Synthetic Physics as Supervision: Learning Real-World Physical Reasoning in Vision-Language Models

Swastik R [1] **Natesha B V** [1]

Abstract

Vision-language models (VLMs) remain unreliable on visually grounded physics reasoning in real-world media, and human-labelled physics supervision is expensive. We study whether *synthetic* simulator scenes alone can supply an effective training signal. We fine-tune Qwen3-VL-30B on $14{,}597$ rigid-body and fluid scenes with free-text answers read directly from simulator state, with no human annotation and no local GPU required. On PhysBench Test (n=9,786), this synthetic-only supervised fine-tuning (SFT) lifts accuracy from 40.7% to 47.6% (a $+6.9$ percentage-point, or $+6.9$pp, gain), improves $27/39$ subtasks, and yields a $+20.1$pp gain on the Scene domain. A follow-up Group Relative Policy Optimization (GRPO) stage with simulator-verifiable rewards preserves this aggregate gain ($+7.0$pp over baseline) while adding targeted improvements on the un-trained `general:relationships` domain, and a data-fidelity probe at both stages shows the gains track the physical quality of the synthetic signal. On a vision-essential ScienceQA-Physics probe, prior textbook-physics ability shows *no statistically significant degradation* at this sample size (directional -4.5pp gap, p=0.09; the probe is underpowered and does not rule out a small drop). The full pipeline costs under $30 in hosted-API credits, providing evidence that synthetic physics is a practical and reproducible supervision signal for grounded visual reasoning in VLMs.

[1]Department of Computer Science and Engineering, Indian Institute of Information Technology Raichur, Raichur, Karnataka, India. Correspondence to: Swastik R <cs23b1075@iiitr.ac.in>, Natesha B V <nateshbv@iiitr.ac.in>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Modern vision-language models (VLMs) achieve strong results on broad multimodal benchmarks but remain unreliable on *physics-grounded* visual reasoning such as predicting whether a stack will topple, which of two fluids is more viscous, or where a rolling ball will end up. Closing this gap with human-labelled physics data is expensive: annotation requires domain expertise and scales poorly. Existing physics-VLM benchmarks such as PhysBench (Chow et al., 2025), SeePhys (Xiang et al., 2025), and PhysReason (Zhang et al., 2025) also report a long tail of subtasks where even frontier models leave substantial headroom. This bottleneck motivates a methodological question: can a VLM learn physics from a sandbox? We fine-tune on synthetic scenes of coloured boxes, spheres, and cylinders interacting in a simulator, then test on PhysBench real-world photos and videos. The source data is deliberately stripped down (geometric objects, minimal visual detail, no people, no clutter), while the target data is naturalistic and visually complex. If PhysBench accuracy improves, that is consistent with learning transferable physical regularities rather than matching synthetic appearance.

Concretely, we ask three connected questions. First, can a VLM acquire grounded physical reasoning purely from synthetic simulator scenes, with answers derived from simulator state and no human-annotated physics labels? Second, do such gains transfer from a deliberately toy training distribution to in-the-wild real-world benchmark imagery? Third, can synthetic supervision deliver this capability *without* narrowing the model into a domain specialist, i.e. does the resulting VLM retain its prior, non-physics abilities? We instantiate this as a low-cost, reproducible recipe and ask: what is the cheapest training signal that materially improves real-world visual physics reasoning in a strong VLM? On PhysBench (Chow et al., 2025), our synthetic-only supervised fine-tuning (SFT) of Qwen3-VL-30B (Team, 2025) improves accuracy by $+6.9$ percentage-point (pp) with broad subtask coverage; ScienceQA-Physics shows no statistically significant degradation of prior textbook ability (directional drop, underpowered probe; §4.4). The full method (data, stages, rewards) and per-domain, per-mode results follow in Sections 3–5.

Contributions.

- A reproducible, fully synthetic training pipeline (15,797 scenes used in training: 14,597 for SFT and 1,200 for GRPO, drawn from a 16,617-scene candidate pool after quality filtering, 8 task families; MuJoCo + PhiFlow) that uses no human annotation, no local GPU, and under \$30 of hosted-API credits for all training stages combined (\$4 to \$5 per single SFT epoch).
- Cross-distribution and cross-format evidence: free-text training on synthetic MuJoCo and PhiFlow scenes improves PhysBench by +6.9pp, and the gain pattern follows the task families seen during training. The generated dataset targets a selected subset of physics properties with direct and indirect links to PhysBench, not all PhysBench classes.
- A detailed analysis that separates three effects in the lift: format commitment, latent-skill recovery, and physics capability gains, showing that most of the net improvement is explained by commitment/recovery, with smaller but measurable new-capability gains, and checking (§5.4) whether the observed gain primarily reflects physics competence, not answer-format effects.
- An item-level comparison of the verifiable-reward (GRPO R3) stage against the preceding SFT stage, showing that GRPO does not significantly improve overall accuracy (+0.08pp, within the Wilson confidence interval (CI)) but redistributes errors across mode and domain, with concentrated gains in `general:relationships` and losses in `image&video:property`; we include the full shift pattern in the paper.
- A data-fidelity probe at both the SFT and RL stages showing that physically accurate fluid supervision improves PhysBench, while a deliberately low-fidelity variant causes a drop (qualitatively at SFT and by −1.2pp/−1.3pp at RL), providing evidence that transfer is driven by the physical quality of the synthetic training signal rather than by surface artefacts of the data.
- A capability-preservation check on ScienceQA-Physics (n=425, vision-essential K-12 subset) showing that synthetic visual-physics SFT does *not* significantly degrade prior textbook-physics ability ($z = -1.69$, p=0.09, Wilson 95% CIs overlap), indicating the model remains a generalist VLM rather than a narrow domain specialist after our training pipeline.

2. Related Work

Physics benchmarks for VLMs. We evaluate on two complementary benchmarks. **PhysBench** (Chow et al., 2025) provides 9,786 test items across four physics domains (Property, Relationships, Scene, Dynamics) and three modes (image-only, image&video, text-only "general") and is in-domain to our evaluation target while remaining visually out-of-domain from our synthetic training distribution. **ScienceQA-Physics** (Lu et al., 2022) provides K-12 textbook-style physics QA and is used as a capability-preservation probe (§4.4).

Synthetic data, intuitive physics, and simulator-derived supervision. Synthetic, simulator-rendered scenes have a long history as probes for visual physics reasoning: IntPhys (Riochet et al., 2018) and Physion (Bear et al., 2021) measure intuitive-physics judgments such as object permanence, support, and dynamics prediction on controlled rigid-body clips. CLEVRER (Yi et al., 2020) introduces collision-event reasoning from physics-engine rollouts; ContPhy (Zheng et al., 2024) extends this to continuum bodies (cloth, fluid) for video QA; PhysGen (Liu et al., 2024a) uses rigid-body physics to ground image-to-video generation. A parallel line evaluates whether generated video itself respects physical laws (VideoPhy (Bansal et al., 2024), PhyWorld (Kang et al., 2024)). Closest to our setting, Physics Context Builders (PCBs) (Balazadeh et al., 2024) fine-tunes a VLM on simulator-generated QA pairs and measures a stability-prediction probe. We differ in using free-text `<answer>` labels read directly from simulator state, in covering both rigid-body and continuum-fluid scenes, and in measuring transfer on the broader PhysBench benchmark across 39 subtasks; head-to-head comparison against PCBs and other physics-aware fine-tunings (e.g. physically grounded VLMs for robotic manipulation (Gao et al., 2024)) on the same evaluation harness is an important next step. The broader idea of using physical laws and constraints as a label-free training signal (Stewart & Ermon, 2017) extends here to instruction-tuned VLMs, where both rendered frames and the numeric simulator state are consumed as ground truth, removing caption or QA-pair labeling.

