Rethinking the Simulation vs. Rendering Dichotomy: No Free Lunch in Spatial World Modelling

Dezhi Luo
University of Michigan
ihzedoul@umich.edu

Qingying Gao
Johns Hopkins University
qgao14@jh.edu

Hokin Deng
Carnegie Mellon University
hokind@andrew.cmu.edu

Abstract

Spatial world models, representations that support flexible reasoning about spatial relations, are central to developing computational models that could operate in the physical world, but their precise mechanistic underpinnings are nuanced by the borrowing of underspecified or misguided accounts of human cognition. This paper revisits the simulation versus rendering dichotomy and draws on evidence from aphantasia to argue that fine-grained perceptual content is critical for model-based spatial reasoning. Drawing on recent research into the neural basis of visual awareness, we propose that spatial simulation and perceptual experience depend on shared representational geometries captured by higher-order indices of perceptual relations. We argue that recent developments in embodied AI support this claim, where rich perceptual details improve performance on physics-based world engagements. To this end, we call for the development of architectures capable of maintaining structured perceptual representations as a step toward spatial world modelling in AI.

1 Introduction

Human reasoning is distinguished by its capacity to construct structured simulations of the external world, or world models, that support prediction, planning, and counterfactual inference beyond what is immediately perceived [Johnson-Laird, 1983, Wooldridge and Jennings, 1995, LeCun, 2022]. Among these, **spatial world models** serve as a foundational substrate for structured spatial reasoning: the ability to represent and manipulate spatial relations in a flexible, generalizable manner [Kuipers, 1978, Kosslyn et al., 1979, Hegarty, 2011, Gao et al., 2025]. Recently, spatial world modelling has garnered significant attention as a promising approach for enhancing spatial reasoning capabilities in foundation models [Chaplot et al., 2021, Spies et al., 2024, Gao et al., 2025, Cai et al., 2025]. Yet, the underlying computational mechanisms remain poorly understood, in part because the concept of a spatial world model is often treated as an idealized abstraction rather than an empirically grounded construct.

Current evaluation methods for model-based spatial reasoning tend to rely on task paradigms that are assumed to require internal simulation, such as mental rotation, visual perspective-taking, and mechanical reasoning [Jelassi et al., 2022, Gao et al., 2024, Sun et al., 2024, Zhang et al., 2024, Wang et al., 2025a]. These assumptions are often informed by cognitive science frameworks that describe how humans approach spatial problems [Shepard and Metzler, 1971, Hegarty, 2004]. If a task is *believed* to necessitate mental simulation of spatial transformations, it is categorized as diagnostic of model-based reasoning. However, such paradigms tend to ignore the nuanced nature of the internal mechanisms humans employ to navigate these problems, which await consensus among cognitive scientists.

This paper addresses this important challenge of formulating a clear conceptualization of spatial world models: namely, what kinds of internal representations are required to sustain robust simulations

Accepted at NeurIPS 2025 Workshop on Space in Vision, Language, and Embodied AI.

of the external world's spatial properties? To this end, we first revisit the so-called "simulation vs. rendering" dichotomy in mental imagery research. Examining the case of aphantasia, we argue that the standard formulation of this dichotomy is misconstrued. Rather than treating simulation and rendering as separate processes [Balaban and Ullman, 2025], we propose that both may rely on convergent, higher-order representations that capture the instantiated relations between perceptual objects. The implication is that the representational substrate of spatial world models must be more fine-grained than previously assumed, requiring architectures that go beyond shallow modelling of spatial relations.

2 On the Split Between Simulation and Rendering

Efforts to characterize spatial world models often appeal to how humans appear to solve tasks that presumably require internal simulation. A canonical example is mental rotation, where response times scale with angular disparity, suggesting incremental manipulation of internal spatial representations [Pylyshyn, 1979, Khooshabeh et al., 2013, Hilton et al., 2022]. This has been interpreted as evidence for simulation-based reasoning [Pylyshyn, 1979, Searle and Hamm, 2017]. However, the underlying mechanisms are not directly observable, and mental rotation is typically accompanied by **conscious visual experience**, raising the question of whether such imagery plays a functional role or is merely epiphenomenal [Wexler et al., 1998, Vingerhoets et al., 2002]. To address this, recent proposals distinguish between **physics-based** and **graphics-based** components of mental imagery [Balaban and Ullman, 2025]. On this view, tasks such as rotation, scanning, and manipulation may be solved through *simulations over structured spatial representations, without requiring perceptual rendering*. Rendering may still occur, but it bares no direct relevance to correctly and robustly completing these tasks. Here, we note that this position hinges on a linear interpretation of the functional specialization across streams of visual processing in the human brain, and that it is misguided. Below, we discuss the theoretical claims and corresponding evidence in objection to the above interpretation, and highlight that they instead point toward a **non-linear, hierarchical relationship** between simulation and rendering.

