INTEGRATING SIMULATION AND CHAIN-OF-THOUGHT REASONING IN MULTIMODAL-LANGUAGE MODELS FOR PHYSICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we present a cognitively-inspired model that tackles the question of resource rationality we encounter in physical reasoning, which is a question that is consistently overlooked by the current AI community. Given the fact that various tools like Chain-of-Thought and Physics-Grounded Simulator could be applied to solve physical reasoning tasks, an observation naturally emerges: when to use which? what is the criterion for choosing the best tools at various scenarios? To tackle this question, our model aim to optimize the scheme in when to use which tools in order to reach the best tradeoff between computational costs and accuracy. Our model is able to (1) improve overall accuracy while significantly reducing the computational costs by nearly 50%. (2) found that two methods performs better in physical reasoning tasks from different categories, without being trained on categories labels.

1 INTRODUCTION

Physical reasoning is an important topic in the current AI community, and it is needed to overcome important challenges like ensuring that embodied AI agents act safely in the world (Zheng et al., 2024) (Chow et al., 2025). There are different approaches for how to implement this reasoning. Chain-of-Thought Kojima et al. (2022) Wei et al. (2022) is one dominant tool, used to solve many reasoning tasks, including physical reasoning. The recent SoTA physical reasoning models like Jin et al. (2026) adapted the chain-of-thought as the only reasoning methods when reasoning about physical tasks. However, this might not be how human approach physical reasoning, as in intuitive physics, current approaches suggests that mental simulation could be a way that people approach physical reasoning (Battaglia et al., 2013) (Ullman et al., 2017). Researchers in AI had also adapted this idea of mental simulation into grounding the physical reasoning in Multi-modal Large Language Models (MLLMs) (Liu et al., 2022). Furthermore, recent breakthrough in diffusion model-based video model has shown emerging zero-shot ability as a physics simulator (Wiedemer et al., 2025) and it's ability to push towards intelligence in robotics (Team et al., 2025).

In theory, Chain-of-thought is considered to be computationally cheaper when compared with simulation with physics simulator, but simulation is considered to be more accurate since it is grounded on real physics. This raises an interesting observation that whether each method would be better suited to solve specific types of physical reasoning tasks, thus creating this unique tradeoff between accuracy and computational costs in physical reasoning in MLLMs when different method was used.

To answer this question, we propose a model that not only empirically shows that such tradeoff between accuracy and computational costs exists, but also able to find the best tradeoff between accuracy and computational costs on various physical reasoning tasks, inspired by research in resource rationality in cognitive science.

2 RESOURCE RATIONALITY IN INTUITIVE PHYSICS

People are able to efficiently reason about the physical dynamics of everyday objects. For example, if you saw a ball flying towards you, you might quickly predict where to catch the ball. This

ability of people being able to reason about physics of the world is called intuitive physics. Many competing theories have been proposed to explain this 'intuitive physics', and recent research have suggested that it is likely that humans use a combination of different computations to carry out this reasoning Hartshorne & Jing (2025). Interestingly, previous work in cognitive science have suggested a hybrid model that combines two most competing theories that are mental simulation and heuristic reasoning, which explain this 'intuitive physics' (Smith et al., 2023) (Sosa et al., 2025).

The hybrid models are designed around one core concept: resource rationality. In Smith et al. (2023), they proposed the Integration of Simulation and Rules (ISR) framework based on resource rationality before performing a task. Before performing physical reasoning on a specific task T, we have to decide on a strategy S to answer this task. But, several factors that are associate with this strategy needs to be balanced and considered. There are mainly two factors to consider in the ISR: the metabolic costs C for adapting this strategy, and the expected utility U of this strategy, needs to be considered. Therefore, the ISR framework aims to optimize for the best strategy S^* that uses the resource in the most rational way as the following:

$$S^* = \arg\max E(U(S,T) - C(S,T)) \tag{1}$$

Furthermore, in Sosa et al. (2025), a hybrid model that could be applied when performing a task was proposed. When humans are performing physical reasoning, even though it is possible that they are adapting a simulation strategy when doing a physical reasoning task, some parts in the task that can be resolved by using heuristics rules could be used to save cognitive resource, achieving resource rationality. This idea of resource rationality in intuitive physics of human is useful for many cases, and serves as a motivation for our model. Resource rationality has been shown to be useful in intuitive physics for reducing the amount of computation humans need to perform for physical reasoning, while remaining accurate. This serves as inspiration for our model, which similarly trades off physical simulation against heuristic-based reasoning, to improve model accuracy while also reducing costs."

