Grounded Physics Representations Enable Causal Reasoning in Language Models

Jesse Pokora
Department of Computer Science
University of Wisconsin – Milwaukee
Milwaukee, WI, USA
jpokora@uwm.edu

Tian Zhao
Department of Computer Science
University of Wisconsin – Milwaukee
Milwaukee, WI, USA
tzhao@uwm.edu

***Abstract*—This paper studies the role of explicit physical grounding in causal physics reasoning by comparing grounded, text-only, and tool-augmented language-based systems under controlled conditions. We introduce Grounded-Physics LM, a hybrid architecture that pairs an Isaac-Sim-trained physics transformer encoder (PhysicsFormer) with a lightweight language decoder (DistilGPT-2) via learned adapter layers. PhysicsFormer is trained exclusively on Isaac Sim simulations; only the prefix adapter is aligned on 90% of CLEVRER's validation causal-reasoning questions. On a matched 1,000-question baseline pool drawn from CLEVRER's validation set, our 171M parameter system (DistilGPT-2 + PhysicsFormer + adapter) achieves 69.2% overall accuracy above GPT-4o (59.4%), Gemini 2.0 Flash (59.0%), and Llama-3.3-70B (62.5%), accomplished with less than 1% of their language-model parameters. These results argue that explicit physical-state representations, learned on synthetic Isaac Sim physics (not CLEVRER), enable generalizable causal reasoning in compact models.**

***Index Terms*—symbol grounding, physics reasoning, LLM, grounded representation, causal reasoning, transformer**

I. Introduction

Large language models (LLMs) perform well on many reasoning benchmarks but persistently struggle with physical dynamics and spatial constraints [1], raising the question of whether text-only representations suffice for robust physical reasoning. Harnad's *symbol grounding problem* [2] frames this directly: purely symbolic systems suffer infinite regress, with tokens for "collision" or "trajectory" defined only by statistical co-occurrence rather than by physical reality. Prior physics-grounded systems such as VRDP [3] and ALOE [4] achieve strong CLEVRER [5] performance but require differentiable physics simulators, multi-stage pipelines, or video-supervised dynamics learning. Whether a simpler architecture conditioning a language model (LM) on explicit physics state can match this remains open.

We introduce *Grounded-Physics LM*, a hybrid architecture pairing a physics-simulation-trained encoder (*PhysicsFormer*, trained on NVIDIA Isaac Sim ground-truth state vectors) with a DistilGPT-2 decoder via learned adapter layers [6]. We evaluate against contemporary text-only LLMs on the CLEVRER causal-reasoning benchmark [5] under identical task specifications, isolating the effect of explicit physical grounding from scale, prompting, and tool access. Structured physical-state representations markedly improve causal reasoning, with our compact model competitive with systems orders of magnitude larger and outperforming them on forward prediction.

A. Contributions

Three contributions: (1) **continuous physics-language integration** via learned prefix embeddings [7], in contrast to PIGLeT's symbolic state interfaces [8] or PhyVLLM's video-implicit extraction [9]; (2) **controlled comparison** with each system measured on data it never trained on: Grounded-Physics LM on the disjoint 10% held-out CLEVRER partition (501 scenes, n=1,998, 710/361/927 explanatory/predictive/counterfactual), zero-shot LLM baselines on a matched 1,000-question pool, both single-frame with shuffled MCQ; (3) **empirical evidence** that grounded representations carry predictive physics: 63.4% on held-out predictive questions vs. GPT-4o 23.7% (+39.7 pp, non-overlapping 95% CIs), with text-only models doing well on retrospective attribution but failing on forward simulation.

II. Related Work

A. LLM Limitations and Intuitive Physics

Recent research documents systematic LLM weakness on spatial and physical reasoning. Rodionov *et al.* [10] found that frontier LLMs fail to respect physical constraints in spatial layouts; Williams *et al.* [1] concluded that LLMs lack fundamental spatial awareness; Pang *et al.* [11] showed that even with chain-of-thought prompting, LLMs struggle on physics benchmarks. While reasoning-focused models show improvement, they remain below grounded systems, suggesting text-only refinements do not address the grounding deficit.

