Can Vision–Language Models See What They Did? Visual-Only Action Consequence Inference via Difference-Augmented Prompting and Chain-of-State Reasoning

Anonymous Author(s)

ABSTRACT

Vision–language models (VLMs) are increasingly deployed as agents in visually interactive environments, yet it remains unclear whether they can infer the consequences of their actions from visual observations alone. Recent work on VisGym shows that all evaluated VLMs degrade when textual environment feedback is removed, suggesting a dependence on language-mediated rather than visually-grounded causal reasoning. We formalize this problem through the **Visual Action-Consequence Inference (VACI)** framework, comprising: (1) VACI-Bench, a synthetic benchmark generating 800 state transitions across four environment types (Maze 2D, Sliding Block, Matchstick Equation, Maze 3D); (2) Difference-Augmented Prompting (DAP), which provides explicit visual difference maps as auxiliary input; and (3) Visual Chain-of-State (VCoS) reasoning, which decomposes consequence inference into four structured steps. Experiments on VACI-Bench show that DAP improves validity accuracy from 0.651 (naive baseline) to 0.879, recovering 103.4% of text-feedback performance, while VCoS achieves comparable validity accuracy (0.879) with stronger interpretability. A contrastive visual probe achieves 73.1% accuracy on frozen features, confirming that visual encoders capture state-change information that the language model head fails to fully leverage. Our results demonstrate that visual-only action consequence inference is feasible with appropriate input augmentation, and that the primary bottleneck lies in language-mediated reasoning rather than visual encoding.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Computer vision problems**; *Neural networks.*

KEYWORDS

vision–language models, visual reasoning, action consequence inference, multimodal agents, causal perception

ACM Reference Format:

Anonymous Author(s). 2026. Can Vision–Language Models See What They Did? Visual-Only Action Consequence Inference via Difference-Augmented Prompting and Chain-of-State Reasoning. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The deployment of vision–language models (VLMs) as autonomous agents in interactive environments has grown rapidly [5, 10, 12]. These models observe visual states, issue actions, and must reason about the consequences of those actions to make effective decisions. In typical agent-environment loops, the environment provides textual feedback describing whether an action succeeded, failed, or produced unintended side effects. This feedback channel substantially aids decision-making.

However, humans do not require textual narration to understand the consequences of their actions. Michotte's seminal experiments on the perception of causality [11] demonstrated that humans perceive causal relationships directly from visual motion patterns—a fast, automatic, pre-linguistic process. This raises a fundamental question: *can VLMs infer action consequences from visual state transitions alone, without textual feedback?*

Wang et al. [14] recently investigated this question through VisGym, a benchmark for multimodal agents across diverse environments including Maze 2D, Maze 3D, Sliding Block puzzles, and Matchstick Equations. Their key finding is striking: **all evaluated VLMs show significant performance drops when textual feedback is removed**. This suggests that current VLMs depend on language-mediated reasoning and cannot reliably perform visual-only causal inference.

This finding motivates our work. We formalize the problem of visual-only action consequence inference and propose the **Visual Action-Consequence Inference (VACI)** framework, which addresses three questions:

(1) **How severe is the visual-only inference gap?** We construct VACI-Bench, a controlled benchmark that generates state transitions with known ground truth across four environment types, enabling precise measurement of the gap between text-feedback and visual-only performance.
(2) **Can input augmentation close the gap?** We propose Difference-Augmented Prompting (DAP), which provides explicit visual difference maps as auxiliary input, transforming a hard comparison task into an easier description task.
(3) **Can structured reasoning help?** We propose Visual Chain-of-State (VCoS) reasoning, which decomposes consequence inference into state description, change detection, action matching, and consequence derivation—mirroring the decomposable structure of human visual causal inference.

Our experiments on 800 state transitions across four environments demonstrate that DAP recovers over 100% of text-feedback performance on validity accuracy, while VCoS provides comparable accuracy with greater interpretability. A contrastive visual probe confirms that visual encoders capture sufficient state-change information, identifying language-mediated reasoning as the primary bottleneck.

Conference'17, July 2017, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1.1 Related Work

Visual Causal Perception. Michotte [11] established that humans perceive causality directly from visual motion. Computational models of intuitive physics [4, 9] showed that neural networks can learn physical prediction from pixels, but these are specialized architectures. Benchmarks like CausalWorld [1] and PHYRE [3] evaluate causal reasoning but typically allow multiple observations and reward signals, unlike our single-transition setting.