Reinforcement learning with verifiable rewards. Recent work on reasoning-LLMs replaces preference-labeled rewards with programmatically checkable ones (e.g. unit tests, exact-match math answers) (DeepSeek-AI, 2025; Shao et al., 2024), removing the dependence on human raters. We adopt GRPO (Shao et al., 2024), a critic-free Proximal Policy Optimization (PPO) (Schulman et al., 2017) variant that estimates per-token advantages from a group of on-policy roll-outs, because it is sample-efficient and removes the value-network cost. Concurrent work on REINFORCE-style reinforcement learning from human feedback (RLHF) (Ahmadian et al., 2024) reaches similar conclusions about the cost asymmetry of value-network methods. Our reward is the simulator state itself, so the verifiable-reward criterion is satisfied without human prefer-

ence labels or learned reward models.

Parameter efficient fine-tuning of large VLMs. End-to-end fine-tuning of large multi-modal models is typically expensive in GPU-hours and memory. We use Low-Rank Adaptation (LoRA) (Hu et al., 2022) on Qwen3-VL-30B with rank-16 adapters for SFT and rank-64 adapters for GRPO (~ 0.4–1.7% of backbone parameters), trained on the public Tinker LoRA service in under 12 wall-clock hours total across both SFT rounds and the GRPO run.

3. Method

3.1. Synthetic Data Generation: Toy Primitives Only

The training distribution is deliberately minimalist: every training scene contains 2-5 geometric primitives (boxes, spheres, cylinders, capsules) on a uniform background, with random masses, sizes, colours, initial velocities and lighting. There are no real-world textures, no humans, no animals, no household objects, and no scene clutter. The test distribution (PhysBench) is the opposite, in-the-wild photos and videos of real-world physical events, people, fluids in containers, mechanical apparatus, laboratory setups. The cross-distribution gap is large by construction, whether the model still transfers is what we test.

We generate a candidate pool of 16,617 scenes using two simulators: 12,023 rigid-body scenes from MuJoCo (Todorov et al., 2012) (time-to-collision (TTC), projectile trajectory, pile stability, manipulation, throwing, collision) and 4,594 continuum scenes from PhiFlow (Holl et al., 2020) (with corrections to viscosity, direction, and fluid-level estimators). After quality filtering and de-duplication, 14,597 scenes are used in training: all 12,023 MuJoCo scenes (SFT R1) plus a 2,574-scene PhiFlow + categorical-comparison subset (SFT R2-redo). The remaining PhiFlow scenes were excluded because they did not pass the Gaussian-density quality check or duplicated existing direction or viscosity prompts. PhiFlow simulates fluid as discrete particles; we deposit particle mass onto a 256×256 density grid and apply a SciPy Gaussian filter ($\sigma = 7$ px) so the result reads as a smooth, connected fluid body. Each scene yields (i) a 1 to 3s RGB rollout sampled at 8 frames, (ii) a single "cover" image, and (iii) all simulator state variables that constitute the physical ground truth. From the same simulator state we automatically construct five categorical comparison families (mass/size, viscosity, level, direction, count/viewpoint), which together with the two continuous MuJoCo families (TTC, trajectory) and one stability family yield *eight* training-task families overall.

For each scene we emit a teacher-forced answer in **free-text** form, wrapped in `<answer>...</answer>` tags. Discrete tasks use a short free-text categorical token (e.g. `<answer>left</answer>`, `<answer>fluid b</answer>`, `<answer>stable</answer>`); continuous tasks use a numeric scalar (e.g. time-to-collision in seconds, viscosity rank on $[0, 1]$). No training answer is in A/B/C/D format; the model is never shown the multiple-choice (MCQ) letter convention used by PhysBench evaluation. The supervision is read directly from simulator state, with no human in the loop.

Table 1. Representative (prompt, ground-truth) pairs from selected training-task families (we use eight families overall: two continuous MuJoCo plus six categorical). Prompts are emitted with the `<reasoning>...</reasoning><answer>...</answer>` schema shown; ground-truth answers and rationales are derived directly from simulator state.

TTC (MuJoCo, continuous): *Prompt*: "Two objects are moving in this 0.6-second video. Object A: purple sphere. Object B: white sphere. When will they collide? Give seconds from the start of the video." *GT*: `<answer>0.78</answer>` (closing speed gives 0.22s remaining after clip ends).
Trajectory (MuJoCo, continuous): *Prompt*: "Watch this projectile's initial flight (0.3s shown). Predict the final resting position." *GT*: `<answer>x=1.61, y=0.02</answer>` (parabolic extrapolation from 4-frame clip).
Stability (MuJoCo, categorical): *Prompt*: "Analyze this stack: base orange box (0.26 m), layer 2 blue sphere (0.18 m). Will it remain stable or topple?" *GT*: `<answer>stable</answer>` (max observed displacement 0.0017 m).
Fluid direction (PhiFlow, categorical): *Prompt*: "A body of fluid (blue) is released. In which primary direction does the fluid move? Options: left, right, down." *GT*: `<answer>left</answer>` (centre-of-mass displacement $\Delta x{=}-0.54$, $\Delta y{=}-0.34$ m).
Fluid viscosity (PhiFlow, categorical): *Prompt*: "Fluid A (blue, left) and Fluid B (orange, right) in the same container. Which is more viscous?" *GT*: `<answer>fluid b</answer>` (B remains compact across the sequence; A spreads into curved streams).

Table 1 shows representative (prompt, ground-truth) pairs from each of the eight training-task families.

No data leakage. The training distribution is disjoint from any PhysBench, ScienceQA, or other public physics-benchmark item. The scenes are sampled de-novo from MuJoCo and PhiFlow with randomized geometry, masses, viscosities, lighting, and camera poses. No PhysBench image, video, or question text appears in our training corpus, and the rendering style (single-camera MuJoCo renderer at 640×480; SciPy-smoothed PhiFlow density at 256×256) is visually distinct from the heterogeneous in-the-wild PhysBench imagery (App. A.6, Fig. 10). Additional training scenes covering the categorical-comparison families used in

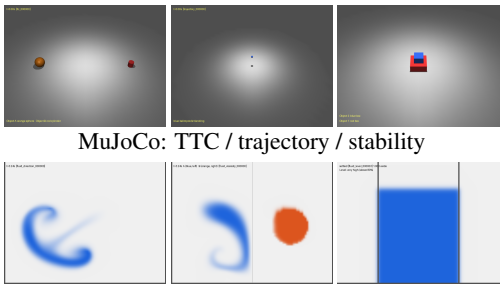


Figure 1. Sample training scenes (one per family). MuJoCo rigid-body scenes (top) and PhiFlow continuum-fluid scenes (bottom). Full per-task grids in App. A.6; PhysBench test scenes in App. A.6, Fig. 10.

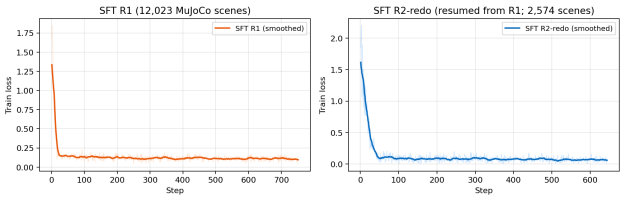


Figure 2. SFT training loss for both rounds. **SFT R1** (orange, 751 steps over 12,023 MuJoCo scenes) converges in one epoch. **SFT R2-redo** (blue, 644 steps over 2,574 scenes) resumes from the R1 weights and is trained on the corrected PhiFlow continuum data plus the eight categorical comparison tasks. Loss values are not directly comparable across rounds because the two corpora differ in token length and label format.

SFT R2-redo are shown in App. A.6, Fig. 9. Generalization across this train-test domain gap is therefore part of what we measure.

