025) (800 four-option questions), and SPAR-Bench view change (Zhang et al., 2025b) on ScanNet and ScanNet++ (261 four-option questions on real images, converted from the numeric questions as in Appendix O). Decoding is greedy (vLLM) with the model asked to output only the option letter; standard scoring asks each question once with its original option order (for SAT, the accuracy over the two orders of each question). Because SAT and ActionEQA are evaluated on these subsets, the standard-scoring base scores differ from those in Table 2. Each trained model is compared with the base model by a paired sign-flip permutation test over questions with a bootstrap 95% confidence interval of the difference.

Training does even out the answer letters: after training they follow the distribution of the correct answers closely (Table 98). Under circular scoring, all four trained models remain significantly above the base model on all three benchmarks (Tables 99 and 100), with larger gains than under standard scoring because the base model's letter preference costs it more; the SFT models, trained without the balancing reward, gain as much as the SFT + GRPO models. The base model's circular score on SPAR-Bench (0.4%) equals the expected score of answering without the content: its answers are

Table 98: Letter-balance control: answer-letter distribution (% of answers), standard scoring.

Benchmark	Model	A	B	C	D
SAT	Correct answers	50.0	50.0		
	Base	62.0	38.0		
	Full mixture, SFT	51.0	49.0		
	Full mixture, SFT + GRPO	49.0	51.0		
	Perceptive × causal dynamics, SFT	52.3	47.7		
	Perceptive × causal dynamics, SFT + GRPO	55.0	45.0		
ActionEQA	Correct answers	24.2	24.0	26.5	25.2
	Base	19.5	32.0	30.0	18.5
	Full mixture, SFT	24.8	21.1	28.0	26.1
	Full mixture, SFT + GRPO	22.8	21.8	31.6	23.9
	Perceptive × causal dynamics, SFT	24.1	23.5	28.5	23.9
	Perceptive × causal dynamics, SFT + GRPO	25.2	19.5	28.0	27.2
SPAR-Bench view change	Correct answers	28.7	23.0	23.0	25.3
	Base	64.0	14.2	17.6	4.2
	Full mixture, SFT	25.7	24.1	27.2	23.0
	Full mixture, SFT + GRPO	19.9	24.9	31.4	23.8
	Perceptive × causal dynamics, SFT	35.2	23.4	21.5	19.9
	Perceptive × causal dynamics, SFT + GRPO	31.8	27.6	21.5	19.2

Table 99: Letter-balance control: accuracy (%) under standard and circular scoring.

Benchmark	Scoring	Base	Full mixture, SFT	Full mixture, SFT + GRPO	Perc. × causal, SFT	Perc. × causal, SFT + GRPO
SAT	Standard	57.3	68.3	67.7	69.7	66.3
	Circular	39.3	62.7	61.3	64.7	58.7
ActionEQA	Standard	43.1	50.2	49.0	49.5	52.0
	Circular	25.6	35.4	34.2	32.8	35.5
SPAR-Bench view change	Standard	26.4	41.8	39.5	54.8	55.9
	Circular	0.4	20.3	19.5	30.7	36.0

Table 100: Letter-balance control: trained minus base, percentage points with bootstrap 95% confidence intervals. Every difference is significant under the paired permutation test ($p \leq 0.006$ under standard scoring; $p < 0.001$ under circular scoring).

Benchmark	Scoring	Full mixture, SFT	Full mixture, SFT + GRPO	Perc. × causal, SFT	Perc. × causal, SFT + GRPO
SAT	Standard	+11.0 [+4.7, +17.7]	+10.3 [+3.7, +16.7]	+12.3 [+6.7, +18.0]	+9.0 [+3.0, +15.0]
	Circular	+23.3 [+14.0, +32.7]	+22.0 [+12.7, +31.3]	+25.3 [+16.7, +34.0]	+19.3 [+11.3, +27.3]
ActionEQA	Standard	+7.1 [+3.0, +11.2]	+5.9 [+1.9, +10.1]	+6.4 [+3.0, +9.8]	+8.9 [+5.5, +12.4]
	Circular	+9.8 [+6.1, +13.4]	+8.6 [+5.1, +12.2]	+7.1 [+4.4, +10.0]	+9.9 [+6.9, +13.0]
SPAR view change	Standard	+15.3 [+7.3, +23.4]	+13.0 [+4.6, +21.1]	+28.4 [+21.1, +35.6]	+29.5 [+22.2, +36.8]
	Circular	+19.9 [+15.3, +24.9]	+19.2 [+14.6, +24.1]	+30.3 [+24.9, +35.6]	+35.6 [+29.9, +41.4]

determined by the option position. The external gains of WOVEN training thus remain significant when a letter preference earns no credit, after SFT and after GRPO; they cannot be attributed to the model having learned to balance its answer letters or to the multiple-choice format.