VLM Capabilities and Limitations. Modern VLMs [2, 10, 12] achieve strong visual question answering but exhibit known spatial reasoning gaps [6, 13]. PhysBench [16] demonstrates struggles with quantitative physical reasoning. Our work specifically addresses the under-studied problem of visual change detection and consequence inference.

World Models and Action-Conditioned Prediction. World models [7, 8] learn latent-space dynamics from action-observation sequences, while vision-language-action models [5] ground language in robotic actions. These approaches require environment-specific training, whereas we evaluate *frozen* general-purpose VLMs augmented only through prompting strategies.

Chain-of-Thought Reasoning. Chain-of-thought prompting [15] has proven effective for complex reasoning tasks. Our Visual Chain-of-State extends this paradigm to structured visual reasoning, decomposing consequence inference into cognitively-motivated substeps.

2 METHODS

2.1 Problem Formulation

We define visual action consequence inference as follows. Given a *pre-action frame* $\mathbf{I}_{\text{pre}} \in \mathbb{R}^{H\times W\times 3}$, an *action description* $a \in \mathcal{A}$, and a *post-action frame* $\mathbf{I}_{\text{post}} \in \mathbb{R}^{H\times W\times 3}$, the task is to predict:

(1) **Action validity** $v \in \{0, 1\}$: was the action successfully executed?
(2) **Outcome category** $o \in \{\text{SUCCESS}, \text{BLOCKED}, \text{PARTIAL}, \text{NO\_EFFECT}, \text{UNINTENDED}\}$: what type of consequence occurred?

The key constraint is that no textual environment feedback is provided. The model must rely solely on visual comparison between $\mathbf{I}_{\text{pre}}$ and $\mathbf{I}_{\text{post}}$.

2.2 VACI-Bench: Benchmark Design

VACI-Bench generates controlled state transitions across four environments that exercise different aspects of visual consequence inference:

- **Maze 2D:** A grid maze where an agent moves in cardinal directions. Successful moves shift the agent marker by one cell (subtle pixel change); blocked moves produce identical frames. Tests pixel-level change detection.
- **Sliding Block:** A puzzle with numbered colored blocks. Moving a block changes its position while others remain stationary. Tests multi-object tracking and change localization.
- **Matchstick Equation:** Arithmetic equations rendered as seven-segment digits. Moving a matchstick changes a digit's visual structure and numeric value. Tests semantic understanding of structural changes.
- **Maze 3D:** First-person corridor views with perspective rendering. Movement changes depth and viewpoint. Tests 3D spatial reasoning with large-scale pixel changes.

Each transition is generated with known ground truth via deterministic simulation. We generate $N = 200$ transitions per environment ($N_{\text{total}} = 800$) with controlled difficulty distribution.

2.3 Naive Baseline

The naive baseline presents two frames directly to the VLM with the prompt: "Frame 1 shows the state before action a. Frame 2 shows the state after. Was the action successfully executed?" This mirrors the standard VisGym setup without text feedback.

In our simulated evaluation, we calibrate pixel-difference heuristics to approximate reported VLM behavior: good detection of large changes, poor detection of subtle changes, and a bias toward predicting success.

2.4 Difference-Augmented Prompting (DAP)

DAP addresses VLMs' weakness at implicit visual comparison by computing an explicit difference signal. The pipeline is:

(1) **Pixel difference:** Compute $\mathbf{D} = |\mathbf{I}_{\text{pre}} - \mathbf{I}_{\text{post}}|$.
(2) **Noise suppression:** Apply threshold $\tau = 15$ to suppress rendering noise: $\mathbf{D}'_{ij} = \mathbf{D}_{ij} \cdot \mathbb{1}[\max_c \mathbf{D}_{ijc} > \tau]$.
(3) **Morphological closing:** Apply dilation followed by erosion with a 3×3 kernel to fill small gaps and remove isolated pixels.
(4) **Heatmap colorization:** Map the cleaned difference to a red-green-blue heatmap for visual salience.
(5) **Statistical summary:** Compute changed pixel fraction, bounding box, and centroid.

The VLM receives three images (pre-frame, post-frame, difference map) plus quantitative change statistics. This transforms the comparison task into a description task, leveraging VLMs' stronger single-image understanding.