2.1 Examining the Spatial Imagery Framework and the Case of Aphantasia

The proposed simulation vs. rendering dichotomy draws particular support from studies of **aphantasia**. Individuals with aphantasia report an inability to generate voluntary visual images, yet often perform comparably to the general population on tasks such as mental rotation [Kay et al., 2024, Balaban and Ullman, 2025]. This has been interpreted as evidence that spatial simulation can operate independently of conscious vision. To explain this, researchers have proposed a distinction between **spatial imagery** and visual or object-based imagery, with the former supporting abstract reasoning about relative positions, distances, and spatial relations without relying on pictorial content. On this account, preserved task performance in aphantasia is attributed to intact spatial imagery mechanisms. Balaban and Ullman [2025] further has broadened this concept to encompass physical simulation for reasoning about motion and interaction. Under this conceptualization, spatial imagery encodes the entire scene—including geometric structure, spatial extent, and other object properties—sufficient to simulate its temporal evolution. According to them, while no rendering occurs, non-physical attributes can still be abstractly represented, akin to an engineer's amodal model. Importantly, this implies that simulations required for capturing the spatial relations of the real-world environment, sufficient for structured physical reasoning, do not demand the computations used to support the **fine-grained perceptual content** in the conscious visual scene that someone without aphantasia would experience when completing tasks.

While the spatial imagery framework offers a plausible account of preserved task performance in aphantasia [Kay et al., 2024], it rests on a notion that remains under-specified. Notably, it is not apparent what exactly constitutes the content of spatial "imagery". Proponents of the framework often suggest that spatial imagery involves abstracted or schematic representations of spatial features, such as location, relation, or shape, that are "modality-neutral" and potentially amodal. Phillips [2025], for example, describes such imagery as "neutral as to whether the location, relation, shape or structure is imagined as seen or touched." Nevertheless, when describing how such abstract "imageries" are deployed in action, an example is given as the "subjects imagine grasping the shape and rotating it". This framing seems to rely on the covert recruitment of embodied sensory modalities without clarifying how the spatial content itself is accessed or represented. Simply put, there does not seem

to be any reason to insist that the spatial component underlying such recruitment must be consciously experienced.

If this commitment to conscious experience in the spatial imagery framework is denied, the account then aligns more closely with theories proposing that aphantasia reflects preserved unconscious mental imagery [Michel et al., 2025]. Yet this position, too, has faced criticism with respect to the notion of "imagery". Some argue that imagery inherently entails fine-grained perceptual content, typically associated with modality-specific sensory representations—for instance, those involving early visual cortex [Scholz et al., 2025]. From that perspective, if individuals with aphantasia lack the capacity to reconstruct such content, it is unclear how spatial imagery, defined in amodal terms, could qualify as imagery at all regardless of its experiential status. Given the dubious nature of the spatial imagery framework with respect to phenomonology, we suggest that a more tractable position may be to accept that individuals with aphantasia rely on spatial representations—whether imagistic or not—to solve tasks such as mental rotation, while lacking the corresponding visual experience [Hinton, 1979, Nanay, 2021]. Determining whether such representations count as "imagery" may ultimately be undecidable, and for the purpose of conceptualizing spatial world models, this ambiguity is largely orthogonal.

2.2 Rethinking the Linear Interpretation

Now that we have clarified the nuances surrounding the aphantasia debate, a key question for understanding the mechanistic basis of spatial world modelling is what kinds of spatial representations underlie task performance in individuals with aphantasia. In this context, Balaban and Ullman [2025] propose that aphantasia serves as a proof-of-concept for their claim that spatial/physical imagery and visual/object imagery, acting as alleged instantiations of simulation and rendering, are subserved by distinct mechanisms, a view grounded in the neuroanatomical dissociation between the **dorsal** and **ventral** visual streams. The dorsal stream is associated with action-oriented spatial processing, while the ventral stream supports object recognition and visual imagery [Goodale and Milner, 1992]. However, this **linear interpretation** of functional specialization is highly contested. Accumulating evidence from inter-stream crosstalk, perceptual integration, and lesion studies suggests that the two streams are not strictly independent, but instead operate within a more distributed and interactive network architecture [Schenk and McIntosh, 2010, de Haan and Cowey, 2011, Milner, 2017]. This is not to say that all functional specializations with respect to the dorsal and ventral streams are non-linear, and what Balaban and Ullman [2025] did acknowledge this nuance and maintain otherwise. However, we highlight that simulation and rendering is indeed one of these examples where this interpretation does not hold, precisely due to the role of spatial representations in conscious vision.