J GENERALIZATION ACROSS MODEL FAMILIES

The effect of WOVEN training holds across model families. The main experiments train Qwen2.5-VL; we repeat the training on Gemma-3-4B-it (Gemma Team, Google DeepMind, 2025), which has a different vision encoder and a different way of handling multiple images, and compare with the untrained Gemma-3-4B-it. Training uses the WOVEN perceptive causal-dynamics subset (forward and inverse dynamics questions, four option orderings, $9{,}750$ training examples) with LoRA (rank 16, $\alpha = 32$, dropout 0.05) on the language model's q/k/v/o projections, vision encoder frozen, multimodal projector trained, learning rate 10^{-4}, 5% warmup, cosine schedule, effective batch size 64, 3 epochs, final checkpoint, with 3 seeds. Evaluation uses greedy decoding (vLLM) with the model asked to output only the option letter, on WOVEN ID ($6{,}228$ questions), WOVEN OOD ($3{,}496$ held-out-scene questions), and SPAR-Bench view change (Zhang et al., 2025b) on ScanNet and ScanNet++ (261 questions with two real images, asking how the camera moved between them, converted to four options as in Appendix O). For each question, the score of the trained model is averaged over the 3 seeds and compared with the base model by a paired sign-flip permutation test over questions, with a bootstrap 95% confidence interval of the difference. Table 102 reports the same training and evaluation on Qwen2.5-VL-3B-Instruct for reference; because the training configuration and the decoding protocol differ from those of the main experiments, these Qwen scores differ from the corresponding scores in the main text.

Table 101: Gemma-3-4B-it: accuracy (%). Δ = trained (mean of 3 seeds) $-$ base, percentage points with bootstrap 95% confidence intervals.

Benchmark	Base	Seed 1	Seed 2	Seed 3	Mean	Δ [95% CI]	p
WOVEN ID	27.9	44.9	44.9	45.5	**45.1**	$+17.2$ $[+15.8, +18.6]$	< 0.001
WOVEN OOD	33.0	44.5	45.7	47.2	**45.8**	$+12.7$ $[+10.7, +14.8]$	< 0.001
SPAR-Bench view change (real images)	28.0	44.4	42.9	41.0	**42.8**	$+14.8$ $[+7.8, +21.8]$	< 0.001

Table 102: The same training on Qwen2.5-VL-3B-Instruct, accuracy (%), for reference.

Benchmark	Base	Mean of 3 seeds	Δ
WOVEN ID	29.8	47.7	$+17.9$
WOVEN OOD	36.2	45.9	$+9.7$
SPAR-Bench view change (real images)	26.4	54.9	$+28.5$

WOVEN training raises Gemma-3-4B-it by 17.2 points on WOVEN ID, 12.7 points on held-out WOVEN scenes, and 14.8 points on real-image camera-motion questions (all $p < 0.001$), with every seed above the base model (Table 101). The effect of WOVEN training holds across model families.

K SHORTCUT CONTROL FOR CROSS-OPERATION TRANSFER: THE UNCHANGED-SCENE OPTION

The transfer from forward dynamics to inverse dynamics and to counterfactual reasoning within an action type does not depend on the unchanged-scene option. Training only on inspective forward dynamics questions (initial view and camera action, pick the resulting view) transfers to inverse dynamics, counterfactual substitution, and counterfactual removal of the same action type (Appendix G.8). Inspective questions contain an option showing the unchanged scene (the initial view, or the text "the camera remains stationary"), which is always wrong in forward dynamics, inverse dynamics, and counterfactual substitution and always correct in counterfactual removal, so part of the transfer to substitution could be a rule about that option rather than structure shared across reasoning types. We remove the option from the forward dynamics, inverse dynamics, and substitution questions, which become three-option questions (chance 33.3%) on which avoiding it no longer helps,

and evaluate removal in its original form, where the unchanged scene is the correct answer and a model that avoids it would fall below the base model.

The model is Qwen2.5-VL-3B-Instruct trained only on inspective forward dynamics (1,020 questions, each in its four option orderings, 4,080 training examples) in one run, with LoRA (rank 16, $\alpha = 32$, dropout 0.05), learning rate 10^{-4}, 3 epochs, final checkpoint, compared with the untrained base model; this configuration differs from that of Appendix G.8, so the gains differ in size. Evaluation uses the WOVEN ID inspective questions, 270 per reasoning type, with greedy decoding (vLLM) and the model asked to output only the option letter; in the static-free versions the remaining options are relabeled A–C in their original order. Differences (forward-only − base) are tested by a paired sign-flip permutation test over questions with a bootstrap 95% confidence interval.

Table 103: Static-free questions (unchanged-scene option removed, three options), accuracy (%).

Reasoning type	Base	Forward-only	Δ [95% CI]	p
Counterfactual substitution	28.5	93.3	$+64.8$ $[+58.5, +71.1]$	< 0.001
Inverse dynamics	43.3	60.7	$+17.4$ $[+12.2, +22.6]$	< 0.001
Forward dynamics (trained)	37.8	87.4	$+49.6$ $[+42.2, +57.0]$	< 0.001

Table 104: Original questions (four options, unchanged-scene option included), accuracy (%).

Reasoning type	Base	Forward-only	Δ [95% CI]	p
Counterfactual substitution	24.8	78.1	$+53.3$ $[+45.9, +60.4]$	< 0.001
Counterfactual removal (answer: unchanged scene)	27.8	55.2	$+27.4$ $[+19.3, +35.6]$	< 0.001
Inverse dynamics	17.4	44.1	$+26.7$ $[+21.5, +32.2]$	< 0.001
Forward dynamics (trained)	28.9	86.3	$+57.4$ $[+50.7, +64.1]$	< 0.001

With the unchanged-scene option removed, the forward-only model still gains 64.8 points on counterfactual substitution and 17.4 points on inverse dynamics (Table 103), and on counterfactual removal, where the unchanged scene is the correct answer, it gains 27.4 points (Table 104). The transfer from forward dynamics to inverse dynamics and to counterfactual reasoning within an action type therefore does not depend on that option.