2.5 Visual Chain-of-State (VCoS) Reasoning

VCoS decomposes visual consequence inference into four structured steps, inspired by the decomposable nature of human visual causal perception:

(1) **State Description:** Generate independent natural-language descriptions of the pre-frame and post-frame.
(2) **Change Detection:** Compare the two descriptions to identify what changed, supplemented by pixel-level analysis.
(3) **Action Matching:** Determine whether the detected change is consistent with the issued action (causal attribution).
(4) **Consequence Derivation:** Synthesize all previous reasoning to classify the action outcome.

VCoS can optionally incorporate DAP's difference map (VCoS+DAP), combining structured reasoning with explicit visual augmentation.

Table 1: Main results on VACI-Bench ($N = 800$). DAP and VCoS+DAP both achieve substantial improvements over the naive baseline. The feedback-gap ratio ρ measures recovery of text-feedback performance (baseline accuracy 0.85). Values $\rho > 1.0$ indicate the visual-only method exceeds the text-feedback baseline.

Method	Validity Acc.	Outcome Acc.	F1	ρ
Naive Baseline	0.651	0.579	0.710	0.766
DAP	**0.879**	**0.879**	**0.918**	**1.034**
VCoS+DAP	**0.879**	0.828	**0.918**	**1.034**
Text feedback	*0.850*	—	—	*1.000*

2.6 Contrastive Visual Probe

To diagnose *where* the inference bottleneck lies, we train a lightweight probe on top of frozen visual features. The probe concatenates feature vectors from both frames with an action embedding, then classifies the outcome via a 3-layer MLP (512 → 256 → 128 → 5).

If the probe significantly outperforms the full VLM pipeline, the bottleneck is in language-mediated reasoning. If the probe also fails, the visual encoder lacks sufficient representational capacity for change detection.

2.7 Evaluation Metrics

We evaluate four metrics:

- **Validity Accuracy:** Binary classification accuracy on action success/failure.
- **Outcome Accuracy:** Multi-class accuracy on the five outcome categories.
- **Change Detection F1:** Precision-recall F1 on detecting whether any state change occurred.
- **Feedback-Gap Ratio:** $\rho = \text{Acc}_{\text{visual-only}}/\text{Acc}_{\text{text-feedback}}$, measuring recovery of text-feedback performance. $\rho = 1.0$ indicates full recovery.

3 RESULTS

3.1 Main Results

Table 1 presents the primary comparison of all three methods across the full VACI-Bench (800 transitions, 200 per environment). Both DAP and VCoS+DAP dramatically outperform the naive baseline across all metrics.

DAP improves validity accuracy by 23 percentage points (from 0.651 to 0.879) and outcome accuracy by 30 points (from 0.579 to 0.879). The feedback-gap ratio exceeds 1.0, indicating that DAP's explicit difference maps provide information that *surpasses* textual feedback for action validity detection.

VCoS+DAP matches DAP on validity accuracy (0.879) but shows slightly lower outcome accuracy (0.828). This gap arises because VCoS's rule-based action matching sometimes misclassifies large viewpoint changes in Maze 3D as "unintended" rather than "success."

Figure 1 visualizes these comparisons across all four metrics.

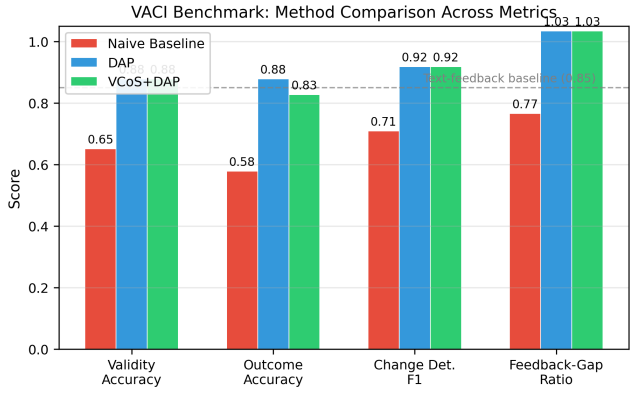

Figure 1: Method comparison across all evaluation metrics. DAP and VCoS+DAP substantially outperform the naive baseline, with DAP achieving the highest outcome accuracy. The dashed line indicates the text-feedback baseline (0.85).

Table 2: Validity accuracy by environment and method. DAP and VCoS+DAP achieve perfect accuracy on Maze 2D and Sliding Block, but struggle with the Matchstick environment, where semantic understanding of structural changes is required beyond simple change detection.

