PHYSLANG: A SMALL DIAGNOSTIC FRAMEWORK FOR LANGUAGE-GROUNDED WORLD MODELING

Noor Mairukh Khan Arnob
University of Asia Pacific, Bangladesh
arnob@uap-bd.edu

Azmine Toushik Wasi
Shahjalal University of Science and Technology, Bangladesh
azminetoushik.wasi@gmail.com

ABSTRACT

World models that understand and respond to natural language instructions are a critical step toward intelligent agents operating in human environments. However, it remains unclear whether current neural world models truly ground language in physical dynamics or merely exploit superficial correlations. We present PHYSLANG, a small, diagnostic benchmark designed for falsification rather than performance maximization, evaluating language grounding in neural world models within a controlled rigid-body physics simulation. PHYSLANG consists of 2,000 short-horizon episodes governed by explicit natural language rules (e.g., "red blocks are heavy"), with evaluation splits targeting compositional recombination and logical contradiction under fixed dynamics. We evaluate five sequence architectures: MLP, GRU, LSTM, Transformer, and Mamba, on predicting future physical states conditioned on language. Our experiments reveal: (i) systematic interactions between architectural inductive biases and language-conditioned dynamics, with recurrent models exhibiting lower short-horizon prediction error; (ii) extreme sensitivity to linguistic perturbations across all architectures (50–296% error increase when language is shuffled), indicating reliance on fragile correlations; and (iii) consistent failure under contradictory rules, where models interpolate between conflicting constraints rather than resolve them. Together, these findings suggest that contemporary neural world models encode language as a weak contextual prior rather than as a causal constraint on physical dynamics. PHYSLANG exposes concrete failure modes in language-grounded world modeling and provides a controlled testbed for studying grounded reasoning beyond benchmark performance. PHYSLANG is available at: https://github.com/Codernob/PhysLang.

1 INTRODUCTION

World models (Ha & Schmidhuber, 2018; LeCun, 2022) aim to learn predictive representations of environment dynamics, enabling agents to reason, plan, and act by simulating future states without direct interaction. Recent foundation world models (Bruce et al., 2024; Agarwal et al., 2025) have demonstrated impressive capabilities in generating realistic, interactive environments from video data. These advances suggest a promising path toward general-purpose agents operating in complex, real-world settings. However, most existing world models are trained and evaluated primarily on perceptual prediction, with limited emphasis on explicit reasoning about abstract or semantic constraints. As a result, a critical question remains largely unexplored: *can world models understand and respond to natural language descriptions of physical rules?* This question motivates a closer examination of language grounding in world modeling.

Despite this progress, the ability to reason about physical dynamics under explicit semantic constraints remains largely untested. Classic world models developed in model-based reinforcement learning emphasize predictive accuracy and planning utility but typically operate without language or rely on implicit correlations in state transitions (Ha & Schmidhuber, 2018; Hafner et al., 2019). More recent

foundation world models scale perceptual prediction through large-scale video pretraining (Bruce et al., 2024; Agarwal et al., 2025), yet their impressive visual realism does not guarantee sensitivity to abstract rules governing physical behavior. In parallel, benchmarks for intuitive physics focus on perceptual plausibility and causal structure but omit language as an active control signal (Bear et al., 2021). This leaves open a fundamental question of whether contemporary world models can treat natural language not merely as context, but as a binding constraint on physical dynamics.

Grounding natural language in world models is essential for human-centered interaction, as it enables agents to follow verbal instructions, adapt to changing rules, and generalize compositionally by recombining known concepts in novel contexts. Moreover, language grounding offers a potential route to interpretability, allowing model behavior to be explained in terms of human-understandable rules rather than opaque latent representations. Despite its importance, existing evaluation benchmarks do not directly test whether language acts as a governing constraint on physical dynamics. Prior work has emphasized visual fidelity (Liang et al., 2025), physical plausibility (Bear et al., 2021; Bansal et al., 2024), or symbolic and text-driven world generation (Hu et al., 2025). These benchmarks either lack interactivity, rely on indirect metrics, or decouple language from continuous physics prediction. As a result, they provide limited insight into whether models truly *understand* language-specified physical rules or merely exploit superficial correlations.

In this work, we propose a small and controlled diagnostic setting to isolate and test language grounding in world models, prioritizing falsification over large-scale realism or state-of-the-art performance (Section 3). We introduce PHYSLANG, a benchmark of 2,000 rigid-body physics episodes in which explicit natural language rules directly determine physical properties such as mass and friction (Section 4). The benchmark includes evaluation splits that systematically test compositional recombination of known rule primitives as well as logical contradiction, allowing us to probe whether models treat language as a governing causal constraint or merely as a weak contextual signal. We conduct a comprehensive diagnostic evaluation of five sequence architectures (Section 5), complemented by ablation studies that quantify sensitivity to linguistic perturbations (Section 6). This controlled setup enables fine-grained analysis of how architectural inductive biases interact with language-conditioned dynamics. Our results reveal consistent and systematic failure modes that remain obscured in existing benchmarks focused on perceptual fidelity or aggregate performance.

2 RELATED WORK

World Models. Learning world models for planning and control has a long history in model-based reinforcement learning, where agents learn predictive representations of environment dynamics to support imagination-based reasoning and decision making (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020). These approaches emphasize compact latent dynamics models that can be rolled forward to evaluate candidate actions, typically prioritizing predictive accuracy and planning utility. More recently, foundation world models have scaled this paradigm by leveraging large-scale video pretraining to generate realistic and interactive environments across diverse domains (Bruce et al., 2024; Agarwal et al., 2025; Decart et al., 2024). Such models demonstrate impressive visual fidelity, temporal coherence, and controllability, suggesting potential as general-purpose simulators. However, their evaluation protocols largely focus on perceptual realism and behavioral plausibility rather than on whether learned dynamics obey explicit physical constraints. Consequently, it remains unclear whether these models internalize abstract rules governing physical behavior, particularly when such rules are specified symbolically or linguistically rather than implicitly through data correlations.

Physics and Physical Reasoning Benchmarks. A complementary line of work has developed benchmarks for physical reasoning and intuitive physics, aiming to assess whether models capture core principles such as object permanence, causality, and collision dynamics. Datasets such as Physion (Bear et al., 2021), IntPhys (Riochet et al., 2020), and PhysBench (Chow et al., 2025) provide controlled environments with carefully designed physical tests that go beyond raw visual prediction. These benchmarks have been instrumental in revealing failure modes in perceptual physics understanding and causal inference. Newton (Campbell, 2025) further extends this direction by targeting interactive foundation world models and evaluating long-term consistency under rigid-body dynamics. However, language plays little to no role in these settings, or is treated only as metadata rather than as an active determinant of dynamics. As a result, these benchmarks do not address whether a model can modify or constrain its physical predictions in response to explicit semantic instructions.