Figure 1 previews one scene per family; full per-task grids and PhysBench test scenes are in Appendix A.6 (Figs. 7–10).

3.2. Supervised Fine-Tuning

We fine-tune Qwen3-VL-30B-A3B-Instruct with LoRA (rank 16, all linear projections) using teacher-forced answers. All training is performed on the public Tinker LoRA service (Thinking Machines, 2025), which exposes a managed sampling and gradient endpoint and removes the need for local GPU infrastructure. We run two SFT rounds in a sequential curriculum (the second resumes from the first):

- **SFT R1**: 12,023 MuJoCo scenes, 1 full epoch (751 steps, $\sim$ 3.7 wall hours). Covers TTC, trajectory, and stability tasks.
- **SFT R2 (low-fidelity fluid, data-fidelity probe)**: resumed from R1 on a low-fidelity PhiFlow corpus in which $\sim$29% of fluid scenes lacked the Gaussian density smoothing described in §3.1 and carried noisy viscosity/direction/level targets. We use this run as an intentional data-fidelity probe: if the model is genuinely learning physics, training on physically less accurate supervision should not improve PhysBench, whereas training on physically accurate supervision should. SFT R2 indeed shows a partial-eval drop/no-improvement on the Dynamics domain, motivating a regeneration of the fluid corpus with corrected physics.
- **SFT R2-redo (corrected physics)**: same recipe as SFT R2 but on the corrected, Gaussian-smoothed fluid data, comprising 2,574 scenes, 644 steps, $\sim$ 3.9 wall hours, again resuming from R1 weights. With physically accurate supervision the model improves on the targeted domains (§4, Tables 3 and 4), completing the data-fidelity contrast and providing evidence that the gain tracks the physical quality of the training signal.

Why a single epoch per round. The SFT R1 training loss bottoms at 0.076 near step $400/751$ and drifts up to 0.136 by end-of-epoch (Fig. 2), already showing mild overfitting on the `<answer>` schema; second- epoch smoke runs confirm rank-16 adapters memorise exact numeric answers on TTC/trajectory. The same effect reappears at RL: an exploratory GRPO Run 2 epoch 2 (475 extra steps) regressed val $65.8\% \rightarrow 64.3\%$ despite training reward 0.84 on the small (n=951) GRPO pool. We therefore keep one epoch per stage.

3.3. Reinforcement Learning with Verifiable Rewards (GRPO)

SFT only sees the final answer tokens, so it pushes the model toward the right string but not toward better intermediate reasoning. We add a third stage using GRPO (Shao et al., 2024) with simulator-verifiable rewards to test whether RL can close that gap. GRPO is critic-free, replacing the value network with a within-prompt baseline. For each prompt we draw G on-policy completions, score each with a scalar reward r_i, and assign every token of completion i the group-normalised advantage

$$\hat{A}_{i,t} = \frac{r_i - \mathrm{mean}(\{r_j\}_{j=1}^{G})}{\mathrm{std}(\{r_j\}_{j=1}^{G})}$$

which is then plugged into the PPO (Schulman et al., 2017) clipped surrogate objective with a Kullback–Leibler (KL) penalty to the SFT R2-redo reference policy.

Reward design. The reward is a weighted sum of two simulator-verifiable terms: a *physics correctness* reward (w_{phys}=0.95) that scores the parsed answer against simulator state (exact-match for MCQ, exponential-decay for continuous targets such as TTC and trajectory), and a small *format* reward (w_{fmt}=0.05)

Table 2. Item-level swap from SFT R2-redo to GRPO R3 (n=9,786 PhysBench Test). "Gained" = R2-redo wrong, GRPO R3 correct; "Lost" = R2-redo correct, GRPO R3 wrong. Rows ordered by |net|.

Mode	Domain	Gained	Lost	Net
general	relationships	276	213	+63
image&video	property	60	129	−69
image&video	dynamics	150	150	0
image&video	scene	90	104	−14
image-only	property	90	65	+25
image-only	dynamics	42	42	0
other (4 cells)		53	50	+3
Total		**761**	**753**	**+8**

that requires the `<reasoning>...</reasoning>` `<answer>...</answer>` schema. No human preference labels are used.

Run 2 (low-fidelity fluid, RL data-fidelity probe). An initial GRPO run resumed from R2-redo on the same low-fidelity PhiFlow corpus used in the SFT R2 probe, providing a parallel data-fidelity contrast at the RL stage. Consistent with the SFT R2 observation, training with physically less accurate supervision yields a -1.2pp movement relative to R1 (46.3% vs. 47.5%) and -1.3pp relative to R2-redo (47.6%), with 481 (4.9%) of completions emitting non-letter free text that the official extractor scored as empty. Recovering these via a free-text-to-option matcher (§5.4) restores most of the gap, so the move is largely a format-collapse artefact. We report it as RL-stage evidence that training-signal quality matters and as a cautionary note on reward shaping under formatting pressure.

Run 3 (corrected, GRPO R3). The reported GRPO model resumes from R2-redo on a corrected curriculum of 1,200 scenes that re-balances the corpus: 300 static-image scenes (preventing image-only regression) plus 900 video scenes covering the directly targeted video task types (TTC, trajectory, stability, fluid, manipulation, throwing). Categorical prompts keep the same free-text `<answer>` schema used in SFT (no MCQ re-wrapping). We use G=4, two groups per batch, temperature 0.7, max tokens 768, KL penalty 0.04, 300 optimisation steps, and rank-64 LoRA adapters (see App. A.4, Figs. 5 and 6 for the full schedule, training-reward trajectory, and per-task data coverage).

What changes after GRPO? GRPO R3 does not significantly improve overall accuracy over SFT R2-redo (47.7% vs. 47.6%, +0.08pp). The matched-item McNemar test confirms this: $\chi^2 = (761 - 753)^2/(761 + 753) \approx 0.042$ ($p \approx 0.84$, 1 dof), well within Wilson 95% CI overlap. At the per-item level, GRPO R3 newly rescues 761 R2-wrong items and newly loses 753 R2-correct items: a near-even swap (net +8 items out of 9,786, +0.08%). The swap is *not* uniformly distributed (Table 2); it concentrates in a few mode–domain cells.

The pattern: GRPO rescues many `general:relationships` items (text-heavy multi-image-option spatial-motion questions R2-redo confuses), at the price of regressing on `image&video:property` (real-world object-attribute/count questions). For example, GRPO *rescues* idx 5099 ("If we want the brown pulley to rotate clockwise, what can we do?"), R2-redo answers "Increase the mass of the brown cube" (×), GRPO produces an explicit force/torque analysis and answers **D** (✓). It *loses* idx 4883 (fluid pass-through), R2-redo correct, GRPO becomes more verbose but commits to a worse spatial inference. Net: GRPO on 1,200 scenes redistributes where the model commits without consistently sharpening physics, so we report the per-domain swap in Table 2 as the main GRPO finding rather than an accuracy lift.

What GRPO sharpens, and where it saturates. Three factors explain why the aggregate move beyond SFT R2-redo is small even though the per-domain swap is informative. Coverage was thin: Run 1's variance-conditioned sampling concentrated rollouts on $\sim 195/15{,}714$ scenes (1.2% of the early GRPO scene pool, which is distinct from the broader SFT candidate pool); Run 2 fixed the concentration (Fig. 6) but only on 951 balanced scenes, and Run 3 trains on 1,200 ($\leq$ 30 rollouts per (mode, task) cell). On top of that, an exploratory Run 2 Epoch 2 (475 extra steps) reached training reward 0.84 while *regressing* val from 65.8 to 64.3%, a clear case of overfitting scene-specific shortcuts on the small fixed pool. And SFT had already absorbed the bulk of the transferable signal, so GRPO can only sharpen items where the SFT policy was already indecisive. Practitioners with fixed compute should likely stop at SFT R2-redo unless the GRPO scene pool can be expanded substantially.