L Consistency between forward and inverse dynamics

After WOVEN training, the forward and inverse queries about the same transition succeed and fail together, whereas for the base model they are independent, so the two reasoning operations draw on one shared mapping between states, actions, and outcomes rather than on two separate skills. Training raises accuracy on both forward dynamics (initial state and action, predict the outcome) and inverse dynamics (initial state and outcome, identify the action); the two could improve independently. We therefore query every transition (s, a_i, s_i') of the WOVEN ID test set twice, by the forward question for a_i and by the inverse question for s_i' (1,134 pairs from 306 sources, all ID transitions that have both questions), and compare the rate at which both are answered correctly (joint) with the rate expected if the two were independent, $P(\text{forward correct}) \times P(\text{inverse correct})$, computed within each action type and weighted by the number of pairs so that differences in difficulty between action types cannot produce a correlation. We also report the cycle rate, the share of transitions for which the inverse question about the outcome the model itself picked in the forward question returns the original action a_i. GRPO values are averaged over 3 GRPO seeds; 95% confidence intervals are by bootstrap over sources (2,000 resamples).

For the base model, success on the forward and the inverse question of a transition is independent (excess -0.1, confidence interval including 0). After training on causal or counterfactual transitions, the two queries of the same transition succeed together more often than independence predicts, for every trained model (excess $+0.4$ to $+4.5$, every confidence interval above 0); the full-mixture models are close to ceiling on both queries (at least 94.8%), which leaves little room for excess. The cycle rate, where the inverse query recovers the action behind the outcome the model itself predicted, rises from 19.0% to 45–94% (Table 105).

Table 105: Forward and inverse consistency on WOVEN ID (%). Joint: both questions of a pair correct; Independent: expected joint rate under independence within action type; Excess: joint − independent; Cycle: the inverse query recovers the action behind the outcome the model predicted; Δ joint: joint − joint of the base model.

Model	Forward	Inverse	Joint	Indep.	Excess [95% CI]	Cycle	Δ joint [95% CI]
Base	21.3	40.4	8.0	8.1	−0.1 [−1.2, +0.9]	19.0	—
Full mixture, SFT	97.1	94.8	93.5	92.1	+1.4 [+0.8, +2.1]	94.3	+85.4 [+83.2, +87.7]
Full mixture, SFT + GRPO	96.6	96.9	94.1	93.7	+0.4 [+0.1, +0.9]	94.4	+86.1 [+83.7, +88.2]
Perceptive × causal, SFT	75.3	76.6	65.5	61.8	+3.7 [+2.6, +4.8]	71.3	+57.5 [+53.6, +61.5]
Perceptive × causal, SFT + GRPO	75.7	77.4	66.9	62.8	+4.1 [+3.2, +4.9]	72.8	+58.8 [+55.1, +62.6]
Inspective × causal, SFT	73.7	75.8	61.9	59.3	+2.7 [+1.6, +3.7]	69.0	+53.9 [+49.6, +58.1]
Inspective × causal, SFT + GRPO	73.9	77.5	64.1	60.8	+3.3 [+2.4, +4.1]	70.5	+56.1 [+51.9, +60.0]
Navigative × causal, SFT	64.9	64.9	52.9	49.0	+3.9 [+2.7, +5.0]	58.6	+44.9 [+40.3, +49.3]
Navigative × causal, SFT + GRPO	63.1	64.3	51.5	47.3	+4.2 [+3.3, +5.1]	57.0	+43.5 [+39.3, +47.9]
Manipulative × causal, SFT	63.3	62.7	53.2	48.6	+4.5 [+3.5, +5.6]	57.1	+45.1 [+40.7, +49.6]
Manipulative × causal, SFT + GRPO	63.5	61.8	51.8	48.4	+3.4 [+2.7, +4.2]	56.3	+43.8 [+39.4, +48.3]
Inspective × counterfactual, SFT	73.7	50.5	40.0	38.0	+2.0 [+1.0, +3.1]	45.0	+32.0 [+28.3, +35.6]
Inspective × counterfactual, SFT + GRPO	72.5	53.4	41.7	39.6	+2.1 [+1.2, +3.0]	46.8	+33.7 [+30.2, +37.2]
Manipulative × counterfactual, SFT	63.2	54.8	46.9	43.7	+3.2 [+2.0, +4.3]	50.9	+38.9 [+34.2, +43.3]
Manipulative × counterfactual, SFT + GRPO	64.9	54.3	47.8	44.8	+3.0 [+2.1, +3.8]	51.3	+39.8 [+35.2, +44.2]