3 Hybrid Simulation and Reasoning Model

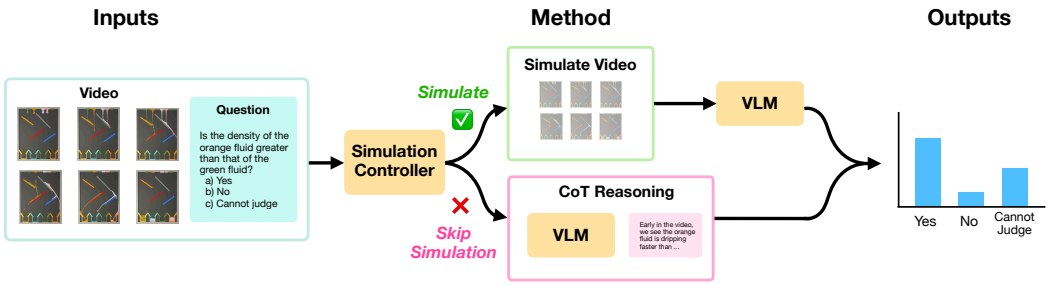

Figure 1: The overview of HSR model. Our HSR model consists of simulation controller, physics simulator-based simulation, and Chain-of-thought reasoner. The inputs consist of a set of video and a question based on this video, which is passed into simulation controller. Depending on the assignment from the simulation controller, either simulation or Chain-of-thought reasoner will be applied to answer the question.

Here we present our model Hybrid Simulation and Reasoning Model (HSR) to incorporate resource rationality into physical reasoning. Consider an arbitrary physical reasoning task that consist of a problem s and a video d that the problem is based on, and two methods of solving physical reasoning tasks: simulator and Chain-of-thought reasoning. First, we determine which of the two methods is best suited to solve this physical reasoning tasks by using a trained simulation controller. Second, we pass the physical reasoning tasks to the method determined by the simulation controller to get an answer for this physical reasoning tasks.

We first process the problem s and the video d into embeddings. For visual embedding, we only use the first frame x_1 of the video d. We used visual embedding model and text embedding model

to generate the corresponding embeddings. Therefore, we will be able to obtain a concatenated embedding $\mathbf{h}$ by:

$$\mathbf{v} = f_{visual}(x_1) \in \mathbb{R}^{512} \qquad \mathbf{u} = f_{text}(s) \in \mathbb{R}^{718} \qquad \mathbf{h} = [\mathbf{u}; \mathbf{v}] \in \mathbb{R}^{1230} \tag{2}$$

We pass the concatenated embedding $\mathbf{h}$ into the simulation controller, as shown in fig 1. The simulation controller is a trained MLP that that process the embedding $\mathbf{h}$ into a logit. We use a 2-hidden-layer MLP ($1230 \to 512 \to 256 \to 1$) with BatchNorm $(512, 256)$, ReLU, and dropout ($p = 0.5$) after each hidden layer to produce a scalar logit ℓ.

Where ℓ is the predicted logit from the model. The final predicted probability p can be recovered by applying sigmoid function on logit:

$$p = \sigma(\ell) = \frac{1}{1 + e^{-\ell}}$$

We define a threshold τ that if $p \leq \tau$ we assign the data entry as Chain-of-thought reasoning, and if $p > \tau$ then we assign the data entry as simulation. Notice that this threshold controls the number of data entries being assigned as simulation, thus controls the computational costs, since simulation is by definition more costly. During the training phase, the model aims to minimize the BCE With Logit Loss, stated as the following:

$$\mathcal{L}_{\text{BCE-logit}}(\ell, y) = -\Big(y \log p + (1 - y) \log(1 - p)\Big) \tag{3}$$

During the training phase, ground truth probability y is provided for all data in the training set. Given that $y \in [0, 1]$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Based on this probability space, we define a random variable $X : \Omega \to \{S, C\}$ where S denotes using simulation and C denotes using chain-of-thought only.

Let $0 \leq p_{\text{sim}}, p_{\text{cot}} \leq 1$. Define p_{sim} and p_{cot} to be the probability of using simulation and the probability of using chain-of-thought only reasoning. For each data entry in the training data, it was answered by using simulation K times, and Chain-of-thought reasoning K times. Therefore:

$$p_{\text{sim}} = \frac{1}{K} \sum_{i=1}^{K} \mathbb{I}\Big[a_{\text{sim}}^{(i)} = a_{\text{gt}}\Big] \in [0, 1] \qquad p_{\text{cot}} = \frac{1}{K} \sum_{i=1}^{K} \mathbb{I}\Big[a_{\text{cot}}^{(i)} = a_{\text{gt}}\Big] \in [0, 1] \qquad i \in [1, K] \tag{4}$$

where $a_{\text{sim}}^{(i)}$ is the answer output by simulation method in ith attempt, $a_{\text{cot}}^{(i)}$ is the answer output by Chain-of-thought method in the ith attempt, and a_{gt} is the ground truth answer. The indicator function allows partially correct answer.