4. Experiments

4.1. Targeted Categories and Mechanism of Impact

Our eight training-task families were designed to load on three of PhysBench's four physics domains. *Direct* targets (training labels mirror the test subtask) are: TTC, trajectory, fluid direction/viscosity/level map to **Dynamics**; stability, mass/size compare, fluid viscosity/level, counting (cardinality of objects) map to **Property** (PhysBench groups counting under `property:number`); viewpoint maps to **Scene**. *Indirect* targets (shared visual primitives or reasoning style, no label-level overlap) include stability mapping to Property:attribute and fluid direction mapping to Relations:motion. The text-only general mode and the Relationships domain were not directly targeted by label-level

Table 3. PhysBench Test accuracy across pipeline stages (n=9,786, T=0, max_tokens 512). Δ is vs. Baseline. Inter-trained-checkpoint gaps (R1 vs. R2-redo vs. GRPO R3) lie within overlapping Wilson 95% CIs and should not be read as significant; only gaps to Baseline are. GRPO Run 2 is the data-fidelity probe trained on the low-fidelity fluid corpus (§3.3). **Bold** marks the best entry per column.

Stage	Overall	img-only	img&vid	general	Δ
Baseline	40.7	58.4	38.8	27.5	n/a
SFT R1	47.5	58.5	50.2	27.2	+6.8
SFT R2-redo	47.6	58.7	**51.5**	23.8	+6.9
GRPO Run 2	46.3	56.3	49.2	26.1	+5.6
GRPO R3	**47.7**	**59.5**	50.1	**27.6**	+7.0

Table 4. Per-task-type accuracy on PhysBench Test. **Bold** marks the best entry per column.

Stage	Dynamics	Property	Relations	Scene
Baseline	37.3	56.6	41.9	26.0
SFT R1	44.4	56.9	43.2	44.8
SFT R2-redo	43.9	**57.9**	42.0	**46.1**
GRPO Run 2	**45.3**	55.6	41.3	41.6
GRPO R3 (ours)	44.3	56.1	**44.7**	45.5

supervision in any family.

Indirectly, fluid-direction training lifts the Relations:motion subtask even though our fluid labels never use relational language; the shared cue is a moving body in a frame. ScienceQA-Physics (§4.4) was not targeted by any family and shows no significant change: we did not train on textbook physics, and that physics did not show up in the test.

4.2. In-Distribution: PhysBench Test (n=9,786)

Table 3 reports stage-wise accuracy on the full PhysBench Test set; Table 4 breaks the same evaluations down by physics domain.

The largest test-set lifts in Table 4 fall on the domains the mapping in §4.1 predicts: **Scene** +20.1pp (from 26.0 to 46.1% at SFT R2-redo, driven by viewpoint and format alignment), **Dynamics** +6.6pp (37.3 to 43.9% at R2-redo, driven by TTC, trajectory, and fluid tasks; SFT R1 alone reaches 44.4%/+7.1pp, and the small R1-to-R2-redo regression on Dynamics is itself within Wilson CI noise), and **Property** +1.3pp (stability, mass/size). At the mode level: image&video gains +12.7pp at R2-redo (the modality our video rollouts match), image-only is flat (+0.3pp at R2-redo, no static-image training), and the un-targeted general mode regresses −3.7pp under SFT.

Synthetic SFT alone delivers the bulk of the gain (+6.9pp overall and +20.1pp on Scene at SFT R2-redo), with the largest moves on image&video (+12.7pp). The text-only "general" mode regresses −3.7pp under SFT and is then recovered by GRPO R3 (restored to 27.6%). GRPO R3 additionally adds +2.7pp on Relationships (42.0→44.7%), a non-targeted domain, at the cost of small movements on Property and Scene; the per-stage swap is in Table 2.

Table 5. Position on the public PhysBench leaderboard (29-model snapshot, 2026-04-01) (Chow et al., 2025), per physics domain. Reference rows (GPT-4o (OpenAI, 2024), InternVL2.5-78B/38B, InternVL2-76B (Chen et al., 2024), NVILA-15B (Liu et al., 2024b)) are reproduced from the public PhysBench leaderboard under each contributor's evaluation conventions; see Fig. 3 for the comparability note.

Model	Prop.	Rel.	Scene	Dyn.
GPT-4o	56.91	64.80	30.15	**46.99**
InternVL2.5-78B	**60.32**	62.13	37.32	46.11
InternVL2.5-38B	58.77	**67.51**	39.04	45.00
NVILA-15B	59.16	42.34	38.78	45.72
InternVL2-76B	57.65	52.43	38.07	40.12
Qwen3-VL-30B (base)	56.60	41.90	26.00	37.30
Ours (R1)	56.90	43.20	44.80	44.40
Ours (R2-redo)	57.90	42.00	46.10	43.90
Ours (GRPO R3)	56.10	44.70	45.50	44.30

4.3. Per-Domain Specialisation

Figure 3 and Table 5 situate our SFT R1, SFT R2-redo, and GRPO R3 checkpoints against the un-tuned Qwen3-VL-30B baseline and several public PhysBench leaderboard entries, broken down by physics domain.

Reference rows are reproduced from the public PhysBench leaderboard under each contributor's evaluation conventions. Two patterns are consistent: on Scene (the dimension our MuJoCo data targets most directly) our trained checkpoints score 44.8–46.1, in the same range as the reference rows; on Relationships, where our training data carries no relational annotation, gains are smaller and mainly emerge after GRPO, while the largest gains concentrate on the directly targeted Property/Dynamics/Scene domains.

Why the Scene domain moves so much. The baseline scores 26.0% on Scene, near the 25% random-guess floor, while the extractor still finds a letter in 96.6% of baseline responses (§5.4); the per-subtask shifts are visualised in Fig. 4. Two SFT effects drive the Scene jump. The model stops hedging across options and commits to one, dropping unparseable "other-wrong" responses from 23% to 7%. On top of that, multi-camera viewpoint training (camera pose as label) directly lifts Scene:`viewpoint` by +23.6pp (n=1008). Two large non-targeted subtask gains, Scene:`air` (+33.3pp, n=42) and Scene:`temperature` (+41.4pp, n=58), look like the same decisiveness effect surfacing latent baseline ability that the extractor previously did not surface; we did not train on air or temperature. The