Language-Conditioned Models. Language conditioning has been explored extensively in embodied robotics, video generation, and planning, demonstrating that models can align actions, trajectories, or visual outputs with textual instructions (Brohan et al., 2023; Yang et al., 2024; Huang et al., 2022). In these systems, language often functions as a high-level goal specification or contextual prompt that guides policy execution or content generation. This paradigm has enabled impressive zero-shot generalization and human-interpretable control, particularly when combined with large pretrained language models. Text2World (Hu et al., 2025) formalizes this idea by evaluating language models on constructing symbolic world representations using PDDL, emphasizing discrete reasoning and planning. However, such approaches abstract away continuous dynamics and physical realism, focusing instead on symbolic correctness. Consequently, they do not test whether neural world models can enforce language-specified physical rules when predicting future states in continuous, physics-governed environments.

Implication for our work. PHYSLANG bridges these lines of work by introducing a controlled physics benchmark in which natural language rules directly govern continuous physical dynamics, enabling explicit tests of language grounding in neural world models. Unlike prior benchmarks that emphasize perceptual realism, symbolic planning, or unconditioned physics, our diagnostic setting isolates whether language functions as a causal constraint on predicted dynamics.

3 PROBLEM FORMULATION

We study a simple physics prediction setting in which objects interact according to fixed physical laws, while their material properties are specified using natural language. In each episode, a short history of object states is observed, along with a language description that defines how certain physical properties (e.g., mass or friction) apply to objects. The model's task is to use both the observed motion and the language description to predict how the scene will evolve over time. This design isolates whether and how language is used to constrain physical dynamics, rather than merely providing contextual or descriptive information.

Task Definition. We formulate language-grounded world modeling as a conditional sequence prediction task. Let $\mathcal{S} \subset \mathbb{R}^d$ denote the state space, where each state $\mathbf{s}_t \in \mathcal{S}$ encodes the physical properties of all objects at timestep t. Let $\mathcal{L}$ denote the space of natural language descriptions. A language-grounded world model is a mapping $\mathcal{M} : \mathcal{S}^T \times \mathcal{L} \to \mathcal{S}^H$ that predicts future states $\hat{\mathbf{s}}_{T+1:T+H}$ given a context sequence $\mathbf{s}_{1:T}$ and a language description $\ell \in \mathcal{L}$:

$$\hat{\mathbf{s}}_{T+1:T+H} = \mathcal{M}(\mathbf{s}_{1:T}, \ell) \tag{1}$$

where T is the context length and H is the prediction horizon.

State Representation. Each state $\mathbf{s}_t$ is a concatenation of per-block features for N objects:

$$\mathbf{s}_t = [\mathbf{b}_t^{(1)}; \mathbf{b}_t^{(2)}; \ldots; \mathbf{b}_t^{(N)}] \in \mathbb{R}^{N \times 9} \tag{2}$$

where each block state $\mathbf{b}_t^{(i)} \in \mathbb{R}^9$ consists of:

$$\mathbf{b}_t^{(i)} = [\underbrace{\mathbf{p}_t^{(i)}}_{\text{position}}; \underbrace{\mathbf{v}_t^{(i)}}_{\text{velocity}}; \underbrace{m^{(i)}}_{\text{mass}}; \underbrace{\mu^{(i)}}_{\text{friction}}; \underbrace{r^{(i)}}_{\text{scale}}] \tag{3}$$

with $\mathbf{p}_t^{(i)}, \mathbf{v}_t^{(i)} \in \mathbb{R}^3$ and $m^{(i)}, \mu^{(i)}, r^{(i)} \in \mathbb{R}$. Crucially, **color information is excluded** from the state representation, so models must learn to associate language descriptions with observed dynamics.

Language Encoding and Training. Language descriptions ℓ are encoded using a frozen sentence transformer (Reimers & Gurevych, 2019): $\mathbf{z}_\ell = \text{SentenceTransformer}(\ell) \in \mathbb{R}^{384}$. We use `all-MiniLM-L6-v2` for computational efficiency. Models are trained to minimize the mean squared error between predicted and ground truth future states:

$$\mathcal{L}(\theta) = \mathbb{E}_{(\mathbf{s},\ell)\sim\mathcal{D}} \left[\frac{1}{H} \sum_{h=1}^{H} \|\hat{\mathbf{s}}_{T+h} - \mathbf{s}_{T+h}\|_2^2 \right] \tag{4}$$

4 PHYSLANG BENCHMARK

Physics Simulation. The benchmark is built on a controlled rigid-body physics environment implemented using Isaac Sim (NVIDIA, 2022), which provides deterministic and high-fidelity simulation of contact dynamics, friction, and collisions. Each episode consists of 3–6 rigid blocks interacting under gravity on a planar surface, with initial positions, velocities, and orientations sampled from fixed bounded distributions to ensure diversity while preserving stability. Objects are visually distinguished by discrete colors (red, blue, green), which serve as the only perceptual cues used by the language rules. Simulations run for 500 timesteps with a fixed integration step of $\Delta t = 0.01$s, yielding short-horizon trajectories where physical effects of rule changes are observable but not trivially predictable. The simulator enforces identical physical laws across all episodes, ensuring that variation in dynamics arises solely from rule-conditioned object properties rather than environmental changes. Figure 1 summarizes the structure of episodes and their associated language annotations.

Table 1: Dataset statistics for PHYSLANG.

Statistic	Value
Total episodes	2,000
Blocks per episode	3–6
Timesteps per episode	500
State dimension	54 (6 blocks $\times$ 9 features)
Context length T	20
Prediction horizon H	10
Language embedding dim	384
Unique rules / Gravity variations	11 / 6

Language Rules. Language rules specify deterministic mappings from object color to latent physical properties, formally defining how semantic attributes constrain the underlying physics. Each rule takes the form of a declarative natural language statement assigning a scalar physical parameter to all objects of a given color, and rules are interpreted as hard constraints rather than probabilistic preferences. Two classes of rules are considered: **(i) mass rules**, which affect momentum transfer and collision outcomes without altering gravitational acceleration, and **(ii) friction rules**, which modulate tangential forces during surface contact. For example, a mass rule such as "red blocks are heavy" assigns a fixed mass value of 15.0 kg to all red objects, while a friction rule such as "blue blocks slide easily" assigns a coefficient of friction $\mu = 0.001$. Rules are introduced dynamically between timesteps 10 and 100 within each episode, creating a temporal dependency where models must infer and propagate the consequences of linguistic constraints over future states. The full rule set and parameter values are enumerated in Appendix A.