We hereby define the ground truth probability y as:

$$y \triangleq \mathbb{P}(X = S) = \frac{p_{\text{sim}}}{p_{\text{sim}} + p_{\text{cot}}},$$

Then, as shown in fig 1, we pass the question and video into simulation or Chain-of-thought reasoning, depending on the decision made by the simulation controller to obtain the final answer to the question. In simulation, a fine-tuned video diffusion model is used as the physics-grounded simulator. We utilize a VLM for Chain-of-thought reasoning and readout phase in simulation.

4 Experiments

We hypothesize that Chain-of-thought and Physics-Grounded Simulator should be applied to specific types of physical reasoning tasks in order to maximize the tradeoff between accuracy and computational costs, and our model is able to catch this tradeoff. We used ContPhy dataset (Zheng et al., 2024) as it contains numerous realistic physical reasoning tasks from variety of categories. For more details about the dataset please refer to Appendix A.1. We analyze the performance of our model by

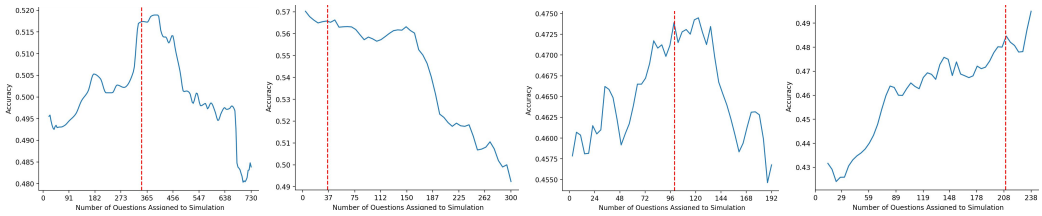

Figure 2: Main results from the HSR model. The **first** figure indicates the accuracy vs computational costs (number of questions assigned to simulation) across all categories. The **second**, **third**, and **fourth** figure indicates the accuracy vs computational costs (number of questions assigned to simulation) in properties, temporal predictive, and counterfactual category. The red dot line indicate the number of questions assigned to simulation determined by HSR model.

looking at how overall accuracy changes as the number of questions assigned to simulation changes, and analyze the dynamics of our model's performance when the threshold τ shifts in each category of the dataset. For additional analysis like comparing the accuracy between our model and SoTA models like Gemini 3 Pro, please refer to Appendix A.3.1. For the model setup and hyperparameter setup please refer to Appendix A.2. When answering a question, the model will be rewarded a score of 1 if all the answers were selected. If the partial correct answers were selected, a score of 0.5 will be rewarded. If none of the correct answers were selected, a score of 0 will be rewarded.

5 Results

5.1 Analysis of Overall Accuracy vs Computational Costs

We start with the analysis of our trained controller in our model. First, we analyze how accuracy changes when the number of questions and videos assigned to simulation changes. As the first figure in fig 5.1 shows, our model is able to identify the near-local maxima of accuracy while optimizing for the computational costs, indicated by the number of questions assigned to simulation, on the accuracy vs computational costs function.

5.2 Analysis of Accuracy vs Computational Costs in Each Categories

As the first figure in fig 5.1 shows, the function of accuracy vs computational costs has a down U-shape functional form, indicating an interesting phenomenon: more simulation does not necessarily improve the overall accuracy. Then a question emerged: are there certain categories that simulation might perform worse? What about chain-of-thought reasoner? Therefore, we compared the accuracy vs the number of questions assigned to simulation in each question category. Our model is preferring less simulation but more chain-of-thought reasoning in the property category, as simulation actually hurts the accuracy when more questions are assigned to simulation. Conversely, in the counterfactual category, our model is preferring more simulation and less chain-of-thought reasoning as simulation constantly improves the accuracy. It is important to emphasize that **category labels are never passed into simulation controller during the training phase**. Therefore, our model being able to pick up the difference at the category level under unsupervised setting strongly indicated that simulation and Chain-of-thought reasoner excels in different categories.

6 Conclusion

In this paper, we introduce Hybrid Simulation and Reasoning Model (HSR) for physical reasoning in MLLMs, and our finding strongly suggested the necessity of having such model to achieve higher accuracy but optimize for computational costs when performing physical reasoning in MLLMs. Furthermore, our finding that simulation and Chain-of-thought reasoner excels in different categories of physical reasoning tasks.

