NEWTPHYS: DISSECTING NEWTONIAN UNDERSTANDING IN FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

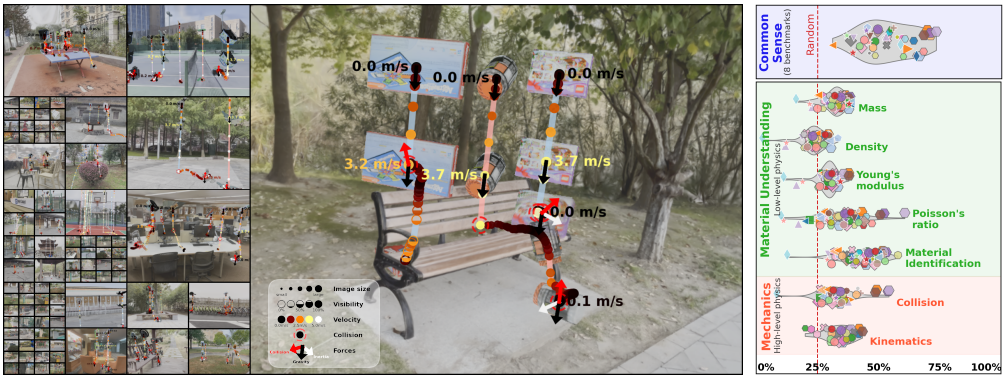


Figure 1: **NewtPhys** scrutinizes *how* vision and vision-language models understand physics, revealing major gaps in Newtonian reasoning. **Left**: Our study is enabled by a new large-scale real-world augmented dataset that captures interactions involving soft and rigid objects across diverse scenes, with a vast number of annotations spanning high-level events and 12 pixel-level physical, kinematic, semantic, and geometric properties. **Right**: NewtPhys enables large-scale benchmarking of 57 VLMs, including 54 open models and 3 frontier closed models, demonstrating that despite reasonable performance on commonsense tasks (top) they perform around chance-level when it comes to physics.

ABSTRACT

Previous work has evaluated physics reasoning in foundation models using synthetic or semi-synthetic scenes and visual question-answering tasks. However, these benchmarks emphasize high-level events and lack the visual fidelity needed to assess low-level Newtonian understanding. We introduce **NewtPhys**, a 4D physically annotated dataset built from multiview images of real-world scenes with physics-grounded simulations. The dataset provides dense, fine-grained annotations across timesteps — including 3D forces and amodal per-pixel quantities covering physics, tracking, semantics and geometry — bridging the gap between simplistic synthetic setups and realistic visual complexity. Using NewtPhys, we systematically evaluate 57 VLMs, including 54 open models and 3 frontier closed models, and 10 VFMs and reveal limitations in low-level physics reasoning. Beyond benchmarking, our study enables future research in physics-grounded vision and next-generation physics-aware evaluation. Code and datasets will be public.

1 INTRODUCTION

Humans develop intuitive Newtonian models early in life (Baillargeon, 1994), enabling prediction of motion, contact, gravity and more, needed to navigate the world and solve complex visual tasks (McCloskey et al., 1983; Carey, 2000). This notion of intuitive physics has strongly influenced machine learning and vision, motivating latent-variable and object-centric models of physical reasoning (Tenenbaum et al., 2011; Battaglia et al., 2013; 2018) as well as visual representations capturing causal structure and physical regularities (Isola et al., 2015). Meanwhile, large-scale vision models, trained without any physical objectives, excel at recognition and open-ended reasoning, raising a fundamental question: do they learn the causal structure of physics, or merely exploit correlations?

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 2: **State of physical understanding.** High-level benchmarks probe general physics in semi-realistic scenes but miss low-level physics; low-level benchmarks use unrealistic toy simulators.

Recent works showed, for example, that generative models violate physical laws (Yuan et al., 2026a), but this does not reveal which physical principle models understand or not. Visual Question Answering (VQA) benchmarks probe spatial awareness (Zhan et al., 2024), 3D reasoning (El Banani et al., 2024), or high-level physics (Chow et al., 2025), yet none evaluates at scale the understanding of fine-grained Newtonian quantities and phenomena such as forces, collisions, gravity, and deformation. We therefore still lack a precise picture of what models understand (or not), where they fail, and whether we have the means to improve them. Existing benchmarks face a trade-off: realistic datasets rely on high-level human judgments (Chow et al., 2025), but they *cannot* provide physics labels, whereas benchmarks with explicit Newtonian annotations rely on simplified synthetic worlds (Yi et al., 2020; Riochet et al., 2021; Bear et al., 2021).

To bridge this gap, we introduce NewtPhys, a 4D benchmark for *low and high-level Newtonian reasoning in realistic settings*. We do so by augmenting real-world scenes with diverse objects using 3DGS (Kerbl et al., 2023) and employ a custom mesh-free Newtonian simulator to generate physically-grounded sequences. Beyond improved visual and physical fidelity, our pipeline captures an unusually rich set of labels, including high-level events such as collision, viewpoint, *etc*., and 12 pixel-aligned maps spanning physics, kinematics, semantics and geometry. Figure 1 illustrates some sequences, with *some* labels overlaid. NewtPhys further provides 141K Visual Question Answering (VQA), across six categories of physical understanding. Together, these annotations enable a large-scale study on 57 Vision-Language Models (VLMs) through VQA and 10 Vision Foundation Models (VFMs) through pixel-level physics probing.

Our study reveals a clear mismatch between general VLM and physical reasoning. While VLMs substantially improved alongside LLMs, their gain in physics is both limited and slower. Models perform well at materials recognition but struggle with granular Newtonian properties such as density or mass, suggesting limited access to the causal mechanics for true physical world modeling. Even frontier models exhibit relatively weak physical and mechanical understanding. This leads us to identify two key factors: (i) physics exposes a failure that general benchmarks mask: VLMs underuse visual evidence (Chen et al., 2024a; Tong et al., 2024), and unlike in everyday VQA, their language priors cannot always compensate, since the LLM decoder itself lacks physical understanding; (ii) popular commonsense benchmarks correlate weakly with physics understanding, which hinders further progress in physics. With our pixel-wise physics labels, we also benchmark 10 vision-only VFMs and introduce *Physics Probing* which measures whether Newtonian signals are accessible from learned representations. Our results suggest that current VFMs are not intrisically physics-grounded and do not reliably capture physically meaningful signals.

Our contributions are threefold. First, we propose a pipeline combining real 3DGS scenes and objects with customized Newtonian simulation to generate videos with rich high- and pixel-level annotations. Second, we release a large-scale benchmark of enhanced visual and physical realism, comprising 0.7M frames and 141K VQA pairs. Third, we dissect the Newtonian physics understanding of 67 VFMs/VLMs (open and closed models), revealing not only *where* existing models fail, but also *why* progress may remain limited. Our simulator, dataset, and benchmark will be released publicly.

2 Related Works

A substantial body of work aims to learn physical laws or latent dynamics from visual input, typically in simplified environments (Wu et al., 2015; 2016). They demonstrate that structured physical representations can be learned, but largely operate in toy 2D/3D with limited visual and physical complexity. Other benchmarks study physical reasoning in richer scenes, including event prediction, counterfactual reasoning, and causal queries (Yi et al., 2020; Riochet et al., 2021; Bear et al., 2021). However, physics annotations in these datasets are still mostly symbolic or event-level.

Though confined to high-level insights, generation benchmarks such as Physics-IQ evaluates physical plausibility of video generation models, with learned or human critics (Motamed et al., 2026), while PISA (Li et al., 2025) or Morpheus (Tragoudaras et al., 2026) focus on gravity effect. Instead, PhysBench (Chow et al., 2025) uses VQA to evaluate high-level physical understanding in real-world photographs or toy simulation data, while NewtonGen (Yuan et al., 2026b) tests qualitative or counterfactual physical reasoning. Yet, these benchmarks lack access to Newtonian quantities and therefore cannot probe low-level understanding to determine whether models represent forces, stresses, or contact patterns internally. Conversely, several studies use simulation to probe low-level physics, including soft-body dynamics (Tung et al., 2023), but in highly unrealistic environments as seen in Figure 2 — especially compared with the realism of NewtPhys in Figure 1.

To our best knowledge, no benchmark provides dense, pixel-aligned supervision of forces and deformations in visually realistic scenes. NewtPhys fills this gap, enabling dissection of Newtonian understanding with rich annotations, at both low- and high-level, and across diverse tasks.

3 The NewtPhys benchmark

NewtPhys, illustrated in Figure 3, augments real-world 3D Gaussian Splatting (3DGS) scenes by spawning objects in a controllable fashion and animating them with a Newtonian simulator. Thanks to their visual fidelity and 3D accuracy, 3DGS reconstructions of real scenes are already extensively used as simulation environments in other domains, notably autonomous driving, for the closed-loop evaluation (NVIDIA et al., 2025) and training (Gao et al., 2025; Li et al., 2026) of driving models. Beyond scale and fidelity, our pipeline uniquely captures both high-level physics events (collisions, motion, camera, *etc*.) and low-level per-pixel ground-truth maps (physics, geometry, kinematics, semantics); thereby allowing to dissect the understanding of Newtonian physics in vision models.

3.1 Dataset creation

To construct realistic data, we optimize 3DGS models (Kerbl et al., 2023) from 53 real-world multiview scenes of DL3DV (Ling et al., 2024) and 109 everyday real objects from Google Scanned Objects (GSO) (Downs et al., 2022). Since splats will serve as particles of a physical simulator, they must share a metric canonical coordinate system, and be fairly dense. This is critical as alignment inconsistencies lead to erroneous physical interactions, while low density triggers inter-penetration. Such calibrated 3DGS models are easy to obtain for GSO objects, which provide textured 3D models, but non-trivial for DL3DV scenes captured with non-calibrated handheld devices. Hence, to maximize the density of scenes, we initialize their 3DGS model with a dense VGGT point cloud (Wang et al., 2025), aligned via Kabsch–Umeyama (Umeyama, 2002) to a *canonical* point cloud obtained from vanilla COLMAP, scaled using metric priors and monocular depth (Lin et al., 2025), and ground-aligned to $Z = 0$. The resulting calibrated scene and objects 3DGS can then be assembled together. Rather than random composition, we define scene-specific regions of interest (ROI) where objects are randomly spawned, enabling scripted interactions (*e.g*., a stiff object falling on a bench, a dozen of fluffy teddy bears colliding with a statue, *etc*.).

Newtonian physics simulation. Unrolling physics is challenging due to 3DGS unstructured and sparse nature, making mesh-based simulators brittle to slight geometric noise. Therefore, we adopt Simplicits (Modi et al., 2024), a high-fidelity mesh-free simulator for deformable dynamics without the need for surface reconstruction. While Simplicits was made for single-object, we extend it to complex scenes through a careful recipe for reduced simulation state. Practically, from the roughly $10^7 - 10^8$ combined scene and object primitives, we sample a controlled set of cubature points via density-aware spatial resampling of the Gaussian mixture, upweighting splats near interaction regions. Object material properties (Young's modulus, Poisson ratio, density, volume), critical for accurate simulation, are estimated from visual cues and metadata and used to learn a custom per-object skinning function with a softness-dependent number of deformation handles.
Our careful design stably simulates large scenes with up to $\sim 10^8$ splats using only $\sim 10^4$ cubature points and $\sim 10^2$ deformation DoFs. For fine-grained interaction we use a 10ms simulation step and render every 40ms (*i.e*., 25fps) via 3DGS rasterization (Kerbl et al., 2023), updating splats through cubature-to-splat mapping. Critically, we modify Simplicits' core engine to capture: (i) point-wise physics forces, kinematics, geometry and semantics – which enables pixel-accurate rasterization of

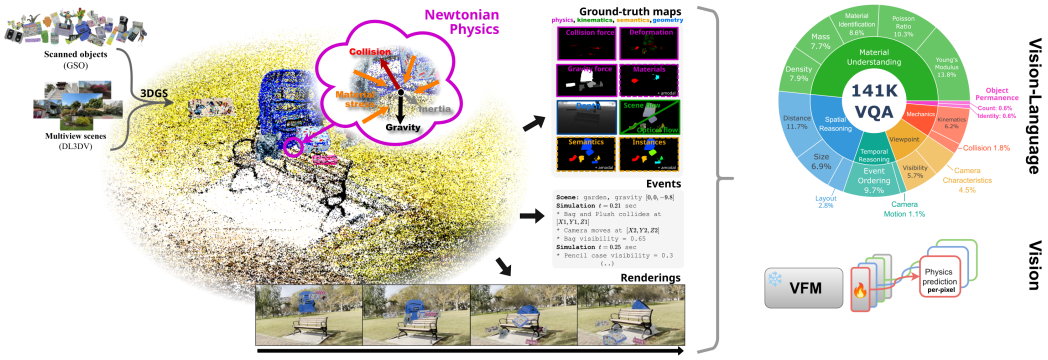


Figure 3: **Dataset construction.** We construct scenarios by spawning up to ten GSO objects (Downs et al., 2022) into DL3DV scenes (Ling et al., 2024), and simulating their 3DGS with a heavily customized Simplicits simulator (Modi et al., 2024), which permits dense capture of *Newtonian dynamics across space and time*. The pipeline renders realistic videos paired with 12 pixel-level ground-truth maps spanning physics, kinematics, semantics and geometry plus frame- and scene-level events such as collisions, forces, semantics, *etc*. The dataset covers a wide range of rigid and deformable materials, as seen with the blue bag compressing upon impact with the bench. In total, NewtPhys includes 11k sequences (730k frames, 8M ground-truth maps) and 141k image&video VQA, which enables diagnosing physics understanding in VLM and VFM.

labels; (ii) a structured simulation state containing collisions, visibility, camera motion, *etc*.– which enables automatic task generation. See Sec. B for details on simulator and material-estimation.

3.2 Dataset overview

Ultimately, NewtPhys dataset employs 53 DL3DV scenes and 109 GSO objects selected for maximum physical and visual diversity, across 11k sequences (1000x562, $\leq$ 10 seconds @25fps), each with up to 10 randomly spawned objects in scene ROIs. To maximize visual diversity, we automatize random camera trajectories while enforcing visible interactions. Further, since a single object may exhibit varying material characteristics, we learn several skinning networks per object, varying properties over the VLM-queried ranges, yielding 333 physically distinct objects models. Besides JSON-encoded high-level events, each frame comes with 12 ground truth maps (including 3 amodal) capturing pixel-level physics (`collisions`, `gravity`, `materials`, and `deformation` capturing material stress), kinematics (`scene flow`, `instance tracking`), semantics (`segmentation`) and geometry (`depth`). Refer to Figure 3 for labels illustration.

Dataset evaluation. Compared to existing datasets, Tab. 1a shows that NewtPhys captures more high-level labels and one order of magnitude more pixel labels while being the only one to leverage augmented real-world scenes. Beyond scale, we evaluate visual fidelity in Tab. 1b with FID and KID metrics w.r.t. the real set of PisaBench (Li et al., 2025), which records falling objects in real scenes, and report the VBench (Huang et al., 2024) performance (mean of the 6 no-reference metrics). NewtPhys outperforms all metrics significantly, demonstrating better visual fidelity. Measuring physical fidelity is much harder but equally important. To that effect, we report the physical plausibility metric of Yuan et al. (2026a) that measures the latent surprise of the VJEPA model over sliding temporal windows. Tab. 1b reports mean of the max and top-3 per-sequence surprises, showing that our dataset is more physically accurate. Beyond increased scale and fidelity, for the first time NewtPhys allows to scrutinize *how* vision models understand Newtonian physics.

Visual Question Answering. Benefitting from NewtPhys's rich labels, we design image&video questions along six axes: **Material understanding** covers material properties (density, Young's modulus, Poisson ratio), object mass and material type; **Mechanics** includes collisions and interactions; **Spatial reasoning** assesses size, distances and scene layout; **Viewpoint** probes visibility and camera-scene relationship. For video, we also consider: **Temporal reasoning** covering event ordering and camera motion, and **Permanence** assessing the key ability to memorize invisible objects.

Dataset	Data				High-level labels				Pixel labels				VQA			
	Type	Scenes	#Vid.	Camera	Material	Events	Kinemat.	Forces	Physics	Kinemat.	Semantics	Geometry	Cat.	Types	#Q.	C.Pair
ContPhy	Sim.	Indoor	2k	Static	◐	◐	✓	-	-	-	Instances	-	3	26	6.5k	✓
IntPhys*	Sim.	Indoor	15k	Static	◐	-	✓	✓	-	-	Instances[a]	Depth	-	-	-	-
IntPhys2*	Sim.	Mixed	506	Dyn.	-	-	-	-	-	-	-	-	-	-	-	-
PhysBench	Sim.&Real†	Mixed†	3k	Static	- †	- †	- †	- †	-	-	-	-	5	19	10k	✗
Physion++	Sim.	Indoor	8k	Static	◐	◐‡	✓	-	-	-	Instances	-	-	-	-	-
Physics-IQ	Real	Indoor	198	Static	-	-	-	-	-	-	-	-	-	-	-	-
PisaBench	Real	Mixed	361	Static	◐	-	-	-	-	-	Instances[b]	-	-	-	-	-
NewtPhys	Sim. (real-world augmented)	Mixed	11k	Mixed	✓	✓	✓	✓	Colision, Deform., Gravity, Material Amod. mat.	Opt. flow, Scene flow	Segment., Instances, Amod. seg., Amod. inst.	Depth	6	48	141k	✓

* For consistency, we report the "possible" set. † Mixing datasets and labels regimes. ‡ Sparse high-level annotations. [a] Not temporally consistent. [b] Pseudo-labels.

(a) Features comparison

Dataset	Visual realism			Physical realism	
	FID ↓	KID ↓	VBench ↑	Max ↓	Top-3 ↓
ContPhy	272.59	0.1665	0.6913	0.4616	0.4469
IntPhys	251.60	0.1556	0.6584	0.4384	0.4277
IntPhys2	215.11	0.0951	0.6877	0.4935	0.4767
PhysBench	216.17	0.1075	0.6799	0.4629	0.4491
Physion++	193.82	0.1185	0.6442	0.4370	0.4254
NewtPhys	**189.68**	**0.0783**	**0.6970**	**0.4287**	**0.4216**
PisaBench	n/a	n/a	0.6914	0.4017	0.3931

(b) Quantitative comparison

Table 1: **Comparison with existing datasets.** Compared to prior datasets, NewtPhys offers (a) features with significantly richer high- and low-level labels and (b) is significantly more *visually and physically* realistic (*cf*. text) across all metrics, bridging the gap to unannotated real datasets.

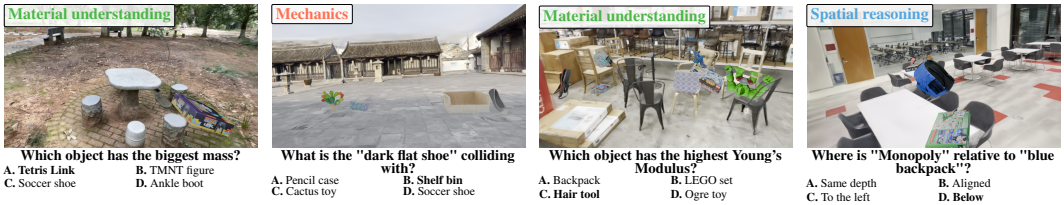


Figure 4: **Example of NewtPhys VQA.** We illustrate examples across spatial reasoning, mechanics, and material understanding. Refer to Sec. D.2 for more examples.