Evaluation Splits. The dataset is partitioned into evaluation splits designed to isolate distinct aspects of language grounding rather than overall predictive accuracy. The training split contains episodes governed by a fixed subset of rules that define the base semantic-to-physical mappings learned by the models. A compositional split evaluates generalization to novel combinations of known primitives, where familiar properties (e.g., mass or friction) are reassigned to previously unseen colors or paired in unseen configurations. A contradiction split probes logical robustness by introducing rules that explicitly negate training-time associations, forcing models to reconcile conflicting semantic constraints with prior experience. Finally, control ablations decouple language from dynamics by either shuffling rule text across episodes or providing incorrect rule descriptions, allowing us to quantify reliance on linguistic input versus visual or statistical cues. Together, these splits enable fine-grained diagnosis of whether language functions as a causal determinant of dynamics or merely as an auxiliary contextual signal.

5 EXPERIMENTS

5.1 MODEL ARCHITECTURES

We evaluate five sequence model architectures spanning distinct computational paradigms, chosen to expose how different inductive biases interact with language-conditioned physical dynamics. Figure 2 provides a schematic overview of the conditioning mechanisms used by each model.

1. MLP Baseline (No Language). The MLP baseline serves as a control model that explicitly excludes language information, processing only flattened sequences of physical states. By removing temporal recurrence and semantic conditioning, this model tests the extent to which future dynamics can be extrapolated from short-horizon motion cues alone. Although incapable of modeling long-

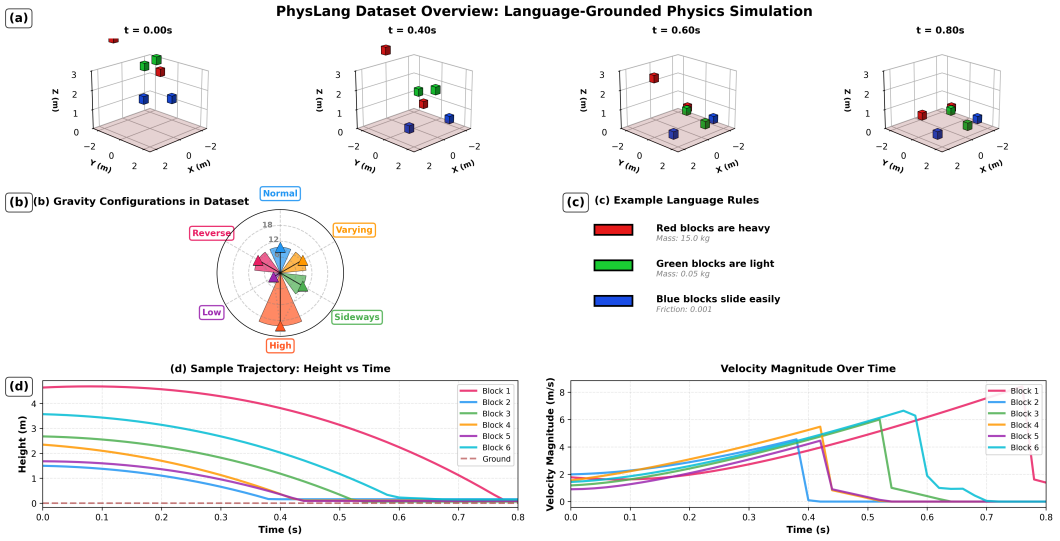

Figure 1: **Overview of the PHYSLANG benchmark.** (a) Temporal evolution of six blocks at t=0.0s, 0.4s, 0.6s, 0.8s, with properties governed by language rules (e.g., "Red blocks are heavy"). (b) Polar visualization of six gravity configurations (normal, reverse, low, high, sideways, varying) with radial axis showing magnitude (0-20 m/s²). (c) Example rule mappings from colors to physical properties, enabling systematic evaluation of controlled compositional recombination. (d) Height trajectories (left) and velocity profiles (right) from a representative six-block episode, demonstrating rule compliance with distinct dynamics for different mass and friction values.

term dependencies or rule changes, the MLP can exploit correlations in velocity, position, and contact patterns to produce reasonable predictions in stable regimes. The MLP baseline processes flattened physical states without access to language information: $\hat{\mathbf{s}}_{T+1:T+H} = \text{MLP}(\text{flatten}(\mathbf{s}_{1:T}))$. It provides a lower bound on performance and helps determine whether language-conditioned models meaningfully exploit semantic input beyond observable kinematics.

2. Recurrent Models: GRU and LSTM. Recurrent neural networks (Rumelhart et al., 1988) provide a natural inductive bias for modeling sequential physical processes through explicit latent state updates. We evaluate both GRU (Cho et al., 2014) and LSTM (Hochreiter & Schmidhuber, 1997) variants to account for differences in gating complexity and memory retention, while keeping the conditioning mechanism identical. Language information is injected at every timestep by concatenating a fixed-dimensional language embedding $\mathbf{z}_\ell$ to the physical state input, encouraging the model to treat language as a persistent contextual variable. The recurrent hidden state thus integrates observations over time while being continuously modulated by linguistic constraints. These models test whether simple recurrent memory is sufficient to internalize and propagate language-specified physical properties through time: $\mathbf{h}_t = \text{RNN}(\mathbf{h}_{t-1}, [\mathbf{s}_t; \mathbf{z}_\ell])$. The final hidden state is decoded to predict future trajectories. $\hat{\mathbf{s}}_{T+1:T+H} = \text{Decoder}(\mathbf{h}_T)$.

3. Transformer. The Transformer encoder (Vaswani et al., 2017) represents physical trajectories as sets of temporally indexed tokens processed through self-attention, enabling flexible long-range dependency modeling. Unlike recurrent models, Transformers do not impose an explicit sequential state update, instead relying on attention mechanisms to relate past states when predicting future dynamics. We condition the Transformer on language via cross-attention, allowing physical state representations to selectively attend to semantic information when forming predictions:

$$\text{CrossAttn}(\mathbf{H}_s, \mathbf{z}_\ell) = \text{softmax}\left(\frac{(\mathbf{H}_s\mathbf{W}_Q)(\mathbf{z}_\ell\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)(\mathbf{z}_\ell\mathbf{W}_V). \tag{5}$$

Here $\mathbf{H}_s \in \mathbb{R}^{T\times d}$ denotes encoded state sequences and $\mathbf{z}_\ell \in \mathbb{R}^{384}$ the language embedding. This design tests whether attention-based architectures can dynamically resolve which aspects of language are relevant to different phases of motion. While Transformers offer greater expressivity and global context access, their lack of an explicit state-space inductive bias may hinder stable propagation of physical constraints over time.