DeepMind's PLATO system provides crucial precedent [12]. PLATO learns intuitive physics from *synthetic video*, representing the world as discrete objects and predicting future states. Unlike Grounded-Physics LM's structured state tensors, PLATO demonstrates that physics intuitions can emerge from visual experience, though it learns perceptual expectations rather than explicit causal reasoning capabilities. Critically, flat baseline models with equal or greater parameters failed to learn physical concepts that PLATO acquired reliably. This finding, that architectural structure matters more than scale

for physical understanding, directly supports our framework. Garrido *et al.* [13] similarly found that intuitive physics emerges from latent-space prediction, while pixel-space and text-based models perform near chance.

B. Physics-Grounded Systems

The move from text-only LLMs toward world-model integration [14] motivates several physics-language hybrids: Mind's Eye [15] uses simulation as an external tool; PIGLeT [8] uses symbolic state interfaces; the General Physics Transformer [16] lacks language integration. Concurrent PhyVLLM [9] extracts physics implicitly from video. Grounded-Physics LM differs by using explicit ground-truth state tensors with learned continuous embeddings, isolating the contribution of physics grounding from perceptual inference.

III. Methodology

We deliberately use DistilGPT-2 (82M) and a similarly compact encoder (50M) rather than larger or more recent architectures. This controls confounders: any observed gains cannot be attributed to pretraining-corpus scale, emergent capabilities, or LM-side architectural innovation, and the comparison against an identical DistilGPT-2 backbone without physics grounding directly quantifies the grounding contribution.

A. Grounded-Physics LM Architecture

Grounded-Physics LM combines a physics-specialized encoder with a language-model decoder (Fig. 1). Physics state tensors $[B, T, N, 35]$ (batch size, timesteps, object count, 35 state dimensions) pass through PhysicsFormer (8-layer Transformer with physics-biased attention) and a 64-token prefix adapter ($768 \to 64 \times 768$); the prefix concatenates with tokenized question embeddings and feeds DistilGPT-2 (6-layer, 82M). State vectors are ground-truth values from NVIDIA Isaac Sim [17] (deterministic rigid-body dynamics with position, velocity, orientation, angular velocity, and static properties), isolating physical grounding from perceptual noise. Prefix embeddings live in DistilGPT-2's embedding space and prepend to the tokenized input so attention can mix physics and language context.

a) PhysicsFormer Encoder: PhysicsFormer adapts the transformer [18] to numerical object-state tensors. Each object at time t has a 35D state vector $\mathbf{s}_i^t$ (position 3, velocity 3, orientation quaternion 4, angular velocity 3, plus 22D static properties: mass, friction, shape). Following the Graph Network Simulator paradigm [19], the encoder consumes the state-delta-augmented vector

$$\mathbf{x}_i^t = \left[\mathbf{s}_i^t;\, \mathbf{s}_i^t - \mathbf{s}_i^{t-1}\right] \in \mathbb{R}^{70}, \tag{1}$$

projected per-object via a shared feedforward $f_{\text{obj}} : \mathbb{R}^{70} \to \mathbb{R}^{768}$ to initial node embeddings $\mathbf{e}_i^t$ for relational reasoning.

b) Relational Modeling and Physics-Biased Attention.: For each ordered pair (i, j) the model computes pairwise physical features including distance, relative velocity, orientation differences, and angular velocity differences. These concatenate to a 12D pairwise feature $\mathbf{f}_{ij}$ mapped via $\mathbf{b}_{ij} = \tanh(W_2\,\mathrm{ReLU}(W_1\mathbf{f}_{ij})) \in \mathbb{R}^H$ to additive attention biases on each head. This introduces an inductive bias favoring physically meaningful interactions; the learned 768D embeddings encode patterns like "about to collide" or "moving as a group."