Further, we implement 48 questions as template functions taking as input a frame or sequence with its ground-truth labels, and outputting a tailored question and four shuffled answers with a triplet of three wrong answers variably distant from the correct one. Running our fully automated pipeline on NewtPhys, we obtain 141K VQA (84K video, 57K image), whose statistics are in Figure 3 (right). Figure 4 shows exemplar questions, while details and more examples are found in Sec. D.

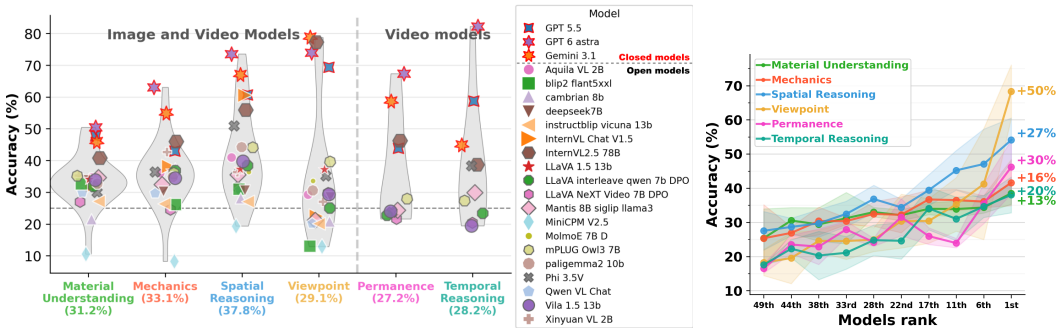
For evaluation with smaller compute budgets, we also release a 15K subset, NewtPhys-15K, with similar question balance, and carefully selected to yield comparable VLM performance.

4 Probing Physics Understanding

Using NewtPhys we study the physical understanding of 67 models; primarily focusing on VQA in Secs. 4.1 to 4.4 since language is a natural proxy for physical reasoning (Bisk et al., 2020), while also probing physics in vision-only models in Sec. 4.5. Our research encompasses 64 open models, spanning a wide range of size (0.4B–78B) and 27 families, fully detailed in table Sec. E, including open models such as LLaVA (Liu et al., 2023), InternVL (Chen et al., 2024c), Molmo (Deitke et al., 2024), Deepseek (Lu et al., 2024a) *etc*., as well as three frontier closed models, GPT 5.5, GPT 6 (OpenAI, 2026), and Gemini 3.1 (Google DeepMind, 2026), to better position our findings. Unless stated otherwise, open models are evaluated on the full 141K NewtPhys, while closed frontiers models use NewtPhys-15K to reduce evaluation cost.

4.1 How do VLMs perform on fundamental physics?

We first assess *to which extent* VLMs grasp physical principles, comparing material understanding and mechanics – critical to interactions and often omitted in benchmarks – with often studied categories. Figure 5a reports performance for each family's largest model. Unsurprisingly, frontier models (GPT, Gemini) dominate the board, although perhaps unexpected InternVL2.5-78B, the largest open model, is a close competitor. Other models typically perform around 25–35% across tasks, a low score that underscores the challenge of NewtPhys w.r.t. most benchmarks where objects of interest are typically large and visually in focus. Predictably, temporal tasks perform lower, with permanence even sitting near chance-level, assessing that this core cognition task still requires a deeper understanding of space and time. We observe also that several models simply fail to understand



(a) Family's largest models (b) Performance per ranks

Figure 5: **Overall VQA performance.** (a) Performance of the largest per family models. The open-model InternVL2.5 78B narrows the gap with closed frontiers models, which dominate the board, while all models show similar trends. Parentheses show avg. open models performance. (b) Average performance of all 54 open models w.r.t. their overall rank shows that physics understanding is progressing much slower. Markers are consistent throughout the study: color identifies model, shape family, size parameter count, and edge color regime. Markers details are in Sec. E.1.

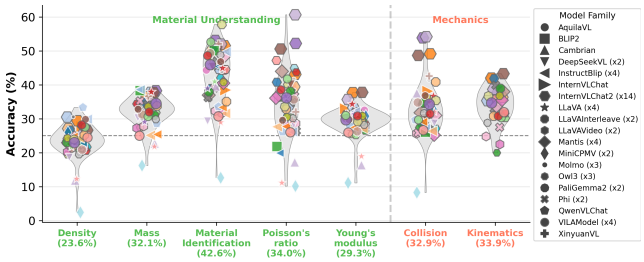


Figure 6: **Detailed physical VQA.** Material understanding is uneven, being dominated by *material identification*, conceptually close to the overwhelming classification task.

viewpoint questions, demonstrating they lack camera-to-scene understanding, leading to sub-random performance, although – similar to temporal tasks – large models (70B+ and close models) are drastically better, presumably because their architectures better capture long-distance correlations.

Perhaps surprisingly, material understanding and mechanics, which testify of core physics principles understanding, are only little lower than spatial reasoning although the latter is overwhelmingly seen in existing vision benchmarks. To investigate why, in Figure 5b we look at 'equally performing' models, scrutinizing all 54 open VLMs, binned by their overall rank on NewtPhys. This highlights two key observations. (i) The best models (rank 1st) perform only +13% and +16% better on material and mechanics than the worse ones (rank 49th), while for other categories their boost goes up to +50%; advocating that *VLMs have poorly progressed in their physical understanding*. (ii) The per-category performance of good models (ranks 11th→1st) is significantly better than those of the largest per family models in Figure 5a, reaching ∼45% for permanence (*vs*. 27.2%) and 40% for material understanding/temporal reasoning (*vs*. 31.2%/28.2%); suggesting that good models are *not well spread* across families and/or that largest models may not perform best. These observations suggest that physics understanding is under-looked by the existing models.

Focusing on physics understanding, Figure 6 shows that sub-category performance is stable across mechanics but not material understanding. Models struggle at estimating *density*, a relatively complex task that requires precise sensing of both volume and material intrinsics, while they perform well on *material identification* which we attribute to the nature of the task – distinguishing materials (*e.g.*, cotton, wood, *etc.*) – conceptually close to the overwhelming classification task in benchmarks. As NewtPhys exposes a greater number and diversity of objects, we dissect performance when varying the objects set in Sec. A.1. It shows that VLMs perform better on softer objects, likely because they expose more visible mechanical cues. Instead, having more objects from $1 \mapsto 10$ little impacts physical understanding, suggesting that even in crowded scenes models identify the objects of interest, but truly struggle at estimating their mechanical or material properties.

4.2 DO VLMS REASON SOLELY FROM LLMS KNOWLEDGE?

A natural question is whether VLM's performance on physical VQAs genuinely reflects visual understanding or merely prior LLM knowledge. Indeed, while Sec. 3 varies object sizes and properties to mitigate such risk, those remain naturally close to their original values to prevent impossible scenarios. Therefore, questions like asking "the mass of a board game" may find a logical answer without need of looking at the visual data. We investigate this with experiments assessing (a) the impact of LLMs knowledge, (b) the ability of VLMs to reason visually.

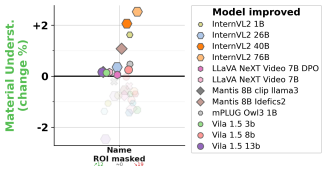


Figure 7: **Masked ROI.** Surprisingly, masking objects improves some models, revealing over-reliance on LLM priors.

(a) To assess LLM bias, we created a 'ROI masked' variation of NewtPhys where regions of interest of questions are masked out (*e.g.*, masking the object whose mass is being asked about). In Figure 7 we report the performance change of 31 video VQA w.r.t. the original VQA. We note that a third of the models *improve* performance when ROI is masked, showing that they *can* solely on LLM knowledge and leakage from the question. Interestingly, most of these "LLM-biased VLMs" are among the large models of our study ($\geq$ 13B params.) which possibly advocates that large models may overly rely on prior knowledge and *answer without looking*.

(b) To further probe the ability to reason visually, we evaluate models on counterfactual questions, therefore *reasoning about unseen events*. Specifically, we use Sec. 3 to generate new sequences, perfectly identical to NewtPhys except for: weaker gravity field or randomly shifting/resizing the object of interest. These sequences serve to evaluate Counterfactual VQA (C-VQA), such as: *Would A collide with B if A was shifted by 1 meter ahead?*. Since not all questions can be expressed counterfactually, we report only valid candidates. From Figure 8, all 54 VLMs perform reasonably well on C-VQA except for 'lower gravity' probably because this is severely OOD for both vision *and* language models – unlikely to have observed non-Earth gravity. Conversely, C-VQA with perturbed objects' location or size is on par or better than factual VQA. Similarly to CoT prompting shown to elicit reasoning in LLMs (Wei et al., 2022), we argue that counterfactual formulations (*What if ..*) may encourage VLMs to reason rather than relying on prior knowledge.

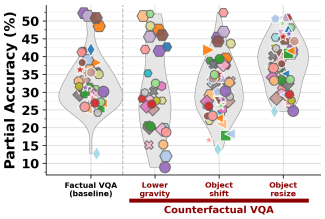
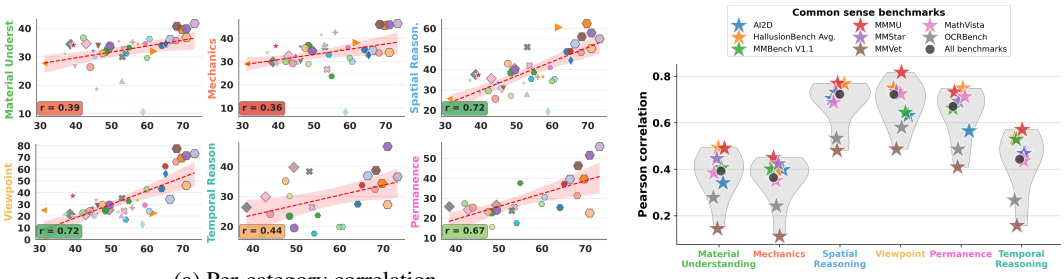


Figure 8: **Counterfactual VQAs.** Models perform reasonably on hypothetical scenarios.

4.3 ARE WE EQUIPPED FOR IMPROVED PHYSICS UNDERSTANDING?

Having observed the physical gap, we question whether the community has the right tools to measure and drive progress in physics understanding. We compare 41 VLM on NewtPhys against eight common-sense benchmarks (Kembhavi et al., 2016; Guan et al., 2024; Liu et al., 2024b; Yue et al., 2024; Chen et al., 2024a; Yu et al., 2024; Lu et al., 2024b; Liu et al., 2024c) using Pearson correlation. While overall NewtPhys performance correlates positively with these benchmarks ($r = 0.67$; *cf.*, Sec. A.2), category-wise diagnosis reveals a sharp divide between low-level physics reasoning and other capabilities. Material understanding and mechanics correlate only weakly ($r = 0.39$ and $r = 0.36$, respectively), compared to spatial reasoning and viewpoint ($r = 0.72$ for both). Thus, current notions of common sense align well with questions on visual attributes (geometry, visibility, *etc.*) but poorly with physics such as forces and collisions. This suggests that existing tools favor general-world understanding, while models still struggle with nuanced low-level physical reasoning.

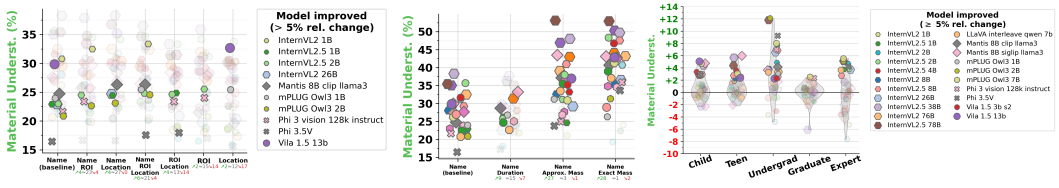
To clarify the picture, we examine benchmark-wise correlations in Figure 9b. Strikingly, all benchmarks weakly correlate ($r < 0.5$) with low-level physics while MMVet (Yu et al., 2024) and OCRBench (Liu et al., 2024c) show even critically low correlation ($r \approx 0.2$), possibly reflecting their object-centric, task-specific nature. For image-only models, some correlations are even negative, suggesting that improvements on those benchmarks favor shortcuts over learning the physical world. On the other hand, MMMU (Yue et al., 2024) and HallusionBench (Guan et al., 2024), which span broader domains and modalities, show stronger correlation with physics ($r \approx 0.45$). This is consistent with the Platonic hypothesis (Huh et al., 2024), which argues that more observations (across data, domain, tasks, *etc.*) may lead to more convergence towards the true model of the world.



(a) Per-category correlation

(b) Per-benchmark correlation

Figure 9: **Common-sense correlation.** (a) Per-category analysis reveals that material understanding and mechanics correlate far less with common sense than other categories, while (b) studying benchmark-level reveals critically low correlations for low-level physics.



(a) Spatial cues

(b) Physical cues

(c) Expert-to-novice prompting

Figure 10: **Effects of cues and formulation.** Across 31 video VLMs, we evaluate (a) spatial cues, which mainly help smaller models; (b) physical cues, which improve most models; and (c) effects of child-to-expert reformulation w.r.t. original NewtPhys. In each plot, stable models are faded out.

From above observations, we conclude that current definitions of "common sense" in the literature are insufficient proxies to unlock true physical understanding.

4.4 Can models perform better on physics?

Without need of finetuning prior works (Shtedritski et al., 2023; Sun et al., 2024) showed that adding cues encourages the VLMs to reason rather than complete. Subsequently, we explore textual/visual cues and question formulation across 31 video VLMs.

Spatial and Physical cues. We evaluate spatial cues by circling the region of interest (`ROI`) or providing its coarse `location`, with or without the object's `name`; yielding 7 VQA variants of $\approx$ 2K questions each. Figure 10a shows only marginal gains confined to smaller models ($\leq$ 8B), suggesting that object's name is sufficient for localization. Conversely, physical cues – `duration`, object's `mass`, or `approximate mass` – improve most models (Figure 10b). Duration helps several strong models, while approx. mass performs similarly to exact mass despite being easier to obtain. Whether these gains improve visual reasoning or merely reinforce language priors remains unclear. See Sec. A.3 for details.

Prompting level. Understanding physical principles may differ from estimating quantities such as density. Infants, for instance, develop a sense of physical plausibility (Baillargeon, 1994) without mastering complex physics metrics. Likewise, VLMs may grasp physics principles yet be limited in their expressivity. Inspired by physics education research distinguishing knowledge novices from expert knowledge (Chi et al., 1981), we explore understanding across a spectrum of expertise ranging from novice to physicist; reformulating questions along five levels (child, teen, undergrad, graduate, expert), each with $\approx$2k VQA. Details are in Sec. D.4.

For video VLMs, Figure 10c reports changes relative to similar questions. Performance consistently peaks at the *Undergrad* level, while expert-like formulations substantially spreads models performance. We hypothesize this reflects training and instruction-tuning biases toward sources such as undergraduate courses and Wikipedia, where much content is written at a comparable level. We also observe that *Graduate* formulations aligns with NewtPhys, showing the smallest change; a reasonable

Supervision type	Model	Objective	Collision	Gravity		Gravity-OOD		Scene Flow
			Fl ↑	mAE↓	magE↓	mAE↓	magE↓	AEE↓
Vision								
Fully-supervised	DeiT III (Touvron et al., 2021)	Classification	48.47	19.34	21.44	15.20	35.43	1.29
	SAM (Kirillov et al., 2023)	Segmentation	54.80	20.73	17.04	16.28	33.98	0.94
	MiDaS (Ranftl et al., 2022)	Depth	54.95	12.12	15.06	8.23	33.80	0.95
Self-supervised	MAE (He et al., 2022)	SSL	28.61	45.69	31.79	42.91	42.50	1.29
	DINO (Caron et al., 2021)	SSL	56.54	13.79	14.92	9.30	33.33	0.94
	DINOv2 Oquab et al. (2024)	SSL	56.52	14.95	14.76	11.02	33.81	0.95
Agglomerative	AM-Radio (Ranzinger et al., 2024)	Distillation	56.96	13.90	13.95	8.25	33.84	0.97
Vision, Language								
Alignment	CLIP (Radford et al., 2021)	Contrastive	53.85	11.30	14.65	7.62	34.35	0.98
	SigLIP (Zhai et al., 2023)	Contrastive	40.91	41.87	27.52	40.19	38.54	1.27
Reconstruction	StableDiffusion (Podell et al., 2024)	Generation	50.39	21.50	21.22	15.64	33.81	1.30

Table 2: **Physics Probing** results for 10 VFMs, grouped by supervision type. Performance is encoded as w o r s e - b e s t.

outcome as our original questions were written by computer vision rather than physicists. This suggests prompt formulation as a simple lever for improving VLM performance on physics VQAs.

4.5 Probing spatial understanding of physics in vision models?

For applications like robotics, pixel-aligned physical predictions, *e.g.* NewtPhys's ground-truth maps (*cf.*, Sec. 3.2), may be more informative. We therefore evaluate Vision Foundation Models (VFMs), large-scale representation learners. Although not designed for physics, their semantic understanding may support physical reasoning and reveal links between visual and physical understanding. Specifically, we introduce *Physics Probing* where lightweight decoders attached to frozen visual VFM encoders learn pixel-level physics maps (*e.g.*, gravity, collision) from our ground truth maps. Given an input frame sequence, the model has to produce pixel-wise quantities for the last frame. We study collision, gravity (magnitude and direction), and scene flow, as they convey key physical cues. For gravity, we additionally evaluate OOD videos with randomized gravity magnitudes in $[0.98m/s^2 - 20m/s^2]$, different from training videos that always use Earth gravity of $9.81m/s^2$.

For experiments, we look at a set of ten prominent VFMs detailed in Sec. E, and report results in Tab. 2 grouping by their learning objective. Among vision models, we observe that *self-supervised* DINO models dominates the board, demonstrating the effectiveness of its training strategy. In the *fully-supervised* group, MiDaS dominates which reflects with its pixel-level depth supervision, a somehow physical task, that transfers to other physical quantities. The *agglomerative* AM-Radio distillation also preserved physical abilities, but has a prohibitive inference time. For VLMs, *alignment* dramatically outperforms *reconstruction*, especially with CLIP encoder, which speculatively could stem from its training data. Across the board, scene flow tends to correlate with physics probing, consolidating the intuition that motion is a viable proxy for physical reasoning.

On Gravity-OOD results, we observe angular prediction performance (mAE) comparable to Gravity, as expected since direction does not change, but we note a severe degradation in magnitude (magE) indicating models severely overfit to Earth gravity – their sole observation point.

Our early empirical results indicate that stronger visual representations may be more transferable to physics, though our observations are preliminary and we encourage future works exploring alternative supervision strategies, as well as other available ground-truth maps in NewtPhys.