Following the conceptualization in the original linear interpretation, the dorsal stream has traditionally been cast as the "zombie" stream—so named for its presumed lack of contribution to conscious vision [Goodale and Milner, 1992, Chalmers, 1997, de Haan and Cowey, 2011, Milner, 2012, Wu, 2014]. However, later accounts challenge this view by proposing that dorsal-stream information related to spatial encoding may in fact play a direct role in shaping visual experience. To begin with, converging evidence from lesion studies and neuroimaging data highlights the involvement of intraparietal regions such as the ventral intraparietal area (VIP) and the lateral intraparietal area (LIP), which encode head- and body-centered reference frames and integrate corollary discharge signals. These dorsal mechanisms help maintain stable spatial representations, including perceived distance, across saccades and object motion, thereby anchoring egocentric spatial frameworks that are essential to the continuity of visual experience [Wu, 2014].

This reappraisal has been substantiated by neural evidence indicating that the high-level regions of the dorsal stream is part of the broader fronto-parietal network that encodes fine-grained perceptual content integral to conscious vision. Rather than merely representing coarse spatial representations to support action-oriented or post-perceptual executive functions, these regions appear to participate directly in constructing the contents of visual experience by interacting with the ventral stream, suggesting that simulation and rendering may not be functionally dissociated. In particular, it has been demonstrated that both prefrontal and posterior parietal cortices can decode object identity from rapidly presented visual stimuli even in the absence of behavioural report, with decoders performing above chance in the posterior parietal cortex (PPC), a region traditionally associated with the dorsal "zombie" stream [Bellet et al., 2022]. This finding indicates that dorsal areas actively encode perceptual information rather than merely mediating visuomotor transformations.

Going further upstream, the dorsolateral prefrontal cortex (DLPFC), a region long implicated in leveraging spatial encodings for action planning and cognitive control, has also been shown to selectively correlate with visually aware information [Lau and Passingham, 2006, Anzulewicz et al., 2019]. Complementary evidence from lesion and studies further supports this view: damage to or deactivation of prefrontal and parietal cortices impairs the integration and maintenance of visual content in awareness [Szczepanski and Knight, 2014, Persaud et al., 2011], while prefrontal ensembles exhibit category- and configuration-selective tuning that evolves with visual experience [Rainer and Miller, 2000, Chan, 2013]. Collectively, these results suggest that simulation and perceptual representation are deeply intertwined within the fronto-parietal network. Rather than functioning as isolated control or sensory systems, the dorsal and ventral streams appear to interact dynamically, integrating spatial encodings with object-level perceptual details to sustain a unified neurocognitive substrate for conscious vision [Panagiotaropoulos, 2024, Rees, 2007].

Together, these findings provide empirical support for higher-order theories (HOT) of consciousness, which posit that meta-representations of perceptual content, supported by fronto-parietal networks, play a central role in shaping perceptual awareness. In particular, relational HOT—including perceptual reality monitoring (PRM) [Lau, 2019] and higher-order state space (HOSS) [Fleming, 2020] models—converge on the view that the representational substrates of higher-order encodings underlies the phenomenal character of experience, including fine-grained spatial properties such as perceived distances and object relations [Fleming and Shea, 2024]. On these views, higher-order representations encode how perceptual states relate to one another within structured quality spaces. Crucially, these higher-order representations are not themselves inherently conscious. Rather, they must be discriminated as either reliable reflections of external reality or internally generated signals (e.g., imagination or noise) in order to be gated into conscious awareness. PRM, for instance, likens this gating mechanism to a discriminator in a generative adversarial network (GAN), whereby higher-order processes evaluate which sensory signals are "real" enough to be experienced [Gershman, 2019, Lau, 2022]. Upon making such a determination, the discriminator generates a pointer that carries information about the location where the totality of visual details within experience is stored. This pointer is then fed back into first-order networks for decoding, thereby rendering the conscious experience [Lau, 2019, Butlin et al., 2023]. In this view, while rendering itself (the decoding process) may not directly contribute to spatial reasoning, the fine-grained geometric structure embedded within higher-order indices that capture complex object-level properties can suffice the kinds of internal simulations underlying spatial world modelling.