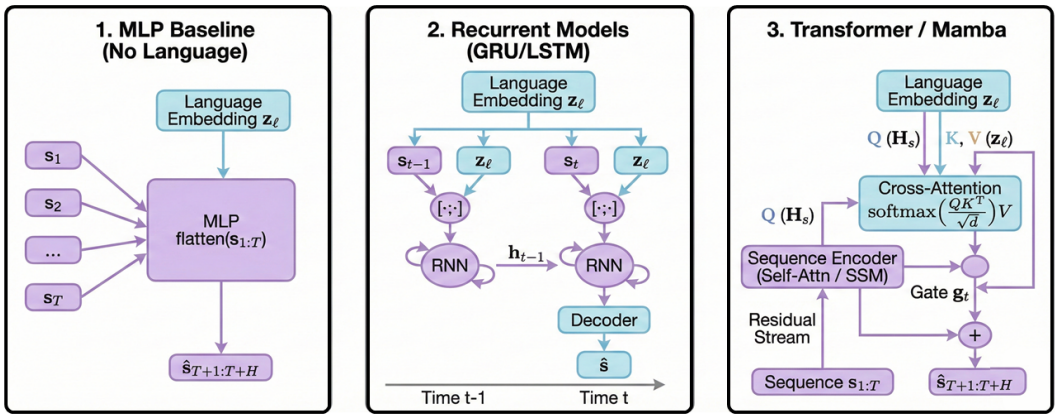

Figure 2: **Model architectures for language-conditioned physics prediction.** Left: MLP baseline without language input. Middle: Recurrent models (GRU/LSTM) with concatenated language embeddings. Right: Transformer/Mamba with cross-attention for language conditioning. All language-conditioned models use frozen sentence transformer embeddings.

Table 2: Main results on PHYSLANG. We report MSE (↓) with 95% confidence intervals across 3 seeds. Δ_{comp} and Δ_{contra} indicate relative performance change from training to compositional and contradiction splits.

Model	Train MSE	Compositional MSE	Contradiction MSE	Δ_{comp} (%)	Δ_{contra} (%)
MLP (no lang.)	$\mathbf{0.320 \pm 0.140}$	0.861 ± 0.084	1.587 ± 0.242	+169.2	+396.6
GRU (Cho et al., 2014)	1.155 ± 0.006	1.673 ± 0.212	2.573 ± 0.993	+44.8	+122.7
LSTM (Hochreiter & Schmidhuber, 1997)	0.563 ± 0.096	$\mathbf{1.066 \pm 0.107}$	$\mathbf{1.956 \pm 0.611}$	+89.5	+247.5
Transformer (Vaswani et al., 2017)	2.577 ± 0.231	1.855 ± 0.416	3.701 ± 2.694	−28.0	+43.6
Mamba (Gu & Dao, 2024)	2.833 ± 0.122	3.273 ± 0.871	4.093 ± 1.059	+15.5	+44.5
Mamba2 (Dao & Gu, 2024)	2.479 ± 0.504	2.743 ± 0.677	3.921 ± 2.036	+10.6	+58.2

4. Mamba. Mamba (Gu & Dao, 2024) is a selective state-space model that combines linear-time sequence processing with input-dependent dynamics, making it a compelling candidate for physical modeling. Its latent state evolves according to learned discretized state-space operators, providing an inductive bias closer to classical dynamical systems than attention-based models: $\mathbf{h}_t = \bar{\mathbf{A}}_t \mathbf{h}_{t-1} + \bar{\mathbf{B}}_t \mathbf{s}_t, \quad \mathbf{o}_t = \mathbf{C}_t \mathbf{h}_t$. . To incorporate language, we interleave Mamba blocks with cross-attention layers and introduce a learned gating mechanism that controls the influence of semantic information at each timestep: $\tilde{\mathbf{s}}_t = (1 - \mathbf{g}_t) \odot \text{Mamba}(\mathbf{s}_t) + \mathbf{g}_t \odot \text{CrossAttn}(\mathbf{s}_t, \mathbf{z}_\ell)$, where $\mathbf{g}_t \in [0, 1]^d$ modulates the influence of language. This design allows the model to modulate the strength of language conditioning as a function of the evolving physical state. Mamba therefore tests whether structured state-space dynamics can support more robust grounding of language-defined physical rules. Comparing its behavior to Transformers and RNNs helps disentangle the roles of recurrence, attention, and continuous state evolution in language-grounded world modeling.

5.2 EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

All models are trained for 100 epochs using AdamW (Loshchilov & Hutter, 2017) with a OneCycleLR schedule, an initial learning rate of 10^{-3}, and gradient clipping at 1.0 to stabilize optimization across architectures. Inputs are normalized to the $[0, 1]$ range using domain-specific scaling to ensure comparable numerical conditioning across physical variables (details in Appendix C). Each model is trained to predict future physical states over a fixed horizon H given a context window of length T, with identical loss functions, batch sizes, and optimization settings to control for confounding factors. Results are averaged over three random seeds to account for stochasticity in initialization and training dynamics. Architectural differences are therefore isolated to representational structure and language-conditioning mechanisms rather than training protocol. This controlled setup allows us to attribute observed performance differences and failure modes to inductive bias and semantic integration rather than optimization artifacts.

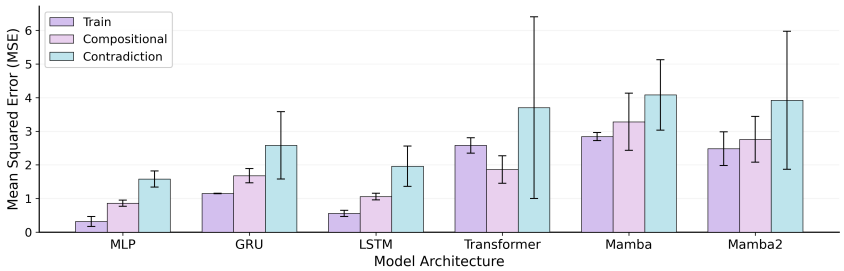

Figure 3: **Prediction error across evaluation splits.** Bar heights indicate mean MSE across 3 seeds; error bars show 95% CI. LSTM achieves the lowest error on compositional and contradiction splits in this setting, while the MLP baseline shows a high generalization drop despite the lowest training error.

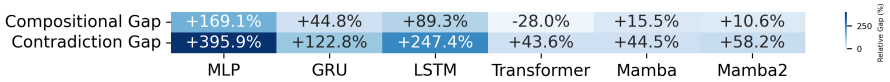

Figure 4: **Generalization gaps across models.** Heatmap showing relative performance degradation (%) from training to compositional and contradiction splits. Darker colors indicate larger gaps. MLP baseline exhibits catastrophic generalization failure (+396.6% on contradiction), while GRU maintains the smallest compositional gap (+44.8%).