5 Discussion

Our findings position NewtPhys as more than a VQA benchmark. Its controlled simulations and dense, temporally aligned annotations enable evaluation and training at both answer and pixel levels, counterfactual testing. Besides, it is designed as an extensible framework, allowing for research to construct new tasks, filter scenes by any physical events or else. Yet, despite being proven richer and more realistic than existing studies, along all measured dimensions, rendering and simulator assumptions introduce a simulation-to-real gap, and controlled human studies could further establish reference performance. Future work should investigate whether training with dense physical supervision improves transfer to real-world dynamics. Ultimately, progress requires models that ground predictions in visual evidence and generalize beyond familiar physical regimes. We hope that NewtPhys drives this progress by making rich physical supervision available to the community.

AI USE STATEMENT

In this work, VQA questions were initially drafted by hand and polished with generative AI tools, and subsequently reviewed and revised by at least two authors. Generative AI tools were also used to suggest questions at different prompting levels; these suggestions were likewise reviewed and revised by at least two authors. We further used generative AI tools to propose physical quantities for GSO objects, which were subsequently checked by two authors, and their validity assessed in the appendix. Generative AI tools were *not* used to create visual content included in the dataset. Additionally, we used generative AI tools to assist with code development; all AI-assisted code was carefully reviewed, verified, and tested for correctness by two authors. As for writing, LLMs models were only used to compact and polish the text. The authors take full responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the assistance of generative AI.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, and Et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 50

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 50

Renée Baillargeon. Physical reasoning in young infants: Seeking explanations for impossible events. *British Journal of Developmental Psychology*, 1994. 1, 8

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*, 2013. 1

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 1

Daniel Bear, Elias Wang, Damian Mrowca, Felix Jedidja Binder, Hsiao-Yu Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin A Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. In *NeurIPS*, 2021. 2

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 50

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, 2020. 5

Susan Carey. The origin of concepts. *Journal of Cognition and Development*, 2000. 1

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 9

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024a. 2, 7

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b. 50

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024c. 5, 50

Michelene TH Chi, Paul J Feltovich, and Robert Glaser. Categorization and representation of physics problems by experts and novices. *Cognitive science*, 1981. 8

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *ICLR*, 2025. 2, 3

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 50

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. In *CVPR*, 2024. 5, 50

Michael Dorkenwald, Nimrod Barazani, Cees GM Snoek, and Yuki M Asano. Pin: Positional insert unlocks object localisation abilities in vlms. In *CVPR*, 2024. 17

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, 2022. 3, 4, 18

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *CVPR*, 2024. 2

Clement Fuji Tsang, Maria Shugrina, Jean Francois Lafleche, Or Perel, Charles Loop, Towaki Takikawa, Vismay Modi, Alexander Zook, Jiehan Wang, Wenzheng Chen, Tianchang Shen, Jun Gao, Krishna Murthy Jatavallabhula, Edward Smith, Artem Rozantsev, Sanja Fidler, Gavriel State, Jason Gorski, Tommy Xiang, Jianing Li, Michael Li, and Rev Lebaredian. Kaolin: A pytorch library for accelerating 3d deep learning research. https://github.com/NVIDIAGameWorks/kaolin, 2024. 17

Hao Gao, Shaoyu Chen, Bo Jiang, Bencheng Liao, Yiang Shi, Xiaoyang Guo, and Yuechuan Pu. Rad: Training an end-to-end driving policy via large-scale 3dgs-based reinforcement learning. In *NeurIPS*, 2025. 3

Google DeepMind. Gemini 3.1. https://deepmind.google, 2026. Proprietary model; no public technical report. 5, 50

Cylingo Group. Xinyuan-vl-2b: A high-performance multimodal large model for the end-side. https://huggingface.co/Cylingo/Xinyuan-VL-2B, 2024. 50

Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. Infinity-MM: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*, 2024. 50

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024. 7

Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *CVPR*, 2024. 17

Antoine Guédon, Diego Gomez, Nissim Maruani, Bingchen Gong, George Drettakis, and Maks Ovsjanikov. Milo: Mesh-in-the-loop gaussian splatting for detailed and efficient surface reconstruction. *TOG*, 2025. 17

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 9

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 4, 19, 51

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *ICML*, 2024. 7

Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H. Adelson. Learning visual groups from co-occurrences in space and time, 2015. 1

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. MANTIS: Interleaved multi-image instruction tuning. *TMLR*, 2024. 50

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 7

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TG*, 2023. 2, 3, 19

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, 2023. 9

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In *ICML*, 2025. 3, 4

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. 50

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 50

Tianyu Li, Li Chen, Caojun Wang, Haochen Liu, Kashyap Chitta, Zhenjie Yang, Yuhang Lu, Naisheng Ye, Yihang Qiu, Yufei Wang, et al. World engine: Towards the era of post-training for autonomous driving. *arXiv preprint arXiv:2606.19836*, 2026. 3

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 3

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *CVPR*, 2024. 50

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024. 3, 4

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 5, 50

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024a. 50

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024b. 7

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 2024c. 7

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024a. 5, 50

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024b. 7

Michael McCloskey, Allyson Washburn, and Linda Felch. Intuitive physics: the straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1983. 1

Vismay Modi, Nicholas Sharp, Or Perel, Shinjiro Sueda, and David IW Levin. Simplicits: Mesh-free, geometry-agnostic elastic simulation. *TOG*, 2024. 3, 4, 17, 18

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *CVPR*, 2026. 3

NVIDIA, Yulong Cao, Riccardo de Lutio, Sanja Fidler, Guillermo Garcia Cobo, Zan Gojcic, Maximilian Igl, Boris Ivanovic, Peter Karkus, Janick Martinez Esturo, Marco Pavone, Aaron Smith, Ellie Tanimura, Michal Tyszkiewicz, Michael Watson, Qi Wu, and Le Zhang. Alpasim: A modular, lightweight, and data-driven research simulator for autonomous driving, 2025. 3

OpenAI. Gpt family. https://openai.com, 2026. Proprietary models; no public technical report. 5, 50

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 9

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 19

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 9

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 9

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2022. 9

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, 2024. 9

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys 2019: A benchmark for visual intuitive physics understanding. *TPAMI*, 2021. 2

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *ICCV*, 2023. 8, 16

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *CVPR*, 2024. 8

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 2011. 1

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024. 2, 50

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 9

Antonios Tragoudaras, Chenyu Zhang, Daniil Cherniavskii, Antonios Vozikis, Thijmen Nijdam, Derck WE Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Evaluating newtonian mechanics in video generative models with real physical systems. In *ICML*, 2026. 3

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. In *NeurIPS*, 2023. 3

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *TPAMI*, 2002. 3

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 3

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 7

Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *NeurIPS*, 2015. 2

Jiajun Wu, Joseph J Lim, Hongyi Zhang, Joshua B Tenenbaum, and William T Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, 2016. 2

Haotian Xue, Yunhao Ge, Yu Zeng, Zhaoshuo Li, Ming-Yu Liu, Yongxin Chen, and Jiaojiao Fan. Point-it-out: Benchmarking embodied reasoning for vision language models in multi-stage visual grounding. *arXiv preprint arXiv:2509.25794*, 2025. 17

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 50

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *ICLR*, 2025. 50

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 *ICLR*, 2020. 2

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*, 2024. 7

Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdzal, and Adriana Romero-Soriano. Inference-time physics alignment of video generative models with latent world models. In *CVPR*, 2026a. 2, 4, 20

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H. Chan. NewtonGen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. In *ICLR*, 2026b. 3

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 7

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 9

Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. A general protocol to probe large vision models for 3d physical understanding. In *NeurIPS*, 2024. 2

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model, 2024. 50

APPENDIX

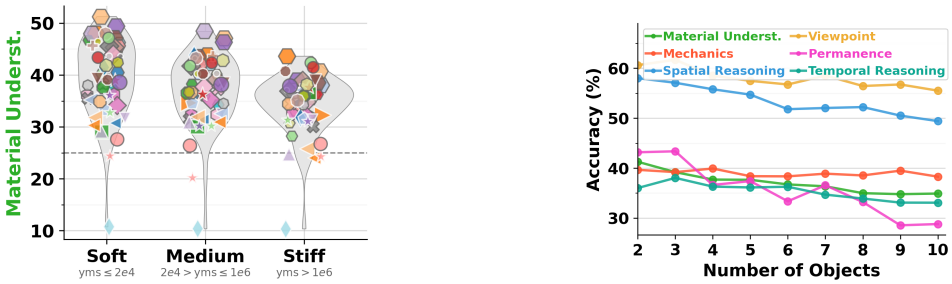
The appendix provides further analyses (Sec. A), along with more details on the Newtonian physics simulator (Sec. B), NewtPhys dataset (Sec. C), and thorough details on VQA taxonomy, examples, counterfactual, expert-to-novice, and automation (Sec. D). Finally, it includes the list of model specifications and detailed markers legend (Sec. E).
We refer to the attached website for better visualization of NewtPhys.

A Further analyses . . . 15
A.1 Performances with varying objects sets . . . 15
A.2 Common sense correlation . . . 16
A.3 Digging the effect of additional cues . . . 16
B Details on Newtonian physics simulation . . . 17
B.1 Solver adaptation. . . . 17
B.2 Estimation of objects' physical properties. . . . 18
B.3 Rendering . . . 19
C Details on NewtPhys dataset . . . 19
C.1 Dataset statistics . . . 19
C.2 Evaluation details . . . 19
C.3 Additional qualitatives. . . . 21
D Details on Visual Question Answering (VQA) . . . 21
D.1 Taxonomy . . . 22
D.2 VQA examples . . . 26
D.3 VQA counterfactual . . . 41
D.4 Expert-to-novice specification . . . 45
D.5 VQA automation . . . 48
E Models specification. . . . 49
E.1 Vision-Language Models . . . 49
E.2 Vision-Foundation Models . . . 50

A FURTHER ANALYSES

A.1 PERFORMANCES WITH VARYING OBJECTS SETS

Benefiting from the rich attributes in NewtPhys, we dissect the performance when varying object's softness or number of objects. Results in Figure 11a indicate that softer objects (*i.e.*, low Young's modulus; YMS) lead to better accuracy w.r.t. scenes with stiff objects. This is partially explained by the exponential growth of the YMS scale, which makes the difference between "stiff" and "super-stiff" visually ambiguous unless a significant mechanical constraint is observed. Conversely, as soft objects deform easily upon contact, it provides a critical visual cue to estimate their material properties like density, Poisson's ratio, *etc*. In Figure 11b we report performance for scenes with varying numbers of objects. While in general performance is relatively stable ($\sim$ 10 rel.% drop) from simple to crowded scenes, some categories like permanence exhibit a bigger 30 rel.% drop showing that memorization



(a) Variation per YMS

(b) Objects variations

Figure 11: **Effect of varying objects sets.** (a) Varying object softness, we notice that VLMs perform better on soft objects. (b) Varying the number of objects shows that some categories are more stable than others.

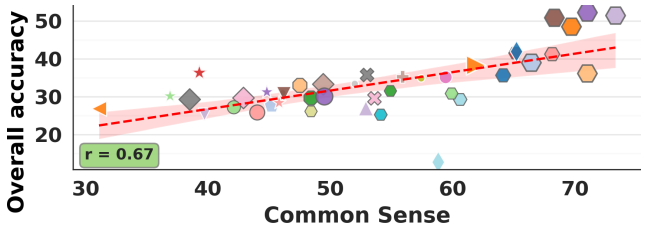


Figure 12: **Common-sense correlation.** Overall accuracy by averaging different benchmarks on common sense. Pearson correlation is 0.67, indicating a positive correlation with the average of the benchmarks on common sense.

of complex scenes is harder. Conversely, we observe a minimal drop for material understanding and temporal reasoning, suggesting that the main challenge indeed lies in the understanding of the material/mechanical properties, rather than in the identification of the objects of interest.

A.2 Common sense correlation

As mentioned in Sec. 4.3, we present here the *overall accuracy* correlation between models performance on NewtPhys and the 8 common-sense benchmarks of the main paper. Overall accuracy correlates well, having a Pearson $r = 0.67$, with models' performance on common-sense benchmarks. The latter are typically used are typically probing visual attributes like object size, position, visibility, and camera characteristics, but omit low-level and mechanical physics, as mentioned in the main work.

Notably, we highlight that for 13 of our 54 open VLMs we could not find their performance on the above mentioned common sense benchmarks. This explains why only 41 VLMs are reported in the common sense analysis in Sec. 4.3.

A.3 Digging the effect of additional cues

In Sec. 4.4 we have explored the addition of spatial of physical cues to improve VQA. We further analyze this in this section.

As evidenced by results in Sec. 4.2, we observe that objects' names already provide a strong cue, although well-functioning VLMs still have to locate objects before estimating their physical properties. To ease this, we introduce spatial cues either in the visual form by circling the Region of Interest (*ROI*) (Shtedritski et al., 2023) or in the textual form by providing coarse *Location* of objects in the image (*e.g.*, 'top-right'). This setup leads to seven variants, each with $\approx$ 2k VQA pairs. While all 31 video VLMs are evaluated, for clarity, in Figure 10a we emphasize only models whose performance improved by at least five percent w.r.t. to the baseline VQA using only objects *Name*. We note that only five families benefit from spatial cues, and models that improved are typically smaller ones

(mainly $\leq$ 8B). This corroborates the observed ability of large VLMs to spatially locate objects in images (Dorkenwald et al., 2024; Xue et al., 2025). Specifically, mPlug 2B and InternVL2 1B get a large $+5$ points boost with all three combined cues. Besides localization, we also explore how physical cues affect physical reasoning, adding to each question hints about *duration* or objects *approximate* or *exact mass*. Results in Figure 10b highlight that almost all models are subsequently boosted. Of note, adding *duration* (*e.g.*, "In this sequence of 3.5 seconds, ...") improves a few of the large and best performing models, an interesting insight as duration is virtually free to provide. Interestingly, providing *approximate mass* (*e.g.*, "... object of approximate 2kg mass ...") brings a similar boost with the *exact mass* (*e.g.*, "... object of 2.35kg ...") though easier to provide. We however highlight that this is not sufficient to assess whether these physical cues help the model to reason visually or simply strengthen priors to the LLM.

B Details on Newtonian physics simulation

In the following, we provide additional details about the physical simulator, mentioned in main paper Sec. 3.1, which constitutes the backbone of our benchmark.

The key complexity of simulating 3DGS $G = \{(\mu_i, \sigma_i, c_i, \alpha_i)\}_{i=1}^{N}$ is that they encode sparse and unstructured information, as opposed to meshes, which are structured 3D data, typically preferred for simulation. We highlight that while prior works (Guédon and Lepetit, 2024; Guédon et al., 2025) demonstrated the ability to optimize meshes with 3DGS, our experiments show that they hardly generalize across scenes and objects, and that even a slightly noisy mesh can produce highly unrealistic simulation with mesh-dependent simulators.

We instead rely on Simplicits (Modi et al., 2024), an extension of the Nvidia Kaolin (Fuji Tsang et al., 2024), which enables mesh-free simulator for time-varying elastodynamics by learning reduced deformation space of complex shapes. We careful extend the latter to allow handling large scenes ($\approx 10^7$ particles) with up to 50 objects.

B.1 Solver adaptation.

Let us consider a single deformable object with particles $x \in \mathcal{X}$. Rather than simulating all particles, Simplicits models a simulation state as a time-varying vector $\mathbf{z}_t$ having much reduced Degrees of Freedom (DoFs) w.r.t. $\mathcal{X}$. The next state is estimated using a Newton-based solver with the following optimization objective:

$$\mathbf{z}_{t+1} = \arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{z} - \bar{\mathbf{z}}_t\|^2_{\mathbf{M}} + h^2 E_{\text{pot}}(\mathbf{z}), \tag{1}$$

where $||\cdot||^2_{\mathbf{M}}$ is the squared norm weighted by the mass matrix $\mathbf{M}$, h the simulation timestep, $\bar{\mathbf{z}}_t$ is the first order predictor for $\mathbf{z}$, and E_{pot} is the potential energy of the system. We note that collision and forces constraints are also added to Equation (1). The mapping $\phi(Z) \mapsto X$ is learned with a small skinning network $\phi(\cdot)$ as-yet undefined, given a fixed number of deformable handles and physical properties. Having no mesh, the mass and potential energy E_{pot} are approximated by *cubature points* randomly sampled in X, *i.e.*, $Q \in X$. Intuitively, cubature points serve as locations for solving the physical forces of the system.

To simulate a full scene we consider only the centers of our 3DGS particles $X = \{\mu_i\}_{i=1}^{N}$ as our physical world, and apply crucial adjustments. Instead of individual object, we define our simulation state as the union of all objects in the scene such that $\mathbf{z}_{\text{sim}} = \{\mathbf{z}_{\text{scene}}, \mathbf{z}_{\text{obj}_1}, \ldots, \mathbf{z}_{\text{obj}_N}\}$. By keeping track of the mapping between particles and objects, the complete 3DGS state can then be updated after each simulation step by applying the relevant $\phi(\cdot)$ mapping for each object: $G = \{\phi_{\text{scene}}(\cdot), \phi_{\text{obj}_1}(\cdot), \ldots, \phi_{\text{obj}_N}(\cdot)\}$. Further, since highly deformable objects intuitively require more DoFs, rather than using a fixed set of handles, we logarithmically vary the number as a function of each object softness. This reduces the rank of $\mathbf{z}_{\text{sim}}$ and significantly lowers the complexity of solving Equation (1). Another source of computational cost is the large number of cubature points needed for accurate simulation. Using only a few cubature points leads to penetration/collision of objects due to lack of physics evaluation basis, while large numbers bring intractable computational costs for our large scenes. Instead, we employ a simple strategy that primarily samples cubature points near the object spawning positions where fine-grained physics interaction is likely to occur.



(a) **YMS shift** (b) **Object scale**

Figure 13: **Robustness to intrinsic perturbations.** We perturb a Young's Modulus and b object scale and observe that VQA accuracy remains largely stable, indicating that our original intrinsics are coherent.

Lastly, to compensate for sparse evaluation basis, each cubature point is modeled as a small sphere rather than a point location.

Assessing correctness of the simulation. By design, simulation requires 3DGS primitives to lie on object surfaces. For objects, because the GSO dataset provides all-around views, the resulting 3DGS reconstructions are highly accurate; leading to only 2.48% error in dimensions upon manual verifications on 20 objects. Instead for the scene reconstructions, the combination of VGGT and COLMAP enables high-density as well as accurate 3DGS reconstruction. Furthermore, our simulations account for sparsity by treating primitives as small spheres (r=3mm), thereby theoretically ensuring tight 6 mm-precision collisions. These choices, together, ensure primitives on the surface, enabling our large-scene, automated simulation pipeline.

We highlight that, without noticeable losses in quality, our adaptation drastically lowers the complexity of each simulation which typically have 10^7 3DGS to a simulation state integrated over $\approx 10^4$ cubature points with only $\approx 10^2$ DoFs. Further, the object-to-particle mapping allows retrieving per-point forces which is crucial for our need of physically-annotated dataset.

B.2 Estimation of objects' physical properties.