This framework assumes a **non-linear, hierarchical relationship between simulation and rendering**: the capacity to simulate physical dynamics and the capacity for conscious visual experience may rely on the same underlying representational substrate—differing only in whether these states are verified and utilized by downstream systems [Rosenthal, 2010, Fleming and Shea, 2024]. From this perspective, broken rendering in aphantasia does not indicate a failure of a dedicated visual stream, but rather a failure in the gating or feedback process: either the discriminator fails to classify a state as a sufficiently fine-grained and reliable index of reality versus imagination, or the downstream consumer system fails to register and decode the output [Michel et al., 2025]. Both failures impact the conscious awareness of mental simulations but not its functional role in solving spatial reasoning tasks. In the case of aphantasia, where performance on tasks demanding model-based spatial reasoning remains intact, it is more likely to be the latter. This is because top-down determination of the reliability of first-order information remains essential for decision-making, even if such processes do not generate a conscious experience. Supporting this view, a recent study of aphantasia patients found that all 12 cases of lesion-induced aphantasia involved damage to regions functionally connected to the left fusiform gyrus, a region strongly implicated in visual mental imagery. Notably, no significant lesions were found in prefrontal cortices, suggesting that higher-order monitoring mechanisms remained intact [Kutsche et al., 2025]. This pattern implies that the deficit in aphantasia may arise from impaired downstream decoding, rather than from the absence of higher-order representations themselves. Taken together, these findings suggest that fine-grained perceptual representations underlying the content of conscious vision are intrinsic to the functional architecture supporting model-based spatial reasoning in humans, going against the alleged dichotomy between simulation and rendering.

3 No Free Lunch in Spatial World Modelling

Rather than treating physical simulation and mental imagery as outputs of dissociable visual systems, the non-linear interpretation locates the critical bottleneck in higher-order re-representation and discrimination processes—mechanisms responsible for indexing and gating perceptual content into conscious experience. This view is consistent with both the broader evidence of dorsal–ventral integration and the specific lesion patterns observed in cases of aphantasia. Crucially, this matters beyond human neuroscience: we propose that the so-called spatial world models in AI may rely on the very same class of structured, fine-grained perceptual representations that underlie conscious visual experience in humans.

In short, there is *no free lunch* in spatial modelling: if human spatial reasoning depends on fine-grained encodings, we should not expect coarse, perceptually abstract approximations to suffice for models aiming to emulate such capacities. This aligns with recent evolutionary accounts proposing that conscious visual experience evolved to stabilize internal simulations—allowing organisms to determine when to commit to a world model and act. On this view, consciousness may have arisen as a by-product mechanism that gates simulations with sufficient fidelity to guide decision-making in the wild [Fleming and Michel, 2025]. Importantly, we do not claim that the specific computational implementations underlying human spatial cognition constitute the only viable path toward such capabilities in artificial systems. Nevertheless, they remain the only empirically validated example of an architecture that demonstrably supports such capacities, and hence are worth considering as a principled reference point for developing spatially-grounded AI [Cassenti et al., 2022, Zador et al., 2022, Li et al., 2024, Luo et al., 2025a]. Below, we show that analogous directions have begun to emerge across recent advances in AI and outline future prospects for such paradigms.

3.1 Language Models Are Not Spatially Competent

Multimodal language models (MLLMs) [Li et al., 2023, Liu et al., 2024, Gemini, 2023] acquire rich visual priors by aligning linguistic and visual data, enabling them to generate coherent spatial descriptions without direct training on visuospatial tasks [Ashutosh et al., 2025, Sharma et al., 2024, Deng, 2025]. These models demonstrate impressive performance in perception and high-level reasoning [Fu et al., 2023, Yang et al., 2025], yet they continue to struggle with tasks that require model-based spatial reasoning, including mental rotation, perspective-taking, and mechanical reasoning [Gao et al., 2024, Sun et al., 2024, Zhang et al., 2024, Luo et al., 2024, Li et al., 2024, 2025, Sun et al., 2025, Wang et al., 2025a, Cai et al., 2025], indicating a lack of structured spatial representations and dynamic transformation mechanisms essential for genuine spatial understanding. Although it has been proposed that, as models scale across tasks and modalities, their latent spaces converge toward shared statistical abstractions of the external world [Huh et al., 2024], this convergence cannot occur meaningfully in the spatial domain without vehicles capable of encoding perceptually rich representations.

3.2 Implicit Models for Embodied Control

Building physics engines as explicit spatial world models has been a holy grail in robotics. MuJoCo enabled precise physics and DRL/LfD breakthroughs in manipulation with encouraging sim-to-real transfer; GPU-based physics engines like Isaac Gym and Genesis scale this via massive parallelism [Todorov et al., 2012, Kumar and Todorov, 2015, Rajeswaran et al., 2017, Gupta et al., 2019, Zhu et al., 2019, 2020, Makoviychuk et al., 2021, Zhou et al., 2024b]. Yet excelling in physics simulation rarely yields structured real-world understanding; policies often lack sufficient models for counterfactuals, and concept grounding, leading to brittleness in open-ended, visually complex settings. In addition, physics-engine–only model-predictive control is brittle: model–reality gaps, partial observability, and heavy contact-dynamics compute force short-horizon, over-tuned policies that transfer poorly [Zhang et al., 2025, Pezzato et al., 2025].

