

# VISUAL GENERATION UNLOCKS HUMAN-LIKE REASONING THROUGH MULTIMODAL WORLD MODELS

Jialong Wu<sup>1,2\*</sup>, Xiaoying Zhang<sup>2</sup>, Hongyi Yuan<sup>2</sup>, Xiangcheng Zhang<sup>1,2\*</sup>, Tianhao Huang<sup>1</sup>, Changjing He<sup>1</sup>, Chaoyi Deng<sup>1,2\*</sup>, Renrui Zhang<sup>2</sup>, Youbin Wu<sup>2</sup>, Mingsheng Long<sup>1</sup>

<sup>1</sup>Tsinghua University <sup>2</sup>ByteDance Seed

<https://thuml.github.io/Reasoning-Visual-World>

## ABSTRACT

Humans construct internal world models and reason by manipulating the concepts within these models. Recent advances in AI, particularly chain-of-thought (CoT) reasoning, approximate such human cognitive abilities, where world models are believed to be embedded within large language models. Expert-level performance in formal and abstract domains such as mathematics and programming has been achieved in current systems by relying predominantly on verbal reasoning. However, they still lag far behind humans in domains like physical and spatial intelligence, which require richer representations and prior knowledge. The emergence of unified multimodal models (UMMs) capable of both verbal and visual generation has therefore sparked interest in more human-like reasoning grounded in complementary multimodal pathways, though their benefits remain unclear. From a world-model perspective, this paper presents the first principled study of when and how visual generation benefits reasoning. Our key position is the *visual superiority hypothesis*: particularly for tasks grounded in the physical world, visual generation more naturally serves as world models, whereas purely verbal world models encounter bottlenecks arising from representational limitations or insufficient prior knowledge. Theoretically, we formalize internal world modeling as a core component of CoT reasoning and analyze distinctions among different forms of world models. Empirically, we identify tasks that necessitate interleaved visual-verbal CoT reasoning, constructing a new evaluation suite, *VisWorld-Eval*. Controlled experiments on a state-of-the-art UMM show that interleaved CoT significantly outperforms purely verbal CoT on tasks that favor visual world modeling, but offers no clear advantage otherwise. Together, this work clarifies the potential of multimodal world modeling for more powerful, human-like multimodal AI.

## 1 INTRODUCTION

Humans construct internal mental models of the external world that represent objects and concepts, along with their relationships, structures, and operational mechanisms ( Craik, 1967; Forrester, 1971). These models support reasoning and decision-making by enabling mental simulation, allowing individuals to anticipate the outcome of actions without actually taking them (Gentner, 2001). For example, if a glass of water is spilled on the table, people can rapidly mentally simulate the ensuing events: the water falling downward, spreading across the surface, and potentially dripping onto the floor. Such predictions lead them to quickly move valuable items away or reach for a towel. Beyond physical systems, mental models also extend to domains where relational structures can be simulated, such as mathematics and logic (Johnson-Laird, 1983; Lakoff & Núñez, 2000), making them fundamental to how humans understand and interact with all aspects of the world.

Cross-disciplinary researchers in philosophy, psychology, cognitive science, and related fields have a long history of developing computational models of human mental models (Norman, 2014). Among them, artificial intelligence (AI) shares a core ambition of building machines that reason like people. Although debates remain, recent breakthroughs, especially in large language models (LLMs) and chain-of-thought (CoT) reasoning, have made a substantial step towards approximating

\*Work done at ByteDance Seed.

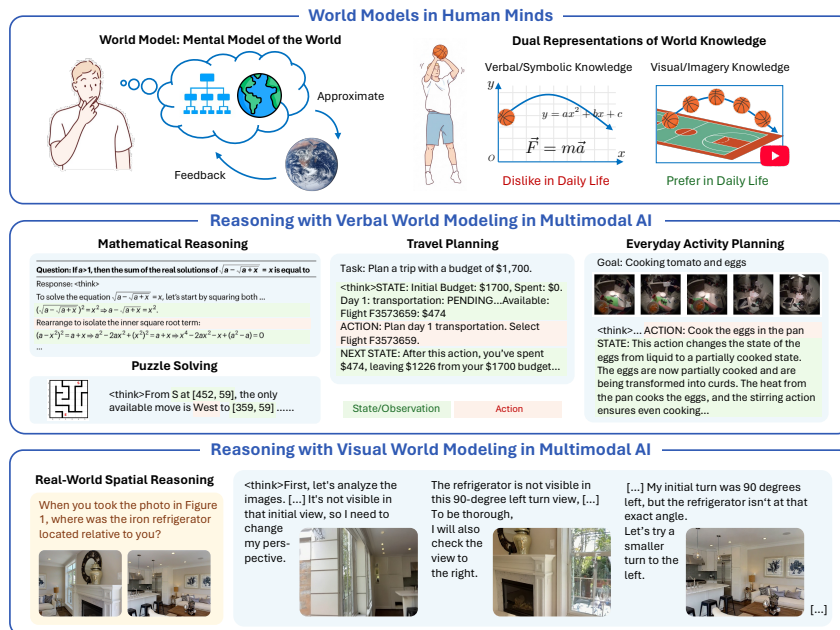


Figure 1: Overview of a world-model perspective on multimodal reasoning. Unified multimodal models enable visual generation for visual world modeling, supporting human-like reasoning through complementary verbal and visual channels. Examples of reasoning with verbal world modeling are adapted from Guo et al. (2025a); Du et al. (2025); Chen et al. (2025); Zhang et al. (2025).

human reasoning grounded in *mental models of the world*, often referred to as **world models** (Ha & Schmidhuber, 2018; LeCun, 2022) in the AI literature. During chain-of-thought reasoning, LLMs explore, reflect, and backtrack within the structured solution space, guided by world knowledge acquired through large-scale pre-training. These capabilities have already driven progress in diverse domains, including programming (Guo et al., 2025a), mathematics (Trinh et al., 2024; Guo et al., 2025a), scientific discovery (Swanson et al., 2025), clinical medicine (Tu et al., 2025), and robotics (Mon-Williams et al., 2025).

Such reasoning capabilities have also been extended to multimodal AI systems, particularly vision language models (VLMs) (Hurst et al., 2024; Bai et al., 2025b; Guo et al., 2025b; Yao et al., 2025). These systems typically incorporate visual inputs by aligning visual representations with the embedding space of LLMs, resulting in reasoning that remains primarily constrained to a linguistic space. In contrast, human mental models operate over multiple forms of mental representations. Dual-coding theory (Paivio, 1990) suggests that the mind processes information through two complementary codes: verbal and imagery (particularly visual) representations. These pathways can function independently but often collaborate to support reasoning. Indeed, visual imagery has been shown to have advantages over words in memory encoding and retrieval (Landy & Goldstone, 2007); and individuals with aphantasia, who lack the ability to visualize mental imagery, exhibit worse performance on tasks such as visual search (Monzel & Reuter, 2024). These evidence from psychology and cognitive science therefore suggest that the absence of a dedicated visual information pathway may explain why current multimodal AI systems excel in formal and abstract domains dominated by verbal world knowledge, yet continue to fall far short of human performance on tasks involving physical and spatial reasoning (Schulze Buschoff et al., 2025; Cai et al., 2025), which fundamentally depend on **visual world modeling**.

Next-generation multimodal AI systems are evolving to be built upon unified multimodal models (UMMs) (Team, 2024; Wu et al., 2025a; Wang et al., 2024b; Deng et al., 2025), which seamlessly integrate both verbal and visual generation capabilities. The newly introduced visual generation component offers the potential to explicitly realize visual world modeling, a critical element of **multimodal world models** in human-like reasoning that current systems largely lack. This naturally makes us ponder: *Can current UMMs truly leverage their visual generation capability to enhance reasoning and thereby narrow the performance gap between multimodal AI and humans?* A growing

body of preliminary research (Li et al., 2025b; Zou et al., 2025; Liang et al., 2025; Zhou et al., 2025b; Gu et al., 2025) has begun exploring this question from different perspectives. However, the findings so far remain inconclusive. Reported empirical results are mixed, showing no consistent trends that visual generation reliably improves reasoning performance. Moreover, the evaluation tasks used in current studies are designed heuristically, lacking a principled basis for understanding when and how visual generation can meaningfully contribute to multimodal reasoning.

In this paper, we present the first principled study of when and how visual generation benefits reasoning from a **world-model perspective** (see Figure 1).

