Let's Simulate Frame-by-Frame: In-Context Physical Simulations with Vision-Language Models

YingQiao Wang [*1] **Eric J. Bigelow** [*1 2] **Tomer Ullman** [1 3]

Abstract

In recent years, multi-modal Vision-Language Models (VLMs) have improved substantially in their ability to generate realistic images. This raises important questions about what sort of representation these models have of the world, in particular, how they represent physical objects and their motion over time. We adopt an experimental paradigm from prior work in cognitive science to study physical reasoning. To improve the physical simulation ability of VLMs, we propose a novel method inspired by in-context reasoning and the psychology of mental simulation, which we call Chain-of-Time simulation. In our experiments, we find that a state-of-the-art VLM is able to simulate into the future, but with great errors. This performance is substantially improved when the Chain-of-Time simulation is used, and in this case we also find a human-like bias where a simulation slows down the longer a simulation is run for.

1. Introduction

Recent developments in multi-modal Vision Language Models (VLMs) show impressive capabilities in generating complex, realistic images. But despite their realism, these images can have distinct flaws, and fail to capture real-world structure that is obvious to humans. Understanding the inner workings of VLMs, as well as their uni-modal counterpart, Large Language Models (LLMs), has become a major topic in contemporary AI research (Dang et al., 2024; Chang et al., 2024).

LLMs have been suggested to have rich world models which can represent different concepts and entities. Prior work has examined various dimensions of these world models and related algorithmic capabilities, such as color (Abdou et al., 2021), space (Patel & Pavlick, 2022), spatial reasoning (Yamada et al., 2023), and planning (Li et al., 2023). Comprehensive benchmarks such as PhysBench (Chow et al., 2025) and WM-ABench (Gao et al., 2025) test VLMs on a wide array of physical simulation capabilities.

In this work, we use an experimental paradigm adopted from cognitive science to study how VLMs reason about the spatio-temporal properties of objects in motion. We present VLMs with a sequence of frames describing the past motion of an object in a simple 2D world, and task them with simulating the future motion of the object. We find that ordinary VLMs struggle with this task, and present a novel technique to address this, that we term *Chain-of-Time Simulation*. This technique is inspired by a combination of Chain-of-Thought reasoning in LLMs, and the cognitive process of mental simulation in humans. Our *Chain-of-Time* method leads to a significant boost in the ability of VLM to simulate physical motion. Further, we find that when using Chain-of-Time Simulation, the VLM shows a trend of slowing down time the longer its simulation runs for, which mirrors findings in human psychology.

2. Theories of Mental Simulation in Humans

People are able to efficiently reason about the physical dynamics of everyday objects. For example, if you saw a glass of water begin to fall off of a table, you might quickly and intuitively predict what sequence of events will happen next. Many competing theories have been proposed to explain this 'intuitive physics', and it is likely that humans use a combination of different computations to carry out this reasoning (Hartshorne & Jing, 2025). Within this, one leading theory is that people make robust inference over natural scenes by using a 'mental physics engine' Battaglia et al. (2013); Ullman et al. (2017). On this proposal, people carry out a mental simulation of a physical scene akin to the computations used by engineered systems designed to simulate real-time dynamics in games and animations.

[*]Equal contribution [1]Department of Psychology, Harvard University [2]CBS-NTT Program in Physics of Intelligence, Harvard University [3]Center for Brain Science, Harvard University. Correspondence to: YingQiao Wang <yingqiaowang@g.harvard.edu>, Eric Bigelow <ebigelow@g.harvard.edu>.

ICML 2025 Workshop on Assessing World Models.