Recent vision foundation models (VFMs) has led to remarkable features such as extracting structured features from high-resolution images, improving reasoning and action [Siméoni et al., 2025]. Their scaling-driven, modality-agnostic latents align with language and proprioception, enabling multimodal fusion, generative imagination, and possibly spatiotemporal structure [Luo et al., 2025b, Huh et al., 2024, Chern et al., 2025, Raugel et al., 2025]. These representational benefits translate directly into improved control. Integrating vision foundation models as perceptual modules within robotic

learning pipelines has yielded dramatic performance gains—not only in basic manipulation tasks but also in long-horizon planning and generalization [Majumdar et al., 2023]. A striking example is the VIP framework, which uses vision-based embeddings to encode task rewards for DRL, effectively reshaping policy optimization through perceptual guidance [Ma et al., 2022]. Such approaches enable more efficient policy learning and produce agents whose behaviour transfers exceptionally well from simulation to the real world [Shah and Kumar, 2021, Ma et al., 2023, Chen et al., 2023, Nair et al., 2022]. This body of work has led to the emergence of a broader paradigm often referred to as visual pretraining for manipulation, in which vision encoders pretrained on internet-scale image and video data are reused to accelerate downstream control. These pretrained representations serve as implicit world models—embedding structured spatial priors that enhance sample efficiency, generalization, and robustness in policy learning [Du and Mordatch, 2019, Du et al., 2023, Zhou et al., 2024a, 2025].

Together, in the context of embodied AI, it is evident that both explicit and implicit world models have proven indispensable. Simulation engines such as MuJoCo, Isaac Gym, and Genesis provide detailed physical environments for training agents or model-predictive control [Todorov et al., 2012, Makoviychuk et al., 2021, Zhou et al., 2024b], but these remain insufficient on their own for achieving robust, real-world performance. By contrast, implicit world models, exemplified by frameworks such as VIP and R3M, leverage pretrained visual representations for control generalizations [Ma et al., 2022, Nair et al., 2022].

3.3 Video Models for Action Imagination

A central question is whether models can, much like humans during navigation, perceive, model, and render spatially relevant tasks without language. On the perceptual side, vision foundation models (e.g., VGG, DINO, CLIP) already exhibit strong transfer across diverse downstream tasks [Mei et al., 2025, Siméoni et al., 2025]. Recent work such as VGGT grounds large vision backbones in a 3D reconstruction objective, yielding visual–geometry priors that capture fine structural detail and boost performance on spatially demanding tasks [Wang et al., 2025b]. Inspired by chain-of-thought in language models, we hypothesize that *chain-of-frames* rollouts (i.e., video generation) can enable step-by-step spatiotemporal reasoning. Early evidence comes from Veo 3, which generates next-scene predictions conditioned on video inputs to tackle visual reasoning tasks such as Sudoku, mazes, and navigation [Wiedemer et al., 2025]. We can leverage pretrained video diffusion models to synthesize video-based plans for actions [Du et al., 2023, Yang et al., 2024]. We see an exciting opportunity in new approaches such as Genex and Genie-3 [Lu et al., 2024a,b, Parker-Holder and Fruchter] and see huge potential in them to mature into brain-like implicit world models, surpassing 2D pixel representations by internalizing scene geometry, dynamics, and affordances. Humans "run movies in the mind" before acting, using vivid imagery to simulate candidate futures. Analogously, we see that video models can generate action rollouts that score and refine action sequences, supplying policies with strong visuomotor plans. We hypothesize that this imagination-as-video strategy could let machines reason about space as we do.

4 Conclusion

This paper revisited the long-standing debate between simulation and rendering by drawing on insights from both cognitive neuroscience and embodied AI. We argued that human spatial reasoning is unlikely to rely on purely schematic or amodal simulations; instead, it depends on perceptually rich content that stabilizes internal models of the world. Recent developments in embodied AI and robotics reinforce this view: models that leverage large-scale vision encoders to provide high-fidelity embeddings, serving as implicit world models, could outperform those trained solely through physics-based simulation. These findings mirror how humans integrate detailed perceptual information into mental simulations, drawing on shared representational structures rather than separate systems. Closing the gap between human and artificial spatial reasoning may therefore depend on leveraging rich perceptual grounding in structured spatial representations.

References

Anna Anzulewicz, Justyna Hobot, Marta Siedlecka, and Michał Wierzchoń. Bringing action into the picture. how action influences visual awareness. *Attention, Perception, & Psychophysics*, 81(7): 2171–2176, 2019.

Kumar Ashutosh, Yossi Gandelsman, Xinlei Chen, Ishan Misra, and Rohit Girdhar. Llms can see and hear without any training, 2025. URL https://arxiv.org/abs/2501.18096.

Halely Balaban and Tomer D Ullman. Physics versus graphics as an organizing dichotomy in cognition. *Trends in Cognitive Sciences*, 2025.