M SELECTION AND SEED ROBUSTNESS OF THE BEST-SUBSET GAINS

The best-subset gains on representative benchmarks of every capability domain are neither an artifact of selecting the best of 22 candidates nor of a single training seed. The controlled-subset result reports, for each benchmark, the best of the 11 subsets after SFT and after GRPO (22 candidates); taking the maximum of many candidates can produce a gain of a few points by selection alone, and a single training run can produce a gain by seed variation alone. To check both, we ran two additional GRPO seeds for every subset (3 in total) and selected two representative benchmarks for each capability domain of §4: SAT and ViewSpatial (spatial), ActionEQA and Robo2VLM (embodied), InPhyRe and CLEVRER (physical), WorldPrediction and WM-ABench (procedural and world-model prediction), and TempCompass and TemporalBench (long) (temporal). Scores are recomputed from the model outputs under each benchmark's official protocol (Table 106); GRPO scores are per-unit means over the 3 GRPO seeds, so the best candidate and its gain can differ from the single reported checkpoint in Table 3. The selection check compares the best gain among the 22 candidates with the distribution of the best gain that selection alone produces: per evaluation unit, the 22 differences from the base model are sign-flipped jointly, the 22 gains are recomputed, and their maximum is taken (10,000 flips); Holm correction over the 22 candidates is reported as well. The seed check evaluates the best subset after GRPO separately for each of its 3 seeds against the base model (paired sign-flip test, 10,000 flips).

Table 106: Representative benchmarks, evaluation units, and scoring.

Domain	Benchmark	Units	Scoring
Spatial	SAT	150	circular: a question counts only if both answer orders are correct
Spatial	ViewSpatial	5,712	question accuracy
Embodied	ActionEQA	3,782	question accuracy
Embodied	Robo2VLM	6,676	question accuracy
Physical	InPhyRe	19,981	question accuracy
Physical	CLEVRER	12,000	per-option accuracy (official stratified sample)
Procedural / world-model	WorldPrediction	825	question accuracy
Procedural / world-model	WM-ABench	4,807	question accuracy, resampled by question group
Temporal	TempCompass	5,536	question accuracy
Temporal	TemporalBench (long)	5,485	dual-order binary accuracy: a caption counts only if both answer orders are correct

On all ten benchmarks, the best-subset gain exceeds what selecting the best of 22 candidates produces by chance and remains significant after Holm correction (Table 107), and the best subset improves significantly over the base model in each of its three GRPO seeds (Table 108).

Table 107: Best of 22 candidates versus the max-selection null (accuracy %, gains in points). Perc., Insp., Mani.: perceptive, inspective, manipulative; cf: counterfactual. Null: 95th percentile of the best gain under selection alone; $p_{\max}$: p under the max-selection null; p_{Holm}: Holm-corrected over the 22 candidates.

Benchmark	Base	Best candidate	Best gain [95% CI]	Null	$p_{\max}$	p_{Holm}
SAT	37.3	Perc. × causal, SFT	+26.0 [+17.3, +34.7]	+10.4	< 0.001	0.002
ViewSpatial	35.3	Mani. × cf, SFT + GRPO	+3.6 [+2.7, +4.5]	+1.3	< 0.001	0.002
ActionEQA	43.4	Mani. × causal, SFT	+11.5 [+9.9, +13.1]	+2.0	< 0.001	0.002
Robo2VLM	33.5	Mani. × cf, SFT	+7.6 [+6.6, +8.6]	+1.2	< 0.001	0.002
InPhyRe	22.0	Insp. × cf, SFT	+10.3 [+9.7, +10.9]	+0.8	< 0.001	0.002
CLEVRER	58.2	Insp. × cf, SFT + GRPO	+4.6 [+3.8, +5.4]	+1.0	< 0.001	0.002
WorldPrediction	34.8	Mani. × causal, SFT + GRPO	+6.9 [+3.0, +10.8]	+4.9	0.003	0.009
WM-ABench	31.0	Mani. × cf, SFT	+4.6 [+3.4, +6.0]	+1.6	< 0.001	0.002
TempCompass	67.7	Mani. × causal, SFT	+2.2 [+1.3, +3.1]	+1.2	< 0.001	0.002
TemporalBench (long)	16.6	Mani. × cf, SFT + GRPO	+8.1 [+7.2, +9.1]	+1.2	< 0.001	0.002

Table 108: Best subset after GRPO: gain over the base model per GRPO seed (points). Every per-seed gain has $p < 0.001$ under the paired sign-flip test except WorldPrediction seeds 1 and 2 ($p = 0.011$ and $p = 0.001$).

Benchmark	Subset	Seed 1	Seed 2	Seed 3	Mean
SAT	Perc. × causal	+20.7	+24.0	+27.3	+24.0
ViewSpatial	Mani. × cf	+3.3	+3.1	+4.3	+3.6
ActionEQA	Mani. × causal	+11.6	+11.4	+9.5	+10.8
Robo2VLM	Mani. × cf	+8.1	+7.0	+6.8	+7.3
InPhyRe	Insp. × cf	+10.2	+9.8	+10.6	+10.2
CLEVRER	Insp. × cf	+5.1	+4.6	+4.2	+4.6
WorldPrediction	Mani. × causal	+5.5	+7.2	+8.1	+6.9
WM-ABench	Mani. × cf	+3.9	+4.5	+2.6	+3.7
TempCompass	Mani. × causal	+1.9	+2.4	+2.3	+2.2
TemporalBench (long)	Mani. × cf	+5.7	+7.0	+11.7	+8.1