While the scene itself is kinematic (*i.e.*, non-deformable) the objects need tailored skinning functions for deformation. We follow (Modi et al., 2024), first densifying 3DGS models to compensate for 3DGS only capturing visible shell, and then learning skinning functions from small individual networks (10 MLP layers) given the densified 3DGS centers (*i.e.*, rest pose) and physics properties. The training objective minimizes the elastic energy w.r.t. the rest pose while enforcing orthogonality of the skinning weights. We highlight that GSO (Downs et al., 2022) does *not* provide physical quantities[1]. Therefore, for each object, we estimate the possible Young modulus, Poisson ratio, and density by querying GPT-5 with 4 object views, while also querying for names of the visible materials. Object volumes are then computed with a greedy Monte Carlo from our metric-scaled 3DGS, which allows accurate estimation of their masses. All properties are manually verified.

Assessing properties correctness. In order to assess the correctness of properties, we manually verified vendor properties by scraping the web for 20 GSO objects. This further check shows that our estimated properties yield an average error of 14.58% for mass (which accounts for both dimension and density), and 36.09% for volume – so that our properties are reliably close. The volume error can be explained by the cubic nature of the metric, and the Monte-Carlo volume estimation approximating discrete points as small spheres, which biases volumes of small objects. Furthermore, to assess the impact of our intrinsics, we perturb the Young's Modulus (YMS) in Figure 13a and scale in Figure 13b, of all models on subsets of NewtPhys. For both our 'original' intrinsics, we observe that the modification leads to coherent VQA accuracy that does not change significantly under perturbation.

[1]Note that while GSO mentions the release of physical labels, the absence of the latter was confirmed via email by the Google GSO team.

B.3 RENDERING

At each time step, the Simplicits routine resolves all physics constraint with a Newton solver while saving only the resulting particles states.

RGB rendering. To render the simulation, we update 3DGS positions and orientations from the current simulation state $\mathbf{z}_t$, using the objects skinning functions, and simply render with the vanilla 3DGS rasterizer (Kerbl et al., 2023).

Ground-truth maps. With some engineering efforts, we capture individual forces (material stress, gravity, collision, *etc*.) along with other kinematics, semantics and geometric labels, by modifying the Newton solver routine in Simplicits.

Rendering per-pixel forces from camera perspective is complex. We do so by duplicating the 3DGS renderers, obtaining one renderer per force to extract (and one for RGB). We then use each *force renderer* as a proxy to store the per-point force value of each Gaussian mapped to spherical harmonics (*i.e*., ultimately to RGB) after careful binarization of the Gaussians' opacity[2]. Importantly, forces renderers do *not* use the skinning functions, as the latter are only valid in the spatial domain. Instead, we use a nearest-neighbour cubature to 3DGS mapping. Subsequently, each 3DGS particle is rendered as having the force value of the closest cubature point. Given that our objects are relatively small in size and uniformly covered by cubature points, the resulting approximation is found to be negligible. A similar process is followed to render kinematics, semantics, and geometry maps.

Events recording. Alongside RGB and physical maps, events like collisions, camera motion, object visibility, *etc*. are capture at each rendering steps, and encoded as a JSON file which later allows, along with the physical maps, for automatic scripting of Visual Question Answering (VQA).

Ground truth consistency. Since labels are rendered directly from the simulator state used to generate the observations as described above, *they are always perfectly aligned with the observed dynamics*. Thus, the labels remain consistent with the simulated scene, independently of the simulator's intrinsic physical accuracy.

C DETAILS ON NEWTPHYS DATASET

C.1 DATASET STATISTICS

We report the statistics of NewtPhys dataset in Figure 14. Compared to prior physics datasets seen in Figure 2, the statistics of our dataset exhibit more cluttered scenes (top, '*Objects per frame*') from moving cameras, and ≈15 distinct collisions per sequence on average (top center). The objects span a wide range of semantics (top right, '*Object type*' pie chart) and materials (right, '*Material*' pie chart) and are simulated with a wide range of physical properties (middle), notably including highly deformable objects with low Young's Modulus (middle left, '*Object Young's Modulus*'). These statistics the high diversity and variability of scenes and dynamics.

C.2 EVALUATION DETAILS

We detail here the experimental protocol and metrics used to evaluate the fidelity of NewtPhys in Tab. 1b.

FID / KID. To compute FID/KID metrics, for each dataset we extract up to 32 individual frames per sequence. Using (Parmar et al., 2022) we evaluate both metrics at 299x299 using 10,000 random frames w.r.t. PisaBench (real set) using the same number of frames.

VBench. We utilize VBench Huang et al. (2024) and report in Tab. 1b the average of all 6 no-reference metrics (aesthetic quality, background consistency, dynamic degree, imaging quality, motion smoothness, subject consistency). To do so, we use 500 random videos per dataset, resized at

[2]We highlight that binarization is required. Failure to do so, would lead to integration of multiple force values along a single ray, making per pixel invalid, ergo, erroneous physical maps.

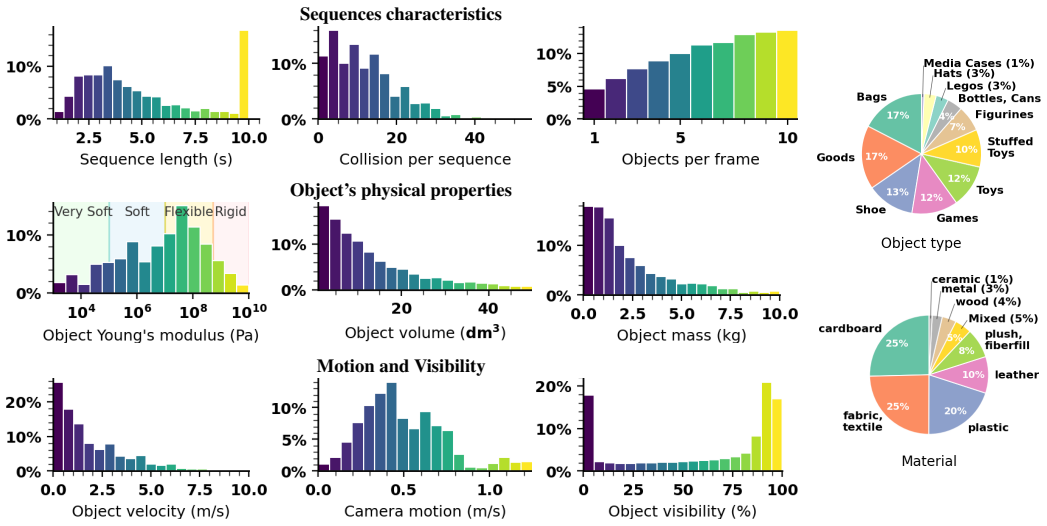


Figure 14: **Detailed datasets statistics.** We report here additional dataset statistics, highlighting the variability of NewtPhys dataset along various axes of study.

Dataset	Visual realism			Physical realism		
	FID ↓	KID ↓	VBench ↑	Mean ↓	Max ↓	Top-3 ↓
ContPhy	272.59	0.1665	0.6913	0.4228	0.4616	0.4469
IntPhys	251.60	0.1556	0.6584	0.4018	0.4384	0.4277
IntPhys2	215.11	0.0951	0.6877	0.4274	0.4935	0.4767
PhysBench	216.17	0.1075	0.6799	0.4237	0.4629	0.4491
Physion++	193.82	0.1185	0.6442	**0.3873**	0.4370	0.4254
NewtPhys	**189.68**	**0.0783**	**0.6970**	0.4105	**0.4287**	**0.4216**
PisaBench	n/a	n/a	0.6914	0.3888	0.4017	0.3931
Physics-IQ	221.18	0.0909	0.6820	0.3905	0.4527	0.4404

Table 3: **Detailed dataset metrics.** Compared to Tab. 1b, we additionally report the mean VJEPA surprise and Physics-IQ. We refer to text for further details.

512x512 resolution (preserving aspect ratio via crop-center if needed), and crucially reinterpreting the videos at 30 fps, *without any temporal interpolation* (*i.e.*, duplicating or dropping frames if needed). Some datasets have less than 500 videos: PhysicsIQ-verified (198) and PisaBench (361) have less than 500 videos, so we use all the ones available. We also measured the metrics with more videos, but found 500 to be enough, leading to similar results, while minimizing the gap with datasets having fewer videos.
Of note, because VBench measures 'dynamic degree' which also captures camera motion, some datasets like PisaBench are somehow penalized because they use a static camera.

VJEPA surprise. For physical realism, we leverage VJEPA Surprise (Yuan et al., 2026a). The latter is computed by sampling temporal windows, and measuring the surprise (the difference between the latent predicted using temporally-masked frames, and the latent observed with full frames). Authors demonstrated that VJEPA surprise correlates with physical plausibility and can be reliably used as a signal for improvement.
We use the implementation of Yuan et al. (2026a), relying on a ViT-Huge, and process videos at 256x256 resolution (using aspect-ratio resize and crop-center), with a temporal window of size $W = 48$ (temporal stride of 8) using the 24 first frames as context for the prediction. Crucially, like for VBench, we reinterpret videos at 30 fps, *without any temporal interpolation* (*i.e.*, duplicating or dropping frames if needed). While undocumented, we found crucial to reinterpret videos correctly at 30 fps as datasets have a wide range of fps (roughly, $5 \rightarrow 120$) since no reinterpretation induces dramatically different temporal horizon with a fixed temporal window size that greatly affect the surprises. We also note that our choice of 30 fps somehow penalizes NewtPhys (originally at 25 fps) since no-interpolation reinterpretation duplicates frames which somehow impact our physical plausibility. Yet, we found this to be the most fair choice for all datasets.

In the main table, for space reasons, we report only the average of the top-3 and maximum per-sequence surprises. We argue, these two metrics best capture how physically realistic is a sequence, as they express the maximum surprise (*i.e.*, physical inconsistency) observed by the VJEPA model. For completion, in Tab. 3 we also report the *mean* of the per-sequence VJEPA surprises. Notably, we consider this metric to be noisier because: (a) sequences have dramatically different lengths (from 1 sec. to 30 sec.), (b) surprise peaks upon objects occlusion or reappearance, which penalize datasets like NewtPhys having complex and cluttered scenes, with many occlusions. Additionally, in Tab. 3 we also report metrics on the real Physics-IQ verified data, for completeness. We highlight its high surprise score because VJEPA measures body dynamics but not optical effects (shadows, reflections, *etc*.) or changes related to thermodynamics, fluids, which are predominant in Physics-IQ and produce a stronger surprise.

C.3 Additional qualitatives.

Figure 15 shows additional exemplar sequences with physical annotations. For better forces visualization, we highlight that renderings are taken from stationary cameras and faded gray. **We refer to the website for full quality illustration of simulation sequences.**

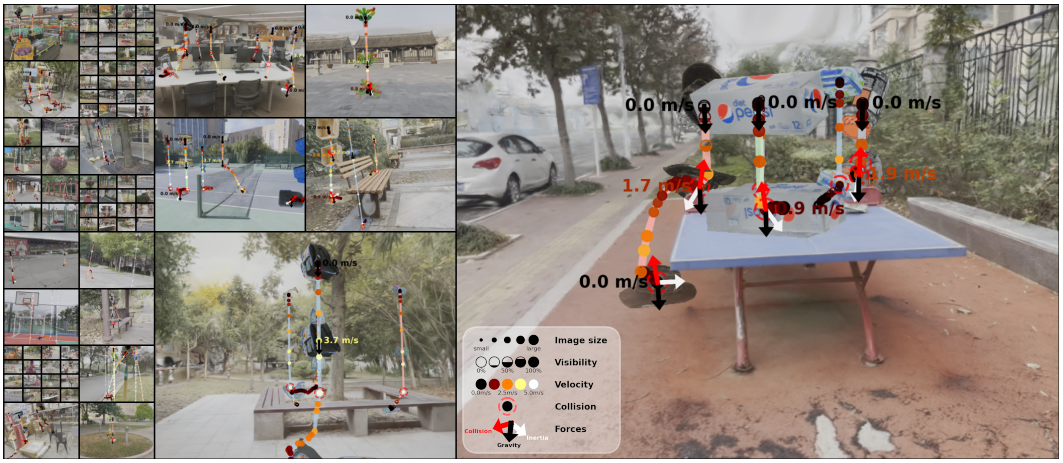


Figure 15: **NewtPhys additional sequences**. We display sequences from our NewtPhys benchmark, though rendered from a stationary viewpoint for visualization purpose. Dynamics are highlighted as overlaid (cf. inset legend).

D Details on Visual Question Answering (VQA)

In the following, we detail the taxonomy used for NewtPhys VQA (Sec. D.1), along with exemplar VQA with visuals (Sec. D.2), and details on our automation process (Sec. D.5).

D.1 TAXONOMY

We design multiple-choice questions (MCQ) having 4 answers, grouped into five high-level categories. Each question is instantiated from a template by replacing placeholders (*e.g.*, `<OBJECT>`, `<OBJECT_1>`, `<OBJECT_2>`) with scene-specific object names. We also define task splits: *single*-frame (answerable from one frame) and *multi*-frame (requires temporal evidence across a sequence). We try to use different scenes for each question to showcase the variety of the dataset.

Table 4: **VQA templates**. Sample VQA question IDs grouped by category and sub-category.

Category	Sub-Category	Question ID and Prompt
Spatial Reasoning	Distance	**F_DISTANCE_OBJECT_OBJECT** **Single Image:** Based on the image, what is the real-world distance between <OBJECT_1> and the <OBJECT_2>? **Multi Image:** Considering all frames, what is the real-world distance between <OBJECT_1> and the <OBJECT_2> in the last frame?
		F_CLOSEST_OBJECT_OBJECT **Single Image:** Based on the image, which object is closest to the <OBJECT> in real-world distance? **Multi Image:** Considering all frames, which object is closest to the <OBJECT> in real-world distance in the last frame?
		F_DISTANCE_OBJECT_CAMERA_DISTANCE **Single Image:** Based on the image, what is the real-world distance between the <OBJECT> and the camera? **Multi Image:** Considering all frames, what is the real-world distance between the <OBJECT> and the camera in the last frame?
		F_CLOSEST_OBJECT_CAMERA **Single Image:** Based on the image, which object is the closest to the camera in real-world distance? **Multi Image:** Considering all frames, which object is the closest to the camera in real-world distance in the last frame?
	Size	**F_SIZE_OBJECT** **Single Image:** Based on the image, what are the real-world dimensions of the <OBJECT>? **Multi Image:** Considering all frames, what are the real-world dimensions of the <OBJECT> in the last frame?
		F_SIZE_OBJECT_BIGGER **Single Image:** Based on the image, which single object has the biggest real-world volume? **Multi Image:** Considering all frames, which single object, visible in the last frame, has the biggest real-world volume?
	Layout	**F_LAYOUT_POSITION_OBJECT_OBJECT** **Single Image:** From the camera's perspective, where is the <OBJECT_1> relative to the <OBJECT_2> in the image? **Multi Image:** Considering all frames, from the camera's perspective, where is the <OBJECT_1> relative to the <OBJECT_2> in the last frame?
Mechanics	Kinematics	**F_KINEMATICS_SPEED_OBJECT** **Multi Image:** Considering all frames, what is the real-world speed of the <OBJECT> at the time of the last frame?
		F_KINEMATICS_ACCEL_OBJECT **Multi Image:** Considering all frames, what is the real-world magnitude of acceleration of the <OBJECT> at the time of the last frame?
		F_KINEMATICS_DISTANCE_TRAVELED_INTERVAL **Multi Image:** Considering all frames, what is the real-world displacement of the centroid of the <OBJECT> from the first to the last frame?
		F_KINEMATICS_SYSTEM_STABILITY **Multi Image:** Analyzing the motion trend across the sequence, which statement best describes the system's state at the final frame?

Continued on next page

Category	Sub-Category	Question ID and Prompt
	Collision	**F_COLLISION_OBJECT_OBJECT_FRAME_SINGLE** **Single Image:** Based on the image, which visible object is the <OBJECT> colliding with?
		F_COLLISION_OBJECT_OBJECT_FRAME_MULTI **Multi Image:** In which frame is the <OBJECT> most likely colliding with another object?
		F_COLLISION_OBJECT_SCENE_FRAME_MULTI **Multi Image:** In which frame is the <OBJECT> most likely colliding with the static scene?
Material Understanding	Mass	**F_MASS_OBJECT** **Single Image:** Based on the image, what is the mass of the <OBJECT>? **Multi Image:** Considering all frames, what is the mass of the <OBJECT>?
		F_MASS_HEAVIEST_OBJECT **Single Image:** Based on the image, which visible single object has the greatest mass? **Multi Image:** Considering all frames, which single object, visible in the last frame, has the greatest mass?
		F_MASS_LIGHTEST_OBJECT **Single Image:** Based on the image, which visible single object has the least mass? **Multi Image:** Considering all frames, which single object, visible in the last frame, has the least mass?
	Density	**F_PHYSICS_PROPERTY_DENSITY_OBJECT** **Single Image:** Based on the image, what is the estimated mean density of the <OBJECT>? **Multi Image:** Considering all frames, what is the estimated mean density of the <OBJECT>?
		F_PHYSICS_PROPERTY_DENSITY_OBJECT_RELATIVE **Single Image:** Based on the image, which visible object has the highest effective density? **Multi Image:** Considering all frames, which object, visible in the last frame, has the highest effective density?
	Young Modulus	**F_PHYSICS_PROPERTY_YOUNG_MODULUS_OBJECT_METRIC_PREFIX** **Single Image:** Based on the image, what is the Young's modulus of the <OBJECT>? **Multi Image:** Considering all frames, what is the Young's modulus of the <OBJECT>?
		F_PHYSICS_PROPERTY_YOUNG_MODULUS_OBJECT_SIMILAR **Single Image:** Based on the image, which visible object has a Young's Modulus most similar to that of the <OBJECT>? **Multi Image:** Considering all frames, which object, visible in the last frame, has a Young's Modulus most similar to that of the <OBJECT>?
		F_PHYSICS_PROPERTY_YOUNG_MODULUS_OBJECT_SIMILAR_NON_TECHNICAL **Single Image:** Based on the image, which visible object has a softness most similar to that of the <OBJECT>? **Multi Image:** Considering all frames, which object, visible in the last frame, has a softness most similar to that of the <OBJECT>?
		F_PHYSICS_PROPERTY_YOUNG_MODULUS_HIGHEST **Single Image:** Based on the image, which visible object exhibits the highest Young's Modulus? **Multi Image:** Considering all frames, which object, visible in the last frame, exhibits the highest Young's Modulus?
		F_PHYSICS_PROPERTY_YOUNG_MODULUS_HIGHEST_NON_TECHNICAL **Single Image:** Based on the image, which visible object is the stiffest? **Multi Image:** Considering all frames, which object, visible in the last frame, is the stiffest?
		F_PHYSICS_PROPERTY_YOUNG_MODULUS_OBJECT_HIGH_LEVEL **Single Image:** Based on the image, which attribute best describes the <OBJECT> in terms of deformability? **Multi Image:** Considering all frames, which attribute best describes the <OBJECT>, visible in the last frame, in terms of deformability?
	Poisson Ratio	**F_PHYSICS_PROPERTY_POISSON_RATIO_OBJECT** **Single Image:** Based on the image, what is the Poisson ratio of the <OBJECT>? **Multi Image:** Considering all frames, what is the Poisson ratio of the <OBJECT>?