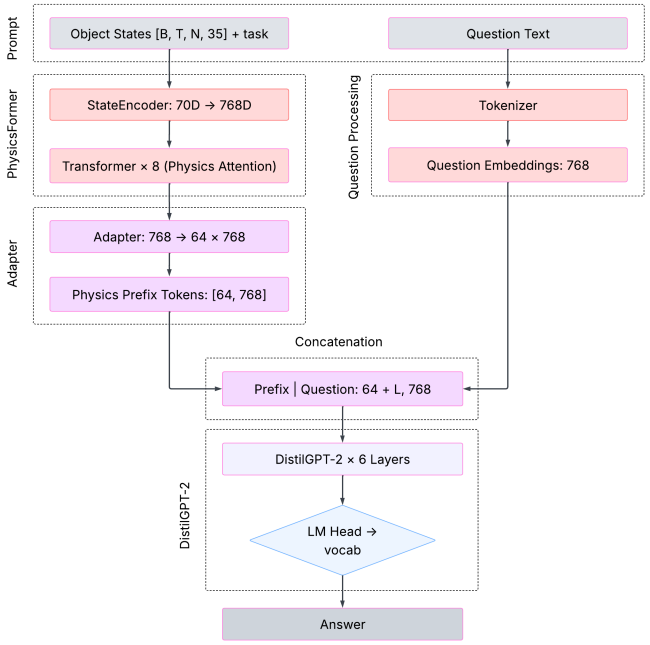

Fig. 1: Grounded-Physics LM Architecture. Object states $[B, T, N, 35]$ are processed by PhysicsFormer: a state encoder that projects the 70D augmented state (state + state-delta) to 768D embeddings, which pass through an 8-layer Transformer with physics-biased attention. The Adapter projects pooled physics embeddings to 64 prefix tokens of dimension 768. These are concatenated with question embeddings from the tokenizer and processed by DistilGPT-2's 6-layer decoder.

c) Adapter Architecture: The adapter projects pooled physics embeddings $768 \to 64 \times 768$ and reshapes the output into 64 prefix tokens that condition DistilGPT-2 generation [7], matching its embedding space.

d) Training Methodology: Unlike video-based systems (ALOE, PLATO), Grounded-Physics LM uses **supervised learning** throughout. PhysicsFormer is trained with next-state prediction and schema classification under AdamW with a 13-level curriculum (gravity/free-fall, collisions, stacking, complex dynamics), using RMSNorm, SwiGLU, and RoPE. Levels 12 and 13 add **causal training** (object dropout with intervention loss) and **counterfactual training** (contrastive "what if" learning). Adapter training uses supervised QA pairs in three phases with progressive unfreezing to prevent catastrophic forgetting.

e) Preventing Modality Collapse: To prevent the trained model from ignoring physics [20], [21], we add a physics-usage contrastive loss penalizing similarity between real and zeroed-physics prefix tokens. The trained adapter reaches mean cosine -0.85 between the two, i.e. near-opposite directions when physics changes.

B. Experimental Design

We evaluate on CLEVRER [5] using ground-truth scene annotations at frame 64 to construct Grounded-Physics LM's state tensors; baseline LLMs receive the same quantities as text. We focus on the three causal categories (explanatory, predictive, counterfactual; all 2–4-option MCQ, random baseline 25%). Descriptive questions are excluded because the LLM scene description lists each object verbatim as "{color} {material} {shape}", reducing items like "how many cyan objects?" to substring lookup against the prompt rather than physical reasoning. Grounded-Physics LM does not see these strings: color, material, and shape are one-hot dimensions inside the 35D state vector, not text tokens. Causal questions are immune to this leak, so restricting to causal categories measures physics reasoning rather than text-retrieval competence.

a) Primary evaluation pool: Each system is evaluated strictly on data it never trained on, with a sample size large enough to support tight Wilson 95% confidence intervals (CIs). *Grounded-Physics LM* is evaluated on the disjoint 10% held-out CLEVRER partition (501 scenes with scene_index $\in [14{,}499, 14{,}999]$, n=1,998 valid causal items, MCQ shuffled), which is the same partition used for all ablations (Section V-B), so that the primary numbers and ablations are computed on the same train/test split. Per-type counts are $710/361/927$ explanatory/predictive/counterfactual. *LLM baselines*: a matched n=1,000 zero-shot pool drawn from CLEVRER validation. LLMs never train on CLEVRER, so any 1,000-question subset is unbiased for them; we use the natural-distribution pool ($397/163/440$ explanatory/predictive/counterfactual) because it tracks CLEVRER's intrinsic question mix. For tighter predictive CIs, we additionally run a per-LLM n=1,000 predictive-only supplement.

b) Fair Comparison Protocol: Both pipelines evaluate on a **single frame** (frame 64, the midpoint of each 5-second 128-frame video). Grounded-Physics LM consumes a $[1, N, 35]$ state tensor (positions, velocities, mass, shape, color, material); LLMs receive the same quantities as text (attributes; positions (x, y, z); velocity direction and speed). Neither receives explicit collision-event labels or temporal trajectory information.

c) Representational asymmetry: The single-frame protocol is *not* a purely symmetric test. PhysicsFormer is trained on Isaac Sim sequences with explicit state-delta features (Eq. 1), meaning its encoder has an inductive bias for extrapolating future state from a position–velocity snapshot that text-only LLMs do not. Text-only models receive the same numerical quantities, but as a token sequence rather than as a structured state vector that a next-state-prediction objective has optimized over. This comparison therefore isolates *representational format with associated training objective* alongside physics grounding.