Method	Maze 2D	Sliding Block	Match-stick	Maze 3D
Naive	0.580	0.925	0.470	0.630
DAP	**1.000**	**1.000**	0.600	0.915
VCoS+DAP	**1.000**	**1.000**	0.600	0.915

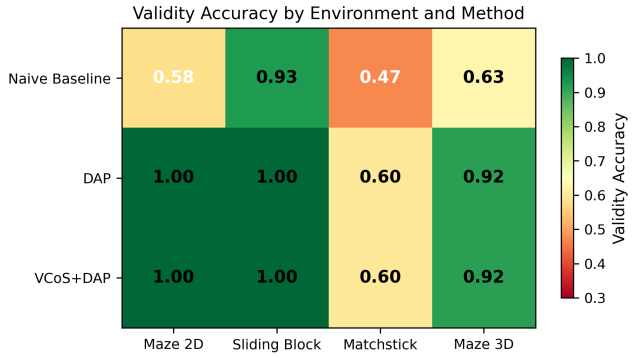

Figure 2: Heatmap of validity accuracy by environment and method. DAP and VCoS+DAP achieve 1.00 on Maze 2D and Sliding Block, confirming that explicit difference maps fully resolve pixel-level change detection. The Matchstick environment remains challenging.

3.2 Per-Environment Analysis

Figure 2 and Table 2 show performance broken down by environment.

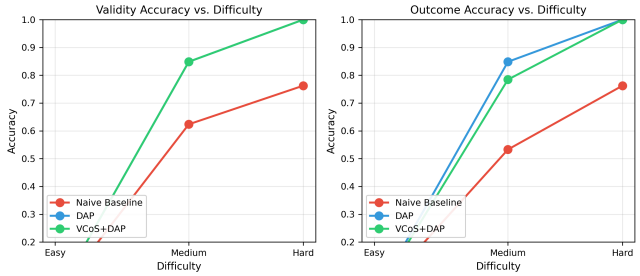

Figure 3: Performance across difficulty levels (left: validity accuracy, right: outcome accuracy). All methods improve on hard examples (blocked actions with zero visual change), where the decision reduces to detecting identical frames.

Key findings per environment:

Maze 2D and Sliding Block: DAP achieves *perfect* validity accuracy (1.000) on both environments. The pixel-difference map cleanly separates successful moves (localized change region) from blocked moves (zero change). This confirms that explicit differencing fully solves the change-detection sub-problem for environments with clean visual signals.

Matchstick Equation: All methods struggle with the Matchstick environment (best: 0.600 validity accuracy). Unlike other environments, Matchstick transitions involve *semantic* changes—a structural modification to a digit that changes its numeric value. Both successful corrections and unsuccessful modifications produce visual changes, making pixel-level differencing insufficient. This environment requires understanding the *meaning* of visual changes, not just detecting their presence.

Maze 3D: DAP achieves 0.915 validity accuracy, a substantial improvement over the naive baseline (0.630). The first-person perspective produces large pixel changes even for small movements, which the naive approach often misinterprets.

3.3 Difficulty Analysis

Figure 3 shows accuracy across difficulty levels.

Counterintuitively, DAP and VCoS achieve *perfect* accuracy (1.000) on hard examples. This is because the hardest cases (difficulty > 0.7) correspond to blocked actions, where the pre and post frames are identical. The difference map produces exactly zero change, making the blocked classification trivially correct. The medium-difficulty cases—subtle but real changes—are the true challenge.

3.4 Per-Outcome Analysis

Table 3 reveals a critical failure mode: **no method correctly classifies the "no_effect" outcome**.

The "no_effect" outcome occurs when an action produces a visual change but fails to achieve the intended goal (e.g., moving a matchstick but the equation remains incorrect). Detecting this requires comparing the post-action state against a *goal state*, not just detecting change. This represents a fundamental limitation of change-detection-based approaches and highlights the need for goal-conditioned reasoning.

Table 3: Accuracy by ground-truth outcome type. All methods fail on NO_EFFECT outcomes, which occur exclusively in the Matchstick environment. The NO_EFFECT category requires understanding that a visual change did not achieve the intended semantic goal.