6 Experimental Findings

6.1 Main Results

✓ **Architectural Inductive Bias and Accuracy.** Table 2 and Figure 3 reveal systematic performance differences across architectures that cannot be explained by optimization or data scale. In this controlled, short-horizon regime, *recurrent models consistently achieve lower prediction error than attention-based and state-space alternatives*. Notably, the LSTM attains the lowest error on both the compositional and contradiction splits (MSE 1.066 and 1.956), outperforming Transformer and Mamba variants by more than 40%, suggesting that explicit recurrent state updates provide a strong prior for propagating language-conditioned physical constraints.

✓ **Limits of Purely Kinematic Extrapolation.** Although the MLP baseline achieves the lowest training error, its performance degrades sharply under distribution shift, exhibiting the largest generalization gaps across all test splits. As shown in Figure 4, error increases dramatically on the compositional and contradiction splits, indicating that kinematic extrapolation alone is insufficient when physical properties are modified by language rules. *This failure demonstrates that access to language is necessary but not sufficient for robust generalization*, as models must also internalize how semantic rules causally influence dynamics.

✓ **Performance of Attention-Based and State-Space Models.** Attention-based and state-space architectures exhibit higher variance and overall error in this setting, suggesting a mismatch between their inductive biases and the demands of short-horizon physical prediction. The Transformer's negative compositional gap is indicative of underfitting, likely due to its reliance on global attention rather than incremental state updates. Mamba similarly shows elevated error, implying that structured state-space dynamics combined with language-conditioned gating do not, by themselves, guarantee stable semantic integration.

✓ **Failure Under Contradictory Language Rules.** Across all architectures, the contradiction split yields the highest prediction errors, making it the most challenging evaluation setting. *Models consistently fail to adapt when language rules explicitly conflict with training-time associations*, instead producing predictions that interpolate between incompatible dynamics. This behavior suggests that language is often treated as a soft contextual cue rather than as a hard constraint governing physical evolution, even in simple and fully observable environments.

6.2 Ablation Studies

✓ **Language Ablation Study.** To quantify how much models rely on language, we evaluate with perturbed linguistic inputs (Table 3). Figure 5 visualizes the degradation patterns. Language sensitivity, defined as $(\text{MSE}_{\text{shuffled}} - \text{MSE}_{\text{normal}})/\text{MSE}_{\text{normal}}$, reveals important patterns. *LSTM shows the highest sensitivity* (295.9%), indicating strong coupling between linguistic input and dynamics prediction, though this may also reflect brittleness to perturbations. *Transformer architecture shows*

the lowest sensitivity (51.0%), potentially indicating underfitting to language. The normal progression of → shuffled → wrong rules confirms that models distinguish between random noise and semantically incorrect information.

Table 3: Language ablation results. "Normal" uses correct language; "Shuffled" uses randomly permuted text; "Wrong Rule" uses mismatched descriptions. Language sensitivity measures relative degradation from normal to shuffled.

Model	Normal	Shuffled	Wrong Rule	Language Sensitivity
Transformer	2.577 ± 0.231	3.884 ± 1.541	4.237 ± 0.490	$51.0\% \pm 26.7\%$
Mamba	2.833 ± 0.122	5.221 ± 0.866	5.959 ± 1.063	$84.4\% \pm 14.7\%$
Mamba2	2.479 ± 0.504	4.808 ± 1.856	5.371 ± 0.571	$93.3\% \pm 18.1\%$
GRU	1.155 ± 0.006	3.130 ± 0.372	3.662 ± 0.865	$170.9\% \pm 13.1\%$
LSTM	$\mathbf{0.563 \pm 0.096}$	2.221 ± 0.308	2.747 ± 0.453	$\mathbf{295.9\% \pm 33.7\%}$

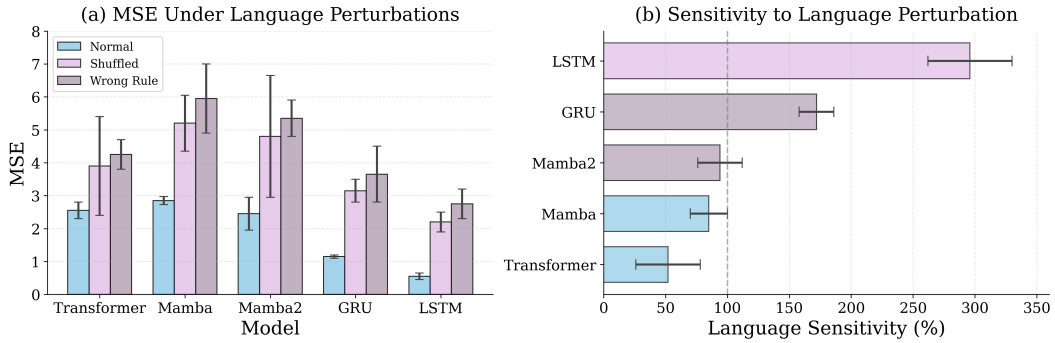

Figure 5: **Language ablation analysis.** (Left) MSE under different language perturbations: normal (correct), shuffled (random word order), and wrong rule (semantically incorrect). (Right) Language sensitivity quantifying relative degradation. LSTM exhibits the highest sensitivity (295.9%), indicating the strongest language-physics coupling.

✓ **Statistical Significance.** Table 4 reports paired statistical tests comparing each model to the MLP baseline. All language-conditioned models show statistically significant differences ($p < 0.05$) with large effect sizes (Cohen's $d > 0.8$).

Table 4: Statistical significance analysis comparing language-conditioned models against MLP baseline on the compositional test split. Cohen's d effect sizes and paired t-test p-values are reported.

Model	vs Baseline Δ	Cohen's d	p-value	Significant?
Transformer	-115.6%	-5.82 (large)	0.0097	✓
Mamba	-280.4%	-6.70 (large)	0.0073	✓
Mamba2	-218.7%	-6.17 (large)	0.0087	✓
GRU	-94.4%	-7.27 (large)	0.0063	✓
LSTM	-23.9%	-3.28 (large)	0.0296	✓

✓ **Computational Efficiency.** Table 5 compares computational requirements. Despite having the most parameters (3.6M), LSTM achieves competitive training times. Mamba2 requires significantly more VRAM (278 MB) due to structured state space computations.

7 Discussion

We designed our experiments to directly probe whether language acts as a governing constraint on physical dynamics or merely as a contextual cue during prediction. The observed differences across architectures and evaluation splits reveal how inductive biases shape this interaction between semantics and temporal state evolution. We now interpret these findings to explain why certain models benefit from language conditioning in controlled settings, while others fail to generalize under compositional or contradictory rules.

Why recurrent models perform better in this regime? The sequential and causal structure of rigid-body physics aligns well with the inductive biases of recurrent architectures, which enforce

Table 5: Computational costs for training on PHYSLANG.