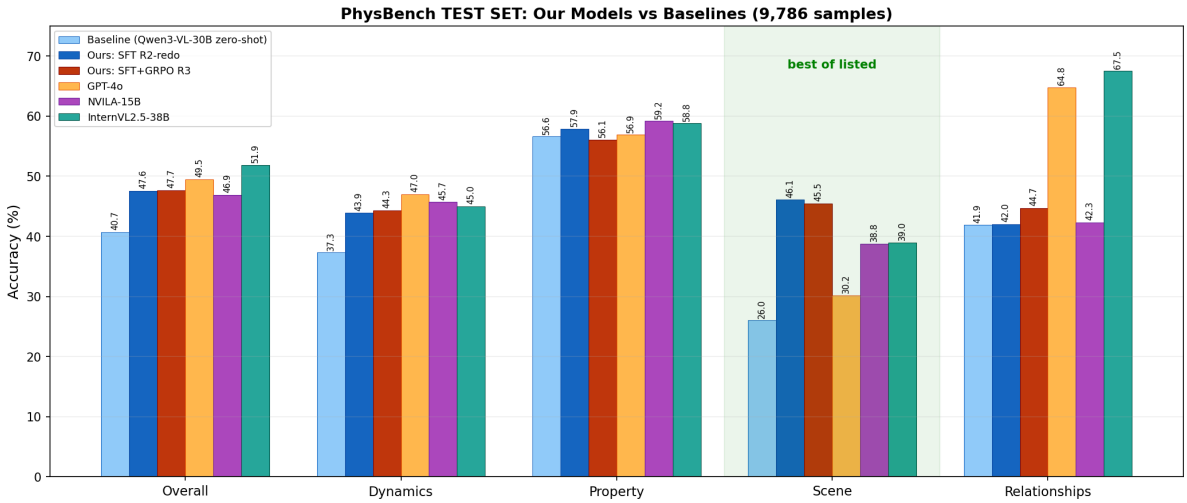


Figure 3. PhysBench Test accuracy (n=9,786) for the un-tuned Qwen3-VL-30B baseline, our PhysSim-VLM checkpoints (SFT R1, SFT R2-redo, GRPO R3), and three public reference VLMs from the PhysBench leaderboard (GPT-4o (OpenAI, 2024), NVILA-15B (Liu et al., 2024b), InternVL2.5-38B (Chen et al., 2024)). Our synthetic-only training recipe lifts overall accuracy by +6.9pp at SFT R2-redo (+7.0pp at GRPO R3) and Scene by +20.1pp. Reference rows are reproduced from the public PhysBench leaderboard and the "best of listed" annotation in the Scene column refers to the four reference rows shown; for transparency we note that each contributor's prompting and extraction conventions can affect cross-row comparability.

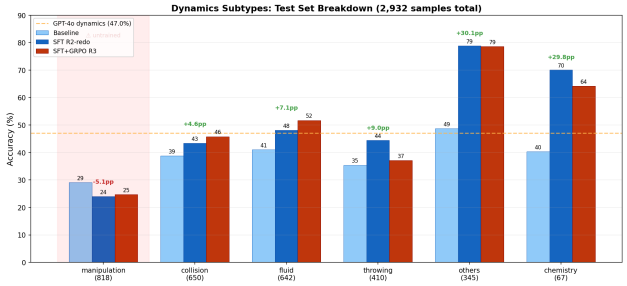


Figure 4. Per-subtask accuracy on PhysBench **Dynamics** subtypes (n=2,932). Bars: baseline (light blue), SFT R2-redo (blue), SFT+GRPO R3 (red); dashed line: GPT-4o Dynamics (47.0%; Table 5 reports 46.99%). Targeted families (`collision`, `fluid`, `throwing`) gain +4.6 to +9.0pp; the un-trained `others` and `chemistry` buckets show large rises (+30.1pp, +29.8pp) that look like latent-skill recovery (§5.4). The small `manipulation` regression (−5.1pp, untrained subtype) is consistent with curriculum coverage.

Scene-domain lift is therefore part true new skill (viewpoint), part recovery of latent baseline knowledge.

4.4. Capability Preservation: ScienceQA-Physics

Since we trained on visual-physics simulator scenes only, we use ScienceQA-Physics as a *capability-preservation probe*, not a transfer claim: did SFT narrow the model into a visual-physics specialist that forgets its prior K-12 textbook ability? On the vision-essential subset (n=425, ≥ 1 image), SFT R2-redo scores 79.53% (338/425, Wilson 95% CI (Wilson, 1927) $[75.44, 83.09]$) vs. baseline 84.00% (357/425, $[80.21, 87.18]$); a two-proportion z-test gives $z = -1.69$ (p=0.09), so the directional −4.5pp gap is *not* statistically significant and the two CIs overlap. We read this cautiously: SFT **does not significantly degrade** prior textbook-physics ability at this sample size, but the probe is underpowered ($\sim$ 50% power for a true −4.5pp effect), so the result is consistent with both no harm and a small drop. A larger off-distribution evaluation and a broader visual question answering (VQA) preservation check (beyond physics-only ScienceQA) are needed to upgrade this to a positive generalist-VLM claim; we note this in §6.

5. Analysis

5.1. Failure-Mode Taxonomy

Table 6 buckets every one of the 9,786 predictions per checkpoint into exactly one of nine response-shape categories. The dominant failure mode at every stage is *off-by-one MCQ*, a parseable single-letter response in {A,B,C,D} that disagrees with the gold letter by exactly one position in option order (23 to 27% of all responses, 43 to 46% of all failures); the gap is one of discrimination between adjacent options, not absent physics knowledge. SFT sharply collapses the "verbose-but-wrong-with-physics" mode (6.2% to 0.6%): the model stops producing long, fluent, wrong explanations. It also introduces a new failure mode, terse wrong answers (1.5% to 10.7%), as the model adopts the short teacher-forced answer style.

Table 6. Failure-mode taxonomy on PhysBench Test (n=9,786 per column). Numbers are absolute counts; percentages of the column total in parentheses. Buckets are assigned in priority order (off-by-1 MCQ first, then terse-wrong-letter on remaining items, then verbose categories), so each item appears in exactly one row but the *shape categories* are not mutually exclusive in the underlying behaviour (e.g. a terse single letter that is off-by-one is counted as off-by-1 MCQ, not terse-wrong-letter).

Mode	Base.	R1	R2-redo	GRPO R3
correct	3984 (40.7)	4649 (47.5)	4662 (47.6)	**4670 (47.7)**
off-by-1 MCQ	2685 (27.4)	2244 (22.9)	2315 (23.7)	2268 (23.2)
other wrong	2256 (23.1)	891 (9.1)	725 (7.4)	924 (9.4)
wrong w/ physics	611 (6.2)	101 (1.0)	57 (0.6)	183 (1.9)
terse wrong letter	149 (1.5)	1014 (10.4)	1048 (10.7)	974 (10.0)
terse wrong	1 (0.0)	873 (8.9)	979 (10.0)	766 (7.8)
empty response	89 (0.9)	0 (0.0)	0 (0.0)	0 (0.0)
unparseable	11 (0.1)	14 (0.1)	0 (0.0)	1 (0.0)
refusal	0	0	0	0

Table 7. Response-length and physics-vocabulary statistics on PhysBench Test. Density = curated physics-vocabulary token fraction $\times 100$.

Checkpoint	Mean len.	Physics dens.	Δ_{c-w}
Baseline	129	1.36	-0.20
SFT R1	13	2.57	-0.39
SFT R2-redo	5	2.80	-0.38
GRPO R3	18	2.87	-0.45

5.2. Chain-of-Thought Length and Physics Vocabulary

Chain-of-thought prompting (Wei et al., 2022) is the standard lever for eliciting step-by-step reasoning from instruction-tuned LLMs. We analyse the post-SFT response length and lexical content as a proxy for whether SFT changes the *shape* of model reasoning, not just the final letter.

Response length and physics-vocabulary density move in opposite directions (Table 7). Mean response length collapses from 129 to 5 words at SFT R2-redo, recovering to 18 after GRPO R3: the model adopts the curt teacher-forced style, so whatever physics it uses sits in the weights, not in visible chain-of-thought. Per-token physics-vocabulary density meanwhile rises monotonically from 1.36% to 2.87% at GRPO R3. On *wrong* answers, density is slightly higher than on correct ones ($\Delta_{c-w} < 0$), so fluent physics language is not a positive indicator of correctness here; this is a hallucination pattern apparently inherited from the teacher labels.

5.3. Calibration Without Logprobs

Because the Tinker sampling API does not expose token logprobs, we use the response-text proxies in Tables 8 and 9. The baseline has a strong letter-prior bias toward "A" (37.2% vs. ground-truth 25.6%); SFT flattens this prior and GRPO R3 over-corrects toward "D". The hedging-language gap collapses ($+6.79$ to -0.24pp at SFT R2-redo, $+0.28$pp at GRPO R3): the model becomes more decisive but slightly over-confident on wrong answers, a known SFT side effect that GRPO partially repairs. The breadth check (Table 9) shows SFT R2-redo improving 27 of 39 PhysBench subtasks and regressing only 5, so the gain is spread out, not driven by a few subtasks.