- Theoretically, we rigorously bridge the concepts of world models and reasoning. (1) **World model formulations:** We formulate multimodal world models to approximate the underlying *multi-observable Markov decision processes* (MOMDP) of tasks, and define two fundamental capabilities of world models, namely *world reconstruction* and *world simulation*. (2) **World model-based reasoning:** To realize world models for reasoning, we introduce three reasoning formulations. Two rely solely on verbal CoTs through *implicit or verbal world modeling*, while the third interleaves verbal and visual CoTs that explicitly incorporate visual generation as a form of *visual world modeling*. (3) **The visual superiority hypothesis:** Under this framework, we analyze the distinctions among different world models, highlighting the richer informativeness and complementary prior knowledge afforded by visual world modeling. These insights motivate our central hypothesis that visual world modeling is superior for certain tasks, particularly those grounded in the physical world.
- Empirically, we validate these insights through a series of controlled experiments. (4) **The VisWorld-Eval suite:** We identify and design tasks that specifically isolate and demand each atomic world model capability, forming a new evaluation suite to facilitate future research. This suite, *VisWorld-Eval*, collects seven tasks spanning both synthetic and real-world domains. (5) **Empirical evaluation:** Experiments with a state-of-the-art UMM (Deng et al., 2025) on VisWorld-Eval reveal findings consistent with our insights and theoretical analysis. In tasks where verbal world modeling suffers from representational bottlenecks or insufficient prior knowledge, interleaved CoT delivers substantial performance improvements. By contrast, it offers no clear advantages in tasks such as mazes and Sokoban, whose simple states do not require explicit visual world modeling. We further conduct dedicated analyses, including evidence revealing emergent implicit world modeling in the maze task.

## 2 A WORLD MODEL PERSPECTIVE ON MULTIMODAL REASONING

Inspired by the aforementioned connections between human cognition and artificial intelligence, we formalize our world-model perspective on multimodal reasoning (see Figure 2) in this section.

### 2.1 FORMULATION: MULTIPLE OBSERVATIONS OF THE WORLD

The world of a specific task can be formulated as a **multi-observable Markov decision process** (MOMDP)  $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, \Phi, \mathcal{O}_\phi, e_\phi)$ , where  $\mathcal{S}$  denotes the state space,  $\mathcal{A}$  the action space,  $p$  the transition function,  $\Phi$  the parameter space of observation functions,  $\mathcal{O}_\phi$  the observation space, and  $e_\phi$  the observation function. Each  $s \in \mathcal{S}$  represents the underlying state of the world, which is typically hidden and not directly observable. Instead, it can be perceived through different instantiations of observations (hereafter also referred to as *views*) (Huh et al., 2024), given by  $o = e_\phi(s) \in \mathcal{O}_\phi$ , parameterized by  $\phi \in \Phi$ . As illustrated in Figure 2a, such views can span multiple modalities—for example, visual observations corresponding to different camera poses, or verbal descriptions expressed with different emphases or styles. When an action  $a \in \mathcal{A}$  is applied to the current state, the world transits according to the dynamics  $s' \sim p(s'|s, a)$  and yields new observations.

### 2.2 ATOMIC CAPABILITIES OF WORLD MODELS

A world model, analogous to human mental models, is then expected to support two fundamental capabilities (LeCun, 2022), illustrated in Figure 2b. The first is called **world reconstruction**. Humans are remarkably skilled at mentally reconstructing the structure of an environment from only a few partial observations (Yin et al., 2025), grounded in their prior knowledge of the world. Such

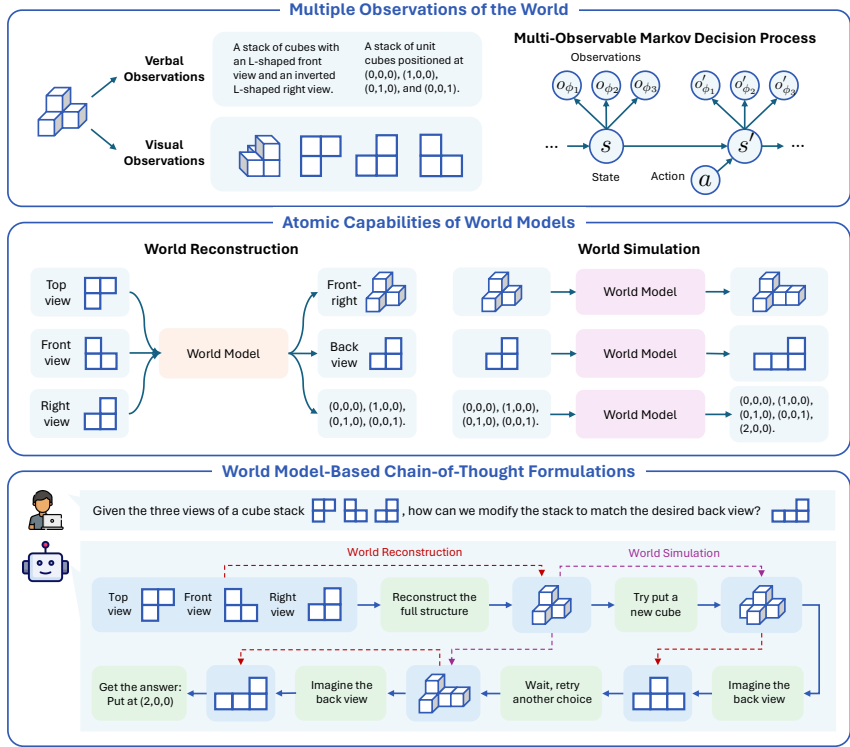


Figure 2: Formulation of chain-of-thought reasoning with internal world modeling, which explicitly maintains an evolving sequence of observations across multiple modalities, generated through atomic world model capabilities.

mental reconstruction allows them to imagine novel views of the same underlying state, supporting skills such as mental rotation. Formally, the perception component of a world model encodes  $n$  observations from limited views into an internal representation:  $\hat{s} = \text{enc}(o_{\phi_1}, \dots, o_{\phi_n}) \approx s$ . This representation approximates the true state, and can then be decoded to generate an unseen observation:  $\hat{o}_{\phi_{n+1}} = \text{dec}(\hat{s}, \phi_{n+1}) \approx e_{\phi_{n+1}}(s)$ , providing an internal “experience” of *navigating* the world. In modern generative models, including UMMs, since their latent representations are not explicitly defined, the world reconstruction capability can be realized through end-to-end novel view generation:

$$p_{\theta}(o_{\phi_{n+1}} \mid o_{\phi_1}, \dots, o_{\phi_n}), \tag{1}$$

which implicitly learns the internal representations required to synthesize the new view.

The second capability is **world simulation**. Humans can mentally simulate how the world evolves into the future, supporting reasoning and decision-making, either purely in their minds or with external aids such as a scratchpad. Formally, this corresponds to the prediction component of a world model, which predicts the transition of the current state and action:  $\hat{s}' \sim \text{pred}(\hat{s}, a)$ , providing an internal “experience” of *interacting* with the world. Similarly, for modern generative models, this capability is more typically realized through predictions of future observations:

$$p_{\theta}(o_{t+1} \mid o_{\leq t}, a_{\leq t}). \tag{2}$$

In our new evaluation suite, we deliberately curate tasks that specifically demand each capability, allowing us to independently validate its contribution to multimodal reasoning (see Section 3).

### 2.3 DELIBERATE REASONING WITH WORLD MODELING ACROSS MODALITIES

We then formalize how world-modeling capabilities within multimodal models contribute to reasoning. Given a question  $Q$  and input images  $I$ , the chain-of-thought reasoning process of a multimodal AI system can be expressed as a sequence of intermediate steps (or thoughts)  $R = \tau_1, \tau_2, \dots, \tau_H$ , followed by the answer  $A$ . Although this general formulation treats each reasoning step  $\tau_i$  as an

unconstrained, free-form operation, our world model perspective suggests that humans reason by prediction and planning, and each step inherently manipulates the underlying world observations of the problem (Wang et al., 2025a; Copet et al., 2025; Zhang et al., 2025). We therefore refine the reasoning formulation as  $\tau_i = (r_i, o_i)$  to explicitly incorporate an evolving sequence of observations:

$$R = (r_1, o_1), (r_2, o_2), \dots, (r_H, o_H), \quad (3)$$

where  $r_i$ <sup>1</sup> denotes a logical reasoning step based on the accumulated context, typically expressed in text, and  $o_i$  denotes the observation generated at that step. Specifically, the input images serve as the initial observation  $o_0 = I$ , and subsequent observations are generated from previous reasoning and observations, by invoking atomic world modeling capabilities: world reconstruction (Eq. 1) and world simulation (Eq. 2), where reasoning steps imply actions  $a$  and view transformations  $\phi$ , as illustrated in Figure 2c.

This formulation is modality-agnostic, allowing observations—and thus world modeling—to arise across various modalities. We focus specifically on verbal and visual observations, motivated by dual-coding theory in human cognition and by the fact that UMMs are equipped to generate both. This yields several concrete CoT instantiations. Specifically, **verbal world modeling** produces purely verbal CoTs, with  $o_i$  as verbal descriptions, whereas **visual world modeling** produces verbal-visual interleaved CoTs, with  $o_i$  as generated images. In addition, prior work has discovered that language models can implicitly learn world models with emergent internal representations of board-game states without explicit supervision (Li et al., 2023). Motivated by this, we also consider **implicit world modeling**, in which no explicit observation is generated ( $o_i = \emptyset$ )<sup>2</sup>.