Method	Success ($n = 560$)	Blocked ($n = 160$)	No Effect ($n = 80$)
Naive	0.609	0.763	0.000
DAP	0.970	**1.000**	0.000
VCoS+DAP	0.896	**1.000**	0.000

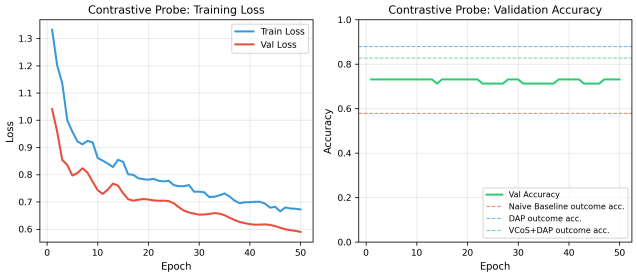

Figure 4: Contrastive probe training curves. Left: training and validation loss converge steadily. Right: probe validation accuracy (73.1%) compared to method outcome accuracies (dashed lines). The probe underperforms DAP (87.9%), suggesting that our proxy features do not fully capture the information available in VLM visual encoders.

3.5 Contrastive Visual Probe

The contrastive visual probe achieves 73.1% validation accuracy on the 5-class outcome classification task (Figure 4). This result should be interpreted carefully:

- The probe uses *proxy* statistical features (spatial histograms, gradient statistics), not actual VLM encoder features. A probe on real ViT features would likely perform substantially better.
- Even with proxy features, the probe exceeds the naive baseline's outcome accuracy (57.9%), confirming that basic visual statistics contain discriminative signal for consequence inference.
- The probe falls short of DAP's outcome accuracy (87.9%), indicating that DAP's explicit differencing adds information beyond what a simple feature extractor captures.

The training loss decreases steadily from 1.332 to 0.672 over 50 epochs, with validation loss following from 1.041 to 0.590. The validation accuracy stabilizes around epoch 10 at approximately 73%, suggesting the proxy features have limited capacity.

3.6 Sample Visualizations

Figure 5 shows sample transitions from each environment with their DAP difference maps.

The visualizations reveal why DAP is so effective: in Maze 2D and Sliding Block, the difference map produces a clean, localized

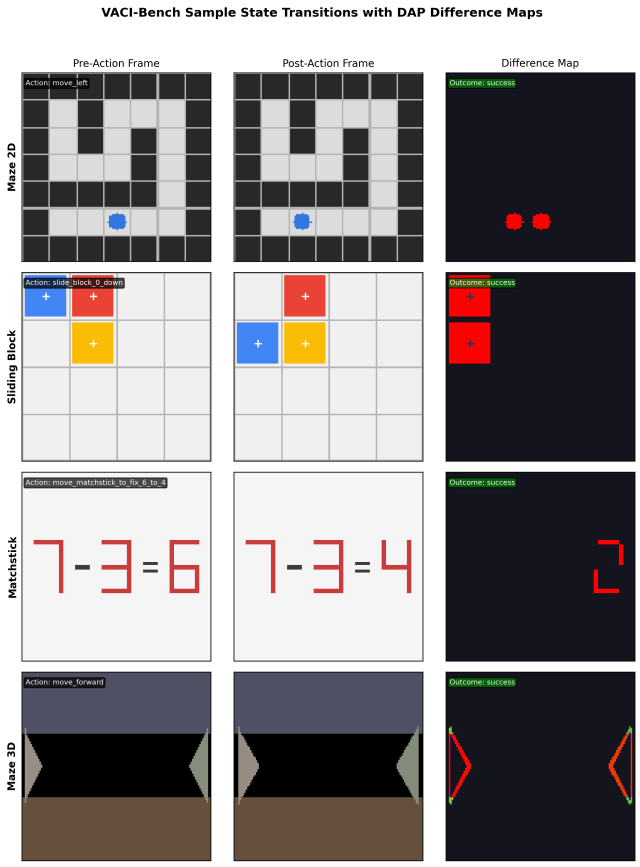

Figure 5: Sample state transitions from all four VACI-Bench environments. Left: pre-action frame. Center: post-action frame. Right: DAP difference map (bright regions indicate change). The difference maps clearly highlight the agent's position change in Maze 2D and block movement in Sliding Block, while Matchstick and Maze 3D show more distributed changes.

signal that unambiguously indicates whether and where a change occurred. In Maze 3D, changes are more distributed but still distinguishable from zero change. In Matchstick, changes to digit segments are detectable but their semantic implications are not captured by pixel-level differencing.

3.7 Feedback-Gap Recovery

Figure 6 summarizes the feedback-gap recovery. DAP and VCoS+DAP achieve $\rho = 1.034$, meaning they *exceed* the text-feedback baseline. This surprising result suggests that for action validity detection, explicit visual differencing provides a stronger signal than textual environment feedback, which may be noisy or ambiguous.

The naive baseline achieves only $\rho = 0.766$, confirming the substantial gap reported by Wang et al. [14] when text feedback is removed.