Continued on next page

Category	Sub-Category	Question ID and Prompt
		F_PHYSICS_PROPERTY_POISSON_RATIO_OBJECT_SIMILAR **Single Image:** Based on the image, which visible object has a Poisson ratio most similar to that of the <OBJECT>? **Multi Image:** Considering all frames, which visible object has a Poisson ratio most similar to that of the <OBJECT> visible in the last frame?
		F_PHYSICS_PROPERTY_POISSON_RATIO_OBJECT_SIMILAR_NON_TECHNICAL **Single Image:** Based on the image, which visible object acts most like the <OBJECT> in terms of how it bulges sideways when squeezed? **Multi Image:** Considering all frames, which visible object acts most like the <OBJECT>, visible in the last frame, in terms of how it bulges sideways when squeezed?
		F_PHYSICS_PROPERTY_POISSON_RATIO_HIGHEST **Single Image:** Based on the image, which visible object exhibits the largest Poisson ratio? **Multi Image:** Considering all frames, which object, visible in the last frame, exhibits the largest Poisson ratio?
		F_PHYSICS_PROPERTY_POISSON_RATIO_HIGHEST_NON_TECHNICAL **Single Image:** Based on the image, which visible object bulges out the most when you press on it? **Multi Image:** Considering all frames, which object, visible in the last frame, bulges out the most when you press on it?
		F_PHYSICS_PROPERTY_POISSON_HIGH_LEVEL **Single Image:** Based on the image, if the <OBJECT> were compressed vertically, how would its horizontal dimensions change? **Multi Image:** Considering all frames, if the <OBJECT>, visible in the last frame, were compressed vertically, how would its horizontal dimensions change?
	Material Identification	**F_MATERIAL_IDENTIFICATION_SIMILAR_OBJECT** **Single Image:** Based on the image, which visible object is made of a material most similar to that of the <OBJECT>? **Multi Image:** Considering all frames, which visible object is made of a material most similar to that of the <OBJECT>?
		F_MATERIAL_IDENTIFICATION_OBJECT_LEVEL_1 **Single Image:** Based on the image, what material is the <OBJECT> made of? **Multi Image:** Considering all frames, what material is the <OBJECT> made of?
		F_MATERIAL_IDENTIFICATION_OBJECT_LEVEL_2 **Single Image:** Based on the image, what material is the <OBJECT> made of? **Multi Image:** Considering all frames, what material is the <OBJECT> made of?
		F_MATERIAL_IDENTIFICATION_OBJECT_LEVEL_3 **Single Image:** Based on the image, what material is the <OBJECT> made of? **Multi Image:** Considering all frames, what material is the <OBJECT> made of?
View Point	Visibility	**F_VISIBILITY_OBJECT** **Single Image:** Based on the image, which of these objects is visible? **Multi Image:** Considering all frames, which of these objects is visible in the last frame?
		F_OCCLUSION_PERCENTAGE_OBJECT **Single Image:** Based on the image, how much of the <OBJECT> is occluded? **Multi Image:** Considering all frames, how much of the <OBJECT> is occluded in the last frame?
		F_VISIBILITY_OBJECT_COUNT **Single Image:** Based on the image, how many objects are visible? **Multi Image:** Considering all frames, how many objects are visible in the last frame?
	Camera Characteristics	**F_VIEWPOINT_CAMERA_ANGLE** **Single Image:** From the camera's perspective, what is the viewing direction relative to the horizon? **Multi Image:** Considering all frames, from the camera's perspective, what is the viewing direction relative to the horizon in the last frame?

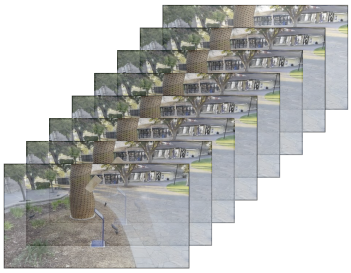
Continued on next page

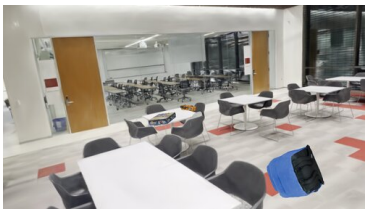
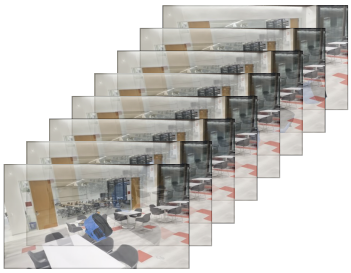
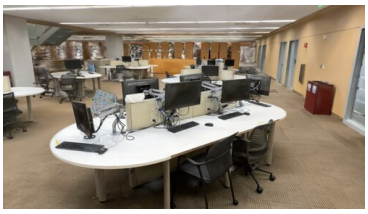
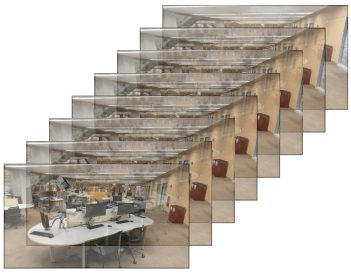
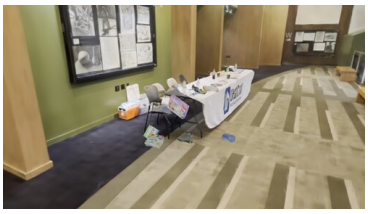
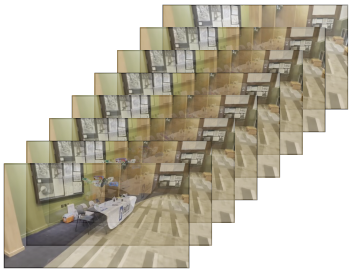
Category	Sub-Category	Question ID and Prompt
		F_FOCAL_LENGTH_CLASS **Single Image:** Based on the image, which focal-length class best matches the perspective observed? **Multi Image:** Considering all frames, which focal-length class best matches the perspective observed?
Temporal	Event Ordering	**F_TEMPORAL_SEQUENCE_IMAGES** **Multi Image:** Given the four unordered frames (A, B, C, D) of the same scene, what is the correct temporal ordering of the events?
		F_TEMPORAL_PREDICTION_NEXT_IMAGE_GRANULARITY_1 **Multi Image:** Which individual frame (A, B, C, or D) is most likely to display an event that occurred after the provided frame sequence?
		F_TEMPORAL_PREDICTION_NEXT_IMAGE_GRANULARITY_2 **Multi Image:** Which individual frame (A, B, C, or D) is most likely to display an event that occurred after the provided frame sequence?
		F_TEMPORAL_PREDICTION_NEXT_IMAGE_GRANULARITY_5 **Multi Image:** Which individual frame (A, B, C, or D) is most likely to display an event that occurred after the provided frame sequence?
		F_TEMPORAL_PREDICTION_PREVIOUS_IMAGE **Multi Image:** Which individual frame (A, B, C, or D) is most likely to display an event that occurred before the provided frame sequence?
		F_TEMPORAL_PREDICTION_MISSING_IMAGE **Multi Image:** Which individual frame (A, B, C, or D) is most likely to display an event that occurred during the time span of the provided frame sequence?
	Camera Motion	**F_CAMERA_MOTION_DIRECTION** **Multi Image:** Across the frame sequence, what is the predominant direction of the camera's motion?
		F_CAMERA_ZOOM_BEHAVIOR **Multi Image:** Across the frame sequence, how does the camera's zoom level change?
Persistence	Identity	**F_PERSISTENCE_OBJECT_PRESENT** **Multi Image:** Considering all frames, which object was visible but disappeared in the last frame?
		F_PERSISTENCE_OBJECT_DISAPPEAR **Multi Image:** Considering all frames, which visible object disappeared and does not reappear in the last frame?
	Counting	**F_PERSISTENCE_OBJECT_TOTAL_COUNT** **Multi Image:** Considering all frames, how many objects are present at the time of the last frame, including those currently hidden or out of frame?
		F_PERSISTENCE_OBJECT_TOTAL_COUNT_HIDDEN **Multi Image:** Considering all frames, how many objects are present at the time of the last frame, but not visible?

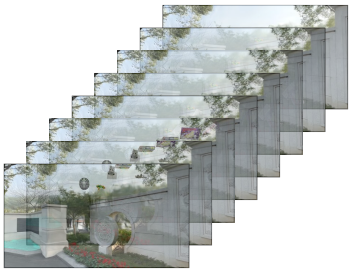
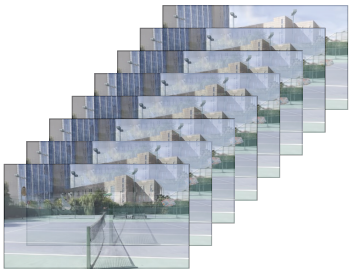
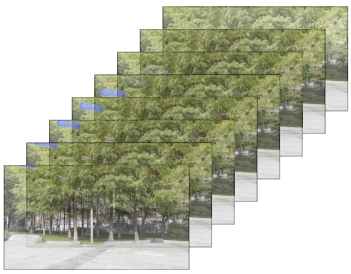
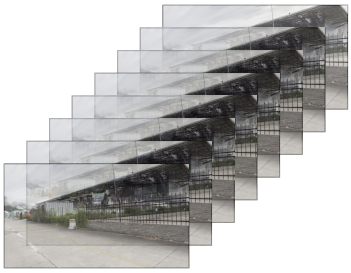
D.2 VQA EXAMPLES

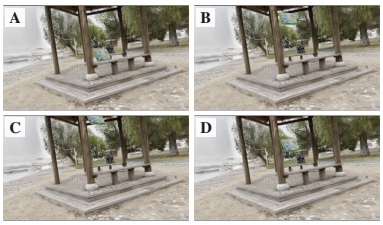
Below, we present examples of the actual VQA questions used to evaluate the 54 VLMs. For each question_id, we show one representative single-image question and one representative multi-image question. In total, the benchmark contains 85 distinct questions, spanning both single-image and multi-image settings. Some questions are defined only for single-image inputs or only for multi-image inputs; this is intentional and reflects the original design of the VQA benchmark.

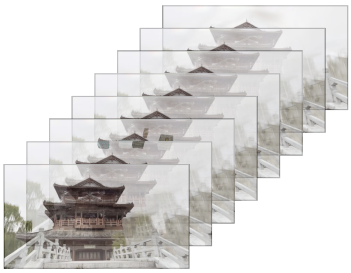
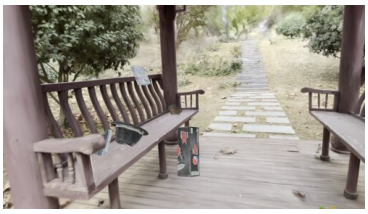
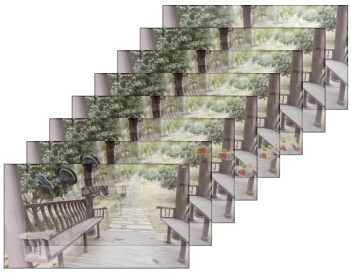
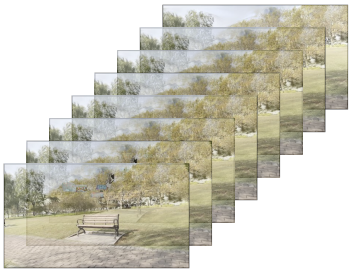
Table 5: **Dataset Samples** Examples of questions and answers of the VQA in the NewtPhys dataset for all question IDs.

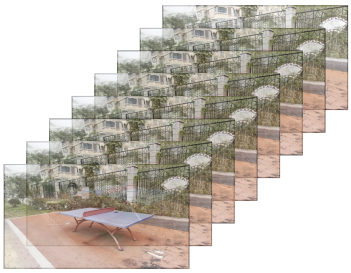
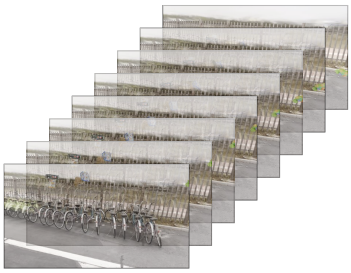
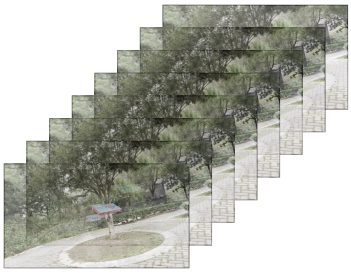
ID	Type	Visual Input	Task & Logic
F_CLOSEST_OBJECT_OBJECT	**Spatial Reasoning** (Distance)		**Image:** Based on the image, which object is closest to the "orange Transformers action figure" in real-world distance? A) Teenage Mutant Ninja Turtles action figurine B) white soccer shoe C) **colored cactus balancing toy** D) small blue painted bucket
			**Image sequences:** Considering all frames, which object is closest to the "orange Transformers action figure" in real-world distance in the last frame? A) Teenage Mutant Ninja Turtles action figurine B) white soccer shoe C) **colored cactus balancing toy** D) small blue painted bucket
F_DISTANCE_OBJECT_CAMERA_DISTANCE	**Spatial Reasoning** (Distance)		**Image:** Based on the image, what is the real-world distance between the "shelf bin" and the camera? A) 1.32 meters B) 2.05 meters C) **2.79 meters** D) 0.58 meters
			**Image sequences:** Considering all frames, what is the real-world distance between the "shelf bin" and the camera in the last frame? A) 1.32 meters B) 2.05 meters C) **2.79 meters** D) 0.58 meters

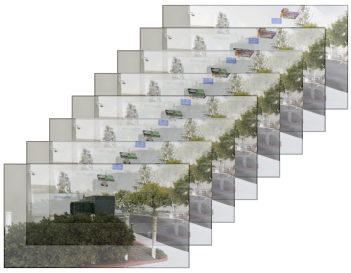
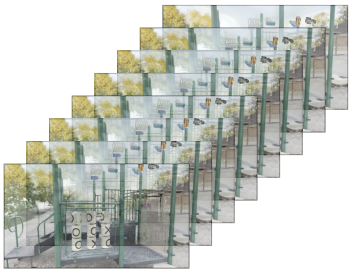
ID	Type	Visual Input	Task & Logic
F_CLOSEST_OBJECT_CAMERA	**Spatial Reasoning** (Distance)		**Image:** Based on the image, which object is the closest to the camera in real-world distance? A) white soccer shoe B) CLUE board game box C) **blue backpack** D) Teenage Mutant Ninja Turtles action figurine
			**Image sequences:** Considering all frames, which object is the closest to the camera in real-world distance in the last frame? A) white soccer shoe B) CLUE board game box C) **blue backpack** D) Teenage Mutant Ninja Turtles action figurine
F_SIZE_OBJECT	**Spatial Reasoning** (Size)		**Image:** Based on the image, what are the real-world dimensions of the "turquoise mermaid-design backpack"? A) 0.62m x 0.53m x 0.43m B) **0.41m x 0.35m x 0.28m** C) 0.74m x 0.63m x 0.51m D) 0.83m x 0.71m x 0.57m
			**Image sequences:** Considering all frames, what are the real-world dimensions of the "turquoise mermaid-design backpack" in the last frame? A) 0.62m x 0.53m x 0.43m B) **0.41m x 0.35m x 0.28m** C) 0.74m x 0.63m x 0.51m D) 0.83m x 0.71m x 0.57m
F_SIZE_OBJECT_BIGGER	**Spatial Reasoning** (Size)		**Image:** Based on the image, which single object has the biggest real-world volume? A) Teenage Mutant Ninja Turtles action figurine B) **228-piece purple LEGO Friends boxed set** C) retail box of fruit snacks (10 pouches) D) white soccer shoe
			**Image sequences:** Considering all frames, which single object, visible in the last frame, has the biggest real-world volume? A) Teenage Mutant Ninja Turtles action figurine B) **228-piece purple LEGO Friends boxed set** C) retail box of fruit snacks (10 pouches) D) white soccer shoe

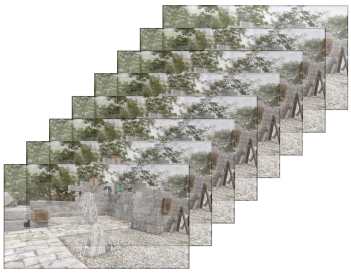
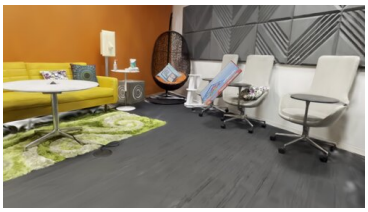
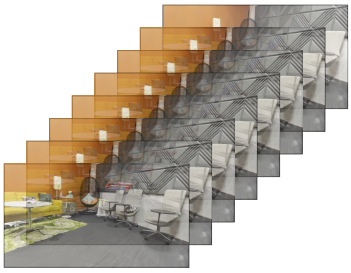
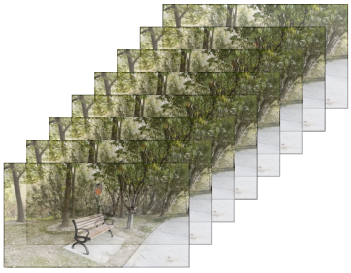
ID	Type	Visual Input	Task & Logic
F_LAYOUT_POSITION_OBJECT_OBJECT	**Spatial Reasoning** (Layout)		**Image:** From the camera's perspective, where is the "motorcycle helmet" relative to the "gray raccoon toy" in the image? A) to the right B) **to the left** C) in front D) horizontally aligned
			**Image sequences:** Considering all frames, from the camera's perspective, where is the "motorcycle helmet" relative to the "gray raccoon toy" in the last frame? A) to the right B) **to the left** C) in front D) horizontally aligned
F_KINEMATICS_SPEED_OBJECT	**Mechanics** (Kinematics)		**Image: No question provided/possible**
			**Image sequences:** Considering all frames, what is the real-world speed of the "12-pack Fresca soda carton" at the time of the last frame? A) 7.71 m/s B) 1.95 m/s C) **3.87 m/s** D) 5.79 m/s
F_KINEMATICS_ACCEL_OBJECT	**Mechanics** (Kinematics)		**Image: No question provided/possible**
			**Image sequences:** Considering all frames, what is the real-world magnitude of acceleration of the "brown boat shoe" at the time of the last frame? A) **9.79 m/s^2** B) 5.05 m/s^2 C) 0.32 m/s^2 D) 14.53 m/s^2
F_KINEMATICS_DISTANCE_TRAVELED_INTERVAL	**Mechanics** (Kinematics)		**Image: No question provided/possible**
			**Image sequences:** Considering all frames, what is the real-world displacement of the centroid of the "blue sport-design kid backpack" from the first to the last frame? A) 2.7 meters B) 0.5 meters C) 3.7 meters D) **1.6 meters**