## 2.4 THE VISUAL SUPERIORITY HYPOTHESIS

Contemporary LLMs and VLMs have achieved impressive performance in structured and abstract domains, such as mathematics and programming, largely driven by large-scale language-centric pre-training and verbal chain-of-thought post-training. Although these models have accumulated extensive verbal and symbolic knowledge, their understanding of the visual world remains limited when trained under purely verbal supervision. As a result, they continue to struggle with tasks grounded in basic physical and spatial intuition that even young children naturally master (Schulze Buschoff et al., 2025; Cai et al., 2025).

Visual world modeling is therefore essential for endowing multimodal AI with complementary forms of information and knowledge. In terms of **informativeness**, while verbal and symbolic representations capture high-level semantic abstractions, they often suffer from ambiguity and representational bottlenecks. In contrast, visual observations are more concrete and information-rich, directly encoding physical properties such as motion and spatial relationships. This provides precise, fine-grained grounding for reasoning about the complex real world, particularly in spatial and physical tasks. In terms of **prior knowledge**, visual world knowledge is inherently complementary to symbolic knowledge. Humans and animals acquire much of this knowledge (e.g., physical interactions and spatial transformations) through perception, largely independent of language. Consequently, humans naturally represent and communicate such knowledge visually—for example, by sketching an approximate parabolic trajectory without performing explicit calculations. This suggests that different aspects of world knowledge are concentrated in different data modalities, and learning from large-scale generative modeling of visual data can thereby expand the effective knowledge landscape available for multimodal reasoning.

We formalize and justify these insights through theoretical analysis in Appendix C. Drawing on these analyses, we present our central hypothesis regarding when and how visual generation benefits reasoning, thereby helping narrow the gap between multimodal AI and human capabilities.

**The Visual Superiority Hypothesis:** In multimodal reasoning tasks grounded in the physical world, visual generation as a world model yields representations that are more informative and knowledge-rich than those produced by verbal world models.

<sup>1</sup>We use  $i$  to index reasoning steps in order to distinguish them from the true time step  $t$  of the underlying MOMDP. The twos are not generally aligned, as we may include branching and backtracking in the reasoning.

<sup>2</sup>In practice, strictly distinguishing implicit from verbal world modeling can be difficult, because there are often partial descriptions of the current state in the reasoning part  $r_i$ . In this work, we treat verbal world modeling as explicitly expressing world states or observations in text, such as coordinates or symbolic matrices.

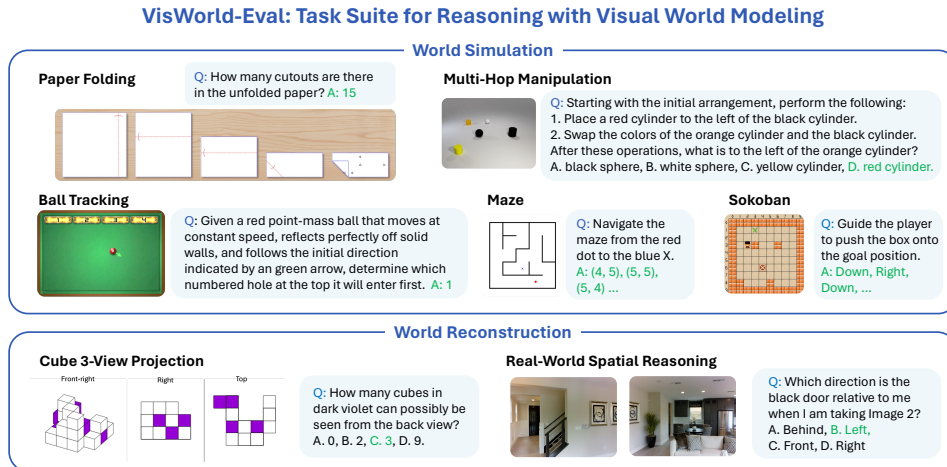


Figure 3: The VisWorld-Eval suite for assessing multimodal reasoning with visual world modeling. VisWorld-Eval comprises seven tasks spanning both synthetic and real-world domains, each designed to isolate and demand specific atomic world-model capabilities.

### 3 VISWORLD-EVAL: TASK SUITE FOR REASONING WITH VISUAL WORLD MODELING

While prior work has primarily designed evaluation tasks heuristically, we principledly evaluate multimodal reasoning across tasks designed to specific world model capabilities. Building on related benchmarks, we identify and curate a total of seven tasks spanning both synthetic and real-world domains, forming an evaluation suite tailored to assess reasoning with visual world modeling. All tasks are framed as question answering with concise, verifiable answers, and performance is measured by answer accuracy. We refer to this suite as *VisWorld-Eval*, and summarize it in Figure 3.

For each task, we construct SFT data by designing different CoT patterns that incorporate implicit, verbal, or visual world modeling, thereby enabling controlled comparative evaluations. Task descriptions, data construction pipeline, and examples are presented in Appendix D.1.

## 4 EXPERIMENTAL RESULTS

In this section, we show that visual world modeling boosts multimodal reasoning through two atomic capabilities: world simulation (Section 4.1) and world reconstruction (Section 4.2). We also identify tasks in which it is unhelpful (Section 4.3), where implicit or verbal world modeling is sufficient.

**Evaluation protocol.** To investigate the benefits of visual generation in multimodal reasoning, we evaluate post-trained UMMs, rather than the zero-shot performance of base models. To the best of our knowledge, no open-source model has been natively optimized for interleaved verbal-visual generation for reasoning. Even commercial closed-source models currently exhibit fundamental limitations in generating visual intermediate reasoning steps (Liang et al., 2025; Zhou et al., 2025b). Focusing on post-trained models, therefore, provides a more meaningful estimate of the upper bound for multimodal reasoning performance, while reducing confounding effects arising from insufficient pre-training due to limited interleaved data availability or quality.

**Model training.** We adopt BAGEL (Deng et al., 2025), a state-of-the-art open-source unified multimodal model, as our base model. Training details of SFT and RLVR are presented in Appendix D.2.

#### 4.1 VISUAL WORLD SIMULATION BOOSTS MULTIMODAL REASONING

**Main results.** Figure 4 summarizes the performance of SFT-trained UMMs under different chain-of-thought formulations across all tasks. We observe that interleaved CoT with visual world modeling significantly outperforms its purely verbal counterparts on three world simulation tasks: paper folding,

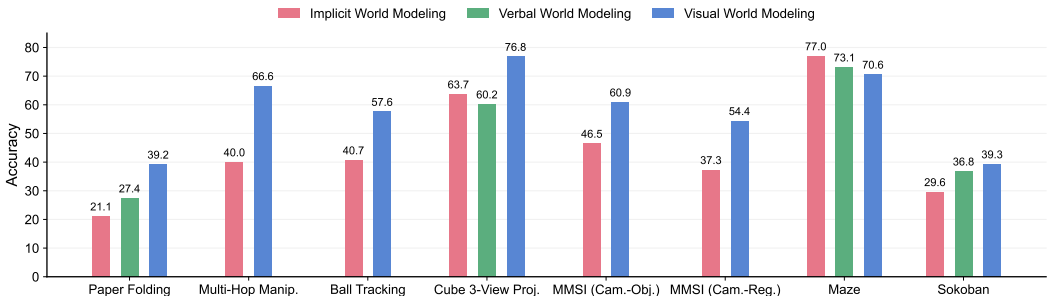


Figure 4: Performance of SFT-trained UMMs with different world model-based chain-of-thought formulations on VisWorld-Eval. Refer to Table 4 for zero-shot performance of advanced VLMs.

multi-hop manipulation, and ball tracking. These gains are attributed to both the richer expressiveness and stronger prior knowledge afforded by the visual modality. In particular, it is difficult for models to precisely ground object coordinates and perform arithmetic operations without external tools in tasks such as multi-hop manipulation and ball tracking, with the latter being especially challenging. Thus, verbal world modeling is inappropriate and omitted in these tasks. This exacerbates ambiguity and hallucinations in purely verbal reasoning. Similarly, in paper folding, although models can track the states of holes, it remains difficult to completely depict the paper contour during unfolding. Moreover, as showcased in Figure 14 and 16, the spatial transformation involved in paper unfolding critically relies on an understanding of geometric symmetry, which can be more naturally learned from visual data like images and videos.

**Sample efficiency.** To further demonstrate the stronger prior knowledge embedded in the visual modality, we experiment comparing the sample efficiency of verbal and visual world modeling on the paper folding task. As shown in Figure 5a, reasoning with visual world modeling exhibits substantially higher sample efficiency, achieving performance comparable to verbal world modeling while using more than  $4\times$  less SFT data.

## 4.2 VISUAL WORLD RECONSTRUCTION BOOSTS MULTIMODAL REASONING

**Main results.** As shown in Figure 4, multimodal reasoning tasks that rely on world reconstruction capabilities also benefit substantially from visual world modeling. In the cube 3-view task, predicting a novel view of stacked cubes, denoted symbolic character matrices, suffers from limited prior knowledge, whereas visually rotating objects has been a rich experience during pre-training with large-scale Internet videos. For MMSI tasks, fully describing a novel view of a realistic scene using text alone is similarly ill-suited as in the previous subsection, and we also discover hallucinations in pure verbal reasoning, which lacks grounding to visual generation. We do not observe consistent improvements on other positional-relationship subtasks in MMSI-Bench, except camera-object and camera-region, which we attribute to current UMM’s limitations in both spatial understanding during verbal reasoning and generation quality in visual world modeling. Full quantitative results and qualitative examples are provided in Appendix E. We expect these limitations to be mitigated in future work with stronger base models.