Model	Parameters	Train Time (min)	Epoch Time (s)	Peak VRAM (MB)
MLP	678K	0.6	0.2	87
Transformer	990K	1.4	0.6	87
Mamba	674K	1.4	0.7	87
Mamba2	676K	2.3	1.1	278
GRU	1.3M	0.9	0.4	87
LSTM	3.6M	1.4	0.5	124

incremental state updates and temporal locality. In contrast, bidirectional attention in Transformers may introduce spurious dependencies between non-adjacent timesteps, which is suboptimal for short-horizon physical prediction. GRU and LSTM hidden states effectively act as temporal smoothers, stabilizing prediction under small perturbations, as reflected in smoother learning curves (Appendix D). We stress that this advantage is specific to the controlled, short-horizon setting studied here and may not generalize to longer or more complex regimes.

The language grounding paradox. Models with higher sensitivity to linguistic perturbations also achieve better absolute performance, indicating that **strong coupling between language and dynamics is beneficial** in this benchmark. The near-monotonic alignment between sensitivity and accuracy rankings suggests that effective grounding requires language to exert a measurable causal influence. However, increased sensitivity also introduces brittleness under distributional shift, exposing a trade-off between semantic influence and robustness. This tension motivates future work on structured or modular language conditioning rather than global semantic injection.

Contradiction as a diagnostic signal. Uniformly poor performance on the contradiction split should not be interpreted as failure of belief revision, as models are not trained to update or arbitrate rules online. Instead, this split diagnoses whether models can override learned semantic–physical associations when presented with conflicting constraints. The observed interpolation between incompatible dynamics suggests that language is encoded as a soft contextual prior rather than a binding causal rule. Addressing this limitation likely requires explicit representations of causal structure and rule hierarchies (Schölkopf et al., 2021).

Implications for foundational world models. Large gaps between training and compositional performance, even under controlled recombination of familiar rule primitives, indicate that compositional generalization remains fragile. These failures arise despite deterministic dynamics, full observability, and simple language, suggesting that scale alone is unlikely to resolve them. While derived from a small diagnostic benchmark, our findings expose failure modes that are likely amplified in larger settings. We therefore position PHYSLANG as a hypothesis-generating tool for stress-testing language grounding in future foundation world models. Our findings imply that such models require explicit mechanisms for disentangling linguistic structure from latent dynamics, rather than relying on implicit pattern matching. In particular, architectures should support modular rule composition and verifiable constraint enforcement to prevent semantic drift under distributional shifts.

8 CONCLUDING REMARKS

We introduced PHYSLANG, a controlled diagnostic benchmark for evaluating language grounding in neural world models. Our results show that even in simple, fully observable physics environments, current architectures exhibit sharp limitations in grounding language, particularly under compositional recombination and contradictory rules. In this short-horizon setting, recurrent models achieve lower prediction error than attention-based alternatives, while all models display high language sensitivity without corresponding semantic robustness. We position PHYSLANG as a lightweight falsification tool for probing language–dynamics coupling, intended to reveal failure modes rather than define a leaderboard.

Promising future directions include integrating explicit physical priors, structured rule representations, and causal intervention mechanisms to improve robustness under semantic conflict. Future work will also explore neuro-symbolic and program-induction approaches to enable explicit rule extraction and verification from language. Scaling the benchmark to partially observable and stochastic environments may further test generalization and long-horizon reasoning. Finally, incorporating human-in-the-loop evaluation could help assess whether learned representations align with human physical intuitions and linguistic semantics.

REFERENCES

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

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

Anthony Brohan et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

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

Spruce Campbell. Newton - a small benchmark for interactive foundation world models. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*, 2025. URL https://openreview.net/forum?id=xlp6P6qaRW.

Kyunghyun Cho, Bart van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Q6a9W6kzv5.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, pp. 10041–10071. PMLR, 2024.

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer, 2024. URL https://oasis-model.github.io/.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Mengkang Hu, Tianxing Chen, Yude Zou, Yuheng Lei, Qiguang Chen, Ming Li, Yao Mu, Hongyuan Zhang, Wenqi Shao, and Ping Luo. Text2World: Benchmarking large language models for symbolic world model generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26043–26066, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1337. URL https://aclanthology.org/2025.findings-acl.1337/.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pp. 9118–9147. PMLR, 2022.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Ao Liang, Lingdong Kong, Tianyi Yan, Hongsi Liu, Wesley Yang, Ziqi Huang, Wei Yin, Jialong Zuo, Yixuan Hu, Dekai Zhu, et al. Worldlens: Full-spectrum evaluations of driving world models in real world. *arXiv preprint arXiv:2512.10958*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

NVIDIA. Isaac sim, 2022. URL `https://developer.nvidia.com/isaac-sim`.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2020.

D.E. Rumelhart, G.E. Hinton, and R.J. Williams. *Learning Internal Representations by Error Propagation*, pp. 399–421. Elsevier, 1988. ISBN 9781483214467. doi: 10.1016/b978-1-4832-1446-7.50035-2. URL `http://dx.doi.org/10.1016/B978-1-4832-1446-7.50035-2`.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

A PHYSICS RULES

Table 6 provides the complete list of physics rules used in PHYSLANG. Each rule maps a color attribute to a specific physical property value. During training, models observe trajectories governed by subsets of these rules and must learn to generalize to novel combinations (compositional split) or contradictory assignments (contradiction split). The rules are designed to create clearly distinguishable dynamics; for example, heavy blocks ($m = 15.0$ kg) exhibit distinct interaction and momentum-transfer dynamics during collisions, while slippery blocks ($\mu = 0.001$) slide farther after contact.

Table 6: Complete list of physics rules in PHYSLANG. Rules define mappings from block colors to physical properties, creating language-grounded dynamics that models must learn to predict.

Rule Key	Natural Language	Property	Value
red_heavy	"Red blocks are heavy"	mass	15.0 kg
blue_heavy	"Blue blocks are heavy"	mass	15.0 kg
green_light	"Green blocks are light"	mass	0.05 kg
red_light	"Red blocks are light"	mass	0.05 kg
blue_light	"Blue blocks are light"	mass	0.05 kg
red_slippery	"Red blocks slide easily"	friction	0.001
green_slippery	"Green blocks slide easily"	friction	0.001
blue_slippery	"Blue blocks slide easily"	friction	0.001
red_sticky	"Red blocks are sticky"	friction	2.0
green_sticky	"Green blocks are sticky"	friction	2.0

B GRAVITY VARIATIONS

To increase environmental diversity and test model robustness, PHYSLANG includes six gravity configurations (Table 7). These variations create qualitatively different dynamics: reversed gravity causes blocks to float upward, low gravity produces slower, more floaty motion, and sideways gravity induces lateral acceleration. The varying gravity condition uses a time-dependent multiplier $\alpha(t) \in [0.5, 1.5]$ that oscillates sinusoidally, requiring models to adapt to changing dynamics within a single episode.