N HUMAN AUDIT OF THE TEST SPLITS

The labelled answers of WOVEN's test questions are valid for at least 98% of the questions, 92.5% of the rollouts carry out the scripted action fully and 98.5% at least partly, and 98.7% of the perturbation edits follow their instruction without other changes. WOVEN's questions and answers are produced automatically from generated rollouts, so three things could be wrong without being visible in model accuracy: a labelled answer may not be the correct one, a rollout may not carry out the scripted action or may show an implausible or discontinuous process, and an edited option in the perturbation split may not follow its edit instruction or may change more than instructed. Two annotators, who are not authors, independently audited a stratified sample of the test splits in three parts through a web page with one item per screen, both seeing the same items in the same fixed, shuffled order; every choice is stored server-side in a hash-chained log per annotator. The sample holds 250 questions from 228 sources of the in-distribution and held-out-scene test splits: 200 main questions stratified by action type (exogenous 43, inspective 43, manipulative 43, perceptive 36, navigative 35; exogenous questions are physical-modeling questions, and for the other action types the questions cycle over every reasoning type the action type has) and 50 perturbation questions (inspective 25, manipulative 25; geometric and appearance edits split evenly), each with three edited options.

In Part 1 (250 questions), the annotator answers each question as a model does, without seeing the labelled answer, states a confidence (guessing, fairly sure, sure), and says whether another option could also be correct; answers can be changed until the page is confirmed, after which it is locked. In Part 2 (100 rollout videos, 20 per action type), the annotator sees the video the question was generated from, with its start frame, frames, and scripted action (for exogenous items, the described event), and judges whether the scripted action is carried out (yes, partly, no), whether the final state is a plausible result of it, whether the video is continuous (nothing appears, vanishes, jumps, or changes shape without cause), and, for exogenous items, whether the process follows the stated physical principle. In Part 3 (50 perturbation questions, 150 edits), the annotator sees the unedited image beside the three edited options with their edit instructions and judges whether all three edits were carried out as

instructed with nothing else changed, and if not, which edits have a problem. Parts 2 and 3 follow Part 1 because the videos and the unedited images reveal the answers.

Table 109: Human audit, Part 1: blind answering (accuracy %, 250 questions).

	Annotator 1	Annotator 2	Pooled [95% CI]
All questions	94.0	93.6	93.8 [91.3, 95.6]
Main questions (200)	94.0	93.5	93.8
Perturbation questions (50)	94.0	94.0	94.0
Exogenous (43)	95.3	93.0	94.2
Inspective (43)	95.3	93.0	94.2
Manipulative (43)	95.3	93.0	94.2
Perceptive (36)	91.7	91.7	91.7
Navigative (35)	91.4	97.1	94.3
Answered with "sure"	97.5 (204)	96.2 (212)	96.9

Table 110: Human audit, Part 2: rollout videos (% of 100 videos).

Question	Annotator 1	Annotator 2	Pooled [95% CI]	Agreement
Scripted action carried out (yes)	94	91	92.5 [88.0, 95.4]	92%
Scripted action carried out (yes or partly)	99	98	98.5 [95.7, 99.5]	—
Final state plausible	99	98	98.5 [95.7, 99.5]	97%
Video continuous	98	97	97.5 [94.3, 98.9]	97%
Follows the stated physical principle (exogenous, 20)	100	100	100 [91.2, 100]	100%

Table 111: Human audit, Part 3: edited images.

	Annotator 1	Annotator 2	Pooled [95% CI]
Edits carried out as instructed with nothing else changed	149/150	147/150	98.7% [96.6, 99.5]
Questions with all three edits correct	49/50	47/50	96.0%

In Part 1 the annotators answered blind with 93.8% accuracy (Table 109); they chose the same option on 92.0% of the questions (Cohen's $\kappa = 0.89$) and both answered correctly on 225 questions. On 5 questions (2.0% [0.9, 4.6]) both chose the same option different from the labelled answer, which bounds the rate of invalid labelled answers from above. In Part 2, 92.5% of the rollouts carry out the scripted action fully and 98.5% at least partly, 98.5% end in a plausible state, 97.5% are continuous, and all exogenous events follow the stated physical principle (Table 110); both annotators judged the scripted action as not carried out on 1 of the 100 videos. In Part 3, 98.7% of the edits follow their instruction without other changes and the annotators agreed on whether all three edits of a question were correct for 96% of the questions (Table 111), so the perturbation split tests what it is designed to test.

O PROSPECTIVE VALIDATION OF THE TRAINING RECIPE

Goal. The recipe in §5 is distilled from transfer results on the 26 benchmarks. The prospective test fixes training mixtures by the recipe before training and compares them with other mixtures of the same size on questions from external benchmarks that were not used to derive the recipe. Every comparison asks how a fixed training budget should be allocated: all mixtures start from the same base model and are trained with the same protocol, so the base model enters none of the comparisons; its scores are reported for reference.