**Effects of task difficulties.** Figure 5b analyzes performance on the cube 3-view projection task across varying sizes of input cube stacks. We observe a consistent advantage of reasoning with visual world modeling over verbal world modeling across all difficulty levels. Notably, for cube stacks of size six—out of the training distribution—visual world modeling still yields approximately a 10% performance improvement.

**World model fidelity.** Modern AI models are known to exhibit hallucinations along their reasoning trajectories, even when producing correct final answers (Liang et al., 2025). We therefore evaluate the fidelity of world modeling in the cube 3-view projection task by comparing ground-truth views with the intermediate views generated verbally or visually during reasoning. To focus on structural correctness, we compare only the shapes of the views and completely ignore color information. Even under this relaxed evaluation setting, Figure 5b shows that verbal world modeling exhibits dramatically low fidelity, with scores degrading to near zero.

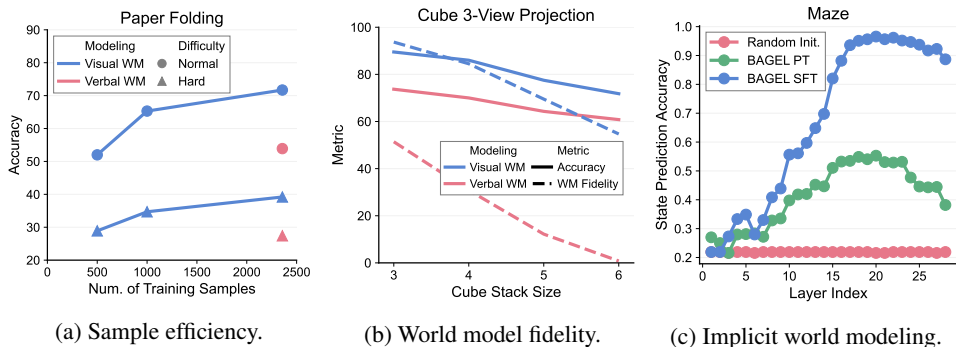


Figure 5: Model analysis: (a) Performance of UMMs on the paper-folding task with varying numbers of SFT samples. (b) Performance of UMMs on the cube 3-view projection task with increasing sizes of input cube stacks, evaluated using both answer accuracy and world-model fidelity. (c) Prediction accuracy of masked point coordinates in CoTs using representations extracted from different layers of different UMMs, revealing emergent internal world representations. PT = Pre-trained.

### 4.3 VISUAL WORLD MODELING IS UNHELPFUL FOR CERTAIN TASKS

**Main results.** (Un)surprisingly, we do not observe notable improvements on grid-world tasks, including maze and Sokoban. In the maze tasks, reasoning with implicit world modeling—without explicitly tracking coordinates—achieves the best performance with a slight advantage. These results are consistent with recent empirical findings (Du et al., 2025). We argue that this is also well explained by our world model perspective. In these tasks, state tracking is relatively simple, typically requiring the maintenance of only one or two two-dimensional coordinates, which can be adequately handled through verbal reasoning alone. Furthermore, in the maze task, we hypothesize that such world modeling can be implicitly encoded in the model’s hidden representations (Li et al., 2023), which helps explain the competitive performance of verbal reasoning without explicit coordinate tracking.

**Demystifying implicit world modeling.** To validate this hypothesis, we probe the internal representations of models, as illustrated in Figure 12. We consider the same architecture, BAGEL, with three different sets of weights: a randomly initialized model, the pre-trained model, and the model supervised fine-tuned on CoT data in the implicit world modeling format, in which special tokens mask all explicit point coordinates during the reasoning process. For each model, we extract the hidden representations of these special tokens at each layer. We then train multilayer perceptrons (MLPs) on these representations to predict the underlying true point coordinates.

Figure 5c reports the prediction accuracy on a validation set. As expected, the randomly initialized model completely fails to internally track point states, achieving only random-guess accuracy on  $5 \times 5$  mazes. In contrast, the pre-trained model (Deng et al., 2025) already exhibits emergent representations that are predictive of maze states. Notably, we observe a non-monotonic trend across layers: prediction accuracy increases from lower layers (which capture low-level features) to middle layers, and then decreases toward the final layers, which are likely specialized for next-token prediction. Finally, supervised fine-tuning on domain-specific data, despite providing no explicit coordinate supervision, substantially enhances this internal predictability, achieving near-perfect accuracy. These in-depth results help explain our main experimental findings: as the model already possesses the capability for implicit world modeling, it does not necessarily benefit from explicit verbal world modeling, let alone more complex forms of visual world modeling.

### 4.4 COMPARISON WITH VLMS: DO UMMs COMPROMISE VERBAL REASONING CAPABILITIES?

One may argue that UMMs are typically trained with a stronger emphasis on visual generation (Deng et al., 2025), which could compromise verbal reasoning capabilities, and bias comparisons in favor of visual world modeling. To address this concern, we compare with a pure VLM baseline, Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), which shares the same Qwen 2.5 LLM base model, with BAGEL. We fine-tune Qwen2.5-VL on the same verbal CoT datasets used in the previous subsections

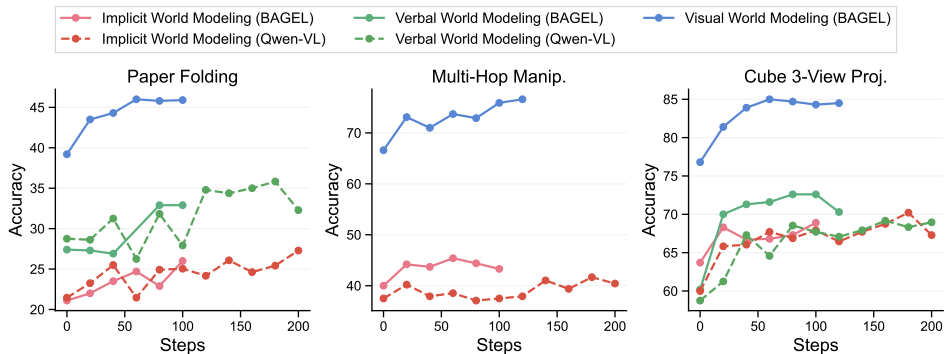


Figure 6: Performance of RLVR-trained VLMs and UMMs with different world-model-based CoT formulations across three tasks. The starting points indicate SFT-trained VLMs and UMMs.

and evaluate it on three representative tasks: paper folding, cube 3-view projection, and multi-hop manipulation.

**Results.** As shown by the starting points in Figure 6 (see also Figure 13 in the Appendix), the SFT performance of Qwen2.5-VL with implicit and verbal world modeling is comparable to that of BAGEL, without exhibiting significant advantages. It still lags behind BAGEL that leverage visual world modeling. These results indicate that our findings arise from the inherent advantages of visual world modeling rather than from compromised verbal reasoning capabilities in UMMs.

#### 4.5 RL ENHANCES VARIOUS CoTs, YET DOES NOT CLOSE THE GAP

Reinforcement learning from verifiable rewards (RLVR) has been a major driver of recent progress in reasoning models equipped with verbal chain-of-thoughts, achieving strong performance across domains such as mathematics (Guo et al., 2025a). While Figure 4 shows a clear advantage of reasoning with visual world modeling after SFT, RLVR may further incentivize emergent reasoning behaviors that improve verbal CoTs. We thus conduct comparative RLVR experiments across different world model-based CoT formulations on three representative tasks.

**Results.** Figure 6 presents the learning curves under RLVR for different models. We observe consistent improvements during RLVR for different CoT formulations. However, the performance gap persists. We also find that VLMs and UMMs generally perform similarly with verbal CoTs. These results suggest that the superiority arises from inherent advantages of the world modeling approach, rather than insufficient post-training. Notably, RL enhances reasoning with visual world modeling, even though only the verbal generation components of interleaved CoTs are directly optimized. We envision that the full potential of interleaved CoTs will be further released with the development of RL algorithms tailored for verbal-visual interleaved generation.

## 5 DISCUSSIONS

By bridging concepts from human cognition and artificial intelligence, we revisit the mechanisms underlying human reasoning and the central role of world models. This provides a new perspective on the use of visual generation for multimodal reasoning, highlighting its potential to serve as visual world models that complement the verbal world models embedded in LLMs. For the first time, this perspective is studied in a principled manner, through theoretical formulations that bridge world models and reasoning, as well as through empirical evaluations whose results are well explained by and strongly support the proposed insights. We hope this work provides early evidence for the central role of multimodal world models in general-purpose AI, in which complementary verbal and visual knowledge emerges from generative modeling across modalities. This work also helps address longstanding questions about the synergistic effects between generation and reasoning, and more broadly contributes to the development of more human-like AI that thinks and acts in the physical world (Gupta et al., 2021). We discuss limitations and future work in Appendix B.