Table 7: Gravity variations in PHYSLANG. Each configuration create distinct trajectory patterns that models must learn to predict.

Key	Description	Vector (m/s^2)
normal	Standard Earth gravity	$(0, 0, -9.81)$
reverse	Reversed (upward) gravity	$(0, 0, +9.81)$
low	Reduced gravity	$(0, 0, -3.0)$
high	Increased gravity	$(0, 0, -20.0)$
sideways	Horizontal gravity	$(9.81, 0, 0)$
varying	Time-varying magnitude	$\alpha(t) \cdot (0, 0, -9.81)$

C IMPLEMENTATION DETAILS

Normalization. To ensure stable training across features with different scales, states are normalized per feature type. Positions are scaled by the workspace bounds, velocities are clipped to prevent outliers from dominating gradients, and physical properties (mass, friction, scale) are normalized to their respective valid ranges:

$$\tilde{\mathbf{p}} = \mathbf{p}/5.0, \quad \tilde{\mathbf{v}} = \text{clip}(\mathbf{v}, -10, 10)/10.0, \quad \tilde{m} = \text{clip}(m, 0.05, 15.0)/15.0 \tag{6}$$

$$\tilde{\mu} = \text{clip}(\mu, 0.001, 2.0)/2.0, \quad \tilde{r} = \text{clip}(r, 0.08, 0.35)/0.35 \tag{7}$$

Model Hyperparameters. Table 8 summarizes the architectural choices for each model. We kept hyperparameters relatively consistent across architectures to ensure fair comparison, with hidden dimensions chosen to yield comparable parameter counts where possible. The MLP uses a 4-layer architecture with expanding-then-contracting width (256→512→512→256), while recurrent and attention models use 2 layers. All models use dropout of 0.2 for regularization.

Table 8: Model hyperparameters. Hidden dimensions and layer counts were chosen to balance expressivity with computational efficiency.

	MLP	**GRU**	**LSTM**	**Transformer**	**Mamba**
Hidden dim	256/512	256	256	128	128
Layers	4	2	2	2	2
Heads	–	–	–	4	–
State dim	–	–	–	–	16/64
Dropout	0.2	0.2	0.2	0.2	0.2

D Additional Visualizations

This section provides additional figures that complement the main experimental results.

Property Distributions. Figure 6 shows the distribution of physical properties across PhysLang. The mass distribution (panel a) exhibits a bimodal pattern reflecting the heavy/light rule assignments, with peaks at 0.05 kg and 15.0 kg. Similarly, friction coefficients (panel b) cluster around the slippery (0.001) and sticky (2.0) values. Block scales (panel c) are uniformly distributed within the valid range, and initial velocity magnitudes (panel d) follow a Chi distribution consistent with Gaussian velocity components.

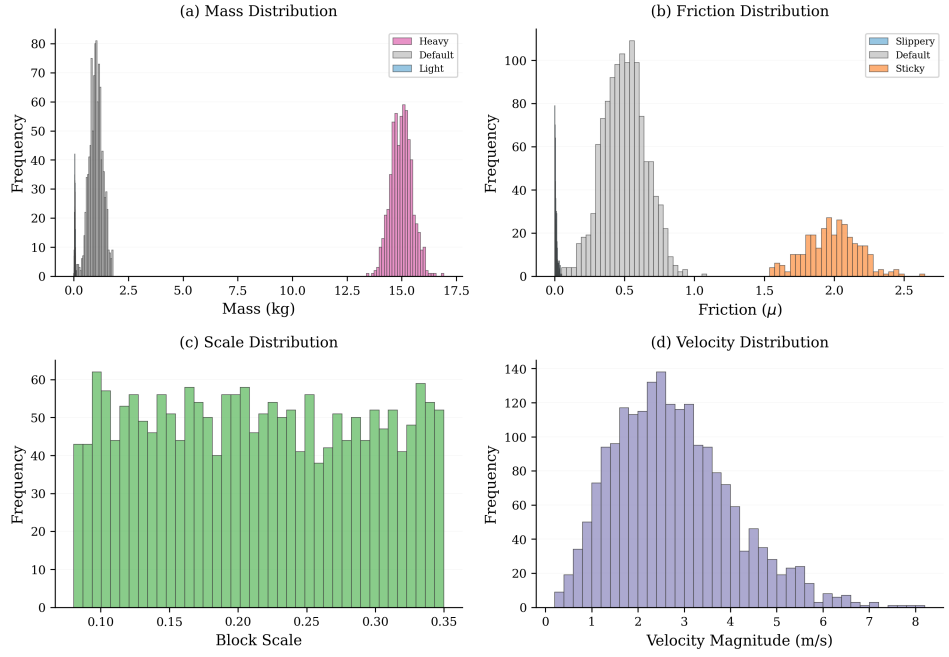

Figure 6: Distribution of physical properties across PhysLang: (a) mass showing bimodal heavy/light pattern, (b) friction coefficient distribution, (c) block scale distribution, (d) initial velocity magnitudes.

Learning Dynamics. Figure 7 visualizes training and validation loss curves across all models. Several patterns emerge that correlate with final test performance. MLP converges fastest but exhibits

early overfitting, with the train-validation gap widening after epoch 20. Recurrent models (GRU, LSTM) show smooth, stable convergence with consistent improvement throughout training. Attention-based models display higher variance, particularly Transformer which shows oscillatory behavior that may explain its underfitting on the training distribution.

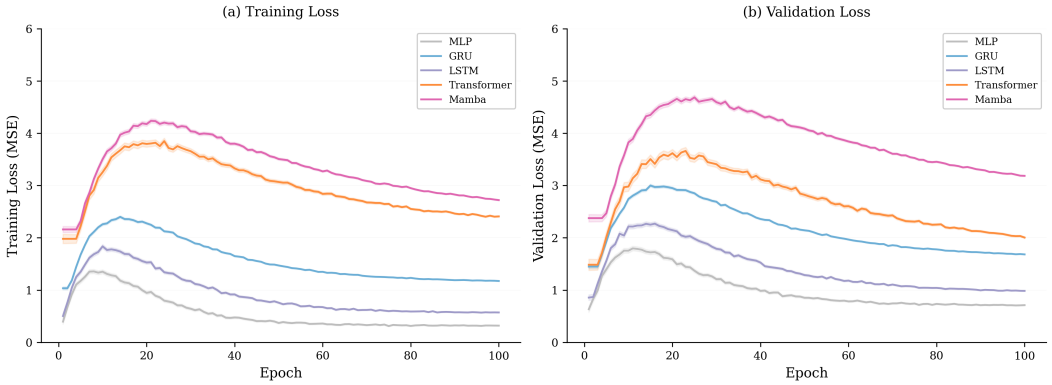

Figure 7: Training dynamics across model architectures. (Left) Training loss curves showing convergence patterns. (Right) Validation loss curves. Shaded regions indicate ± 1 standard deviation across 3 random seeds. The widening gap between train and validation for MLP foreshadows its poor generalization.