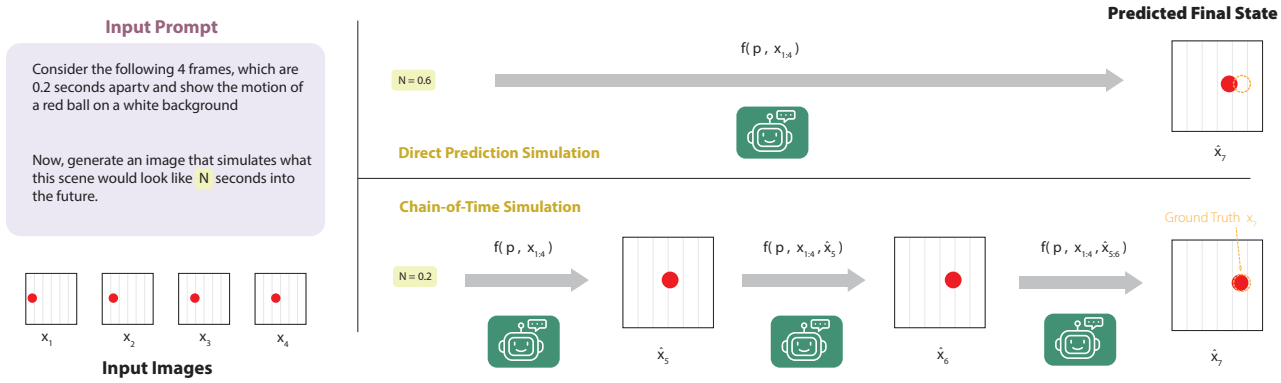

Figure 1. For our experimental paradigm, we give VLMs a sequence of input images and a prompt instructing the model to simulate into the future for a specified length of time (Left). As a baseline, Direct Prediction (Right, Top) directly predicts the final world state. We propose a novel method, Chain-of-Time Simulation, which instead generates a sequence of images preceding the predicted final state.

Again, while this is not the only proposal for how people carry out intuitive physics (and see for example (Ludwin-Peery et al., 2021)), it has found support in cognitive science, computational modeling, development, and neuroscience (Fischer et al., 2016; Gerstenberg & Stephan, 2021; Allen et al., 2021; Fischer, 2021).

We formulate the mental physics engine in terms of a statistical model as follows:

Suppose that we want to simulate the physical dynamics of a scene that lasts T seconds, after we observe the scene for t second. The mental game engine can be thought of as a function, ϕ, that takes in the current state of the world and applies dynamic rules to output a distribution over future states of the world. To simplify here, we can consider a deterministic transition such that:

$$X_T|X_{T-1} = \phi(X_{T-1})$$

For each state t, a noise parameter σ will be added to account for the noise in perception and simulation due to the complexity of the real-world scenario and potential noise in perception, turning the transition into a probabilistic transition. When the simulation of physical dynamics terminates and we obtain the distribution of X_T.

Notice that the states at timestep t is Markovian if a physics engine is used. This Markov Chain will form a working memory space that the following distribution could describe:

$$p(X_T, X_{T-1}....X_0) = p(X_T|X_{T-1})....p(X_1|X_0)p(X_0)$$

It is worth pointing out another potential flaw in the framework: how fine-grained the step t is would potentially impact the hypothesis space of the distribution $X_T, X_{T-1}, ..., X_0$, causing the working memory space to contain varied information about the physical dynamics it is trying to simulate, depending on how fine-grained the step t is. Therefore, the fine-grainness of t could potentially impact the prediction accuracy, which we will investigate in section 5.

This step-by-step process of human physical simulation is useful for many cases, and serves as a motivation for how VLMs may be made to reason about physical scenes.

3. Chain-of-Time Simulation

Motivated by studies of intuitive physics and mental simulation in humans, we propose a novel method for improving physical reasoning in VLMs with in-context simulations, which we call *Chain-of-Time Simulation* (Figure 1). The goal of physical simulation in VLMs is as follows: given a sequence of input images up to a given time T $x_{t=0...T}$, generate a new image $\hat{x}_{T+k}$ that accurately depitcs what the scene will look like k time steps (or "frames") into the future. Chain-of-Time Simulation involves two prompts (provided in Appendix A): first, a Simulation Instruction Prompt that, along with a sequence of input images, instructs the model to simulate an image s frames (or equivalently, s seconds) into the future. Clearly, s must be smaller than k. In our experiments, we use $k = 0.8$ s and $s \in \{0.2\text{ s}, 0.4\text{ s}\}$. After the VLM generates a single image, we continue with our Simulation Follow-up Prompt, which instructs the model to generate another image simulated an additional s frames into the future. As a baseline for this task, we construct a Direct Prediction Simulation prompt, which instructs the VLM to directly predict $\hat{x}_{t+k}$ given $x_{t=1...4}$. Note that this is equivalent to Chain-of-Time simulation with only a single timestep, i.e. $s = k$.