Joachim Bellet, Marion Gay, Abhilash Dwarakanath, Bechir Jarraya, Timo Van Kerkoerle, Stanislas Dehaene, and Theofanis I Panagiotaropoulos. Decoding rapidly presented visual stimuli from prefrontal ensembles without report nor post-perceptual processing. *Neuroscience of consciousness*, 2022(1):niac005, 2022.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M Fleming, Chris Frith, Xu Ji, et al. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

Zhongang Cai, Yubo Wang, Qingping Sun, Ruisi Wang, Chenyang Gu, Wanqi Yin, Zhiqian Lin, Zhitao Yang, Chen Wei, Xuanke Shi, et al. Has gpt-5 achieved spatial intelligence? an empirical study. *arXiv preprint arXiv:2508.13142*, 2025.

Daniel N Cassenti, Vladislav D Veksler, and Frank E Ritter. Editor's review and introduction: Cognition-inspired artificial intelligence, 2022.

David J Chalmers. *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks, 1997.

Annie W-Y Chan. Functional organization and visual representations of human ventral lateral prefrontal cortex. *Frontiers in psychology*, 4:371, 2013.

Devendra Singh Chaplot, Deepak Pathak, and Jitendra Malik. Differentiable spatial planning using transformers. In *International conference on machine learning*, pages 1484–1495. PMLR, 2021.

Tao Chen, Megha Tippur, Siyang Wu, Vikash Kumar, Edward Adelson, and Pulkit Agrawal. Visual dexterity: In-hand reorientation of novel and complex object shapes. *Science Robotics*, 8(84): eadc9244, 2023.

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 2025.

Edward HF de Haan and Alan Cowey. On the usefulness of 'what'and 'where'pathways in vision. *Trends in cognitive sciences*, 15(10):460–466, 2011.

Hokin Deng. Reinforcement learning versus natural language programs: Where is flexible planning and problem solving in natural intelligence coming from? *OSF*, 2025.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

Stephen M Fleming. Awareness as inference in a higher-order state space. *Neuroscience of consciousness*, 2020(1):niz020, 2020.

Stephen M Fleming and Matthias Michel. Sensory horizons and the functions of conscious vision. *Behavioral and Brain Sciences*, pages 1–53, 2025.

Stephen M Fleming and Nicholas Shea. Quality space computations for consciousness. *Trends in Cognitive Sciences*, 28(10):896–906, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv: 2306.13394*, 2023.

Qingying Gao, Yijiang Li, Haiyun Lyu, Haoran Sun, Dezhi Luo, and Hokin Deng. Vision language models see what you want but not what you see. *arXiv preprint arXiv:2410.00324*, 2024.

Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, et al. Do vision-language models have internal world models? towards an atomic evaluation. *arXiv preprint arXiv:2506.21876*, 2025.

Gemini. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv: 2312.11805*, 2023.

Samuel J Gershman. The generative adversarial brain. *Frontiers in Artificial Intelligence*, 2:18, 2019.

Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019.

Mary Hegarty. Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, 8(6): 280–285, 2004.

Mary Hegarty. The cognitive science of visual-spatial displays: Implications for design. *Topics in cognitive science*, 3(3):446–474, 2011.

Christopher Hilton, Leonie Raddatz, and Klaus Gramann. A general spatial transformation process? assessing the neurophysiological evidence on the similarity of mental rotation and folding. *Neuroimage: Reports*, 2(2):100092, 2022.

Geoffrey Hinton. Imagery without arrays. *Behavioral and Brain Sciences*, 2(4):555–556, 1979.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL `https://arxiv.org/abs/2405.07987`.

Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press, 1983.

Lachlan Kay, Rebecca Keogh, and Joel Pearson. Slower but more accurate mental rotation performance in aphantasia linked to differences in cognitive strategies. *Consciousness and cognition*, 121:103694, 2024.

Peter Khooshabeh, Mary Hegarty, and Thomas F Shipley. Individual differences in mental rotation. *Experimental psychology*, 2013.

Stephen M Kosslyn, Steven Pinker, George E Smith, and Steven P Shwartz. On the demystification of mental imagery. *Behavioral and Brain Sciences*, 2(4):535–548, 1979.

Benjamin Kuipers. Modeling spatial knowledge. *Cognitive science*, 2(2):129–153, 1978.

Vikash Kumar and Emanuel Todorov. Mujoco haptix: A virtual reality system for hand manipulation. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 657–663. IEEE, 2015.

Julian Kutsche, Calvin Howard, William Drew, Matthias Michel, Alexander L Cohen, Michael D Fox, and Isaiah Kletenik. Visual mental imagery and aphantasia lesions map onto a convergent brain network. *medRxiv*, pages 2025–05, 2025.