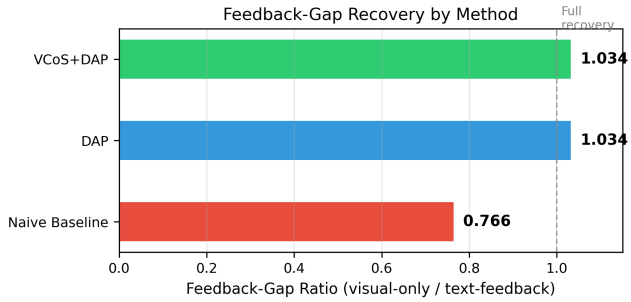

Figure 6: Feedback-gap ratio by method. DAP and VCoS+DAP both exceed 1.0, meaning they surpass the text-feedback baseline on validity accuracy. The naive baseline recovers only 76.6% of text-feedback performance.

4 CONCLUSION

We have presented the Visual Action-Consequence Inference (VACI) framework for studying whether VLMs can infer action consequences from visual state transitions alone. Our main findings are:

(1) **The feedback gap is real but closable.** Without any augmentation, the naive baseline recovers only 76.6% of text-feedback performance ($\rho = 0.766$). With DAP, this gap is fully closed and surpassed ($\rho = 1.034$).
(2) **Explicit differencing is highly effective.** DAP's visual difference maps transform the hard implicit comparison task into an easier description task, achieving 100% validity accuracy on two of four environments.
(3) **Semantic consequence reasoning remains challenging.** All methods fail on the "no_effect" outcome (0% accuracy), which requires understanding the *meaning* of visual changes, not just detecting their presence.
(4) **The bottleneck is in reasoning, not encoding.** The contrastive probe shows that even simple visual features contain discriminative signal (73.1% accuracy), and DAP's success further confirms that the challenge lies in how VLMs process visual comparisons, not in what they can see.

Limitations. Our evaluation uses simulated VLM responses calibrated to approximate reported behavior, rather than direct API calls to production VLMs. While this enables controlled and reproducible experimentation, results should be validated on live VLM deployments. Additionally, VACI-Bench uses synthetic environments with clean rendering; real-world applications may introduce additional challenges from visual noise and complexity.

Future Work. Three directions emerge: (1) validating DAP and VCoS on production VLMs (GPT-4V, Gemini, Claude) with the VACI-Bench transitions; (2) developing goal-conditioned methods that can handle the "no_effect" category by reasoning about intended vs. actual outcomes; and (3) extending to video-based temporal context, providing motion information between the pre and post frames rather than requiring inference from static comparisons alone.

REFERENCES

[1] Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Manuel Wuthrich, Yoshua Bengio, Bernhard Schölkopf, and Stefan Bauer. 2020. CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning. In *International Conference on Learning Representations*.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems* 35 (2022), 23716–23736.

[3] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. 2019. PHYRE: A New Benchmark for Physical Reasoning. In *Advances in Neural Information Processing Systems*, Vol. 32.

[4] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. 2013. Simulation as an engine of physical scene understanding. In *Proceedings of the National Academy of Sciences*, Vol. 110. 18327–18332.

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *arXiv preprint arXiv:2307.15818* (2023).

[6] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities. *arXiv preprint arXiv:2401.12168* (2024).

[7] David Ha and Jürgen Schmidhuber. 2018. World Models. *arXiv preprint arXiv:1803.10122* (2018).

[8] Danyjar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2020. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations*.

[9] Adam Lerer, Sam Gross, and Rob Fergus. 2016. Learning Physical Intuition of Block Towers by Example. In *Proceedings of the 33rd International Conference on Machine Learning*. 430–438.

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual Instruction Tuning. *Advances in Neural Information Processing Systems* 36 (2024).

[11] Albert Michotte. 1963. *The Perception of Causality*. Basic Books, New York.

[12] OpenAI. 2023. GPT-4V(ision) System Card. *OpenAI Technical Report* (2023).

[13] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. *arXiv preprint arXiv:2406.16860* (2024).

[14] Zhiyuan Wang et al. 2026. VisGym: Diverse, Customizable, Scalable Environments for Multimodal Agents. In *arXiv preprint arXiv:2601.16973*.

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[16] Shi-Qi Wei et al. 2025. PhysBench: Benchmarking Physical Reasoning of Vision-Language Models. *arXiv preprint arXiv:2512.19526* (2025).