Chain-of-Time Simulation is inspired by two bodies of prior literature: the cognitive science of mental simulation (described above in Section 2) and in-context reasoning in LLMs. In-context reasoning methods with LLMs coerce the model to spell out intermediate reasoning steps in its output stream, before giving a final answer. This may be through prompting, as in Chain-of-Thought reasoning (Kojima et al., 2022), or through specialized training regimes (Guo et al., 2025; Jaech et al., 2024). Various theories have been developed to try to explain precisely why and how these methods work (Wang et al., 2022), and in some cases a model's intermediate reasoning tokens may not align with its final answer (Turpin et al., 2023).

Similar to Chain-of-Time simulation, prior works have proposed in-context reasoning methods for VLMs, which use images instead of language to represent individual reasoning steps. However, our method differs from these works in a few critical ways. Hu et al. (2024) proposes a method to solve simple reasoning problems with a VLM, such as geometry and spatial reasoning, and individual steps involve interleaved images and text outputs. (Xu et al., 2025) proposes a method for planning where a VLM generates sequential images to solve tasks such as maze navigation; their approach requires additional training. By contrast, the goal of Chain-of-Time simulation is to improve physical simulation with VLMs, where "steps" in a chain correspond to segments of time. Further, like Hu et al. (2024), our method can be applied to out-of-the-box VLMs with no additional training.

4. Experiments

We hypothesize that by using Chain-of-Time simulation, VLM models will be able to achieve better accuracy than when using direct prediction. We test this on a simple physical reasoning task in which a VLM must predict the position of an object, with accuracy measured in RMSE (Square Root Mean Squared Error) between the actual location of the ball and the VLM's prediction of it. We aim to determine if the precision of the Chain-of-Time simulation influences the simulation's accuracy, by lowering the RMSE between the predicted trajectory and ground truth trajectory.

4.1. Experimental Setup

Stimuli Design To measure a model's accuracy in physical reasoning, we designed an experiment that involves classical psycho-physical stimuli, and asked VLMs to perform a simulation and prediction task.

The stimuli we used generally resembles stimuli in previous studies of intuitive physics (Smith & Vul, 2013), (Bass et al., 2021) ,(Gerstenberg & Stephan, 2021), involving the motion of an object in a 2D plane. The specific stimulus shows a 2D red ball rolling on a white background. There is no friction, and no visible boundaries on the white surface. The white surface is flat and featureless, allowing the red ball to roll without obstructions, as shown in Figure 1. We varied the speed at which the ball rolled over the white surface with three different speeds: 120, 300, and 500 (in units of pixels/second).

Experimental Procedure We used OpenAI's GPT (gpt-image-1 [1]) as the VLM model in our experiment. As described above, the stimulus will run for 0.8 seconds. The objective of the task is to predict the future position of the ball after 0.8 seconds.

At the start of each trial, the model was given 5 frames of the stimulus, showing the scene at 0, 0.2, 0.4, 0.6, and 0.8 seconds. Given the 5 frames, the model was asked to generate the first simulated frames, following the Initial Simulation Prompt we listed in **Appendix A: Prompt**. The rest of the frames were generated following the Simulation Follow-Up Prompt we listed in **Appendix A: Prompt**.

Sampling Details We designed two Chain-of-Time simulations with different precisions. The first one is named "Chain-of-Time 0.2s", meaning the model generate the simulated frame for every 0.2 seconds, for 4 times after the video began. The 4 frames generated by the model describe the position of the red ball 0.2, 0.4, 0.6, and 0.8 seconds after the last frame provided to the model. The second one is named "Chain-of-Time 0.4s", which is the same as "Chain-of-Time 0.2s", but the model will generate the simulated frame every 0.4 seconds for 2 times.

In addition, we have a control process termed "Direct Prediction". In this method, we asked the model to directly generate the simulated frame 0.8 seconds after the last frame we provided to the model.

For all three simulations types, we will run each simulation type on each stimulus 5 times (N=5).

5. Results

5.1. Accuracy Analysis

As described in Section 4, we measured the model's performance under three different simulations we defined: Chain-of-Time 0.2s, Chain-of-Time 0.4s, and Direct Prediction. As Table 1 shows, the Chain-of-Time 0.2s has the best overall RMSE across all three speeds, beating Chain-of-Time 0.4s and Direct Prediction. When collapsed across different speed, as figure 2 shows,the ΔX prediction error reached the lowest level when Chain-of-Time 0.2s was used, indicating that Chain-of-Time 0.2s was able to reduce the error in predictions.