Table 8. Calibration proxies on PhysBench Test (full n=9,786). Letter-bias = $\mathrm{KL}(\text{pred} \parallel \text{GT})$; calibration gap = $P(\text{hedge} \mid \text{wrong}) - P(\text{hedge} \mid \text{correct})$.

Model	Letter A%	Letter-bias	Calib. gap
Baseline	37.2	0.0626	$+6.79$
SFT R1	25.7	0.0034	$+0.51$
SFT R2-redo	29.0	0.0139	-0.24
GRPO R3	21.1	0.0432	$+0.28$

Table 9. Subtask-improvement breadth across all 39 PhysBench subtasks: a generalisation-vs-overfit proxy.

Checkpoint	Improved	Regressed	Mean Δ	Median Δ
SFT R1	25	7	$+10.97$	$+6.90$
SFT R2-redo	27	5	$+11.58$	$+6.15$
GRPO R3	26	5	$+8.64$	$+6.95$

5.4. Format vs. physics, and parser sensitivity

A natural worry is that the $+6.9$pp lift reflects the model learning the A/B/C/D convention, not physics. MCQ format learning is ruled out by construction: the training corpus contains *no multiple-choice questions and no A/B/C/D labels*; every answer is a free-text simulator-state string (e.g. "the red sphere arrives first") inside `<answer>` tags, so the model never observes an option letter during training. The baseline already scores 40.7% precisely because the last-letter extractor already finds a letter in 96.6% of its verbose responses—format extraction was fully functional before SFT. A residual concern remains: SFT may teach a "commit to one answer" prior that makes extraction *more reliable* without improving physics competence. A strict extractor (leading option letter or "Answer: X") moves SFT scores up by only 0.7–1.2pp while dropping the verbose baseline by -19.8pp, confirming the SFT gain is largely extractor-insensitive; full numbers are in Appendix A.5. The intended shuffled-physics control exceeded our budget; we leave it open. Overall, SFT R1 contributes most parser-sensitive rescues, R2-redo adds narrower corrections, and GRPO R3 mainly redistributes errors.

6. Limitations

This study focuses on transfer from synthetic scenes to PhysBench and uses ScienceQA-Physics (n=425) only as a capability-preservation probe, which is underpowered to

detect small effects; a positive generalist-VLM claim requires a broader, non-physics VQA evaluation that we leave to future work. We also do not run the clean causal control (SFT on the same schema with shuffled-physics labels) that would fully separate format-commitment from new physics capability; this control exceeded our compute budget and is the most important missing experiment. The training curriculum targets selected physics families (with rigid-body MuJoCo scenes outnumbering filtered PhiFlow continuum scenes roughly 4:1), so improvements are strongest in domains with direct or indirect coverage. We evaluate one base VLM and one main GRPO configuration, leaving broader architectural and RL hyperparameter sweeps for future work, and LoRA-rank and reward-weight sweeps were not exhaustive (we used rank 16 at SFT and rank 64 at GRPO based on early stability runs). Cost is reported in Tinker hosted-API credits; the \$4–\$5 per SFT epoch corresponds to roughly 4–5 Tinker wall-clock hours per epoch, and all stages combined total under 12 wall-clock hours.

7. Conclusion and Future Work

Synthetic scenes generated with physically consistent supervision improve real-world visual physics reasoning in VLMs at low cost. Synthetic-only SFT delivers a +6.9pp PhysBench lift with broad subtask coverage and a +20.1pp gain on Scene, and a follow-up GRPO stage preserves this aggregate gain (+7.0pp over baseline) while adding targeted improvements on text-heavy `general:relationships` questions the SFT model misses. A data-fidelity probe at both stages further shows that the gains track the physical quality of the synthetic signal: physically accurate fluid supervision improves PhysBench, while a deliberately low-fidelity variant of the same corpus does not. Together, these results support synthetic physics as a practical, reproducible supervision signal for grounded visual reasoning without manual labeling.

Future work has four directions. First, expand the synthetic curriculum beyond rigid-body and continuum-fluid scenes to relational and temporal families (e.g. multi-object causal chains, deformables, soft contact) that directly target the PhysBench Relationships domain currently weakest in our coverage. Second, run the deferred shuffled-physics SFT control and explore process-level verifiable rewards (per-frame state, intermediate predicates) so RL improves aggregate accuracy rather than only redistributing errors. Third, broaden off-distribution evaluation beyond the underpowered ScienceQA-Physics probe to larger physics suites (e.g. SeePhys, PhysReason), to free-form, non-MCQ physics QA, and to a broader generalist VQA probe to upgrade capability-preservation into a positive generalist-VLM claim. Fourth, run the same recipe on additional base VLMs and a lightweight sim-to-real mix to map when synthetic-only supervision saturates.

Reproducibility. The synthetic training data are publicly available on the Hugging Face Hub: the full corpus at https://huggingface.co/datasets/Swastikr/PhysSim-VLM-Dataset and the corrected SFT R2 split at https://huggingface.co/datasets/Swastikr/PhysSim-VLM-SFT-R2-Data. We additionally release the synthetic data generation scripts (MuJoCo and PhiFlow), the eight task-family prompt templates, the LoRA adapter configurations for SFT R1, SFT R2-redo and GRPO R3, the PhysBench / ScienceQA-Physics evaluation harness, the standard and strict output extractors used in §5.4, and per-item predictions for every reported checkpoint in the project code repository,[1] so that paired tests (e.g. McNemar, paired bootstrap) and parser-sensitivity reproductions can be re-run independently. Total training cost was under \$30 in hosted-API credits, plus approximately \$12 per full PhysBench Test evaluation sweep; underlying hardware is managed by the Tinker service and is not user-visible.

Acknowledgments

We thank the Thinking Machines team for access to the Tinker managed LoRA fine-tuning service, which enabled all training and evaluation on hosted infrastructure without local GPUs. We also thank the Indian Institute of Information Technology Raichur for institutional support, and the AI4Physics @ ICML 2026 reviewers for thorough and constructive feedback that shaped the camera-ready version.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning, specifically the physical reasoning capabilities of vision-language models. The core contribution is a low-cost, fully synthetic training pipeline that requires no human-annotated data and no local GPU, making physics-grounded VLM research more accessible to resource-constrained research groups. The training data and generation code are publicly released to support reproducibility and follow-on research.

We do not foresee immediate negative societal consequences specific to this work. The fine-tuned model is not deployed in any safety-critical system, and the improvements demonstrated are confined to a constrained physics-reasoning benchmark. Potential downstream misuse of improved physical reasoning in VLMs (e.g. in autonomous systems) is a concern for the broader field and is mitigated by the modest

[1] Code repository: https://github.com/QuantumByte-01/PhysSim-VLM (released with the camera-ready).

scale of the gains reported here. We release all artifacts under permissive open licenses to encourage transparent research use.

References

Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., Pietquin, O., Üstün, A., and Hooker, S. Back to basics: Revisiting REINFORCE style optimization for learning from human feedback in LLMs. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Balazadeh, V., Ataei, M., Cheong, H., Khasahmadi, A. H., and Krishnan, R. G. Physics Context Builders: A modular framework for physical reasoning in vision-language models, 2024.

Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.-W., and Grover, A. VideoPhy: Evaluating physical commonsense for video generation, 2024.

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R. T., Holdaway, C., Tao, S., Smith, K., Sun, F.-Y., Fei-Fei, L., Kanwisher, N., Tenenbaum, J. B., Yamins, D. L. K., and Fan, J. E. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Chow, W., Mao, J., Li, B., Seita, D., Guizilini, V., and Wang, Y. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations (ICLR)*, 2025.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Gao, J., Sarkar, B., Xia, F., Xiao, T., Wu, J., Ichter, B., Majumdar, A., and Sadigh, D. Physically grounded vision-language models for robotic manipulation, 2024.