Baseline models include GPT-4o, Claude Sonnet 4, Claude 4.5 Sonnet, Gemini 2.0 Flash, DeepSeek-V3, Qwen3-235B (22B active MoE), Qwen2.5-7B, and Llama-3.3-70B. All baselines were evaluated zero-shot with greedy decoding (temperature=0) using identical prompt formats on 1,000 causal questions.

IV. Training Data

```
[INPUT: State Tensor [1, 3, 35]]
Frame 64 of 128, 3 objects
35D per object: pos, vel, quat, mass,
  radius, color, shape, friction, rest.

Object 0 (blue rubber cube):
  pos: [3.380, -1.140, 0.200]
  vel: [-2.960, 0.480, 0.0]
  mass: 1.0, radius: 0.36
Object 1 (brown rubber sphere):
  pos: [0.860, -0.750, 0.200]
  vel: [0.0, 0.0, 0.0]
  mass: 1.0, radius: 0.30
Object 2 (yellow rubber cylinder):
  pos: [0.670, 5.700, 0.200]
  vel: [-0.280, -2.700, 0.0]
  mass: 1.0, radius: 0.33

[INPUT: Question + Options]
Question: Without the blue cube, which
of the following will happen?
Options:
  A) The sphere and the cylinder collide
  B) The brown object and the cylinder collide
  C) The cylinder exits to the right
  D) The sphere exits to the left

[OUTPUT] Answer: C
```

Fig. 2: Counterfactual example showing the verbatim input of Grounded-Physics LM and its correct output. Issac Sim uses a right-handed coordinate system, where +Y is left.

A. PhysicsFormer Pre-training (Isaac Sim, no CLEVRER)

PhysicsFormer is pre-trained on NVIDIA Isaac Sim physics: $\approx 44{,}400$ episodes across 37 schemas in 11 curriculum groups of increasing complexity (gravity/free fall, collisions, stacking, rolling/sliding, projectiles, dominos, scattering, physics variety, articulated, rotation, complex dynamics; sequences up to 128 frames at 60 Hz). Levels 12–13 add causal training (object dropout with intervention loss) and counterfactual training (contrastive "what if" learning). PhysicsFormer uses 39,280 Isaac-Sim QA samples across 41 question types with a 90%/10% internal split. No CLEVRER data, neither scenes nor questions, is used for PhysicsFormer training.

B. Adapter Alignment on CLEVRER (with 10% held-out split)

The prefix adapter is aligned on CLEVRER's validation set because Isaac-Sim QA does not produce CLEVRER's causal types. We train on 90% of CLEVRER's 21,378 causal MCQ items (scenes $\sim$10k–14,498) and reserve the disjoint 10% ($\sim$501 scenes, scene indices $\sim$14,499–14,999, n=1,998 valid) as a generalization test the adapter never sees. Training uses a language-modeling loss on the correct-choice text with per-epoch MCQ-shuffling; Phase-1 trains the prefix with LLM frozen, Phase-2 adds LoRA on attention, Phase-3 merges LoRA and enables the physics-usage contrastive loss throughout. Fig. 2 shows the input format for Grounded-Physics LM.

V. Experimental Results

A. Key Findings

Fig. 3 reports per-question-type accuracy with each system measured on data it never trained on: Grounded-Physics LM on the disjoint 10% held-out CLEVRER partition; LLM baselines on a matched zero-shot n=1,000 pool, with predictive numbers from a per-LLM n=1,000 predictive-only supplement that narrows the LLM CIs ($\pm$3 pp) without changing the comparison's fairness. All systems run single-frame at frame 64 with shuffled MCQ.