[1] openai.com/index/image-generation-api/

Speed	Chain-of-Time 0.2s	Chain-of-Time 0.4s	Direct Prediction
120 Pixels / Second	**78 ± 15**	143 ± 47	143 ± 96
300 Pixels / Second	**102 ± 46**	133 ± 92	110 ± 102
500 Pixels / Second	**48 ± 22**	126 ± 54	144 ± 141

Table 1. The Mean RMSE between predicted trajectory and ground truth trajectory, by simulation types and speeds, with 95% CI.

The findings indicate that the Chain-of-Time simulation was able to increase the prediction accuracy, and the precision of Chain-of-Time influences the accuracy of the predicted positions.

5.2. Predicted Positions Analysis

As shown in Table 1, both Chain-of-Time methods were able to obtain the lowest RMSE at 500 pixels/second speed, demonstrating that under 500 pixels/seconds, models were able to produce the most reliable prediction among all three speeds we provided to the model. Therefore, we zoomed in to analyze the predicted positions from all three simulations under the 500 pixels/second speed.

We performed linear regression over the difference between the predicted horizontal positions (X positions) and the ground truth horizontal positions, and time step over data of Chain-of-Time 0.2s, Chain-of-Time 0.4s, and Direct Prediction. Since Direct Prediction only predicts positions once, We assume that the simulation runs under constant speed within a trial, and the previous positions would be inferred by evenly dividing the predicted trajectory into four parts.

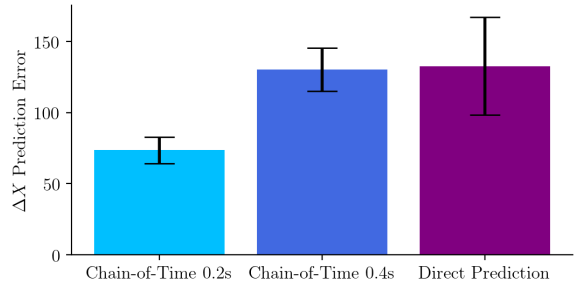

Figure 2. Prediction errors for all three simulation types, averaged across all data. Prediction error is measured by taking the average difference between the ground truth displacements (i.e. change in x-location between frames) and the predicted displacement. Error bars are 95% CI.

As Figure 3 shows, the ΔX for the Direct Prediction and Chain-of-Time 0.4s are always below 0, meaning that both simulation types suffered left bias during the simulation. Chain-of-Time 0.2s has positive ΔX and its confidence interval covers 0 before 0.5 seconds, meaning that Chain-of-Time could simulate with right bias or no bias before 0.5 seconds. But Chain-of-Time 0.2s exhibited left bias after around 0.5 seconds too.

This indicates that the model tends to underestimate how far the object moved forward. Furthermore, both Chain-of-Time 0.2s and Chain-of-Time 0.4s exhibited a downward trend, meaning that the left bias grew over time, meaning that the distance the ball traveled simulated by the model decreases over time. This indicates that the simulation somehow slowed down within the model, leading to prediction errors.

Both left bias and the slowing down effect were consistent with human behavior, as (Wang & Ullman, 2025) demonstrated that under the same experimental paradigm, the predicted positions reported by humans exhibited left bias, and the left bias grew over time, causing the distance between the predicted horizontal positions and ground truth horizontal positions to increase negatively.

6. Discussion

Our results suggest that while VLMs inherently have some degree of physical simulation ability, accuracy degenerates significantly when simulating further into the future. Our Chain-of-Time method seems to greatly improve this ability, particularly with long simulations. When using this method, we also find a surprising similarity between VLM behavior and patterns observed with humans: in both cases, time be-

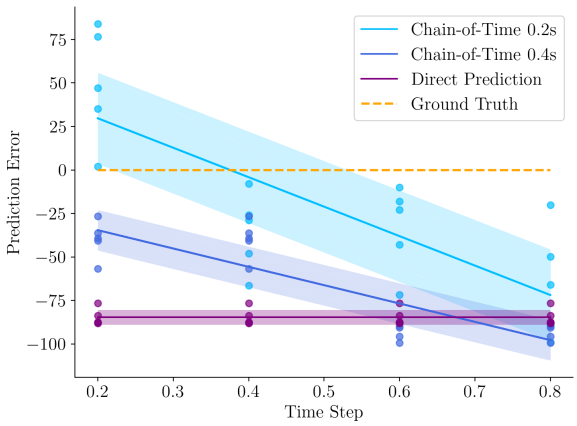

Figure 3. Accuracy in predicting the ball's X-location as a function of the number of time steps simulated (speed: 500 pixels/sec). Negative values in Prediction Error represent model-predicted positions are to the left of the ground truth, and and positive value represent right-bias. Shaded areas represent 95% CI.