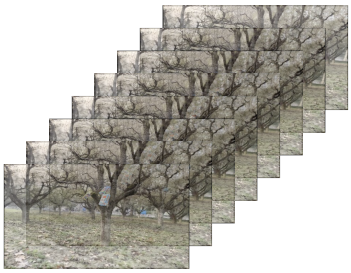
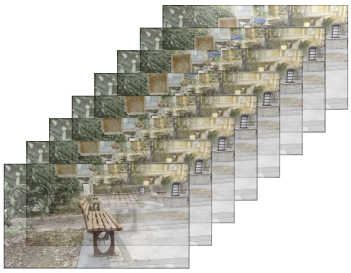
ID	Type	Visual Input	Task & Logic
F_KINEMATICS_SYSTEM_STABILITY	**Mechanics** (Kinematics)		**Image: No question provided/possible**
			**Image sequences:** Analyzing the motion trend across the sequence, which statement best describes the system's state at the final frame? A) Cyclic: All objects returned to their initial position B) **Stable: The system has stopped** C) Unstable: The system is currently moving D) Invisible: All objects have moved out of the frame entirely
F_COLLISION_OBJECT_OBJECT_FRAME_SINGLE	**Mechanics** (Collision)		**Image:** Based on the image, which visible object is the "dark ballet flat shoe" colliding with? A) **shelf bin** B) white soccer shoe C) colored cactus balancing toy D) empty purple and green pencil case
			Image sequences: No question provided/possible
F_COLLISION_OBJECT_OBJECT_FRAME_MULTI	**Mechanics** (Collision)		**Image: No question provided/possible**
			**Image sequences:** In which frame is the "12-pack carton of diet pepsi cans" most likely colliding with another object? A) A B) B C) C D) **D**
F_COLLISION_OBJECT_SCENE_FRAME_MULTI	**Mechanics** (Collision)		**Image: No question provided/possible**
			**Image sequences:** In which frame is the "boxed My Monopoly board game" most likely colliding with the static scene? A) **A** B) B C) C D) D

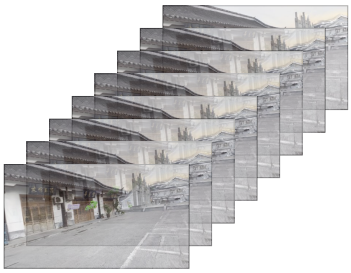
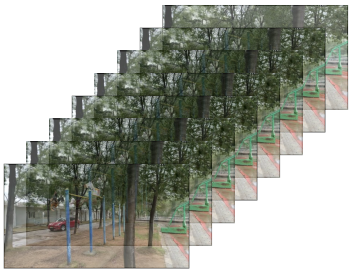
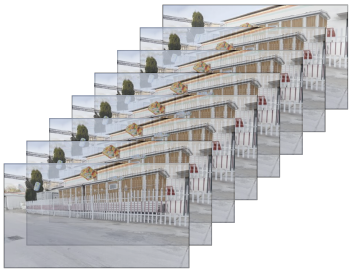
ID	Type	Visual Input	Task & Logic
F_MASS_OBJECT	Material Understanding (Mass)		**Image:** Based on the image, what is the mass of the "army-design lunch bag"? A) 2.00 kgs B) **3.12 kgs** C) 4.24 kgs D) 0.89 kgs
			**Image sequences:** Considering all frames, what is the mass of the "army-design lunch bag"? A) 2.00 kgs B) **3.12 kgs** C) 4.24 kgs D) 0.89 kgs
F_MASS_HEAVIEST_OBJECT	Material Understanding (Mass)		**Image:** Based on the image, which visible single object has the greatest mass? A) boxed Frozen board game B) Teenage Mutant Ninja Turtles action figurine C) woven straw fedora hat D) **12-pack carton of Pepsi MAX cans**
			**Image sequences:** Considering all frames, which single object, visible in the last frame, has the greatest mass? A) boxed Frozen board game B) Teenage Mutant Ninja Turtles action figurine C) woven straw fedora hat D) **12-pack carton of Pepsi MAX cans**
F_MASS_LIGHTEST_OBJECT	Material Understanding (Mass)		**Image:** Based on the image, which visible single object has the least mass? A) 12-pack Fresca soda carton B) turquoise insulated lunch bag C) boxed white Creationary board game D) **black and white ASICS golf shoe**
			**Image sequences:** Considering all frames, which single object, visible in the last frame, has the least mass? A) 12-pack Fresca soda carton B) turquoise insulated lunch bag C) boxed white Creationary board game D) **black and white ASICS golf shoe**

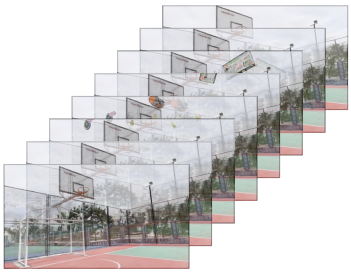
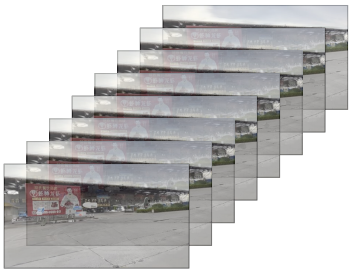
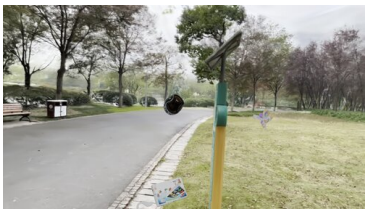
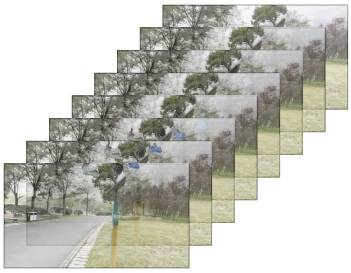
ID	Type	Visual Input	Task & Logic
F_PHYSICS_PROPERTY_DENSITY_OBJECT	**Material Understanding** (Density)		**Image:** Based on the image, what is the estimated mean density of the "motorcycle helmet"? A) **110.0 kg/m^3** B) 19.4 kg/m^3 C) 79.8 kg/m^3 D) 49.6 kg/m^3
			**Image sequences:** Considering all frames, what is the estimated mean density of the "motorcycle helmet"? A) **110.0 kg/m^3** B) 19.4 kg/m^3 C) 79.8 kg/m^3 D) 49.6 kg/m^3
F_PHYSICS_PROPERTY_DENSITY_OBJECT_RELATIVE	**Material Understanding** (Density)		**Image:** Based on the image, which visible object has the highest effective density? A) Teenage Mutant Ninja Turtles action figurine B) **12-pack carton of Pepsi MAX cans** C) white soccer shoe D) roll-along turtle toy
			**Image sequences:** Considering all frames, which object, visible in the last frame, has the highest effective density? A) Teenage Mutant Ninja Turtles action figurine B) **12-pack carton of Pepsi MAX cans** C) white soccer shoe D) roll-along turtle toy
F_PHYSICS_PROPERTY_YOUNG_MODULUS_OBJECT_METRIC_PREFIX	**Material Understanding** (Young Modulus)		**Image:** Based on the image, what is the Young's modulus of the "boxed Aggravation board game"? A) 0,032 MPa B) 3,15 MPa C) 0,315 MPa D) **315,00 MPa**
			**Image sequences:** Considering all frames, what is the Young's modulus of the "boxed Aggravation board game"? A) 0,032 MPa B) 3,15 MPa C) 0,315 MPa D) **315,00 MPa**

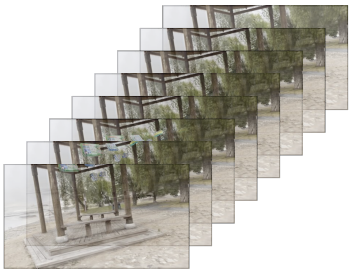
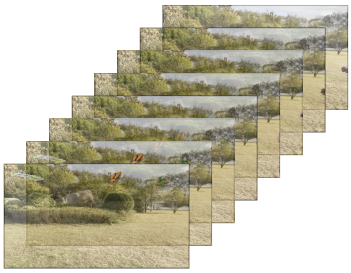
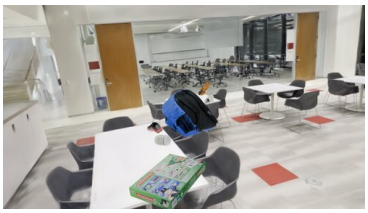
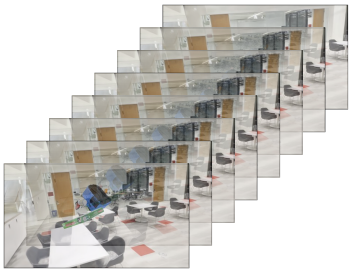
ID	Type	Visual Input	Task & Logic
F_PHYSICS_PROPERTY_YOUNG_MODULUS_OBJECT_SIMILAR	**Material Understanding** (Young Modulus)		**Image:** Based on the image, which visible object has a Young's Modulus most similar to that of the "boxed Clue board game"? A) **boxed purple Operation board game** B) Teenage Mutant Ninja Turtles action figurine C) orange Transformers action figure D) box of Samsung C406S cartridge
			**Image sequences:** Considering all frames, which object, visible in the last frame, has a Young's Modulus most similar to that of the "boxed Clue board game"? A) **boxed purple Operation board game** B) Teenage Mutant Ninja Turtles action figurine C) orange Transformers action figure D) box of Samsung C406S cartridge
F_PHYSICS_PROPERTY_YOUNG_MODULUS_OBJECT_SIMILAR_NON_TECHNICAL	**Material Understanding** (Young Modulus)		**Image:** Based on the image, which visible object has a softness most similar to that of the "jar of Twinlab protein powder"? A) painted toy castle made of stacked blocks B) boxed LIFE board game C) **futuristic dinosaur robot figure** D) boxed Connect 4 Launchers game
			**Image sequences:** Considering all frames, which object, visible in the last frame, has a softness most similar to that of the "jar of Twinlab protein powder"? A) painted toy castle made of stacked blocks B) boxed LIFE board game C) **futuristic dinosaur robot figure** D) boxed Connect 4 Launchers game
F_PHYSICS_PROPERTY_YOUNG_MODULUS_HIGHEST	**Material Understanding** (Young Modulus)		**Image:** Based on the image, which visible object exhibits the highest Young's Modulus? A) boxed Balderdash board game B) **light gray file organizer rack** C) black and white ASICS golf shoe D) Teenage Mutant Ninja Turtles action figurine
			**Image sequences:** Considering all frames, which object, visible in the last frame, exhibits the highest Young's Modulus? A) boxed Balderdash board game B) **light gray file organizer rack** C) black and white ASICS golf shoe D) Teenage Mutant Ninja Turtles action figurine

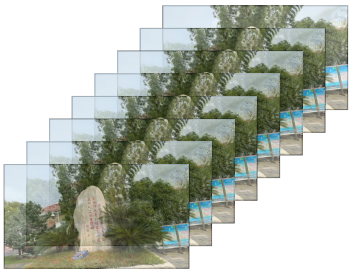
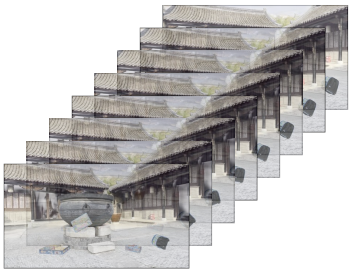
ID	Type	Visual Input	Task & Logic
F_PHYSICS_PROPERTY_YOUNG_MODULUS_HIGHEST_NON_TECHNICAL	**Material Understanding** (Young Modulus)		**Image:** Based on the image, which visible object is the stiffest? A) gift box B) **motorcycle helmet** C) Teenage Mutant Ninja Turtles action figurine D) white soccer shoe
			**Image sequences:** Considering all frames, which object, visible in the last frame, is the stiffest? A) gift box B) **motorcycle helmet** C) Teenage Mutant Ninja Turtles action figurine D) white soccer shoe
F_PHYSICS_PROPERTY_YOUNG_MODULUS_OBJECT_HIGH_LEVEL	**Material Understanding** (Young Modulus)		**Image:** Based on the image, which attribute best describes the "boxed Aggravation board game" in terms of deformability? A) **Rigid (Holds shape perfectly)** B) Flexible (Bendable but tough) C) Soft (Deformable like stiff foam) D) Very Soft (No resistance, like a plush toy)
			**Image sequences:** Considering all frames, which attribute best describes the "boxed Aggravation board game", visible in the last frame, in terms of deformability? A) **Rigid (Holds shape perfectly)** B) Flexible (Bendable but tough) C) Soft (Deformable like stiff foam) D) Very Soft (No resistance, like a plush toy)
F_PHYSICS_PROPERTY_POISSON_RATIO_OBJECT	**Material Understanding** (Poisson Ratio)		**Image:** Based on the image, what is the Poisson ratio of the "jar of Twinlab protein powder"? A) 0.10 B) 0.20 C) 0.40 D) **0.30**
			**Image sequences:** Considering all frames, what is the Poisson ratio of the "jar of Twinlab protein powder"? A) 0.10 B) 0.20 C) 0.40 D) **0.30**

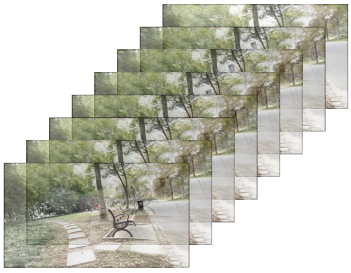
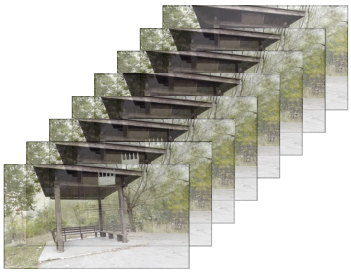
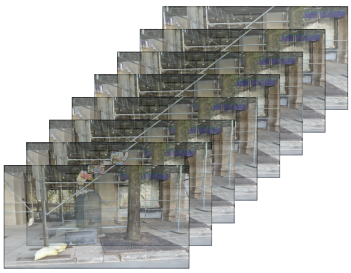
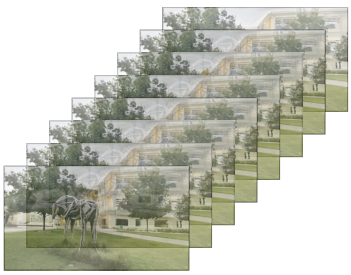
ID	Type	Visual Input	Task & Logic
F_PHYSICS_PROPERTY_POISSON_RATIO_OBJECT_SIMILAR	**Material Understanding** (Poisson Ratio)		**Image:** Based on the image, which visible object has a Poisson ratio most similar to that of the "boxed SLIDERS board game"? A) white soccer shoe B) **boxed Connect 4 Launchers game** C) small blue painted bucket D) Teenage Mutant Ninja Turtles action figurine
			**Image sequences:** Considering all frames, which visible object has a Poisson ratio most similar to that of the "boxed SLIDERS board game" visible in the last frame? A) white soccer shoe B) **boxed Connect 4 Launchers game** C) small blue painted bucket D) Teenage Mutant Ninja Turtles action figurine
F_PHYSICS_PROPERTY_POISSON_RATIO_OBJECT_SIMILAR_NON_TECHNICAL	**Material Understanding** (Poisson Ratio)		**Image:** Based on the image, which visible object acts most like the "shelf bin" in terms of how it bulges sideways when squeezed? A) Teenage Mutant Ninja Turtles action figurine B) white soccer shoe C) small blue painted bucket D) **empty blue pencil case**
			**Image sequences:** Considering all frames, which visible object acts most like the "shelf bin", visible in the last frame, in terms of how it bulges sideways when squeezed? A) Teenage Mutant Ninja Turtles action figurine B) white soccer shoe C) small blue painted bucket D) **empty blue pencil case**
F_PHYSICS_PROPERTY_POISSON_RATIO_HIGHEST	**Material Understanding** (Poisson Ratio)		**Image:** Based on the image, which visible object exhibits the largest Poisson ratio? A) **empty horse-design pencil case** B) gift box C) brown teddy bear toy D) boxed Monopoly Hotels board game
			**Image sequences:** Considering all frames, which object, visible in the last frame, exhibits the largest Poisson ratio? A) **empty horse-design pencil case** B) gift box C) brown teddy bear toy D) boxed Monopoly Hotels board game

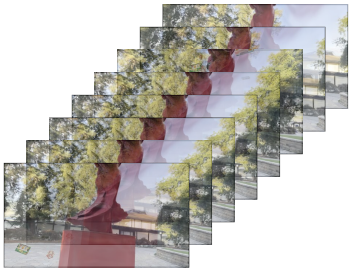
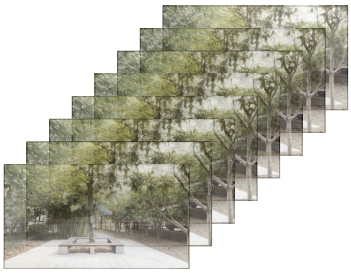
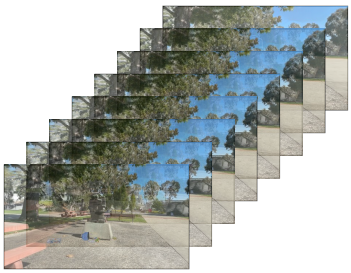
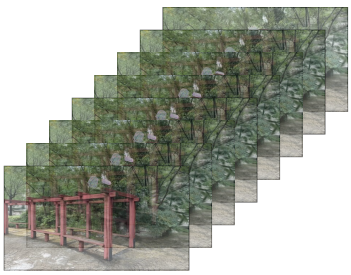
ID	Type	Visual Input	Task & Logic
F_PHYSICS_PROPERTY_POISSON_RATIO_HIGHEST_NON_TECHNICAL	**Material Understanding** (Poisson Ratio)		**Image:** Based on the image, which visible object bulges out the most when you press on it? A) white-striped purple pencil case B) white soccer shoe C) Teenage Mutant Ninja Turtles action figurine D) **green ogre toy figurine**
			**Image sequences:** Considering all frames, which object, visible in the last frame, bulges out the most when you press on it? A) white-striped purple pencil case B) white soccer shoe C) Teenage Mutant Ninja Turtles action figurine D) **green ogre toy figurine**
F_PHYSICS_PROPERTY_POISSON_HIGH_LEVEL	**Material Understanding** (Poisson Ratio)		**Image:** Based on the image, if the "boxed LIFE board game" were compressed vertically, how would its horizontal dimensions change? A) **It would expand sideways a moderate amount** B) It would barely change width C) It would contract inwards D) It would bulge out significantly
			**Image sequences:** Considering all frames, if the "boxed LIFE board game", visible in the last frame, were compressed vertically, how would its horizontal dimensions change? A) **It would expand sideways a moderate amount** B) It would barely change width C) It would contract inwards D) It would bulge out significantly
F_MATERIAL_IDENTIFICATION_SIMILAR_OBJECT	**Material Understanding** (Material Identification)		**Image:** Based on the image, which visible object is made of a material most similar to that of the "Make-a-Match memory game box"? A) Teenage Mutant Ninja Turtles action figurine B) outback hat C) blue insulated lunch box with dinosaurs D) **1000-piece yellow LEGO brick assortment boxed set**
			**Image sequences:** Considering all frames, which visible object is made of a material most similar to that of the "Make-a-Match memory game box"? A) Teenage Mutant Ninja Turtles action figurine B) outback hat C) blue insulated lunch box with dinosaurs D) **1000-piece yellow LEGO brick assortment boxed set**