a) Overall accuracy: Grounded-Physics LM reaches 69.2% on the held-out partition (n=1,998, Wilson $[67.2, 71.2]$), +6.7 pp above the best LLM baseline (Llama-3.3-70B 62.5%, $[59.5, 65.4]$, non-overlapping CIs) and above every frontier LLM (GPT-4o 59.4%, Claude 4.5 Sonnet 59.6%, Gemini 2.0 Flash 59.0%), with $400\times$ fewer LM parameters.

b) Predictive reasoning: Grounded-Physics LM scores 63.4% (n=361, Wilson $[58.3, 68.2]$): +10.9 pp over the best LLM (Qwen2.5-7B 52.5% on the n=1,000 predictive supplement) with non-overlapping Wilson 95% CIs $[58.3, 68.2]$ vs. $[49.4, 55.6]$. GPT-4o collapses to 23.7% ($[21.2, 26.4]$); within that single model, the explanatory-vs-predictive gap is $\sim$45 pp (68.5% vs. 23.7%), suggesting retrospective attribution and forward prediction rely on different mechanisms.

c) Counterfactual and explanatory reasoning: Grounded-Physics LM scores 79.4% explanatory (n=710, Wilson $[76.3, 82.2]$), modestly above the strongest LLM (Claude 4.5 Sonnet 75.8%, $[71.4, 79.8]$); the 95% CIs overlap, so the explanatory column is best read as comparable to top LLMs. Counterfactual: 63.6% (n=927, $[60.5, 66.7]$) is bracketed by GPT-4o no-tools at 66.4% ($[61.8, 70.6]$) and Llama-3.3-70B at 65.5% ($[60.9, 69.7]$), with overlapping CIs. Critically, zero-physics on this same held-out partition drops explanatory accuracy to 16.2% and counterfactual to 2.4% (Section V-B), so the model uses physics features even on categories where the LLM gap is smaller.

B. Ablation: Grounding Contribution

We conduct three ablations (Table I) on the held-out validation subset (501 scenes, n=1,998 valid questions never seen during adapter training).

a) Zero-Physics.: Zeroing physics state tensors at encoder input (preserving adapter and DistilGPT-2) collapses accuracy from 69.2% to $\mathbf{7.2}\%$, far below random chance. Predictive (1.9%) and counterfactual (2.4%) fall to floor. The -62.0 pp overall gap is significant at $p < 0.001$.

b) Zero-Prefix + Shuffle: the "true text-only" baseline.: Zeroing only the 64 adapter prefix tokens while shuffling MCQ choice order yields accuracy of $\mathbf{65.6}\%$. Relative to this baseline, the physics prefix contributes +3.6 pp overall (p=0.015), with the largest per-type effect on predictive (+7.7 pp, p=0.041). The small prefix-only Δ shows that most of the physics signal is already baked into the LoRA-modified attention weights during training.

c) Asymmetry reveals trust.: The key evidence for genuine physics integration is the *asymmetry* between the two ablations: zero-physics (7.2%) is far worse than zero-prefix-shuffled (65.6%), a 58.4 pp gap. The adapter's prefix tokens are highly differentiated (cosine $= -0.85$ between real and zeroed); when they encode misleading "no physics" signals the LM trusts and follows that guidance, ruling out the interpretation that the model merely ignores the prefix.

Condition	Overall	Expl.	Pred.	Ctr.
Grounded-Physics LM (shuffled)	69.2	79.4	63.4	63.6
Zero Physics	7.2	16.2	1.9	2.4
Zero Prefix + Shuffle	65.6	75.9	55.7	61.5
Δ Physics contribution	+62.0	+63.2	+61.5	+61.3
Δ Prefix contribution	+3.6	+3.5	+7.7	+2.1

TABLE I: Ablation study (%) on held-out subset (n=1,998). All physics effects significant at $p < 0.001$.

C. Object Complexity Scaling: 15-Object Benchmark

We stress-test scene complexity by augmenting CLEVRER validation scenes to 15 objects (vs. the standard 3–6) under the same protocol. Results show Grounded-Physics LM degrades from 69.2% to 63.1% overall on the 15-object 1K pool (-6.1 pp), mid-pack among LLMs. **The predictive advantage persists at scale**: Ours $64.6\% \pm 3.4$ vs. DeepSeek-V3 $53.8\% \pm 3.9$ vs. Llama-3.3-70B $48.8\% \pm 4.0$ (non-overlapping CIs), while Claude 4.5 Sonnet collapses to 16.9% and GPT-4o to 26.2%.

a) Inverse Complexity Scaling.: Counterintuitively, several frontier LLMs *improve* on the harder 15-object scenes (Anthropic no-tools variants: +12.8 pp, +11.1 pp). This is analogous to inverse-scaling where capacity degrades performance through a "distractor task" mechanism.