## REFERENCES

- Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chatopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.
- Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *NeurIPS*, 2024.
- Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.
- Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.
- Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.
- 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.
- DeLong Chen, Theo Moutakanni, Willy Chung, Yejin Bang, Ziwei Ji, Allen Bolourchi, and Pascale Fung. Planning with reasoning using vision language world model. *arXiv preprint arXiv:2509.02722*, 2025.
- Jade Copet, Quentin Carbonneaux, Gal Cohen, Jonas Gehring, Jacob Kahn, Jannik Kossen, Felix Kreuk, Emily McMilin, Michel Meyer, Yuxiang Wei, et al. Cwm: An open-weights llm for research on code generation with world models. *arXiv preprint arXiv:2510.02387*, 2025.
- Kenneth James Williams Craik. *The nature of explanation*, volume 445. CUP Archive, 1967.
- Google DeepMind. Genie 3: A new frontier for world models. 2025.
- Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.
- Yifan Du, Kun Zhou, Yingqian Min, Yue Ling, Wayne Xin Zhao, and Youbin Wu. Revisiting the necessity of lengthy chain-of-thought in vision-centric reasoning generalization. *arXiv preprint arXiv:2511.22586*, 2025.
- Yunhai Feng, Jiaming Han, Zhuoran Yang, Xiangyu Yue, Sergey Levine, and Jianlan Luo. Reflective planning: Vision-language models for multi-stage long-horizon robotic manipulation. *arXiv preprint arXiv:2502.16707*, 2025.
- Jay W Forrester. Counterintuitive behavior of social systems. *Theory and decision*, 2(2):109–140, 1971.
- Debre Gentner. *Mental Models, Psychology of*, pp. 9683–9687. 12 2001. ISBN 9780080430768. doi: 10.1016/B0-08-043076-7/01487-X.

- Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. *arXiv preprint arXiv:2510.27492*, 2025.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 2025a.
- Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025b.
- Meng-Hao Guo, Xuanyu Chu, Qianrui Yang, Zhe-Han Mo, Yiqing Shen, Pei-lin Li, Xinjie Lin, Jinnian Zhang, Xin-Sheng Chen, Yi Zhang, et al. Rbench-v: A primary assessment for visual reasoning models with multi-modal outputs. *arXiv preprint arXiv:2505.16770*, 2025c.
- Ziyu Guo, Renrui Zhang, Hongyu Li, Manyuan Zhang, Xinyan Chen, Sifan Wang, Yan Feng, Peng Pei, and Pheng-Ann Heng. Thinking-while-generating: Interleaving textual reasoning throughout visual generation. *arXiv preprint arXiv:2511.16671*, 2025d.
- Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, et al. Can we generate images with cot? let’s verify and reinforce image generation step by step. In *CVPR*, 2025e.
- Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature Communications*, 2021.
- David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.
- Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 2025.
- Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. In *ICML*, 2022.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *ICML*, 2024.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Michael Igoevich Ivanitskiy, Rusheb Shah, Alex F. Spies, Tilman Räuher, Dan Valentine, Can Rager, Lucia Quirke, Chris Mathwin, Guillaume Corlouer, Cecilia Diniz Behn, and Samy Wu Fung. A configurable library for generating and manipulating maze datasets. *arXiv preprint arXiv:2309.10498*, 2023.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.
- PN Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, 1983.
- George Lakoff and Rafael Núñez. *Where mathematics comes from*, volume 6. New York: Basic Books, 2000.
- David Landy and Robert L Goldstone. How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(4):720, 2007.
- Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

- Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025a.
- Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. In *ICML*, 2025b.
- Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *ICLR*, 2023.
- Yongyuan Liang, Wei Chow, Feng Li, Ziqiao Ma, Xiyao Wang, Jiageng Mao, Jiu-hai Chen, Jiatao Gu, Yue Wang, and Furong Huang. Rover: Benchmarking reciprocal cross-modal reasoning for omnimodal generation. *arXiv preprint arXiv:2511.01163*, 2025.
- Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. In *NeurIPS*, 2025.
- Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *CVPR*, 2025.
- Ruaridh Mon-Williams, Gen Li, Ran Long, Wenqian Du, and Christopher G Lucas. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence*, 2025.
- Merlin Monzel and Martin Reuter. Where’s wanda? the influence of visual imagery vividness on visual search speed measured by means of hidden object pictures. *Attention, Perception, & Psychophysics*, 86(1):22–27, 2024.
- Donald A Norman. Some observations on mental models. In *Mental models*, pp. 7–14. Psychology Press, 2014.
- Allan Paivio. *Mental representations: A dual coding approach*. Oxford university press, 1990.
- Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jiali Wang, Zhiyang Xu, Jiu-hai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.
- Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 2020.
- Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 2025.
- Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.
- Weikang Shi, Aldrich Yu, Rongyao Fang, Houxing Ren, Ke Wang, Aojun Zhou, Changyao Tian, Xinyu Fu, Yuxuan Hu, Zimu Lu, et al. Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning. *arXiv preprint arXiv:2510.14958*, 2025a.
- Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, et al. Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. *arXiv preprint arXiv:2509.24897*, 2025b.
- Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 2025.

- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, et al. Game-rl: Synthesizing multimodal verifiable game data to boost vlms’ general reasoning. *arXiv preprint arXiv:2505.13886*, 2025a.
- Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *ICCV*, 2025b.
- Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 2024.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 2025.
- Kangrui Wang, Pingyue Zhang, Zihan Wang, Yaning Gao, Linjie Li, Qineng Wang, Hanyang Chen, Chi Wan, Yiping Lu, Zhengyuan Yang, et al. Vagen: Reinforcing world model reasoning for multi-turn vlm agents. In *NeurIPS*, 2025a.
- Ruoyao Wang, Graham Todd, Ziang Xiao, Xingdi Yuan, Marc-Alexandre Côté, Peter Clark, and Peter Jansen. Can language models serve as text-based world simulators? In *ACL*, 2024a.
- Siting Wang, Luoyang Sun, Cheng Deng, Kun Shao, Minnan Pei, Zheng Tian, Haifeng Zhang, and Jun Wang. Spatialviz-bench: Automatically generated spatial visualization reasoning tasks for mllms. *arXiv preprint arXiv:2507.07610*, 2025b.
- Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.
- Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *CVPR*, 2025a.
- Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideoopt: Interactive videoopts are scalable world models. In *NeurIPS*, 2024.
- Jialong Wu, Shaofeng Yin, Ningya Feng, and Mingsheng Long. Rlvr-world: Training world models with reinforcement learning. In *NeurIPS*, 2025b.
- Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025.
- Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let’s think only with images. *arXiv preprint arXiv:2505.11409*, 2025.
- Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025a.
- Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025b.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Chi Chen, Haoyu Li, Weilin Zhao, et al. Efficient gpt-4v level multimodal large language model for deployment on edge devices. *Nature Communications*, 2025.
- Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshige Yan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV’25*, 2025.

- Kai Zhang, Xiangchao Chen, Bo Liu, Tianci Xue, Zeyi Liao, Zhihan Liu, Xiyao Wang, Yuting Ning, Zhaorun Chen, Xiaohan Fu, et al. Agent learning via early experience. *arXiv preprint arXiv:2510.08558*, 2025.
- Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *CVPR*, 2025a.
- Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. In *NeurIPS*, 2025b.
- Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025a.
- Yiyang Zhou, Haoqin Tu, Zijun Wang, Zeyu Wang, Niklas Muennighoff, Fan Nie, Yejin Choi, James Zou, Chaorui Deng, Shen Yan, et al. When visualizing is the first step to reasoning: Mira, a benchmark for visual chain-of-thought. *arXiv preprint arXiv:2511.02779*, 2025b.
- Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, and Ziwei Liu. Uni-mmmu: A massive multi-discipline multimodal unified benchmark. *arXiv preprint arXiv:2510.13759*, 2025.

## A RELATED WORK

**World models.** The field of world models is rapidly evolving, yet remains far from reaching consensus on definitions or methodologies. Although psychology and cognitive science suggest that human mental models rely on compact representations that discard irrelevant details, how to scale approaches capable of learning such abstract representations (Schrittwieser et al., 2020; Hansen et al., 2022; LeCun, 2022) to arbitrary domains and modalities is still unclear. Consequently, most current techniques preserve complete information of observations, either through reconstructable latent representations (Ha & Schmidhuber, 2018; Hafner et al., 2025) or directly at the level of raw data. Prominent examples include modern video generation world models (DeepMind, 2025; Agarwal et al., 2025; Alonso et al., 2024; Wu et al., 2024) which capture concrete pixel-level dynamics. In contrast, language inherently provides a higher level of abstraction, making it more similar to human mental representations (Wang et al., 2024a; Wu et al., 2025b; Wang et al., 2025a; Yin et al., 2025; Chen et al., 2025). This motivates the promise of unified multimodal models that generate both languages and visuals as a new direction for building more human-like world models.