Training. Qwen2.5-VL-3B-Instruct, single-stage SFT with LoRA (rank 16, $\alpha = 32$, dropout 0.05), learning rate 10^{-4}, 5% warmup, 3 epochs. The WOVEN training set is partitioned into the 11 controlled subsets of §4, one per combination of action type and reasoning family. Each model is trained on 1,500 WOVEN items drawn from these subsets, each item with its four option orderings, about 7,000 training examples in total. The five mixtures differ only in how the 1,500 items are allocated across the 11 subsets (Table 112). Each mixture is trained with 8 seeds, where the seed determines which items are drawn and the training order, 40 models in total.

Table 112: Prospective test: number of training items per subset in each mixture (of 1,500).

Subset	Recipe	Uniform	Temporal-heavy	Camera-heavy	Exogenous-heavy
Perceptive × causal dynamics	197	137	45	245	45
Inspective × causal dynamics	197	137	45	245	45
Navigative × causal dynamics	196	137	45	46	45
Manipulative × causal dynamics	196	137	45	46	45
Inspective × counterfactual	226	136	44	245	45
Manipulative × counterfactual	226	136	44	46	45
Exogenous × physical modeling	54	136	44	46	548
Perceptive × temporal	52	136	297	245	45
Inspective × temporal	52	136	297	245	45
Navigative × temporal	52	136	297	46	45
Exogenous × temporal	52	136	297	45	547

Mixtures. The target questions of the main comparisons (Set A below) ask for the outcome of an agent's action or for the action behind an observed change, that is, they call for causal and counterfactual reasoning about agent-driven transitions. By item (1) of the recipe, the Recipe mixture allocates 82% of the budget to the six agent-action subsets of the causal and counterfactual families; by item (3), temporal and passive-physical supervision does not transfer to such questions, so the remaining five subsets keep 52–54 items each. Each comparison mixture represents another way of selecting supervision: Camera-heavy selects by matching the action (the camera-motion and object-inspection subsets, across reasoning families); Temporal-heavy selects the wrong reasoning operation (the temporal subsets); Uniform makes no selection (equal shares over the 11 subsets); Exogenous-heavy is dominated by passive physical events.

Evaluation questions. All questions are drawn from external benchmarks, MMSI-Bench (Yang et al., 2025b), VLM4D (VLM4D Team, 2025), PhysBench (Chow et al., 2025), SPAR-Bench (Zhang et al., 2025b), MVBench (MVBench Team, 2023), IntPhys 2 (IntPhys 2 Team, 2025), and CameraBench (Lin et al., 2025), none of which is among the 26 benchmarks used to derive the recipe. Decoding is greedy (vLLM), and the model is asked to output only the option letter. The questions are grouped by what each comparison requires of them. Recipe versus Temporal-heavy, Uniform, and Exogenous-heavy require only that the questions call for causal or counterfactual reasoning about an agent's action (Set A). The other direction of item (3) requires questions about passive physical events (Set B). Recipe versus Camera-heavy asks whether matching the reasoning operation or matching the action matters more, so it is run on questions whose action Camera-heavy matches, with the action fixed to camera motion (Set C). Item (4) requires static questions that involve no change (Set D).

*Set A, agent-action questions (*1,673*).* All require the outcome of an action or the motion behind a change; the action type is not restricted (Table 113). The score is the accuracy on each subset, averaged with equal weight over the five subsets. The PhysBench questions come from a sample of 1,001 PhysBench questions stratified by category.

*Set B, passive-physics questions (*1,174*).* Every subset shows objects that move, collide, or persist without any agent acting on them (Table 114). MVBench items use 8 frames sampled uniformly from each video; IntPhys 2 items ask "Is what happens in this video physically possible?" with the answers Yes / No. The score is the accuracy over all questions.

*Set C, camera-motion questions (*7,592*).* In every subset the action is the camera's own motion, which corresponds to WOVEN's perceptive actions, and the reasoning is to infer how the camera moved from the observations (inverse dynamics; Table 115). The 74 MMSI-Bench questions also belong to Set A. For SPAR view change, the original questions ask for the camera's translation and rotation as numbers and are converted to four-option questions: the correct option is the motion the camera actually made (a turn of at least 30°, or, with a turn of at most 10°, a translation of at least 0.6 m in a single direction), and the three wrong options are motions that clearly did not happen, including the opposite of the correct one. The CameraBench questions are the yes/no questions of the CameraBench VQA release, grouped by the camera motion they ask about, with 8 frames sampled uniformly from each video. The score is the accuracy over all questions.

Table 113: Prospective test: Set A.

Source	Subset	Questions	Example
MMSI-Bench	Motion (Cam.)	74	"The images are taken continuously from a first-person perspective. In which direction are you moving?"
VLM4D (real videos)	object motion direction	1,111	"Which way does the person turn?"
VLM4D (real videos)	motion direction from a stated viewpoint	260	"From the camera perspective, which direction are the cyclists moving toward?"
PhysBench	relationships · motion	145	"Moving in the direction indicated by the arrow in the picture, which of the following options are you most likely to see?"
PhysBench	dynamics · manipulation	83	"Which of the following options presents the correct sequence to use the spoon to pick the objects in the green bowl and put them in the green pot?"

Table 114: Prospective test: Set B.

