
Will Visual World Models Replace Large Language Models?

Odysseus Kaloidas
UC Berkeley

Abstract

Visual world models have a strong case in robotics: a machine acting in the physical world needs to learn from what it sees and from the consequences of its actions. The case for replacing large language models (LLMs) is less clear. Much reasoning expressed in language concerns relationships and rules that further visual experience may do little to clarify. Vision can enrich the concepts a model learns, but better physical prediction does not automatically produce better conceptual reasoning. The more plausible prospect is that visual world models complement language models. How far their benefits extend beyond physical tasks remains an empirical question, not a consequence of learning from more modalities.

1 A Strong Case for Robotics

A robot asked to carry a glass across a crowded table needs more than a description of the task. It needs to locate the glass, anticipate collisions and judge how its movements will affect what it holds. Learning from images, video and physical interaction makes immediate sense here. Why should the same advance replace a model used to interpret a rulebook, work through a mathematical argument or reason about a software specification?

That is the gap in the replacement argument. Visual world models could substantially improve AI's ability to act in physical environments without producing an equally large improvement in reasoning about concepts. Their promise in robotics gives a reason to extend language models, but does not by itself give a reason to replace them.

A visual world model predicts how an environment changes from images or video, sometimes given a proposed action. It does more than recognize objects in a picture: it can help anticipate what happens next. World models need not be visual, but the visual case is the subject here. Its most direct contribution is to bring additional kinds of experience into learning, especially experience that text describes incompletely.

V-JEPA 2 illustrates this direction. It learns representations from video, and its action-conditioned version supports prediction and robotic planning [1]. This is a substantive technical achievement. Yet demonstrating physical prediction and control does not establish how much the same approach improves a model's ability to reason from written premises.

2 What Vision Adds, and What Text Already Supplies

Language often presents a problem after someone has selected its relevant features. A rulebook specifies what counts as a legal move. A software specification states conditions an implementation must satisfy. A mathematical definition fixes relationships between abstract objects. These descriptions can still demand difficult reasoning, but their difficulty need not come from missing visual information.

Suppose two clauses in a rulebook appear to conflict. One permits an action; another restricts it under an exception. Settling the case requires understanding the scope of the clauses and applying them

consistently. Seeing more examples of people moving objects may contribute little to that particular difficulty. The obstacle is working through the relationships already stated.

This is not an argument that text contains everything worth knowing. The glass on the table has a position, shape and physical behavior that the instruction leaves unspecified. Nor does a complete description make a problem easy: all the rules of chess can fit into a short text without making good play straightforward. The point is that missing information and difficult inference are different reasons to fail.

For this reason, the amount of sensory data a system receives is not a measure of its value for every task. More observations can reveal a regularity; they can also repeat details irrelevant to the question. How much vision enriches language-based reasoning depends on which useful relationships it teaches, and whether the model could otherwise learn and apply them.

3 Could Seeing Change How a Model Thinks?

The strongest case for transfer concerns learning, rather than supplying another image at the moment of answering. A model might learn containment by watching objects enter and leave boxes, then use that concept to reason about nested groups in a written problem. Even when the problem states every membership relation, earlier visual experience could have helped the model acquire a useful abstraction.

This is a possible route to improvement, not evidence that it is better than learning the same relationship through language. It matters because the absence of a need for new visual input does not establish that visual training was useless. The question is whether such training produces better reasoning from the same premises.

Research on grounding gives this possibility a serious foundation. Bender and Koller question what linguistic form alone can supply, while Bisk and colleagues examine what language learning gains from experience beyond text [2; 3]. CLIP demonstrates the usefulness of learning relationships between images and text [6]. These works motivate broader experience; they do not settle the size of its contribution to abstract reasoning.

The expectation developed here is therefore limited but consequential: gains may be much larger where physical information is missing than where a task mainly requires applying concepts and rules already available. That expectation needs evidence across tasks. It should not be turned into a claim that vision can never help mathematics or that concepts arise only through words.

4 From Physical Intelligence to Replacement

In his May 2026 interview, Yann LeCun argues that current LLMs will not reach human-level intelligence and gives language an interface role in a future system. He also acknowledges the usefulness of LLMs in mathematics and coding [4]. His proposal is broader than adding a camera: it includes predictive representations, planning and different internal computation [5].

The disagreement concerns the step from the limits of text-only learning in the physical world to the displacement of language models more generally. A better way to plan or search could help with a written problem without relying on vision. Learning from actions can also occur in software environments. If a successor improves through several changes at once, its success would not show that visual learning supplied the decisive improvement in conceptual reasoning.

A robot may ultimately combine language, visual prediction and action in one model. Yang and colleagues' position on video for decision-making explicitly considers video alongside language [7]. Such integration could make today's separate components unnecessary. It would not establish that learning from language had ceased to contribute, or that stronger visual capabilities alone had replaced the reasoning previously carried out by an LLM.

Visual world models answer an important need: learning to anticipate events in environments where words omit much of what matters. That is ample reason to develop them. For many problems posed in language, however, the hard part is drawing the right consequences from what is already described. Until transfer between these demands is established more broadly, complementing LLMs is the more convincing prospect. The glass and the disputed rule both require intelligence, but they need not be waiting for the same advance.

References

- [1] Mido Assran, Adrien Bardes, David Fan, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL <https://arxiv.org/abs/2506.09985>. arXiv:2506.09985.
- [2] Emily M. Bender and Alexander Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, 2020. doi: 10.18653/v1/2020.acl-main.463. URL <https://aclanthology.org/2020.acl-main.463/>.
- [3] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience Grounds Language. arXiv:2004.10151, 2020. URL <https://arxiv.org/abs/2004.10151>.
- [4] Annelies Gamble. AI for the real world: A conversation with Yann LeCun. Mixture of Experts, May 12, 2026. URL <https://www.mixtureofexperts.co/p/ai-for-the-real-world-a-conversation>. Firsthand interview.
- [5] Yann LeCun. A path towards autonomous machine intelligence, 2022. URL <https://openreview.net/forum?id=BZ5a1r-kVsf>. Position paper, version 0.9.2.
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision, 2021. URL <https://arxiv.org/abs/2103.00020>. arXiv:2103.00020.
- [7] Sherry Yang, Jacob C Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Position: Video as the new language for real-world decision making. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 56465–56484. PMLR, 2024. URL <https://proceedings.mlr.press/v235/yang24z.html>.