Hakwan Lau. Consciousness, metacognition, & perceptual reality monitoring. *PsyArXiv*, 2019.

Hakwan Lau. *In consciousness we trust: The cognitive neuroscience of subjective experience*. Oxford University Press, 2022.

Hakwan C Lau and Richard E Passingham. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences*, 103(49): 18763–18768, 2006.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *CONFERENCE*, 2023.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, et al. Core knowledge deficits in multi-modal language models. *arXiv preprint arXiv:2410.10855*, 2024.

Yijiang Li, Bingyang Wang, Tianwei Zhao, Qingying Gao, Hokin Deng, and Dezhi Luo. Evaluating multi-modal language models through concept hacking. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Taiming Lu, Tianmin Shu, Junfei Xiao, Luoxin Ye, Jiahao Wang, Cheng Peng, Chen Wei, Daniel Khashabi, Rama Chellappa, Alan Yuille, et al. Genex: Generating an explorable world. *arXiv preprint arXiv:2412.09624*, 2024a.

Taiming Lu, Tianmin Shu, Alan Yuille, Daniel Khashabi, and Jieneng Chen. Generative world explorer. *arXiv preprint arXiv:2411.11844*, 2024b.

Dezhi Luo, Haiyun Lyu, Qingying Gao, Haoran Sun, Yijiang Li, and Hokin Deng. Vision language models know law of conservation without understanding more-or-less. *arXiv preprint arXiv:2410.00332*, 2024.

Dezhi Luo, Yijiang Li, and Hokin Deng. The philosophical foundations of growing ai like a child. *arXiv preprint arXiv:2502.10742*, 2025a.

Grace Luo, Trevor Darrell, and Amir Bar. Vision-language models create cross-modal task representations, 2025b. URL `https://arxiv.org/abs/2410.22330`.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. In *International Conference on Machine Learning*, pages 23301–23320. PMLR, 2023.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36: 655–677, 2023.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

Zhiting Mei, Ola Shorinwa, and Anirudha Majumdar. Geometry meets vision: Revisiting pretrained semantics in distilled fields, 2025. URL `https://arxiv.org/abs/2510.03104`.

Matthias Michel, Jorge Morales, Ned Block, and Hakwan Lau. Aphantasia as imagery blindsight. *Trends in Cognitive Sciences*, 2025.

A David Milner. Is visual processing in the dorsal stream accessible to consciousness? *Proceedings of the Royal Society B: Biological Sciences*, 279(1737):2289–2298, 2012.

A David Milner. How do the two visual streams interact with each other? *Experimental brain research*, 235(5):1297–1308, 2017.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

Bence Nanay. Unconscious mental imagery. *Philosophical Transactions of the Royal Society B*, 376 (1817):20190689, 2021.

Theofanis I Panagiotaropoulos. An integrative view of the role of prefrontal cortex in consciousness. *Neuron*, 112(10):1626–1641, 2024.

Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. URL `https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/`. Blog post.

Navindra Persaud, Matthew Davidson, Brian Maniscalco, Dean Mobbs, Richard E Passingham, Alan Cowey, and Hakwan Lau. Awareness-related activity in prefrontal and parietal cortices in blindsight reflects more than superior visual performance. *Neuroimage*, 58(2):605–611, 2011.

Corrado Pezzato, Chadi Salmi, Elia Trevisan, Max Spahn, Javier Alonso-Mora, and Carlos Hernández Corbato. Sampling-based model predictive control leveraging parallelizable physics simulations. *IEEE Robotics and Automation Letters*, 2025.

Ian Phillips. Aphantasia reimagined. *Noûs*, 2025.

Zenon W Pylyshyn. The rate of "mental rotation" of images: A test of a holistic analogue hypothesis. *Memory & cognition*, 7(1):19–28, 1979.

Gregor Rainer and Earl K Miller. Effects of visual experience on the representation of objects in the prefrontal cortex. *Neuron*, 27(1):179–189, 2000.

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

Joséphine Raugel, Marc Szafraniec, Huy V. Vo, Camille Couprie, Patrick Labatut, Piotr Bojanowski, Valentin Wyart, and Jean-Rémi King. Disentangling the factors of convergence between brains and computer vision models. *Manuscript*, page 1–15, 2025. preprint PDF, 15 pages. URL: `https://scontent-lga3-2.xx.fbcdn.net/v/t39.2365-6/533250329_1837870417159359_3262810533067032733_n.pdf?âĂę`.

Geraint Rees. Neural correlates of the contents of visual awareness in humans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):877–886, 2007.

David Rosenthal. How to think about mental qualities. *Philosophical Issues*, 20:368–393, 2010.

Thomas Schenk and Robert D McIntosh. Do we have independent visual streams for perception and action? *Cognitive Neuroscience*, 1(1):52–62, 2010.