Source	Subset	Questions	What is asked
MVBench	counterfactual inference	200	which collision happens if an object is removed
MVBench	moving attribute	200	attributes of the objects that move or stay still
MVBench	moving count	200	how many objects of a kind are moving
MVBench	object existence	198	whether moving objects are still present at the end
IntPhys 2	solidity, fixed camera	104	whether the event is physically possible
IntPhys 2	immutability, moving camera	136	whether the event is physically possible
IntPhys 2	continuity, moving camera	136	whether the event is physically possible

Table 115: Prospective test: Set C.

Source	Subset	Questions	What is asked
MMSI-Bench	Motion (Cam.)	74	how the camera moved or rotated across consecutive images
VLM4D	camera motion	6	which way the camera turned or moved in the video
SPAR-Bench	view change, ScanNet / ScanNet++	125 / 136	which motion the camera made between two images
CameraBench	pan	1,200	whether the camera pans left or right
CameraBench	tilt	703	whether the camera tilts up or down
CameraBench	roll	524	whether the camera rolls clockwise or counterclockwise
CameraBench	dolly	753	whether the camera moves forward or backward
CameraBench	truck	863	whether the camera moves laterally left or right
CameraBench	pedestal	220	whether the camera moves up or down
CameraBench	static	628	whether the camera stays still
CameraBench	tracking	1,344	whether the camera follows a moving subject
CameraBench	stability	828	whether the camera motion is smooth or shaky
CameraBench	other camera motion	188	pedestal-only and single-direction movements

Set D, static questions (886). Questions about a single state with no change, from the same benchmarks as Set A: PhysBench property (attribute, color, mass, number) and relationships (depth, distance, location, size), 327 questions; MMSI-Bench Attribute (Appr.), Attribute (Meas.), and the five positional-relationship categories, 559 questions. The two benchmarks are scored separately.

Statistics. Each of the 8 models of a mixture gives one score. Two mixtures are compared by an exact two-sided permutation test over all 12,870 splits of their 16 model scores; $p < 0.05$ is significant. The 95% confidence interval of the difference in means uses Welch's t. One comparison is made per recipe item and direction; no family-wise correction is applied.

Results. Table 116 reports the five comparisons; in every comparison both mixtures score above the untrained base model. Selecting the wrong reasoning operation costs 3.5 points (Temporal-heavy) and selecting passive physical events costs 4.7 points (Exogenous-heavy); making no selection costs 1.5 points (Uniform). On camera-motion questions, Camera-heavy holds more exact-match supervision than Recipe (735 versus 620 items in the camera-motion and object-inspection subsets of the causal and counterfactual families) and still trails by 0.6 points ($p = 0.04$); the two mixtures differ in the remaining budget, which Camera-heavy spends on same-action temporal subsets (490 items) and Recipe on same-operation navigation and manipulation subsets (618 items). On passive-physics questions, Exogenous-heavy scores 0.5 points above Recipe ($p = 0.04$), consistent with the direction of item (3). In the other direction, Exogenous-heavy gains least on Set A (Table 117), consistent with §5.3: passive-physical supervision contributes least to agent-action questions. Its small gain is attributable to the 405 items (27%) it draws from agent-action subsets, whereas §5.3 uses the passive-physical subset alone.

Table 116: Prospective test: results (accuracy %, mean over 8 models; Δ = Recipe − comparison mixture; Item = recipe item tested; Base = untrained model on the same questions).

Comparison mixture	Item	Questions	Base	Recipe	Comp.	Δ	95% CI	p
Temporal-heavy	(1)	Set A	31.1	37.9	34.4	+3.48	[+2.03, +4.92]	0.0002
Uniform	(1)	Set A	31.1	37.9	36.4	+1.49	[+0.18, +2.81]	0.03
Camera-heavy	(1)	Set C	54.1	58.4	57.8	+0.58	[+0.02, +1.14]	0.04
Exogenous-heavy	(3)	Set A	31.1	37.9	33.2	+4.68	[+3.19, +6.17]	0.0002
Exogenous-heavy	(3)	Set B	70.3	71.0	71.5	−0.53	[−1.05, −0.02]	0.04

Item (4). This item concerns whether training helps at all on tasks that involve no transition, so the comparison is with the untrained base model (Table 117). On static questions the five mixtures differ from the base model by -0.2 to $+1.2$ points, and Recipe does not differ significantly from any other mixture ($p \geq 0.37$), whereas the same models gain 2.1 to 6.8 points over the base model on Set A. The PhysBench static questions are the main evidence; the MMSI-Bench static questions are close to chance for the base model and are reported for completeness.

Table 117: Prospective test: static questions versus agent-action questions (accuracy %, mean over 8 models).

Questions	Base	Recipe	Uniform	Temporal-heavy	Camera-heavy	Exogenous-heavy
Set D, PhysBench static (327)	55.7	56.1	55.4	56.5	56.5	56.5
Set D, MMSI-Bench static (559)	28.4	29.0	28.6	29.0	29.0	29.6
Set A (1,673)	31.1	37.9	36.4	34.4	36.3	33.2

Item (2) (supervision with larger state changes gives more robustness) is outside this test: the external benchmarks have no perturbation questions graded by the extent of the state change.

P OFF-THE-SHELF VIDEO PRIORS ON WOVEN

Before training, we ask whether engines pretrained at scale on video reconstruction and future-frame prediction already solve WOVEN, and whether the transition knowledge they hold is directly queryable. We evaluate three frozen off-the-shelf engines on `test_id` under one shared read-out recipe: the engine scores every option of the 4-way MCQ and the argmax is taken (chance = 25%).