Holl, P., Koltun, V., and Thuerey, N. Learning to control PDEs with differentiable physics. In *International Conference on Learning Representations (ICLR)*, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., and Feng, J. How far is video generation from world model: A physical law perspective, 2024.

Liu, S., Ren, Z., Gupta, S., and Wang, S. PhysGen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2024a.

Liu, Z., Zhu, L., Shi, B., et al. NVILA: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024b.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

OpenAI. GPT-4o system card, 2024.

Riochet, R., Castro, M. Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., and Dupoux, E. IntPhys: A framework and benchmark for visual intuitive physics reasoning, 2018.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Stewart, R. and Ermon, S. Label-free supervision of neural networks with physics and domain knowledge. In *AAAI Conference on Artificial Intelligence*, 2017.

Team, Q. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Thinking Machines. Tinker: A managed LoRA fine-tuning service for large language and vision models. Public managed fine-tuning API., 2025. Accessed 2026-04.

Todorov, E., Erez, T., and Tassa, Y. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5026–5033, 2012.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Xiang, K., Li, H., Zhang, T. J., Huang, Y., et al. SeePhys: Does seeing help thinking? benchmarking vision-based physics reasoning, 2025.

Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., and Tenenbaum, J. B. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

Zhang, X., Dong, Y., Wu, Y., et al. PhysReason: A comprehensive benchmark towards physics-based reasoning, 2025.

Zheng, Z., Yan, X., Chen, Z., Wang, J., Lim, Q. Z. E., Tenenbaum, J. B., and Gan, C. ContPhy: Continuum physical concept learning and reasoning from videos. In *International Conference on Machine Learning (ICML)*, 2024.

A. Appendix

A.1. Training-Task Families and PhysBench Mapping

Table 10 lists the eight synthetic training-task families and the PhysBench domains they are designed to load on, expanding the mechanism summary in §4.1. *Direct* targets share label semantics with the test subtask; *indirect* targets share visual primitives or reasoning style but no label-level overlap. The text-only general mode and the Relationships domain are not directly targeted by any family, which is consistent with Relationships remaining among the weaker domains after training.

Table 10. Mapping from the eight synthetic training-task families to PhysBench domains. "Cont." = continuous numeric target; "Cat." = categorical free-text target.

Training family	Type	PhysBench mapping
Time-to-collision	Cont.	Dynamics (direct)
Trajectory	Cont.	Dynamics (direct)
Stability	Cat.	Property (direct); attribute (indirect)
Mass/size compare	Cat.	Property (direct)
Fluid viscosity	Cat.	Dynamics/Property (direct)
Fluid level	Cat.	Dynamics/Property (direct)
Fluid direction	Cat.	Dynamics (direct); Relations:motion (indirect)
Count/viewpoint	Cat.	Property:number, Scene:viewpoint (direct)

A.2. Evaluation Protocol

All PhysBench Test numbers use the full n=9,786 item set across four domains (Property, Relationships, Scene, Dynamics) and three modes (image-only, image&video, text-only "general"), each item a 4-way multiple-choice question. Unless stated otherwise we decode at temperature 0 with a 512-token cap and score with the official PhysBench last-letter A/B/C/D extractor; the strict extractor variant used for the parser-sensitivity analysis is described in §A.5. ScienceQA-Physics uses the vision-essential subset (n=425, ≥ 1 image). For proportions we report Wilson 95% confidence intervals (Wilson, 1927); for the paired SFT-vs-GRPO comparison we use a matched-item McNemar test (§3.3); and for the ScienceQA contrast a two-proportion z-test. No item, image, or question text from any evaluation benchmark appears in the synthetic training corpus.

A.3. Compute and Cost

All training and inference run on the hosted Tinker LoRA service (Thinking Machines, 2025); we use no local GPU, and the exact accelerator type is abstracted by the service. The backbone is Qwen3-VL-30B-A3B-Instruct (a 30B-parameter mixture-of-experts model with $\sim$ 3B active parameters); LoRA adapters are $\sim$ 0.4–1.7% of backbone parameters (rank 16 for SFT, rank 64 for GRPO). Table 11 summarises the per-stage step counts and wall-clock times; the three stages total roughly 12 wall-clock hours and under \$30 in hosted-API credits (\$4 to \$5 per single SFT epoch). Evaluation is billed separately: a full PhysBench Test sweep (n=9,786, concurrency 4) costs approximately \$12 in hosted-API inference credits per checkpoint.

Table 11. Per-stage training cost. Wall-clock times are as reported by the hosted service.

Stage	Steps	Wall time	LoRA rank
SFT R1	751	$\sim$ 3.7 h	16
SFT R2-redo	644	$\sim$ 3.9 h	16
GRPO R3	300	$\sim$ 4.5 h	64
Total	—	$\sim$ 12 h	—

A.4. GRPO Hyperparameters

Table 12. GRPO R3 hyperparameters.

LoRA rank / α	64/32
Group size G	4
Groups per batch	2
Sampling temperature	0.7
Max tokens	768
Min completion tokens	20
KL penalty coefficient	0.04
Reward weight (physics)	0.95
Reward weight (format)	0.05
TTC reward decay	2.0
Trajectory reward decay	1.0
LR warmup steps	50
LR min ratio	0.1
Total optimisation steps	300

GRPO R3 was trained with the configuration in Table 12.

Table 13. Output-parser rescue on PhysBench Test (full n=9,786). Same model outputs, only the extractor differs.

Checkpoint	Standard	Strict	Δ
Baseline	40.71	20.92	−19.79
SFT R1	47.51	**48.24**	+0.74
SFT R2-redo	47.64	**48.88**	+1.24
GRPO R3	47.72	47.66	−0.06

We resume from the SFT R2-redo checkpoint on a corrected 1,200-scene curriculum: 300 static-image scenes (counting, viewpoint, manipulation) and 900 video scenes (TTC, trajectory, stability, fluid, manipulation, throwing). Prompts use the same free-text `<answer>...</answer>` schema as SFT (no MCQ re-wrapping). Total wall time is approximately 4.5 hours.

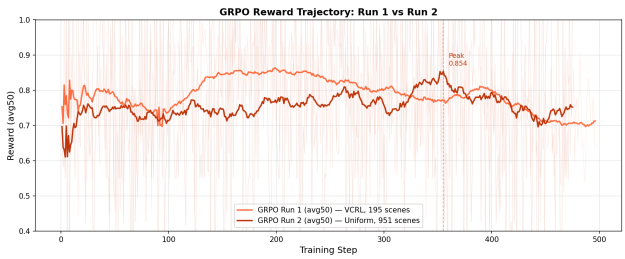


Figure 5. GRPO R3 mean reward per step (raw and 50-step exponential moving average, EMA), compared to the earlier Run 2 trace on uncorrected data. The corrected curriculum stays above the Run 2 baseline throughout training; the variance reduction visible after step ∼150 corresponds to the model converging on a short MCQ-letter answer style for the categorical sub-batch.