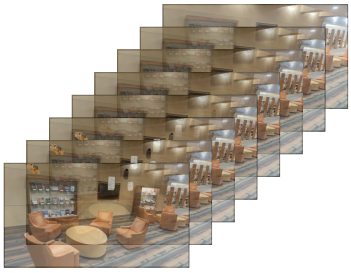
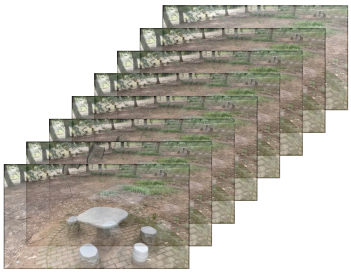
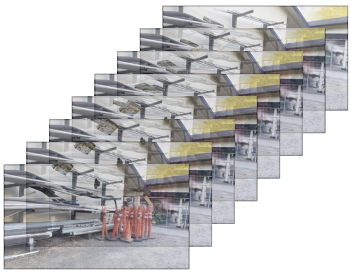
ID	Type	Visual Input	Task & Logic
F_MATERIAL_IDENTIFICATION_OBJECT_LEVEL_1	**Material Understanding** (Material Identification)		**Image:** Based on the image, what material is the "jar of Twinlab protein powder" made of? A) **synthetic polymers** B) vegetal and cellulosic C) skins, textiles and fibers D) inorganic rigid
			**Image sequences:** Considering all frames, what material is the "jar of Twinlab protein powder" made of? A) **synthetic polymers** B) vegetal and cellulosic C) skins, textiles and fibers D) inorganic rigid
F_MATERIAL_IDENTIFICATION_OBJECT_LEVEL_2	**Material Understanding** (Material Identification)		**Image:** Based on the image, what material is the "12-pack carton of diet pepsi cans" made of? A) **processed cellulose** B) ligneous matter C) woven textiles D) cellular synthetics
			**Image sequences:** Considering all frames, what material is the "12-pack carton of diet pepsi cans" made of? A) **processed cellulose** B) ligneous matter C) woven textiles D) cellular synthetics
F_MATERIAL_IDENTIFICATION_OBJECT_LEVEL_3	**Material Understanding** (Material Identification)		**Image:** Based on the image, what material is the "Princess Celestia pony figurine toy" made of? A) metal B) **plastic** C) paper/cardboard D) ceramic
			**Image sequences:** Considering all frames, what material is the "Princess Celestia pony figurine toy" made of? A) metal B) **plastic** C) paper/cardboard D) ceramic

ID	Type	Visual Input	Task & Logic
F_VISIBILITY_OBJECT	**View Point** (Visibility)		**Image:** Based on the image, which of these objects is visible? A) turquoise mermaid-design backpack B) boxed Clue board game C) **boxed My Monopoly board game** D) white spoon
			**Image sequences:** Considering all frames, which of these objects is visible in the last frame? A) turquoise mermaid-design backpack B) boxed Clue board game C) **boxed My Monopoly board game** D) white spoon
F_OCCLUSION_PERCENTAGE_OBJECT	**View Point** (Visibility)		**Image:** Based on the image, how much of the "orange Transformers action figure" is occluded? A) **Fully Visible (>95% visible)** B) Partially Occluded (25-65% visible) C) Slightly Occluded (65-95% visible) D) Severely Occluded (0-25% visible)
			**Image sequences:** Considering all frames, how much of the "orange Transformers action figure" is occluded in the last frame? A) **Fully Visible (>95% visible)** B) Partially Occluded (25-65% visible) C) Slightly Occluded (65-95% visible) D) Severely Occluded (0-25% visible)
F_VISIBILITY_OBJECT_COUNT	**View Point** (Visibility)		**Image:** Based on the image, how many objects are visible? A) 7 B) 4 C) 5 D) **6**
			**Image sequences:** Considering all frames, how many objects are visible in the last frame? A) 7 B) 4 C) 5 D) **6**

ID	Type	Visual Input	Task & Logic
F_VIEWPOINT_CAMERA_ANGLE	**View Point** (Camera Characteristics)		**Image:** From the camera's perspective, what is the viewing direction relative to the horizon? A) **eye level (-15 to 15 degrees)** B) high angle (-60 to -15 degrees) C) bird's-eye (<=-60 degrees) D) low angle (15 to 60 degrees)
			**Image sequences:** Considering all frames, from the camera's perspective, what is the viewing direction relative to the horizon in the last frame? A) **eye level (-15 to 15 degrees)** B) high angle (-60 to -15 degrees) C) bird's-eye (<=-60 degrees) D) low angle (15 to 60 degrees)
F_FOCAL_LENGTH_CLASS	**View Point** (Camera Characteristics)		**Image:** Based on the image, which focal-length class best matches the perspective observed? A) normal (20-60) B) **wide (60-100)** C) ultra-wide (>=100) D) telephoto (<20)
			**Image sequences:** Considering all frames, which focal-length class best matches the perspective observed? A) normal (20-60) B) **wide (60-100)** C) ultra-wide (>=100) D) telephoto (<20)
F_TEMPORAL_SEQUENCE_IMAGES	**Temporal** (Event Ordering)		**Image: No question provided/possible**
			**Image sequences:** Given the four unordered frames (A, B, C, D) of the same scene, what is the correct temporal ordering of the events? A) A-C-B-D B) **A-D-B-C** C) C-D-B-A D) C-A-B-D

ID	Type	Visual Input	Task & Logic
F_TEMPORAL_PREDICTION_NEXT_IMAGE_GRANULARITY_1	**Temporal** (Event Ordering)		**Image: No question provided/possible**
			**Image sequences:** Which individual frame (t5, t6, t7, or t8) is most likely to display an event that occurred after the provided frame sequence (t1, t2, t3, t4)? A) t5 B) t6 C) t7 D) **t8**
F_TEMPORAL_PREDICTION_NEXT_IMAGE_GRANULARITY_2	**Temporal** (Event Ordering)		**Image: No question provided/possible**
			**Image sequences:** Which individual frame (t5, t6, t7, or t8) is most likely to display an event that occurred after the provided frame sequence (t1, t2, t3, t4)? A) t5 B) t6 C) **t7** D) t8
F_TEMPORAL_PREDICTION_NEXT_IMAGE_GRANULARITY_5	**Temporal** (Event Ordering)		**Image: No question provided/possible**
			**Image sequences:** Which individual frame (t5, t6, t7, or t8) is most likely to display an event that occurred after the provided frame sequence (t1, t2, t3, t4)? A) t5 B) **t6** C) t7 D) t8
F_TEMPORAL_PREDICTION_PREVIOUS_IMAGE	**Temporal** (Event Ordering)		**Image: No question provided/possible**
			**Image sequences:** Which individual frame (t5, t6, t7, or t8) is most likely to display an event that occurred before the provided frame sequence (t1, t2, t3, t4)? A) t5 B) **t6** C) t7 D) t8

ID	Type	Visual Input	Task & Logic
F_TEMPORAL_PREDICTION_MISSING_IMAGE	**Temporal** (Event Ordering)		**Image: No question provided/possible**
			**Image sequences:** Which individual frame (t5, t6, t7, or t8) is most likely to display an event that occurred during the time span of the provided frame sequence (t1, t2, t3, t4)? A) **t5** B) t6 C) t7 D) t8
F_CAMERA_MOTION_DIRECTION	**Temporal** (Camera Motion)		**Image: No question provided/possible**
			**Image sequences:** Across the frame sequence, what is the predominant direction of the camera's motion? A) left then up B) backward then down C) **backward** D) up then down
F_CAMERA_ZOOM_BEHAVIOR	**Temporal** (Camera Motion)		**Image: No question provided/possible**
			**Image sequences:** Across the frame sequence, how does the camera's zoom level change? A) **no zoom** B) zoom out C) zoom out then in D) zoom in then out
F_PERSISTENCE_OBJECT_PRESENT	**Persistence** (Object Persistence)		**Image: No question provided/possible**
			**Image sequences:** Considering all frames, which object was visible but disappeared in the last frame? A) **white-striped purple pencil case** B) Teenage Mutant Ninja Turtles action figurine C) white soccer shoe D) small blue painted bucket

ID	Type	Visual Input	Task & Logic
F_PERSISTENCE_OBJECT_DISAPPEAR	**Persistence** (Object Persistence)		**Image: No question provided/possible**
			**Image sequences:** Considering all frames, which visible object disappeared and does not reappear in the last frame? A) small blue painted bucket B) **white jar of proteins** C) Teenage Mutant Ninja Turtles action figurine D) white soccer shoe
F_PERSISTENCE_OBJECT_TOTAL_COUNT	**Persistence** (Object Persistence)		**Image: No question provided/possible**
			**Image sequences:** Considering all frames, how many objects are present at the time of the last frame, including those currently hidden or out of frame? A) 0 B) 2 C) **1** D) 3
F_PERSISTENCE_OBJECT_TOTAL_COUNT_HIDDEN	**Persistence** (Object Persistence)		**Image: No question provided/possible**
			**Image sequences:** Considering all frames, how many objects are present at the time of the last frame, but not visible? A) 2 B) 0 C) **1** D) 3

D.3 VQA COUNTERFACTUAL

Counterfactual Questions are one of the main strengths of Newtphys and of great value for the community. In order to produce these questions, we had to replicate one of the scenes, and apply a small perturbation to the initial conditions of at most one object, denoted `<OBJECT-CF>` in the simulation or the global conditions of the scene *e.g.* gravity. In particular, we identified two perturbation vectors of objects, namely `scale` and `position` for the single object and `gravity` of the scene. We further identified a handful of questions that can be answered logically by some counterfactual question, and they are reported in tables 6, 7 and 8. Note that the question set varies for each perturbed initial condition.

Code implementation-wise, we had two different simulation information, the base one and the counterfactual. The idea is to simply provide the input images from the base simulation to the VLM and answer the question using the information and labels produced by the counterfactual simulation. Because of the intrinsic nature of the task, most pairs of simulations do not produce what we call "interesting pairs". An "Interesting pair" is a pair of simulations that simultaneously satisfy two conditions: 1) it is a possible question to be answered (which is the same condition that needs to be satisfied in the factual questions), and the two simulations produce two different outcomes for the same question. The second condition is much more unlikely to happen, because the initial perturbations are not far from the initial correct condition. Because of this, the number of questions produced during the counterfactual is much less. The following tables outline the questions template, for which

our code fills the <OBJECT-CF> and OBJECT which can be or not the same object. We therefore end up with nearly 2K counterfactual simulations across the 3 perturbation vectors.

Table 6: **Counterfactual gravity**. Counterfactual question templates for perturbation of gravity.

CF_OCCLUSION_PERCENTAGE_OBJECT *(View Point)*	
Single	**Multi**
Under a counterfactual scenario where the gravitational constant is reduced to $g/10$, how much of the <OBJECT> would be occluded?	Under a counterfactual scenario where the gravitational constant is reduced to $g/10$, how much of the <OBJECT> would have been occluded in the last frame?
CF_VISIBILITY_OBJECT_COUNT *(View Point)*	
Single	**Multi**
Under a counterfactual scenario where the gravitational constant is reduced to $g/10$, how many objects would be visible?	Under a counterfactual scenario where the gravitational constant is reduced to $g/10$, how many objects would have been visible in the last frame?
CF_KINEMATICS_SPEED_OBJECT *(Mechanics)*	
Single	**Multi**
–	Under a counterfactual scenario where the gravitational constant is reduced to $g/10$, what would have been the real-world speed of the <OBJECT> at the time of the last frame?
CF_KINEMATICS_ACCEL_OBJECT *(Mechanics)*	
Single	**Multi**
–	Under a counterfactual scenario where the gravitational constant is reduced to $g/10$, what would have been the magnitude of acceleration of the <OBJECT> at the time of the last frame?
CF_KINEMATICS_DISTANCE_TRAVELED_INTERVAL *(Mechanics)*	
Single	**Multi**
–	Under a counterfactual scenario where the gravitational constant is reduced to $g/10$, what would have been the real-world displacement of the centroid of the <OBJECT> from the first to the last frame?
CF_PERSISTENCE_OBJECT_TOTAL_COUNT_HIDDEN *(Persistence)*	
Single	**Multi**
–	Under a counterfactual scenario where the gravitational constant is reduced to $g/10$, how many objects would have been present but not visible in the last frame?

Table 7: **Counterfactual volume**. Counterfactual question templates for perturbation of volume.

CF_SIZE_OBJECT *(Spatial Reasoning)*	
Single	**Multi**
In a counterfactual scenario where the dimensions of the <OBJECT-CF> are scaled by <SCALE>, what would be the real-world dimensions of the <OBJECT> visible in the frame?	In a counterfactual scenario where the dimensions of the <OBJECT-CF> are scaled by <SCALE>, what would have been the real-world dimensions of the <OBJECT> in the last frame?
CF_SIZE_OBJECT_BIGGER *(Spatial Reasoning)*	
Single	**Multi**

In a counterfactual scenario where the dimensions of the <OBJECT-CF> are scaled by <SCALE>, which single object would have the biggest real-world volume?	In a counterfactual scenario where the dimensions of the <OBJECT-CF> are scaled by <SCALE>, which single object would have had the biggest real-world volume?
CF_OCCLUSION_PERCENTAGE_OBJECT *(View Point)*	
Single	**Multi**
–	In a counterfactual scenario where the dimensions of the <OBJECT-CF> are scaled by <SCALE>, how much of the <OBJECT> would have been occluded in the last frame?
CF_VISIBILITY_OBJECT_COUNT *(View Point)*	
Single	**Multi**
–	In a counterfactual scenario where the dimensions of the <OBJECT-CF> are scaled by <SCALE>, how many objects would have been visible in the last frame?
CF_KINEMATICS_DISTANCE_TRAVELED_INTERVAL *(Mechanics)*	
Single	**Multi**
–	In a counterfactual scenario where the dimensions of the <OBJECT-CF> are scaled by <SCALE>, what would have been the real-world displacement of the centroid of the <OBJECT> from the first to the last frame?
CF_COLLISION_OBJECT_OBJECT_FRAME_SINGLE *(Mechanics)*	
Single	**Multi**
In a counterfactual scenario where the dimensions of the <OBJECT-CF> are scaled by <SCALE>, which object would the <OBJECT> be colliding with in the frame?	–
CF_MASS_OBJECT *(Material Understanding)*	
Single	**Multi**
–	In a counterfactual scenario where the dimensions of the <OBJECT-CF> are scaled by <SCALE>x, what would have been the mass of the <OBJECT>?
CF_MASS_HEAVIEST_OBJECT *(Material Understanding)*	
Single	**Multi**
–	In a counterfactual scenario where the dimensions of the <OBJECT-CF> are scaled by <SCALE>, which single object visible in the last frame would have had the greatest mass?
CF_PERSISTENCE_OBJECT_TOTAL_COUNT_HIDDEN *(Persistence)*	
Single	**Multi**
–	In a counterfactual scenario where the dimensions of the <OBJECT-CF> are scaled by <SCALE>, how many objects would have been present but not visible in the last frame?

Table 8: **Counterfactual position** Counterfactual question templates for perturbation of initial position.

CF_CLOSEST_OBJECT_OBJECT *(Spatial Reasoning)*
Single | **Multi**

Assuming that the "<OBJECT-CF>" is moved by <X>, <Y>, and <Z> in the camera reference system, which visible object would be the closest to the <OBJECT> in real-world distance?	Assuming that the "<OBJECT-CF>" at the time of the first frame is moved by <X>, <Y>, and <Z> in the camera reference system, which visible object would have been the closest to the <OBJECT> in real-world distance in the last frame?
CF_CLOSEST_OBJECT_CAMERA *(Spatial Reasoning)*	
Single	**Multi**
Assuming that the "<OBJECT-CF>" is moved by <X>, <Y>, and <Z> in the camera reference system, which visible object would be the closest to the camera?	Assuming that the "<OBJECT-CF>" at the time of the first frame is moved by <X>, <Y>, and <Z> in the camera reference system, which visible object would have been the closest to the camera in the last frame?
CF_LAYOUT_POSITION_OBJECT_OBJECT *(Spatial Reasoning)*	
Single	**Multi**
Assuming that the "<OBJECT-CF>" is moved by <X>, <Y>, and <Z> in the camera reference system, from the camera's perspective, where would the <OBJECT_1> be relative to the <OBJECT_2> in the image?	Assuming that the "<OBJECT-CF>" at the time of the first frame is moved by <X>, <Y>, and <Z> in the camera reference system, from the camera's perspective, where would the <OBJECT_1> have been relative to the <OBJECT_2> in the last frame?
CF_OCCLUSION_PERCENTAGE_OBJECT *(View Point)*	
Single	**Multi**
Assuming that the "<OBJECT-CF>" is moved by <X>, <Y>, and <Z> in the camera reference system, how much of the <OBJECT> would be occluded?	Assuming that the "<OBJECT-CF>" at the time of the first frame is moved by <X>, <Y>, and <Z> in the camera reference system, how much of the <OBJECT> would have been occluded in the last frame?
CF_VISIBILITY_OBJECT_COUNT *(View Point)*	
Single	**Multi**
Assuming that the "<OBJECT-CF>" is moved by <X>, <Y>, and <Z> in the camera reference system, how many objects would be visible?	Assuming that the "<OBJECT-CF>" at the time of the first frame is moved by <X>, <Y>, and <Z> in the camera reference system, how many objects would have been visible in the last frame?
CF_KINEMATICS_DISTANCE_TRAVELED_INTERVAL *(Mechanics)*	
Single	**Multi**
–	Assuming that the "<OBJECT-CF>" at the time of the first frame is moved by <X>, <Y>, and <Z> in the camera reference system, what would have been the real-world displacement of the centroid of the <OBJECT> from the first to the last frame?
CF_COLLISION_OBJECT_OBJECT_FRAME_SINGLE *(Mechanics)*	
Single	**Multi**
Assuming that the "<OBJECT-CF>" is moved by <X>, <Y>, and <Z> in the camera reference system, which visible object would the <OBJECT> be colliding with in the frame?	–
CF_PERSISTENCE_OBJECT_TOTAL_COUNT_HIDDEN *(Persistence)*	
Single	**Multi**
–	Assuming that the "<OBJECT-CF>" at the time of the first frame is moved by <X>, <Y>, and <Z> in the camera reference system, how many objects would have been present but not visible in the last frame?

D.4 EXPERT-TO-NOVICE SPECIFICATION

We outline the queries utilized to test model adaptability across different domain expertise. We focus on a carefully curated set of seven physics-based questions. To systematically evaluate how different models handle varying degrees of complexity, each of the seven original questions is expanded into five distinct variations: child, teenager, undergraduate, graduate, and expert. These 5 levels were carefully selected to represent diverse comprehension levels. The combination of the 5 levels and 7 physical questions leads to a total of 11K questions.

Table 9: **VQA templates for Expert-to-novice**. Sample VQA question IDs with the different phrase at each of the 5 level of complexity.