D. Comparison to Prior Systems

Published neuro-symbolic systems on CLEVRER, VRDP [3] (82.0%), ALOE [4] (79.0%), and NS-DR [5] (72.7%), achieve their accuracy with full 128-frame video plus specialized architectures. Grounded-Physics LM reaches 69.2% on a training-disjoint 10% held-out partition using a single frame, a strictly harder input regime requiring inference from an instantaneous snapshot rather than an observed trajectory.

E. Mechanistic Analysis

To understand *why* physics grounding improves reasoning, we conduct mechanistic analysis using gradient-based saliency and probing.

a) Feature Importance: By gradient magnitude, position contributes most (694.3), followed by velocity (390.2), mass/size (334.1), and friction (253.8). Color (183.1) and shape (133.1) show moderate importance for object identity.

b) Probing Experiments: Linear probing reveals where physics knowledge is encoded:

- **Object count**: Perfectly recoverable from input ($R^2 = 0.998$) but progressively *lost* through transformer layers (pooled: $R^2 = 0.056$). The network compresses object-level detail into relational features.

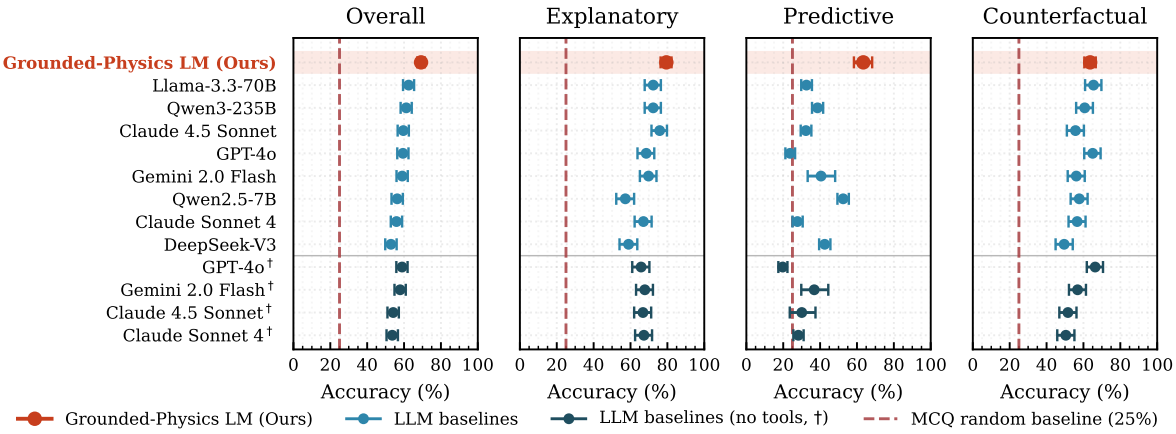

Fig. 3: **Primary cross-system comparison** on CLEVRER causal-MCQ pool: Wilson 95% CIs for Grounded-Physics LM (Ours, n=1,998 valid items on disjoint 10% held-out partition) and all frontier LLM baselines (zero-shot, n=1,000 each) per question type. Per-type n: Ours are $710/361/927$ explanatory/predictive/counterfactual; LLMs are $397/163/440$ explanatory/predictive/counterfactual and additional 1,000 predictive. **Predictive panel is the key finding**: Grounded-Physics LM CI $[58.3, 68.2]$ does not overlap any LLM. LLM scale ranges 7B–671B parameters; Grounded-Physics LM is 173M.