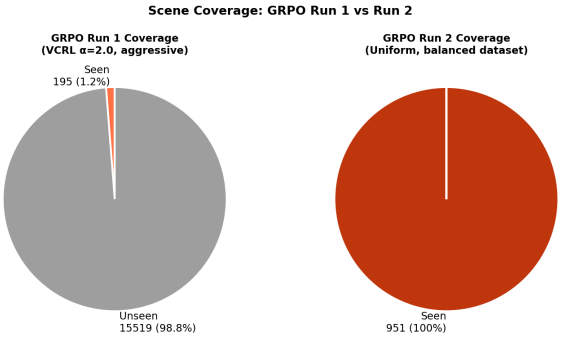
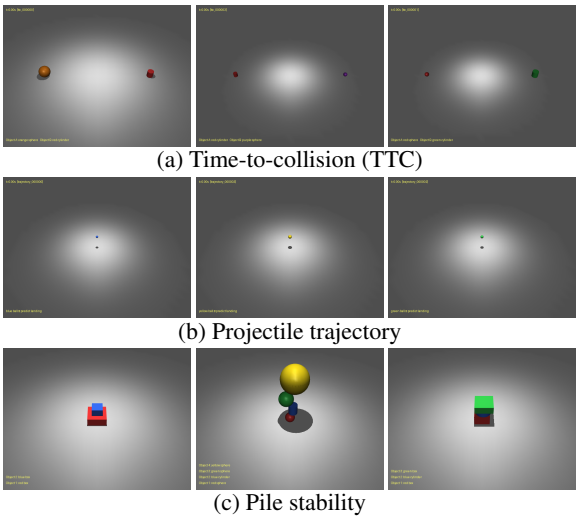


Figure 6. GRPO scene coverage. *Left:* an early variance-conditioned (VCRL, α=2) sampler concentrated rollouts on 195/15,714 scenes (1.2%). *Right:* a uniform balanced sampler over 951 scenes (used for the Run 2 data-fidelity probe and as the precursor to R3). The reported R3 model trains on a re-balanced 1,200-scene curriculum (300 static-image + 900 video) chosen to mirror the PhysBench mode mix.

A.5. Output-Parser Sensitivity

The standard PhysBench MCQ extractor selects the *last* standalone capital letter in the response (regex `\b[A-D]\b`, last match). For SFT outputs of the form "`B. a small cyan ball`" the trailing "`a`" (uppercased to "A") is preferred over the intended leading "B", producing a spurious miss. Table 13 compares the official last-letter extractor to a strict alternative that prefers a leading-position option letter, and otherwise prefers an explicit "`Answer: X`" or "`X.`" anchor, on identical model outputs.

The strict extractor is not a universally better scorer: on the baseline it discards last-letter answers the model accidentally produces in the right position. On the SFT checkpoints, the model emits a leading option letter often enough that the strict extractor recovers approximately 0.7–1.2pp. GRPO R3 remains nearly flat (−0.06pp), consistent with its slight response-length re-expansion (Table 7).



(a) Time-to-collision (TTC)

(b) Projectile trajectory

(c) Pile stability

Figure 7. MuJoCo training scenes used in SFT R1: time-to-collision, projectile trajectory, and pile stability. Each scene is rendered as an 8-frame rollout; first frame shown.

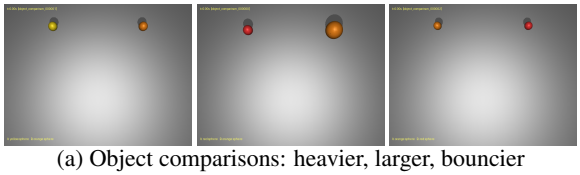
A.6. Dataset Samples

We show representative scenes from each of the eight training-task families: MuJoCo rollouts (Fig. 7), PhiFlow fluid renders (Fig. 8), and categorical comparison grids (Fig. 9). The temporal structure of the video rollouts is illustrated in Fig. 11. The heterogeneous in-the-wild PhysBench Test imagery used for evaluation is shown in Fig. 10, with additional samples from the un-targeted Relationships domain in Fig. 12.

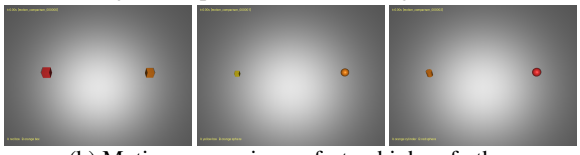
(a) Fluid direction

(b) Viscosity (low / high) and fluid level (high)

Figure 8. PhiFlow training scenes used in SFT R2-redo: flow direction, viscosity comparison, and fluid level. Rendered top-down at 256×256 via particle deposition + Gaussian density filter ($\sigma = 7$ px).



(a) Object comparisons: heavier, larger, bouncier

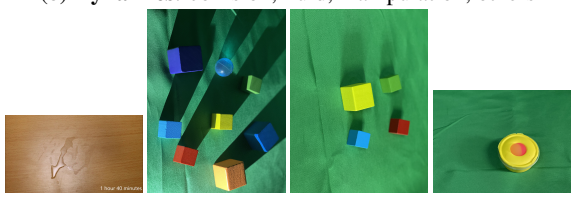


(b) Motion comparisons: faster, higher, farther

Figure 9. Categorical comparison scenes used in SFT R2-redo (object mass/size/elasticity and motion magnitude). The model is asked to identify the correct object from a short free-text answer (e.g. `<answer>red</answer>`); ground truth is read directly from the simulator state.

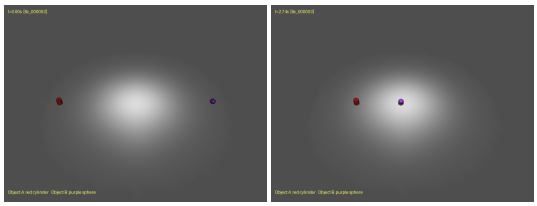
(a) **Property**: attribute, colour, mass, number

(b) **Dynamics**: collision, fluid, manipulation, others

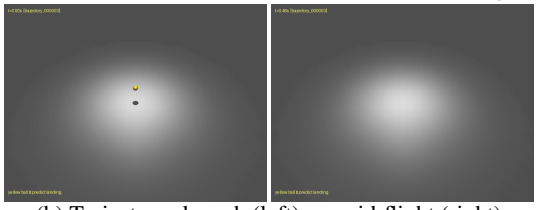


(c) **Scene** and **Relationships**

Figure 10. PhysBench Test scenes spanning all four task domains. Each item poses a 4-way MCQ. Our training distribution overlaps Property, Dynamics, and Scene; Relationships is not directly represented in our training data and remains among the weaker domains.

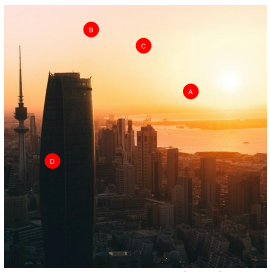


(a) TTC: first frame (left) vs. near-contact frame (right)

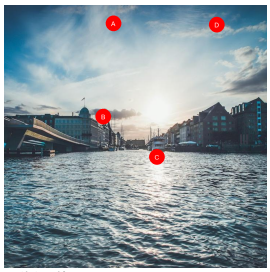


(b) Trajectory: launch (left) vs. mid-flight (right)

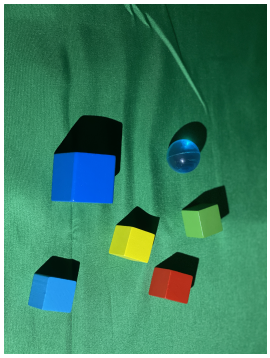
Figure 11. Temporal structure of MuJoCo training rollouts. Each scene is presented to the model as an 8-frame video; here we show two frames from a single time-to-collision scene (top) and a single projectile-trajectory scene (bottom) to illustrate the motion the model must reason over. Ground-truth targets (collision time, landing position) are read directly from simulator state.



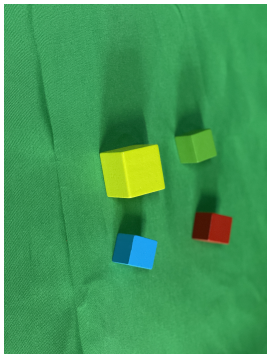
(a) depth



(b) distance



(c) location



(d) size

Figure 12. PhysBench **Relationships** test scenes (depth, distance, location, size). This domain has no label-level overlap with our synthetic training families and remains among the weaker domains after training (§4); we include samples here to illustrate the spatial-relational reasoning these items require.