**Unified multimodal models.** Multimodal understanding (Hurst et al., 2024; Bai et al., 2025b; Guo et al., 2025b) and visual generation (Rombach et al., 2022; Seedream et al., 2025) have long developed in isolation. Recently, there has been growing interest in integrating these two capabilities into a single unified model. This can be straightforwardly achieved by forwarding the representations of vision language models to an external visual generation module (Tong et al., 2025b; Pan et al., 2025). A more unified approach is to model both language and visual modalities within a single backbone. While language is predominantly modeled through autoregressive next-token prediction, the design space of visual modalities spans a wide spectrum, from discrete tokenization with autoregressive (Wang et al., 2024b; Team, 2024; Wu et al., 2025a) or masked modeling (Xie et al., 2025; Guo et al., 2025e), to continuous tokenization with diffusion or flow-based modeling (Zhou et al., 2025a; Ma et al., 2025; Deng et al., 2025). Among these efforts, BAGEL (Deng et al., 2025) is one of the most widely adopted open-source models achieving state-of-the-art performance. Despite substantial progress in building unified multimodal models (UMMs), existing evaluations still primarily assess their understanding and generation capabilities separately. One widely recognized advantage of UMMs lies in leveraging reasoning abilities of handling complex instructions to enhance visual generation or editing (Zhao et al., 2025b; Guo et al., 2025d). Yet when and how visual generation, in turn, enhances reasoning remains insufficiently explored, lacking solid empirical evidence and community consensus.

**Benchmarking visual generation for reasoning.** This paper contributes to a growing line of research on visual generation for reasoning. RealUnify (Shi et al., 2025b) and Uni-MMMU (Zou et al., 2025) design tasks in which generation is expected to enhance reasoning, but report mixed results without revealing clear trends regarding the benefits of visual generation. ROVER (Liang et al., 2025) reveals fundamental limitations of current models in generating meaningful visual reasoning steps, often resulting in minimal or even negative gains in final accuracy. In contrast, MIRA (Zhou et al., 2025b) conducts a sanity test by providing manually annotated visual cues, thereby bypassing the evaluation of visual world modeling capability. While the aforementioned works evaluate zero-shot performance, ThinkMorph (Gu et al., 2025) fine-tunes UMMs to reveal emergent reasoning behaviors but restricts each CoT to a single intermediate image, thereby not fully exploiting the potential of interleaved CoT. Our work distinguishes itself through a world-model perspective that enables a principled investigation, allowing us to both demonstrate and systematically explain when visual generation yields positive gains and when it does not.

## B LIMITATIONS AND FUTURE WORK

This work primarily focuses on spatial and physical reasoning tasks, where multimodal AI systems exhibit a pronounced performance gap relative to humans. Many other tasks proposed in the related literature can also be interpreted through our world model perspective. For example, a prominent class of benchmarks involves visual jigsaw tasks (Shi et al., 2025b; Gu et al., 2025; Liang et al., 2025; Zou et al., 2025), in which input image patches are cropped, masked, or shuffled. Such tasks essentially probe a form of world reconstruction capability, as corrupted images and videos are commonly treated as specific views within the world model literature (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025). Another active area of interest lies in STEM reasoning. Recent work

(Shi et al., 2025a) leverages visual generation for mathematical diagram editing, such as constructing auxiliary geometric lines. This closely resembles how humans use visual sketchpads to support math understanding and reasoning, constructing visual world models of a symbolic system. However, as symbolic representations in mathematics are largely complete, and mathematical reasoning has been extensively optimized in modern LLMs, it remains unclear whether multimodal interleaved CoT can fundamentally break through the performance limit, warranting further investigation.

We do not apply reinforcement learning to the visual generation components of verbal–visual interleaved CoTs (Liu et al., 2025). Prior work has shown that world models themselves can be improved through RLVR (Wu et al., 2025b). As discussed in Section 4.5, developing RL algorithms specifically tailored to interleaved verbal–visual generation may further improve world-model fidelity during reasoning and incentivize the emergence of stronger and intriguing world-modeling capabilities.

The analysis of emergent representations for implicit world modeling in Figure 5c is intriguing but preliminary. We hope this result will rekindle interest in probing approaches (Li et al., 2023) for interpreting the latent representations learned by different models. In particular, we are interested in comparing the internal representations of VLMs and UMMs, as the latter may capture complementary aspects of world knowledge through training for multimodal generation.

Artificial intelligence is increasingly being embodied in the physical world (Gupta et al., 2021). Our work, particularly the visual superiority hypothesis, suggests that learning visual world models is therefore essential for embodied intelligence. Visual world modeling enables embodied agents to better understand their environments, from imagining occluded regions to interpreting user intentions from an egocentric perspective, thereby supporting more reliable and natural everyday services. It also facilitates planning and decision-making by allowing agents to mentally simulate the precise outcomes of potential actions, leading to more effective interaction with the world. Rather than relying on loosely coupled modules (Feng et al., 2025) or performing only single-step reasoning (Zhao et al., 2025a), we envision a future direction in which flexible multimodal world modeling and reasoning, empowered by interleave verbal-visual generation within a unified model, form core components of physical and embodied AI.

## C THEORETICAL ANALYSIS

### C.1 INFORMATIVENESS

In this section, we present the rigorous versions of our world-model-based chain-of-thought formulations and analyze them using an information-theoretic approach.

#### C.1.1 PROBLEM SETUP AND ASSUMPTIONS

Given a question  $Q$  and input images  $I$ , multimodal reasoning generates a chain-of-thought process  $R$ , followed by a final answer  $A$ . We explicitly formulate the reasoning process as an interleaving of logic reasoning steps and observations of the underlying MOMDP defined in Section 2.1:  $R = (r_1, o_1), (r_2, o_2), \dots, (r_H, o_H)$  where  $H$  denotes the (fixed) CoT length. For notation convenience, we denote the input image(s) as the initial observation  $o_0$ .

We assume that each MOMDP observation function admits a two-stage decomposition:  $e_\phi(s) = g_{\phi_m}(f_{\phi_s}(s))$ ,  $\Phi = \Phi_s \times \Phi_m$ , where the inner modality-agnostic mapping  $f_{\phi_s}$  (parameterized by  $\phi_s \in \Phi_s$ ) extracts a *slice* of the underlying state  $s$ , retaining only partial state information, and the outer modality-specific mapping  $g_{\phi_m}$  (parameterized by  $\phi_m \in \Phi_m$ ) renders the extracted slice into a particular observation modality.

Under this decomposition, we assume that reasoning across different modalities of observations shares a common underlying *oracle* reasoning process:

$$\begin{aligned} & p(Q, \bar{s}_0, r_1, \bar{s}_1, \dots, r_H, \bar{s}_H, A) \\ &= p(Q) \left[ \prod_{i=1}^H p(r_i | \bar{s}_{0:i-1}, r_{1:i-1}, Q) p(\bar{s}_i | \bar{s}_{0:i-1}, r_{1:i}, Q) \right] p(A | r_{1:H}, \bar{s}_{0:H}, Q), \end{aligned}$$

where  $\bar{s}_i = (s_i, \phi_{s_i}) \in \mathcal{S} \times \Phi_s$  denotes a modality-agnostic *sliced* state. Each logic step  $r_i$  is assumed to reason on sufficient sliced state information:  $p(r_i | \bar{s}_{0:i-1}, r_{1:i-1}, Q) =$

$p(r_i | f_{\phi_{s_0}}(s_0), \dots, f_{\phi_{s_{i-1}}}(s_{i-1}), r_{1:i-1}, Q)$ , and produces actionable outcomes that either (i) transit a previous world state  $s_{j < i}$  via an implicit action  $a_i: \bar{s}_i = (s_i, \phi_{s_j}), s_i \sim p(s_j, a_i)$  or (ii) query the same underlying world state with a new slice  $\phi_{s_i}$ , yielding  $\bar{s}_i = (s_j, \phi_{s_i})$ . The oracle reasoning process is then rendered into a specific modality via  $o_i = g_{\phi_m}(f_{\phi_{s_i}}(s_i))$ . Unless otherwise specified, we abuse notation and use  $s_i$  to denote  $\bar{s}_i = (s_i, \phi_{s_i})$  in the remainder of our analysis.

Given the above oracle CoT generation process, we learn a model  $p_\theta$  whose joint distribution over CoTs and answers factorizes into a reasoning component and a world-modeling component:

$$p_\theta(R, A | Q, I) = p_\theta(r_1, o_1, r_2, o_2, \dots, r_H, o_H, r_{H+1} | r_0, o_0) = \prod_{i=1}^{H+1} p_\theta(r_i | R_i) \prod_{i=1}^H p_\theta(o_i | \tilde{R}_i), \quad (4)$$

where we denote the question as  $r_0$ , the initial observation (input image(s)) as  $o_0$ , and the final answer as  $r_{H+1}$ . The CoT prefixes are defined as  $R_i = (r_0, o_0, r_1, o_1, \dots, r_{i-1}, o_{i-1}), \tilde{R}_i = (r_0, o_0, r_1, o_1, \dots, r_{i-1}, o_{i-1}, r_i)$ .

### C.1.2 THEORETICAL RESULTS

We first establish that the end-to-end answer error admits an upper bound that naturally decomposes into reasoning and world-modeling errors.