- **Collision prediction**: High accuracy (92.5%) at *all* layers; collision-relevant representations are computed early and robustly preserved.
- **Spatial centroid**: Best at encoder ($R^2 = 0.87$), remains decodable through layers ($R^2 \geq 0.60$).

c) From Probing Results to Behavioral Patterns: This profile maps directly onto question-type performance. **Explanatory** questions require identifying specific interactions: the 92.5% collision-prediction accuracy provides exactly the needed feature. **Predictive** questions require extrapolating future trajectories: spatial structure remains decodable ($R^2 \geq 0.60$ through all layers). **Counterfactual** questions require tracking removed-object identity: the R^2 collapse from 0.998 to 0.056 at pooling explains weaker counterfactual performance.

d) Encoder physics signal abstracts across domains.: Running the same probes on out-of-distribution Isaac Sim physics scenes the adapter was never trained on shows encoder-pool R^2 for object count is essentially identical across domains (CLEVRER 0.42 vs. Isaac 0.42). Spatial centroid is recovered at R^2=0.64 on CLEVRER and R^2=0.83 on Isaac Sim. Collision presence is decoded above chance on both domains. This rules out CLEVRER-specific representational capture.

F. Qualitative Error Analysis

Failures cluster into three patterns matching the mechanistic account. **Predictive failures** predict the correct event type but wrong object pair (consistent with spatial structure preserved but object identity compressed). **Counterfactual failures** concentrate on removed-object selection (consistent with collapsed object-count probe). **Explanatory failures** involve multi-cause attribution (mode-collapse on attribution distribution, not physics failure). Across the held-out subset, object-pair swap dominates at 78.1% of verbatim-wrong picks, directly confirming the prefix preserves event detection while losing object discriminability.

VI. Discussion

A. Interpretation Through Symbol Grounding

Our predictive-reasoning gap is consistent with Harnad's symbol grounding hypothesis: text-only LLMs achieve strong accuracy on retrospective questions, where statistical patterns over captioned video descriptions are likely present in pretraining (GPT-4o: 68.5% explanatory) but collapse on forward prediction (GPT-4o: 23.7% predictive). Grounded-Physics LM closes this gap by conditioning on explicit kinematic state rather than on token co-occurrence statistics.

We do not claim that PhysicsFormer's learned representations constitute *sensorimotor* grounding in Harnad's strict sense. A defensible narrow reading is that the encoder simply supplies a better-structured statistical prior keyed to object-level kinematics rather than to text co-occurrence. The ablation asymmetry, predictive-reasoning gap, and question-type-specific probing signatures are each consistent with this narrower reading.

B. Corroborating Evidence

Independent studies corroborate our findings. Using Physics Context Builders on CLEVRER, Lin *et al.* [22] found GPT-4o achieves 62.7% on descriptive questions but falls to just 18.7% on predictive, closely matching our 23.7% measurement. PhysUniBench [23] found even sophisticated models achieve only 17–37% on undergraduate physics. Research on epistemic calibration revealed a "reasoning paradox" where extended reasoning *worsens* performance [24], consistent with smaller models (Qwen2.5-7B at 52.5% predictive) outperforming much larger ones (Qwen3-235B at 38.6%).

C. Limitations

a) CLEVRER-only evaluation.: All behavioral results come from CLEVRER. Encoder-level cross-domain probing partially closes the gap by showing the *physics signal* decodes at comparable rates on non-CLEVRER data; a behavioral out-of-distribution evaluation is the immediate next step.

b) Single-frame representational asymmetry.: PhysicsFormer's state-delta features give it a learned inductive bias for snapshot extrapolation that text-only LLMs do not share. The single-frame comparison therefore isolates representational format alongside physics grounding.

c) Adapter scope limits.: The adapter is trained for MCQ selection, not free-form generation (99.45% "unknown" without choices), and is brittle to paraphrased choice text. These are training-condition scope effects, fixable by mixed-format and paraphrase-augmented training.

d) Baseline prompting.: LLM baselines were evaluated zero-shot with greedy decoding; few-shot, chain-of-thought, and structured-prompt variants could narrow the predictive gap. Independent measurements bracket our baseline, indicating it is not catastrophically under-prompted.

VII. Conclusion

Our 173M-parameter system achieved 69.2% overall accuracy on a disjoint 10% held-out partition, $+6.7$ pp above Llama-3.3-70B while using $400\times$ fewer parameters. The predictive-reasoning gap is the clearest behavioral signature: 63.4% on held-out predictive questions vs. GPT-4o's 23.7%. Zero-physics collapses accuracy to 7.2% (-62.0 pp, $p < 0.001$), confirming the model trusts and follows the physics pathway. Whether interpreted through Harnad's symbol grounding framework or representation-learning principles, explicit physical-state representations enable forward-simulation reasoning in compact models that remains out of reach for much larger text-only systems.