Table 118: Prospective test: per-seed scores (accuracy %). Base model: Set A 31.1, Set B 70.3, Set C 54.1.

Questions	Mixture	s1	s2	s3	s4	s5	s6	s7	s8	Mean
Set A	Recipe	38.7	36.8	38.3	37.0	38.4	38.3	36.3	39.6	37.9
Set A	Uniform	38.4	36.9	35.2	34.5	38.0	36.1	36.2	36.1	36.4
Set A	Temporal-heavy	35.8	35.1	33.0	35.8	35.6	35.3	32.5	32.4	34.4
Set A	Exogenous-heavy	30.6	33.5	34.5	32.2	35.2	33.9	34.4	31.6	33.2
Set B	Recipe	70.3	71.1	71.5	70.7	70.6	71.0	71.3	71.1	71.0
Set B	Exogenous-heavy	71.4	71.2	71.2	70.7	71.8	72.6	71.5	71.6	71.5
Set C	Recipe	58.2	58.3	58.3	58.8	58.5	58.8	58.2	57.7	58.4
Set C	Camera-heavy	57.5	56.8	57.4	58.8	58.0	57.4	58.0	58.3	57.8

Video encoder. V-JEPA2 (ViT-L) (Assran et al., 2025). Non-action reasoning types use its native energy read-out zero-shot; action-conditioned types are read out by a probe trained on the frozen representation at two capacities, *linear* and *attentive*, with an open-vocabulary action space reached through a CLIP text bridge.

Video generator. I2VGen-XL (Zhang et al., 2023) (2.84B denoiser). Options are scored by conditional denoising likelihood over a fixed grid of diffusion steps with shared noise. Its forward-only rollouts cover 7 of the 8 reasoning types (counterfactual substitution has no forward likelihood read-out).

Image generator. OmniGen (Xiao et al., 2024) (3.8B, matched to the video generator's denoiser scale), an image-editing generative model conditioned on the initial frame and the action instruction, scored with the identical likelihood recipe. This engine tests whether an image prior could stand in for a video prior as the WOVEN rollout generator.

Table 119: **Off-the-shelf video priors on WOVEN** (`test_id`, 4-way MCQ accuracy in %, chance = 25). Columns are reasoning types (fwd: forward dynamics; inv: inverse dynamics; cf-rem/cf-sub: counterfactual removal/substitution; out/cued: outcome/cued prediction; t-ord/t-adj: temporal ordering/adjacency). For the encoder, the event cue of cued prediction is text and outside its input space, so outcome and cued prediction share one read-out. Overall for the generator engines spans the seven reasoning types their forward-only read-out covers.

Model	fwd	inv	cf-rem	cf-sub	out	cued	t-ord	t-adj	overall
V-JEPA2 linear	51.3	51.9	**66.5**	20.0	**29.2**	**29.2**	37.5	16.9	39.3
V-JEPA2 attentive	**70.7**	**75.4**	57.7	**35.9**	**29.2**	**29.2**	37.5	16.9	**47.7**
Video generator	19.8	27.7	37.6	—	13.2	13.9	**53.1**	**56.2**	36.4
Image generator	20.0	22.5	24.5	—	13.5	12.5	22.0	21.5	19.5

WOVEN is not trivially solvable from off-the-shelf priors. The strongest read-out reaches 47.7% overall against the 92.3% human baseline, and each engine's profile is partial: the encoder is strong on action-conditioned types (forward 70.7, inverse 75.4) but weak on temporal ones (37.5/16.9), while the video generator shows the complementary profile (temporal 53.1/56.2, forward 19.8).

The transition prior lives in the video prior. At matched denoiser scale and under the identical likelihood recipe, the image-editing generator scores 19.5% overall with no reasoning type above chance. An image prior cannot stand in for a video prior as the rollout generator, which supports building WOVEN on VGM rollouts.

The knowledge is present but not directly accessible. On identical frozen V-JEPA2 features, the attentive probe recovers +19 to +34pp over the linear probe on action-conditioned types, and a random-initialization control (same probe, same training data, randomly initialized encoder) trails the pretrained encoder by 27.6pp on average, positive on all 12 action cells: the gain is attributable to the pretrained representation, and the transition knowledge it holds is recoverable but not linearly accessible. This is the premise behind §4: video-pretrained engines hold the (s, a, s') prior (which is what makes them strong rollout generators for WOVEN) but do not expose it to direct querying, and post-training on (s, a, s') supervision is what installs the access.

Q LIMITATIONS

WOVEN deliberately scopes to world modeling, single-step, vision-grounded, action-conditioned reasoning over (s_t, a, s_{t+1}) triples, and therefore does not directly evaluate the broader world-model lineages: long-horizon procedural planning over abstract action sequences (WorldPrediction Team, 2025), symbolic-environment dynamics inference (Warrier et al., 2025; Chen et al., 2025b), and purely latent-space predictive representations evaluated by downstream control rather than per-cell attribution (Hafner et al., 2023; Assran et al., 2025). These programmes operate over distinct state spaces and admit different yardsticks; extending the cell-level statistical-attribution methodology to them is a natural follow-up but outside the present scope.