**Theorem 1.** *For any observation modality  $m$ , the following inequality holds:*

$$\begin{aligned} \text{KL}(p(A | Q, I) || p_\theta(A | Q, I)) &\leq \text{KL}(p(R, A | Q, I) || p_\theta(R, A | Q, I)) \\ &= \sum_{i=1}^{H+1} \underbrace{\mathbb{E}_p [\text{KL}(p(r_i | R_i) || p_\theta(r_i | R_i))]}_{\text{reasoning errors}} \\ &\quad + \sum_{i=1}^H \underbrace{\mathbb{E}_p [\text{KL}(p(o_i | \tilde{R}_i) || p_\theta(o_i | \tilde{R}_i))]}_{\text{world modeling errors}}. \end{aligned} \quad (5)$$

*Proof.* The first inequality follows from the data processing inequality: marginalizing out  $R$  cannot increase the KL divergence. For the equality, we apply the chain rule for KL divergence together with the CoT factorization in Eq. 4. In particular, substituting the factorizations of  $p(R, A | Q, I)$  and  $p_\theta(R, A | Q, I)$  into  $\text{KL}(p(R, A | Q, I) || p_\theta(R, A | Q, I))$  leads to the stated decomposition.  $\square$

**Remarks.** This decomposition reveals a fundamental trade-off between the informativeness of world models for reasoning and the fidelity of the world model itself. In the case of implicit world modeling, where  $o_i = \emptyset$ , we get rid of the world-modeling error. However, this typically comes at the cost of increased uncertainty and learning difficulty in reasoning, as all state transitions must be implicitly encoded. Empirically, world models that explicitly track the task states, serving as verbal or visual sketchpads, are generally beneficial for reasoning. We dive into the reasoning component of Eq. 5 to elucidate the factors underlying these benefits.

**Theorem 2.** *For any observation modality  $m$ , the reduction in reasoning uncertainty achieved by explicit world modeling satisfies:*

1. *Reasoning uncertainty does not increase:*  $\mathbb{H}(r_i | o_0, r_{0:i-1}) - \mathbb{H}(r_i | R_i) = \mathbb{I}(o_{1:i-1}; r_i | o_0, r_{0:i-1}) \geq 0$ .
2. *Uncertainty reduction is upper-bounded by both (i) the information that observations provide about the underlying states and (ii) the information that the reasoning step requires about those states:*

$$\mathbb{I}(o_{1:i-1}; r_i | o_0, r_{0:i-1}) \leq \min(\mathbb{I}(o_{1:i-1}; s_{1:i-1}), \mathbb{I}(r_i; s_{0:i-1}, r_{0:i-1})). \quad (6)$$

*Proof.* The first property follows the definition and the non-negativity of mutual information.

For the second property, denote the conditioning context as  $C = (o_0, r_{0:i-1})$ . Using the properties of ternary mutual information:  $\mathbb{I}(X; Y; Z) = \mathbb{I}(X; Y) - \mathbb{I}(X; Y | Z)$ . we obtain

$$\begin{aligned} \mathbb{I}(o_{1:i-1}; r_i | C) &= \mathbb{I}(o_{1:i-1}; r_i | C) - \mathbb{I}(o_{1:i-1}; r_i | s_{1:i-1}, C) = \mathbb{I}(s_{1:i-1}; o_{1:i-1}; r_i | C) \\ &= \mathbb{I}(o_{1:i-1}; s_{1:i-1} | C) - \mathbb{I}(o_{1:i-1}; s_{1:i-1} | r_i, C) \leq \mathbb{I}(o_{1:i-1}; s_{1:i-1} | C), \end{aligned} \quad (7)$$

where  $\mathbb{I}(o_{1:i-1}; r_i | s_{1:i-1}, C) = 0$  follows from the conditional independence  $r_i \perp o_{1:i-1} | s_{1:i-1}$ .

Further, due to  $o$  as the deterministic function of  $s$ , we have:

$$\begin{aligned} \mathbb{I}(o_{1:i-1}; s_{1:i-1} | C) &= \mathbb{H}(o_{1:i-1} | C) - \mathbb{H}(o_{1:i-1} | s_{1:i-1}, C) \\ &\leq \mathbb{H}(o_{1:i-1}) - \mathbb{H}(o_{1:i-1} | s_{1:i-1}) = \mathbb{I}(o_{1:i-1}; s_{1:i-1}), \end{aligned}$$

where  $\mathbb{H}(o_{1:i-1} | s_{1:i-1}) = \mathbb{H}(o_{1:i-1} | s_{1:i-1}, C) = 0$ .

Symmetrically, we have:

$$\begin{aligned} \mathbb{I}(o_{1:i-1}; r_i | C) &= \mathbb{I}(s_{1:i-1}; o_{1:i-1}; r_i | C) \leq \mathbb{I}(s_{1:i-1}; r_i | C) = \mathbb{H}(r_i | C) - \mathbb{H}(r_i | s_{1:i-1}, C) \\ &\leq \mathbb{H}(r_i) - \mathbb{H}(r_i | s_{0:i-1}, r_{0:i-1}) = \mathbb{I}(r_i; s_{0:i-1}, r_{0:i-1}), \end{aligned}$$

where  $\mathbb{H}(r_i | s_{0:i-1}, r_{0:i-1}) \leq \mathbb{H}(r_i | s_{1:i-1}, o_0, r_{0:i-1})$  due to data processing inequality.

Combining the two upper bounds proves Eq. 6.  $\square$

**Remarks.** The uncertainty of the target distribution is closely related to sample efficiency and learning difficulty. Consequently, the upper bound on the improvement of reasoning uncertainty (Eq. 6) highlights another trade-off in the choice of observation modality for world modeling. The first term indicates that observations should be sufficiently informative about the underlying latent states. In contrast, the second suggests that they need only preserve the task-relevant aspects of the states required to select appropriate reasoning steps. Excessively detailed observations may be unnecessary and even detrimental, increasing world modeling errors.

**Corollary 1.** *If observations are fully informative about the underlying states, i.e.,  $\mathbb{H}(s_i | o_i) = 0$  for all  $i$ , and the state transition dynamics are deterministic, then explicit world modeling provides no reduction in reasoning uncertainty:  $\mathbb{I}(o_{1:i-1}; r_i | o_0, r_{0:i-1}) = 0$ .*

*Proof.* By Eq. 7, we have

$$\mathbb{I}(o_{1:i-1}; r_i | o_0, r_{0:i-1}) \leq \mathbb{I}(o_{1:i-1}; s_{1:i-1} | o_0, r_{0:i-1}) \leq \mathbb{H}(s_{1:i-1} | o_0, r_{0:i-1}).$$

Under the assumption  $\mathbb{H}(s_0 | o_0) = 0$ , the initial observation  $o_0$  uniquely determines  $s_0$ . Moreover, under deterministic state transitions, the trajectory  $s_{1:i-1}$  is uniquely determined by  $(s_0, r_{1:i-1})$ . Hence,

$$\mathbb{H}(s_{1:i-1} | o_0, r_{1:i-1}) = \mathbb{H}(s_{1:i-1} | s_0, r_{1:i-1}) = 0.$$

Therefore,  $\mathbb{I}(o_{1:i-1}; r_i | o_0, r_{1:i-1}) = 0$ , which proves the corollary.  $\square$

**Remarks.** Corollary 1 shows that in deterministic and fully observable environments, given sufficient data and model capacity, explicit world modeling provides no additional benefit. This theoretical result is consistent with our empirical findings on the simple maze task.

## C.2 PRIOR KNOWLEDGE

In this section, we first derive a generalization bound for transfer learning under distribution shift, and relate it to our perspective on prior knowledge in multimodal reasoning.

### C.2.1 GENERAL TRANSFER LEARNING ANALYSIS

**Problem setup.** A standard transfer learning setup involves a pre-training data distribution  $P$  and the fine-tuning data distribution  $Q$  over samples  $(x, y) \in \mathcal{X} \times \mathcal{Y}$ , and a loss function  $\ell_\theta(x, y) \in [0, 1]$ . Define the population risks  $\mathcal{L}_D(\theta) := \mathbb{E}_{(x,y) \sim D}[\ell_\theta(x, y)]$ ,  $D \in \{P, Q\}$ , and the population minimizers  $\theta_D^* \in \arg \min_{\theta \in \Theta} \mathcal{L}_D(\theta)$ ,  $D \in \{P, Q\}$ . We assume we can obtain  $\theta_P^*$  as the pre-trained

model given sufficient data and optimization. For a radius  $r > 0$ , we then define the fine-tuning constraint set (local neighborhood around the pre-trained model)

$$\Theta_r := \{\theta \in \Theta : \|\theta - \theta_P^*\| \leq r\}.$$

Given  $n$  i.i.d. samples from  $Q$ :  $S = \{(x_i, y_i)_{i=1}^n\}$ ,  $(x_i, y_i) \sim Q$ , the fine-tuned model  $\theta_Q$  minimize empirical risk over  $\Theta_r$ ,  $\widehat{\mathcal{L}}_Q(\theta) := \frac{1}{n} \sum_{i=1}^n [\ell_\theta(x_i, y_i)]$ . Our analysis focus on the *excess risk* on  $Q$ :  $\mathcal{E}_Q(\theta_Q) := \mathcal{L}_Q(\theta_Q) - \mathcal{L}_Q(\theta_Q^*)$ .