A natural extension is **action prediction**: given a desired goal state, what intervention achieves it? This requires *inverse physics* (working backward from outcomes to causes), the core capability underlying goal-conditioned planning in robotics. Future work could extend Grounded-Physics LM with an action prediction head enabling sim-to-sim transfer and edge-deployed onboard robotic systems.

Implementation code, trained model weights, and evaluation scripts are available at https://github.com/uwm-se/PhysicsFormer.

References

[1] S. Williams and J. Huckle, "Easy problems that LLMs get wrong," *arXiv preprint arXiv:2405.19616*, 2024.

[2] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[3] M. Ding, Z. Chen, T. Du, P. Luo, J. B. Tenenbaum, and C. Gan, "Dynamic visual reasoning by learning differentiable physics models from video and language," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[4] D. Ding, F. Hill, A. Santoro, M. Reynolds, and M. Botvinick, "Attention over learned object embeddings enables complex visual reasoning," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[5] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "CLEVRER: Collision events for video representation and reasoning," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=HkxYzANYDB

[6] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *International Conference on Machine Learning*, 2019, pp. 2790–2799.

[7] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 4582–4597.

[8] R. Zellers, A. Holtzman, M. Peters, R. Mottaghi, A. Kembhavi, A. Farhadi, and Y. Choi, "PIGLeT: Language grounding through neuro-symbolic interaction in a 3D world," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 2040–2050.

[9] Y. Zhan, X. Wang, H. Chen, T. Feng, W. Feng, R. Wang, G. Li, Q. Li, and W. Zhu, "PhyVLLM: Physics-guided video language model with motion-appearance disentanglement," *arXiv preprint arXiv:2512.04532*, 2025.

[10] F. Rodionov, A. Eldesokey, M. Birsak, J. Femiani, B. Ghanem, and P. Wonka, "FloorplanQA: A benchmark for spatial reasoning in LLMs using structured representations," *arXiv preprint arXiv:2507.07644*, 2025.

[11] X. Pang, R. Hong, Z. Zhou, F. Lv, X. Yang, Z. Liang, B. Han, and C. Zhang, "Physics reasoner: Knowledge-augmented reasoning for solving physics problems with large language models," *arXiv preprint arXiv:2412.13791*, 2024.

[12] L. S. Piloto, A. Weinstein, P. Battaglia, and M. Botvinick, "Intuitive physics learning in a deep-learning model inspired by developmental psychology," *Nature Human Behaviour*, vol. 6, no. 9, pp. 1257–1267, 2022.

[13] Q. Garrido, N. Ballas, M. Assran, A. Bardes, L. Najman, M. Rabbat, E. Dupoux, and Y. LeCun, "Intuitive physics understanding emerges from self-supervised pretraining on natural videos," *arXiv preprint arXiv:2502.11831*, 2025.

[14] T. Feng, X. Wang, Y. Jiang, and W. Zhu, "Embodied AI: From LLMs to world models," *arXiv preprint arXiv:2509.20021*, 2025.

[15] R. Liu, J. Wei, S. S. Gu, T.-Y. Wu, S. Vosoughi, C. Cui, D. Zhou, and A. M. Dai, "Mind's eye: Grounded language model reasoning through simulation," in *International Conference on Learning Representations*, 2022.

[16] F. Wiesner, M. Wessling, and S. Baek, "Towards a physics foundation model," *arXiv preprint arXiv:2509.13805*, 2025.

[17] NVIDIA Corporation, "NVIDIA Isaac Sim," https://developer.nvidia.com/isaac-sim, 2023, high-fidelity robotics simulation platform built on Omniverse.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[19] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. W. Battaglia, "Learning to simulate complex physics with graph networks," in *International Conference on Machine Learning*, 2020, pp. 8459–8468.

[20] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6904–6913.

[21] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4971–4980.

[22] H. Lin *et al.*, "Physics context builders: A modular framework for physical reasoning in vision-language models," *arXiv preprint arXiv:2412.08619*, 2024.

[23] Y. Wang *et al.*, "PhysUniBench: A comprehensive benchmark for undergraduate physics reasoning," *arXiv preprint*, 2024, evaluates LLMs on undergraduate physics with 17–37% accuracy.

[24] M. Shojaee *et al.*, "Do large language models know what they don't know? evaluating epistemic calibration via prediction markets," *arXiv preprint arXiv:2512.16030*, 2024.