gins to slow down the longer a simulation is run for. We see there being enormous opportunity for further experiments in this vein, perhaps examining other aspects of physical reasoning such as causality and non-simulative (heuristic- or abstraction-based) physical reasoning.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., and Søgaard, A. Can language models encode perceptual structure without grounding? a case study in color. *arXiv preprint arXiv:2109.06129*, 2021.

Allen, K. R., Smith, K. A., Bird, L.-A., Tenenbaum, J. B., Makin, T. R., and Cowie, D. Lifelong learning of cognitive strategies for physical problem-solving: the effect of embodied experience. *bioRxiv*, pp. 2021–07, 2021.

Bass, I., Smith, K. A., Bonawitz, E., and Ullman, T. D. Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, 38(7-8):413–424, 2021.

Battaglia, P. W., Hamrick, J. B., and Tenenbaum, J. B. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110 (45):18327–18332, 2013.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Chow, W., Mao, J., Li, B., Seita, D., Guizilini, V., and Wang, Y. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

Dang, Y., Huang, K., Huo, J., Yan, Y., Huang, S., Liu, D., Gao, M., Zhang, J., Qian, C., Wang, K., et al. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*, 2024.

Fischer, J. The building blocks of intuitive physics in the mind and brain, 2021.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., and Kanwisher, N. Functional neuroanatomy of intuitive physical inference. *Proceedings of the national academy of sciences*, 113(34):E5072–E5081, 2016.

Gao, Q., Pi, X., Liu, K., Chen, J., Yang, R., Huang, X., Fang, X., Sun, L., Kishore, G., Ai, B., et al. Do vision-language models have internal world models? towards an atomic evaluation. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*, 2025.

Gerstenberg, T. and Stephan, S. A counterfactual simulation model of causation by omission. *Cognition*, 216:104842, 2021.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Hu, Y., Shi, W., Fu, X., Roth, D., Ostendorf, M., Zettlemoyer, L., Smith, N. A., and Krishna, R. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR*, 2023.

Ludwin-Peery, E., Bramley, N. R., Davis, E., and Gureckis, T. M. Limits on simulation approaches in intuitive physics. *Cognitive Psychology*, 127:101396, 2021.

Patel, R. and Pavlick, E. Mapping language models to grounded conceptual spaces. In *International conference on learning representations*, 2022.

Smith, K. A. and Vul, E. Sources of uncertainty in intuitive physics. *Topics in cognitive science*, 5(1):185–199, 2013.

Turpin, M., Michael, J., Perez, E., and Bowman, S. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36: 74952–74965, 2023.

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

Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., and Sun, H. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2022.

Wang, Y. and Ullman, T. D. Bullet time: The time distortion during physical reasoning explained by cognitive resource constrain and goal specification. *In Prep*, 2025.

Xu, Y., Li, C., Zhou, H., Wan, X., Zhang, C., Korhonen, A., and Vulić, I. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025.

Yamada, Y., Bao, Y., Lampinen, A. K., Kasai, J., and Yildirim, I. Evaluating spatial understanding of large language models. *arXiv preprint arXiv:2310.14540*, 2023.

A. Prompts

For our Chain-of-Time simulation method, as well as our Direct Prediction baseline, models are provided the following prompt, with different methods (Chain-of-Time 0.2s, 0.4s, and Direct Prediction) varying the {{`number of seconds forward`}} parameter:

Simulation Instruction Prompt

```
Consider the following {{number of inputs}} frames, which show the motion of a red
    ball on a white background. Note that each frame is precisely .2 seconds apart.

Now, please generate an image that simulates what this scene would look like {{
    number of seconds forward}} Seconds into the future.

Make sure that your image is 2d and consists of a single red circle on a solid white
     background. Ensure that the circle is exactly the same size as the input images.
     Assume that there is no friction, the ground is flat, and the ball can pass
    through objects.

{{image sequence}}
```

For Chain-of-Time simulation, we use the following prompt to elicit subsequent simulation steps from the VLM:

Simulation Follow-Up Prompt

```
Now, simulate additional {{number of seconds forward}} seconds into the future.
```