Multi-Dimensional Comparison. Figure 8 presents a radar chart comparing models across six evaluation dimensions: train MSE, compositional MSE, contradiction MSE, language sensitivity, training efficiency, and memory usage. Scores are normalized such that higher values indicate better performance on all axes. LSTM achieves the best overall profile in this setting, dominating on compositional and contradiction handling while maintaining competitive computational efficiency. The MLP baseline shows strength only in training efficiency and train MSE, but its poor generalization metrics result in a collapsed profile on the test-related dimensions. Attention-based models (Transformer, Mamba) occupy an intermediate position but lag on accuracy metrics.

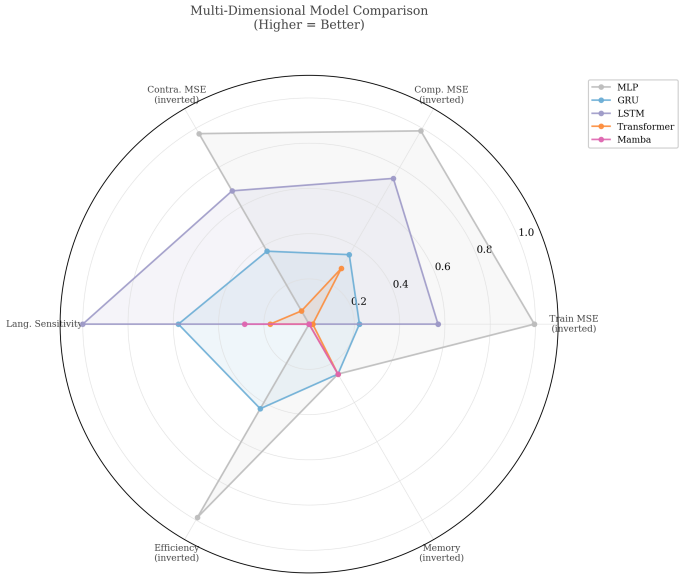

Figure 8: Multi-dimensional model comparison across six evaluation dimensions (higher is better for all axes). LSTM achieves the best overall profile in this setting, excelling in generalization metrics while maintaining efficiency. MLP's profile collapses on test dimensions despite strong training performance.

Computational Efficiency. Figure 9 provides a visual comparison of computational requirements across architectures. Despite having the most parameters (3.6M), LSTM achieves competitive training times due to efficient CUDA implementations. Mamba2 requires significantly more VRAM (278 MB vs. 87 MB for most models) due to its structured state space computations, which may limit scalability to longer sequences or larger batch sizes.

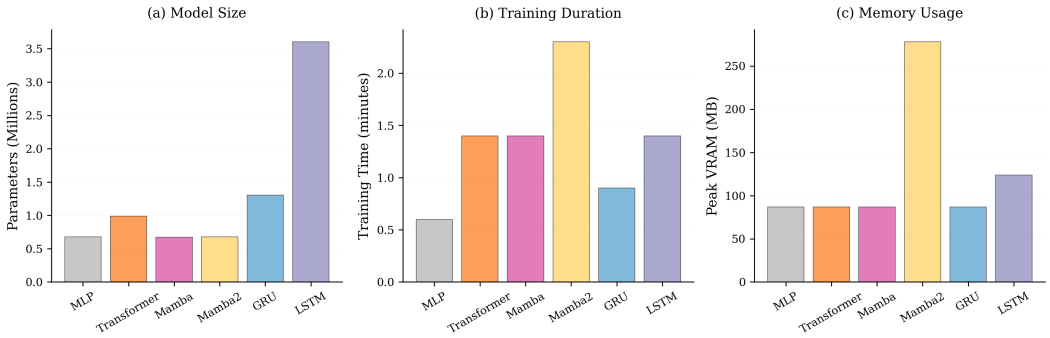

Figure 9: Computational efficiency comparison: (a) parameter counts across architectures, (b) total training time in minutes, (c) peak GPU memory usage. LSTM achieves the best accuracy-efficiency trade-off despite having the most parameters.

Effect Sizes. Figure 10 visualizes Cohen's d effect sizes comparing each language-conditioned model against the MLP baseline on the compositional split. All effects are large ($|d| > 0.8$), confirming that the differences reported in Table 4 are practically significant. The negative values indicate that language-conditioned models have higher absolute MSE than the MLP baseline, but as shown in the main results, they exhibit substantially better generalization patterns.

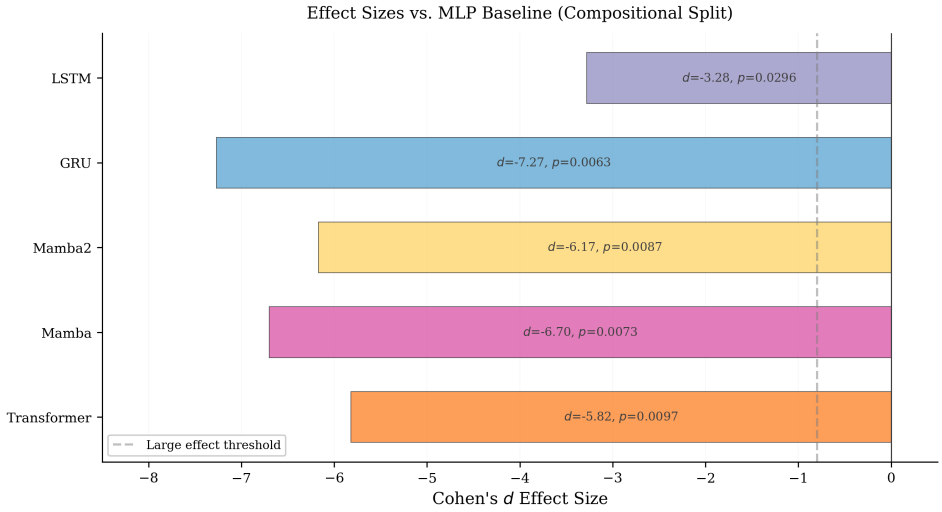

Figure 10: Cohen's d effect sizes comparing language-conditioned models against the MLP baseline on the compositional split. All effects are large ($|d| > 0.8$) with statistical significance ($p < 0.05$).