For future work, we want to study the relationship between scene complexity and the necessity for simulation, as well as explore this model using different base VLM model besides Gemini 3 Pro.

REFERENCES

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

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

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Joshua K Hartshorne and Mengguo Jing. Insights into cognitive mechanics from education, developmental psychology and cognitive science. *Nature Reviews Psychology*, pp. 1–15, 2025.

Jingyi Jin, Shun Zhang, Xiaodong Yang, and Joseph Wang. Post-train cosmos reason 2 for autonomous vehicle video captioning and vqa. `https://nvidia-cosmos.github.io/cosmos-cookbook/recipes/post_training/reason2/video_caption_vqa/post_training.html`, January 2026. NVIDIA Cosmos Cookbook.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M Dai. Mind's eye: Grounded language model reasoning through simulation. *arXiv preprint arXiv:2210.05359*, 2022.

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. *Proceedings of the 38th International Conference on Machine Learning*, 139:8748–8763, 2021. doi: 10.48550/arxiv.2103.00020.

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*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

Kevin Smith, Peter Battaglia, and Joshua Tenenbaum. Integrating heuristic and simulation-based reasoning in intuitive physics. 2023.

Felix A Sosa, Samuel J Gershman, and Tomer D Ullman. Blending simulation and abstraction for physical reasoning. *Cognition*, 254:105995, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemini Robotics Team, Krzysztof Choromanski, Coline Devin, Yilun Du, Debidatta Dwibedi, Ruiqi Gao, Abhishek Jindal, Thomas Kipf, Sean Kirmani, Isabel Leal, Fangchen Liu, Anirudha Majumdar, Andrew Marmon, Carolina Parada, Yulia Rubanova, Dhruv Shah, Vikas Sindhwani, Jie Tan, Fei Xia, Ted Xiao, Sherry Yang, Wenhao Yu, and Allan Zhou. Evaluating gemini robotics policies in a veo world simulator, 2025. URL `https://arxiv.org/abs/2512.10675`.

Tomer D Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 21(9):649–665, 2017.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Matan Ben Yosef, Naomi Ken Korem, and Tavi Halperin. Ltx-video community trainer, 2025.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B Tenenbaum, and Chuang Gan. Contphy: Continuum physical concept learning and reasoning from videos. In *International Conference on Machine Learning*. PMLR, 2024.

A Appendix

A.1 Details about ContPhy Dataset

The ContPhy dataset contains categories like property, counterfactual, and temporal dynamic prediction and under four different physical setting: fluid hourglass, bouncing ball, pulley system, and clothing. We consider two sub-sections in this dataset: Fluid and Ball. There are 150 questions in property category for both sections; There are 104 questions in temporal predictive category and 116 questions in counterfactual category for Fluid section; There are 88 questions in temporal predictive category and 122 questions in counterfactual category for Ball section.

A.2 Model Setup and Hyperparameter Settings

We adapt $K = 5$ as the sampling parameter for our ground truth probability y. The visual embedding models we used is CLIP ViT-B/32 (Radford et al., 2021) and the text embedding model we used is all-MiniLM-L6-V2 (Reimers & Gurevych, 2019). For the simulator used in simulation, we used LTX-Video (HaCohen et al., 2024) as the physics-grounded simulator, and we fine-tuned our simulator using the training data in the ContPhy. The parameter for the fine-tuning the LTX-Video model that we used as the simulator follows the official guideline listed in (HaCohen et al., 2024) and the LTX-Video community trainer (Yosef et al., 2025). For the VLM that was used in Chain-of-thought reasoning method and readout phase in the simulation method, we used Gemini 3 Pro as the VLM (Team et al., 2023).

A.3 Additional Analysis

A.3.1 Accuracy Analysis

We first analyze the performance of our model against Gemini 3 Pro, which is a SoTA model across various multi-modal domains. We notice that our model outperform SoTA model like Gemini 3 Pro on the dataset, producing statistically significant performance boosting.

Models	Fluid	Ball	Average
Gemini 3 Pro	0.48 ± 0.02	0.49 ± 0.02	0.48 ± 0.01
Our Model	$\mathbf{0.53 \pm 0.01}$	0.51 ± 0.01	$\mathbf{0.52 \pm 0.01}$

Table 1: The accuracy comparison between our model and Gemini 3 Pro. The accuracies in two sections, fluid and ball, and the average across two sections, are shown in the table. The statistically significant improvements in accuracies are in **bold**. The standard deviation is included. The result indicates that our model is able to outperform Gemini 3 Pro in Fluid section and overall, and the outperformance is statistically significant.