Christian O Scholz, Merlin Monzel, and Jianghao Liu. Absence of shared representation in the visual cortex challenges unconscious imagery in aphantasia. *Current Biology*, 35(13):R645–R646, 2025.

Jordan A Searle and Jeff P Hamm. Mental rotation: An examination of assumptions. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(6):e1443, 2017.

Rutav Shah and Vikash Kumar. Rrl: Resnet as representation for reinforcement learning. *arXiv preprint arXiv:2107.03380*, 2021.

Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, and Antonio Torralba. A vision check-up for language models, 2024. URL `https://arxiv.org/abs/2401.01862`.

Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171 (3972):701–703, 1971.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL `https://arxiv.org/abs/2508.10104`.

Alex F Spies, William Edwards, Michael I Ivanitskiy, Adrians Skapars, Tilman Räuker, Katsumi Inoue, Alessandra Russo, and Murray Shanahan. Transformers use causal world models in maze-solving tasks. *arXiv preprint arXiv:2412.11867*, 2024.

Haoran Sun, Qingying Gao, Haiyun Lyu, Dezhi Luo, Yijiang Li, and Hokin Deng. Probing mechanical reasoning in large vision language models. *arXiv preprint arXiv:2410.00318*, 2024.

Haoran Sun, Suyang Yu, Yijiang Li, Qingying Gao, Haiyun Lyu, Hokin Deng, and Dezhi Luo. Probing perceptual constancy in large vision language models. *arXiv preprint arXiv:2502.10273*, 2025.

Sara M Szczepanski and Robert T Knight. Insights into human behavior from lesions to the prefrontal cortex. *Neuron*, 83(5):1002–1018, 2014.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

Guy Vingerhoets, Floris P de Lange, Pieter Vandemaele, Karel Deblaere, and Erik Achten. Motor imagery in mental rotation: an fmri study. *Neuroimage*, 17(3):1623–1633, 2002.

Bingyang Wang, Yijiang Li, Qingyang Zhou, Hui Yi Leong, Tianwei Zhao, Letian Ye, Hokin Deng, Dezhi Luo, and Nuno Vasconcelos. Do vision language models infer human intention without visual perspective-taking? towards a scalable" one-image-probe-all" dataset. In *ICML 2025 Workshop on Assessing World Models*, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer, 2025b. URL `https://arxiv.org/abs/2503.11651`.

Mark Wexler, Stephen M Kosslyn, and Alain Berthoz. Motor processes in mental rotation. *Cognition*, 68(1):77–94, 1998.

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, 2025. URL `https://arxiv.org/abs/2509.20328`.

Michael Wooldridge and Nicholas R Jennings. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995.

Wayne Wu. Against division: Consciousness, information and the visual streams. *Mind & Language*, 29(4):383–406, 2014.

Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024.

Shuo Yang, Siwen Luo, Soyeon Caren Han, and Eduard Hovy. Magic-vqa: Multimodal and grounded inference with commonsense knowledge for visual question answering. *arXiv preprint arXiv:2503.18491*, 2025.

Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Toward next-generation artificial intelligence: Catalyzing the neuroai revolution. *arXiv preprint arXiv:2210.08340*, 2022.

John Z Zhang, Taylor A Howell, Zeji Yi, Chaoyi Pan, Guanya Shi, Guannan Qu, Tom Erez, Yuval Tassa, and Zachary Manchester. Whole-body model-predictive control of legged robots with mujoco. *arXiv preprint arXiv:2503.04613*, 2025.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. *arXiv preprint arXiv:2410.17385*, 2024.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024a.

Siyuan Zhou, Yilun Du, Yuncong Yang, Lei Han, Peihao Chen, Dit-Yan Yeung, and Chuang Gan. Learning 3d persistent embodied world models. *arXiv preprint arXiv:2505.05495*, 2025.

Xian Zhou, Yiling Qiao, Zhenjia Xu, Tsun-Hsuan Wang, Zhehuan Chen, Juntian Zheng, Ziyan Xiong, Yian Wang, Mingrui Zhang, Pingchuan Ma, Yufei Wang, and Zhiyang Dou. Genesis: A universal and generative physics engine for robotics and beyond. *URL https://github.com/Genesis-Embodied-AI/Genesis*, 2024b.

Henry Zhu, Abhishek Gupta, Aravind Rajeswaran, Sergey Levine, and Vikash Kumar. Dexterous manipulation with deep reinforcement learning: Efficient, general, and low-cost. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3651–3657. IEEE, 2019.

Henry Zhu, Justin Yu, Abhishek Gupta, Dhruv Shah, Kristian Hartikainen, Avi Singh, Vikash Kumar, and Sergey Levine. The ingredients of real-world robotic reinforcement learning. *arXiv preprint arXiv:2004.12570*, 2020.