ID	Questions
F_PHYSICS_PROPERTY_DENSITY_OBJECT_RELATIVE	**NewtPhys (baseline):** Considering all frames, which object, visible in the last frame, has the highest effective density? **Child:** Looking at all the pictures, if all the visible objects had the same size which one would be the heaviest? **Teen:** Looking at all the pictures, which object visible in the last frame do you think is the heaviest compared to how big it looks? **Undergrad:** Considering the full image sequence, which object visible in the last frame would have the highest mass relative to its volume? **Graduate:** Based on the entire frame sequence, which object visible in the final frame is likely to have the highest effective density (mass per unit volume)? **Expert:** Considering the entire temporal frame sequence, which visible object visible in the final frame has the highest effective density?
F_PHYSICS_PROPERTY_YOUNG_MODULUS_OBJECT_SIMILARF	**NewtPhys (baseline):** Considering all frames, which object, visible in the last frame, has a Young's Modulus most similar to that of the <OBJECT>? **Child:** Looking at all the pictures, which object you can still see at the end is just as hard to bend or squishe as the <OBJECT>? **Teen:** Looking at all the pictures, which object visible in the last frame would feel about as stiff or stretchy as the <OBJECT> if you tried to bend it? **Undergrad:** Considering the full image sequence, which object visible in the last frame appearas to have a similar stiffness (Young's modulus) to the <OBJECT>? **Graduate:** Based on the entire frame sequence, which object visible in the final frame would you expect to have a Young's modulus closest to that of the <OBJECT>? **Expert:** Considering the entire temporal frame sequence, which visible object visible in the final frame would you infer to have a Young's modulus most similar to that of the <OBJECT>?

ID	Questions
F_PHYSICS_PROPERTY_YOUNG_MODULUS_OBJECT_SIMILAR_NON_TECHNICAL	**NewtPhys (baseline):** Considering all frames, which object, visible in the last frame, has a softness most similar to that of the <OBJECT>? **Child:** Looking at all the pictures, which object you can still see at the end feels squishy or hard in the same way as the <OBJECT>? **Teen:** Looking at all the pictures, which object visible in the last frame would feel about as soft or hard as the <OBJECT>? **Undergrad:** Considering the full image sequence, which object visible in the last frame appears to have a similar softness to the <OBJECT>? **Graduate:** Based on the entire frame sequence, which object visible in the final frame would you expect to have a softness most similar to that of the <OBJECT>? **Expert:** Considering the entire temporal frame sequence, which object visible in the final frame would you infer to have a comparable effective softness to the <OBJECT>, based on its deformation or response to interaction?
F_PHYSICS_PROPERTY_YOUNG_MODULUS_HIGHEST	**NewtPhys (baseline):** Considering all frames, which object, visible in the last frame, exhibits the highest Young's Modulus? **Child:** Looking at all the pictures, which object you can still see at the end is the hardest and least bendy? **Teen:** After watching all the frames, which object visible in the last frame seems the stiffest or hardest to bend? **Undergrad:** Considering the full image sequence, which object visible in the last frame appears to be the stiffest? **Graduate:** Based on the entire frame sequence, which object visible in the final frame would you expect to have the highest Young's modulus? **Expert:** Considering the entire temporal frame sequence, which object visible in the final frame would you infer to exhibit the highest Young's modulus, assuming linear elastic behavior under comparable loading?
F_PHYSICS_PROPERTY_POISSON_RATIO_OBJECT_SIMILAR	**NewtPhys (baseline):** Considering all frames, which visible object has a Poisson ratio most similar to that of the <OBJECT> visible in the last frame? **Child:** Looking at all the pictures, which object you can still see at the end is as elastic as the <OBJECT> you see at the end? **Teen:** After watching all the frames, which object visible in the last frame changes its width in a similar way to the <OBJECT> in the last frame when it is squeezed or stretched? **Undergrad:** Considering the full image sequence, which object visible in the last frame shows a similar width change under stretching or compression as the <OBJECT> in the last frame? **Graduate:** Based on the entire frame sequence, which object visible in the final frame would you expect to have a Poisson ratio closest to that of the <OBJECT> visible in the final frame? **Expert:** Considering the entire temporal frame sequence, which object visible in the final frame would you infer to have a Poisson ratio most similar to that of the <OBJECT> visible in the final frame, based on comparable transverse-to-axial strain behavior?

ID	Questions
F_PHYSICS_PROPERTY_POISSON_RATIO_HIGHEST	**NewtPhys (baseline):** Considering all frames, which object, visible in the last frame, exhibits the largest Poisson ratio? **Child:** Looking at all the pictures, which object you can still see at the end spreads out the most on the sides when it is squished? **Teen:** After watching all the frames, which object visible in the last frame gets the widest on the sides when it is squeezed? **Undergrad:** Considering the full image sequence, which object visible in the last frame shows the greatest sideways expansion when compressed? **Graduate:** Based on the entire frame sequence, which object visible in the final frame would you expect to have the largest Poisson ratio? **Expert:** Considering the entire temporal frame sequence, which object visible in the final frame would you infer to exhibit the largest Poisson ratio, based on maximal transverse strain relative to axial strain?
F_PHYSICS_PROPERTY_POISSON_HIGH_LEVELF	**NewtPhys (baseline):** Considering all frames, if the <OBJECT>, visible in the last frame, were compressed vertically, how would its horizontal dimensions change? **Child:** Looking at all the pictures, if you squish the <OBJECT> you see at the end from top to bottom, how would it change? **Teen:** After watching all the frames, if the <OBJECT> in the last frame were pressed down, how would its width change? **Undergrad:** Considering the full image sequence, if the <OBJECT> visible in the last frame were compressed vertically, how would its horizontal size change? **Graduate:** Based on the entire frame sequence, if the <OBJECT> visible in the final frame were subjected to vertical compression, how would its horizontal dimensions be expected to respond? **Expert:** Given the full temporal sequence, if the <OBJECT> visible in the final frame were compressed along the vertical axis, how would its transverse (horizontal) dimensions change as dictated by its Poisson response?

D.5 VQA AUTOMATION

The construction of the VQA dataset is a non-trivial process that transforms the templates described in Section D.1 into the final set of questions illustrated in Section D.2.

The core idea is that each `question_id` corresponds to a dedicated function written in python. This function receives as input the simulation data describing the world state, together with optional `<OBJECT>` parameters required by some templates. The function first checks whether the question is answerable within the given simulation. If the required conditions are satisfied, the correct answer is computed, the corresponding visual evidence is extracted from the simulated frames (either single-frame or multi-frame), and the full metadata associated with the example is stored.

Confounding answer generation. In addition to the correct answer, a set of confounding answers is generated in order to construct the final multiple-choice question. The generation strategy depends on the type of answer:

- **Numerical answers.** Confounders are sampled around the correct value. The sampled values are distributed approximately symmetrically around the ground truth with equal spacing, where the spacing varies across questions to avoid predictable patterns. This produces plausible alternatives that remain close to the correct answer while maintaining sufficient separation.
- **Object-based answers.** When the answer corresponds to an object category or identity, confounders are sampled from objects that are visible in the scene. If no suitable candidates are available, sampling is performed from the global set of objects present in the dataset.

Object visibility criteria. A careful definition of object visibility is required, since some objects may appear in the frame with only a few pixels and would not be perceptible to a human observer. An object is therefore considered visible only if it satisfies the following conditions:

- the projected area in the image exceeds a minimum threshold of **2000 pixels**;

- at least **31%** of the object is visible within the image frame; d

These constraints ensure that objects used for both correct answers and confounding answers are visually meaningful.

Frame sampling. For single-image questions, a single frame is extracted from the simulation. For multi-frame questions, up to **eight frames** are sampled. Frames are selected at approximately uniform temporal intervals while preserving short temporal gaps between consecutive frames in order to maintain motion consistency. In practice, the separation between sampled frames ranges between **one and four frames**, depending on the dynamics of the simulation.

Determinism and parallelization. The entire generation pipeline is deterministic: all random operations are controlled through explicit seeding to guarantee reproducibility. Since each simulation and each question type can be processed independently, the pipeline is highly parallelizable and is implemented as a threaded program where questions are generated concurrently across simulations.

Ambiguity filtering. During the generation process, certain candidate questions may result in ambiguous or unverifiable answers. This can occur, for example, when multiple answers would be valid, when the relevant information is not sufficiently visible in the scene, or when the simulation state does not provide enough evidence to determine a unique answer. In such cases, the corresponding question is automatically flagged as *unanswerable*. All questions marked in this way are discarded during dataset construction and are therefore not included in the final evaluation set. This filtering step ensures that each retained question has a well-defined and uniquely verifiable ground-truth answer.

E Models specification

E.1 Vision-Language Models

In this section, we provide more details about the Visual Language Models that have been scrutinized in the paper.

In particular, our study focuses on what we refer to throughout the paper as "open models" and "closed models". Open models refer to both open-weight models, such as models for which only the weights have been released to the public, no training recipes, training data, or similar. On the contrary, we define open-source models as those that provide not only the model weights but also the training code, training data, and recipes. A few examples of open-weight models are: DeepSeekVL, paligemma2 etc. Among the open source models, instead, we have: Aquila-VL-2B, Molmo, InternVL family, etc. Closed models are usually frontier models, or widely recognized as SOTA by the community, and usually do not release model weights. Those models are allegedly much bigger in parameter count, and their model weights are not (generally) accessible to the public.

Figure 16 shows all the markers used in main paper's plots and their corresponding models.

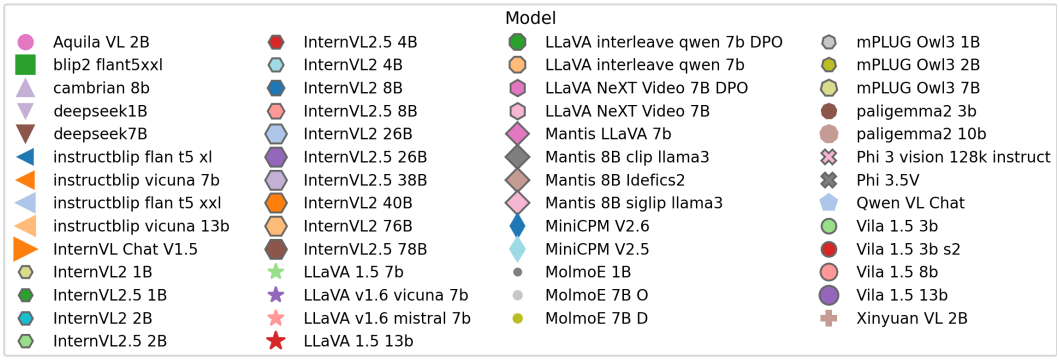


Figure 16: **Open models VLM marker listing.** We report here the complete legend of the VLMs markers used in the main paper.

Tab. 10 also provides a detail listing of the 54 open models VLMs used, spanning across 18 families, and covering both single-image models (*e.g.*, cambrian-8b) and multi-image models (*e.g.*, vila-1.5-3b).

Table 10: **Models metadata.** List of all the VLMs used in this work, with details on type, parameters count, year, and license.

#	Model ID	Family	type	Params (B)	Year	Licence
1	instructblip-flan-t5-xl	InstructBlip (Dai et al., 2023)	image-only	4.023	2023	mit
2	instructblip-flan-t5-xxl	InstructBlip (Dai et al., 2023)	image-only	12.310	2023	mit
3	instructblip-vicuna-7b	InstructBlip (Dai et al., 2023)	image-only	7.914	2023	other
4	instructblip-vicuna-13b	InstructBlip (Dai et al., 2023)	image-only	14.192	2023	other
5	blip2-flant5xxl	BLIP2 (Li et al., 2023)	image-only	12.230	2023	mit
6	llava-1.5-7b-hf	LLaVA (Liu et al., 2023)	image-only	7.063	2023	llama2
7	llava-1.5-13b-hf	LLaVA (Liu et al., 2023)	image-only	13.351	2023	llama2
8	llava-v1.6-mistral-7b-hf	LLaVA (Liu et al., 2024a)	image-only	7.567	2024	apache-2.0
9	llava-v1.6-vicuna-7b-hf	LLaVA (Liu et al., 2024a)	image-only	7.063	2024	llama2
10	deepseek1B	DeepSeekVL (Lu et al., 2024a)	image-only	1.975	2024	other
11	deepseek7B	DeepSeekVL (Lu et al., 2024a)	image-only	7.344	2024	other
12	Xinyuan-VL-2B	XinyuanVL (Group, 2024)	image-only	2.209	2024	apache-2.0
13	Aquila-VL-2B	AquilaVL (Gu et al., 2024)	image-only	2.179	2024	apache-2.0
14	Phi-3-vision-128k-instruct	Phi (Abdin et al., 2024)	general	4.147	2024	mit
15	Phi-3.5V	Phi (Abdin et al., 2024)	general	4.147	2024	mit
16	mPLUG-Owl3-1B-241014	Owl3 (Ye et al., 2025)	general	0.924	2024	apache-2.0
17	mPLUG-Owl3-2B-241014	Owl3 (Ye et al., 2025)	general	1.977	2024	apache-2.0
18	mPLUG-Owl3-7B-241101	Owl3 (Ye et al., 2025)	general	8.073	2024	apache-2.0
19	MiniCPM-V2	MiniCPMV (Yao et al., 2024)	image-only	3.435	2024	N/A
20	MiniCPM-V2.6	MiniCPMV (Yao et al., 2024)	image-only	8.099	2024	N/A
21	Qwen-VL-Chat	QwenVLChat (Bai et al., 2023)	image-only	9.600	2023	Qwen License
22	InternVL-Chat-V1-5-quantable	InternVLChat (Chen et al., 2024c)	image-only	25.514	2024	mit
23	llava-interleave-qwen-7b-hf	LLaVAInterleave (Li et al., 2024)	general	8.141	2024	other
24	llava-interleave-qwen-7b-dpo-hf	LLaVAInterleave (Li et al., 2024)	general	8.141	2024	other
25	vila-1.5-3b	VILAModel (Lin et al., 2024)	general	3.000	2024	cc-by-nc-sa-4.0 (weights); code Apache-2.0
26	vila-1.5-3b-s2	VILAModel (Lin et al., 2024)	general	3.000	2024	cc-by-nc-sa-4.0 (weights); code Apache-2.0
27	vila-1.5-8b	VILAModel (Lin et al., 2024)	general	8.000	2024	cc-by-nc-sa-4.0 (weights); code Apache-2.0
28	vila-1.5-13b	VILAModel (Lin et al., 2024)	general	13.000	2024	cc-by-nc-sa-4.0 (weights); code Apache-2.0
29	cambrian-8b	Cambrian (Tong et al., 2024)	image-only	8.333	2024	apache-2.0
30	paligemma2-3b	PaliGemma2 (Beyer et al., 2024)	image-only	3.033	2024	gemma
31	paligemma2-10b	PaliGemma2 (Beyer et al., 2024)	image-only	9.664	2024	gemma
32	LLaVA-NeXT-Video-7B-DPO-hf	LLaVAVideo (Zhang et al., 2024)	general	7.063	2024	llama2
33	LLaVA-NeXT-Video-7B-hf	LLaVAVideo (Zhang et al., 2024)	general	7.063	2024	llama2
34	MolmoE-1B	Molmo (Deitke et al., 2024)	image-only	1.000	2024	apache-2.0
35	MolmoE-7B-O	Molmo (Deitke et al., 2024)	image-only	7.665	2024	apache-2.0
36	MolmoE-7B-D	Molmo (Deitke et al., 2024)	image-only	8.021	2024	apache-2.0
37	InternVL2-1B	InternVLChat2 (Chen et al., 2024c)	general	0.938	2024	mit
38	InternVL2-2B	InternVLChat2 (Chen et al., 2024c)	general	2.206	2024	mit
39	InternVL2-4B	InternVLChat2 (Chen et al., 2024c)	general	4.147	2024	mit
40	InternVL2-8B	InternVLChat2 (Chen et al., 2024c)	general	8.075	2024	mit
41	InternVL2-26B	InternVLChat2 (Chen et al., 2024c)	general	25.514	2024	mit
42	InternVL2-40B	InternVLChat2 (Chen et al., 2024c)	general	40.069	2024	mit
43	InternVL2-76B	InternVLChat2 (Chen et al., 2024c)	general	76.262	2024	llama3
44	InternVL2_5-1B	InternVLChat2 (Chen et al., 2024c)	general	0.938	2024	mit
45	InternVL2_5-2B	InternVLChat2 (Chen et al., 2024b)	general	2.206	2024	mit
46	InternVL2_5-4B	InternVLChat2 (Chen et al., 2024b)	general	3.713	2024	mit
47	InternVL2_5-8B	InternVLChat2 (Chen et al., 2024b)	general	8.075	2024	mit
48	InternVL2_5-26B	InternVLChat2 (Chen et al., 2024b)	general	25.514	2024	mit
49	InternVL2_5-38B	InternVLChat2 (Chen et al., 2024b)	general	38.388	2024	mit
50	InternVL2_5-78B	InternVLChat2 (Chen et al., 2024b)	general	78.408	2024	other
51	Mantis-8B-Idefics2	Mantis (Jiang et al., 2024)	general	8.403	2024	apache-2.0
52	Mantis-llava-7b	Mantis (Jiang et al., 2024)	general	7.063	2024	apache-2.0
53	Mantis-8B-siglip-llama3	Mantis (Jiang et al., 2024)	general	8.480	2024	llama3
54	Mantis-8B-clip-llama3	Mantis (Jiang et al., 2024)	general	8.355	2024	llama3
55	GPT-5.5	GPT (OpenAI, 2026)	general	unknown	2026	Proprietary
56	GPT-6-astra	GPT (OpenAI, 2026)	general	unknown	2026	Proprietary
57	Gemini-3.1-flash	Gemini (Google DeepMind, 2026)	general	unknown	2026	Proprietary

E.2 Vision-Foundation Models

In our study, we assess ten prominent VFMs, detailed in Tab. 11.

For collision prediction, the objective is to estimate a binary mask of pixel locations where collisions occur. Performance is evaluated using the F1 score.

For gravity prediction, we ask the model to predict, for each pixel, (i) the gravity direction, evaluated using the Mean Angular Error (mAE), and (ii) the magnitude of the gravity force applied to the corresponding object,

evaluated using the absolute Magnitude Error (magE). We note that the predicted magnitude should depend on both the object's mass and the strength of the gravitational field in the environment. The task is non-trivial due to camera rotations and varying object masses.

As motion understanding is a key indicator of physical reasoning (Huang et al., 2024), we additionally investigate the task of scene flow estimation, which is evaluated using the Average Endpoint Error (AEE).

Model	**Architecture**	**Supervision**	**Dataset**
DeiT III	ViT-B/16	Classification	ImageNet-22k
SAM	ViT-B/16	Segmentation	SA-1B
MiDaS	ViT-L/16	Depth	MIX-6
MAE	ViT-B/16	SSL	ImageNet-1k
DINO	ViT-B/16	SSL	ImageNet-1k
DINOv2	ViT-B/14	SSL	LVD-142M
CLIP	ViT-B/16	VLM	WIT-400M
SigLIP	ViT-B/16	VLM	WebLI
AM-Radio	ViT-H/16	Distillation	DataComp-1B
StableDiffusion	UNet	Generation	LAION

Table 11: **Visual Foundation Models**' architecture, supervision types and training datasets