**From distribution shift to parameter drift.** We first derive how the distribution shift relates to the shift of the population minimizer.

**Lemma 1** (Uniform Loss Shift under Total Variation). *For any subset  $\mathcal{S} \subseteq \Theta$ ,*

$$\sup_{\theta \in \mathcal{S}} |\mathcal{L}_Q(\theta) - \mathcal{L}_P(\theta)| \leq \text{TV}(P, Q).$$

*Proof.* Fix any  $\theta \in \mathcal{S}$  and define  $f_\theta(h, a, o') := \ell_\theta(h, a, o') \in [0, 1]$ . By the definition of total variation and the standard inequality for bounded functions,  $|\mathbb{E}_Q[f_\theta] - \mathbb{E}_P[f_\theta]| \leq \text{TV}(P, Q)$ . Taking the supremum over  $\theta \in \mathcal{S}$  yields the claim.  $\square$

**Lemma 2** (Risk Proximity of  $\theta_Q^*$  under  $P$ ).

$$\mathcal{L}_P(\theta_Q^*) \leq \mathcal{L}_P(\theta_P^*) + 2\text{TV}(P, Q). \quad (8)$$

*Proof.* By Lemma 1,  $\mathcal{L}_P(\theta_Q^*) \leq \mathcal{L}_Q(\theta_Q^*) + \text{TV}(P, Q)$ . By optimality of  $\theta_Q^*$  on  $Q$ ,  $\mathcal{L}_Q(\theta_Q^*) \leq \mathcal{L}_Q(\theta_P^*)$ . Applying Lemma 1 again,  $\mathcal{L}_Q(\theta_P^*) \leq \mathcal{L}_P(\theta_P^*) + \text{TV}(P, Q)$ . Chaining the three inequalities proves 8.  $\square$

**Assumption 1** (Local Quadratic Growth / Sharpness of  $\mathcal{L}_P$ ). *There exists  $\mu > 0$  such that for all  $\theta$  in a neighborhood containing  $\theta_Q^*$ ,*

$$\mathcal{L}_P(\theta) \geq \mathcal{L}_P(\theta_P^*) + \frac{\mu}{2} \|\theta - \theta_P^*\|^2.$$

**Lemma 3** (Parameter Drift Controlled by  $\text{TV}(P, Q)$ ). *Under Assumption 1,*

$$\|\theta_Q^* - \theta_P^*\| \leq \sqrt{\frac{4}{\mu} \text{TV}(P, Q)}. \quad (9)$$

*Proof.* By Assumption 1 with  $\theta = \theta_Q^*$ ,  $\mathcal{L}_P(\theta_Q^*) \geq \mathcal{L}_P(\theta_P^*) + \frac{\mu}{2} \|\theta_Q^* - \theta_P^*\|^2$ . Rearranging,  $\frac{\mu}{2} \|\theta_Q^* - \theta_P^*\|^2 \leq \mathcal{L}_P(\theta_Q^*) - \mathcal{L}_P(\theta_P^*)$ . Applying Lemma 2 yields  $\frac{\mu}{2} \|\theta_Q^* - \theta_P^*\|^2 \leq 2\text{TV}(P, Q)$ , and hence 9.  $\square$

**Control of the bias term.** Recall the fine-tuning bias induced by restricting to  $\Theta_r$ :  $\varepsilon_{\text{bias}}(r) := \inf_{\theta \in \Theta_r} \mathcal{L}_Q(\theta) - \mathcal{L}_Q(\theta_Q^*)$ .

**Assumption 2** ( $\mathcal{L}_Q$  is Locally Lipschitz). *There exists  $L_Q > 0$  such that for all  $\theta, \theta' \in \Theta_r$ ,*

$$|\mathcal{L}_Q(\theta) - \mathcal{L}_Q(\theta')| \leq L_Q \|\theta - \theta'\|.$$

**Theorem 3** (Bias Bound via Distribution Shift). *Under Assumption 1 and Assumption 2,*

$$\varepsilon_{\text{bias}}(r) \leq L_Q \left( \sqrt{\frac{4}{\mu} \text{TV}(P, Q)} - r \right)_+, \quad (10)$$

where  $(x)_+ := \max\{x, 0\}$ . In particular, if  $r \geq \sqrt{\frac{4}{\mu} \text{TV}(P, Q)}$ , then  $\varepsilon_{\text{bias}}(r) = 0$ .

*Proof.* If  $r \geq \|\theta_Q^* - \theta_P^*\|$ , then  $\theta_Q^* \in \Theta_r$  and thus  $\inf_{\theta \in \Theta_r} \mathcal{L}_Q(\theta) \leq \mathcal{L}_Q(\theta_Q^*)$ , implying  $\varepsilon_{\text{bias}}(r) = 0$ .

Now consider  $r < \|\theta_Q^* - \theta_P^*\|$ . Let  $\theta_r$  be the projection of  $\theta_Q^*$  onto the closed ball  $\Theta_r$ , i.e.,  $\theta_r := \theta_P + r \cdot \frac{\theta_Q^* - \theta_P^*}{\|\theta_Q^* - \theta_P^*\|}$ . Then  $\theta_r \in \Theta_r$  and  $\|\theta_r - \theta_Q^*\| = \|\theta_Q^* - \theta_P^*\| - r$ . Therefore,

$$\varepsilon_{\text{bias}}(r) = \inf_{\theta \in \Theta_r} \mathcal{L}_Q(\theta) - \mathcal{L}_Q(\theta_Q^*) \leq \mathcal{L}_Q(\theta_r) - \mathcal{L}_Q(\theta_Q^*) \leq L_Q \|\theta_r - \theta_Q^*\| = L_Q (\|\theta_Q^* - \theta_P^*\| - r).$$

Using Lemma 3 to bound  $\|\theta_Q^* - \theta_P^*\|$  completes the proof of 10.  $\square$

**Fine-tuning excess risk bound.** We then arrive at the final result:

**Theorem 4** (Fine-tuning Excess Risk Bound with Shift-Controlled Bias). *Assume Assumptions 1 and 2 and uniform convergence over  $\Theta_r$  holds: with probability at least  $1 - \delta$  over samples  $S$ ,*

$$\sup_{\theta \in \Theta_r} |\mathcal{L}_Q(\theta) - \widehat{\mathcal{L}}_Q(\theta)| \leq \varepsilon_{\text{gen}} = O\left(\sqrt{\frac{\text{Rad}_{Q,n}(\Theta_r) + \log(1/\delta)}{n}}\right),$$

where  $\text{Rad}_{Q,n}(\Theta_r)$  is the Rademacher complexity of the function class  $\{\ell_\theta, \theta \in \Theta_r\}$  with respect to  $Q$  for sample size  $n$ . Then with probability at least  $1 - \delta$ ,

$$\mathcal{E}_Q(\theta_Q) \leq 2\varepsilon_{\text{gen}} + L_Q \left( \sqrt{\frac{4}{\mu} \text{TV}(P, Q) - r} \right)_+. \quad (11)$$

*Proof.* Decompose the excess risk as  $\mathcal{E}_Q(\theta_Q) = \left( \mathcal{L}_Q(\theta_Q) - \inf_{\theta \in \Theta_r} \mathcal{L}_Q(\theta) \right) + \varepsilon_{\text{bias}}(r)$ . The first term is bounded by a standard ERM argument using uniform convergence:  $\mathcal{L}_Q(\theta_Q) - \inf_{\theta \in \Theta_r} \mathcal{L}_Q(\theta) \leq 2\varepsilon_{\text{gen}}$ . The second term is bounded by Lemma 3. Combining the two bounds yields 11.  $\square$

## C.2.2 REMARKS ON MULTIMODAL REASONING

Theorem 4 reveals a trade-off between modality complexity and distribution shift. This general transfer learning analysis can be instantiated in our setting of learning world models and reasoning policies. Specifically, training pairs  $(x, y)$  can be instantiated as  $((o_{0:i}, r_{0:i+1}), o_{i+1})$  for world modeling and  $((o_{0:i}, r_{0:i}), r_i)$  for reasoning, respectively. In the following discussion, we focus on world model learning as an example.

Although visual world models are more informative, they are intrinsically more difficult to learn from scratch due to the high dimensionality and complexity of visual observations. Fortunately, modern AI systems are typically large-scale pre-trained, which endows them with strong prior knowledge and enables faster convergence and improved generalization during downstream post-training. As discussed in Section 2.4, humans tend to represent different aspects of world knowledge through different modalities. Consequently, for a given downstream task, the distribution shift between its transition distribution and that learned during large-scale Internet pre-training can vary substantially across modalities. For example, there are abundant visual demonstrations of paper folding on the Internet, whereas detailed verbal descriptions of folding dynamics are comparatively scarce.

The generalization bound (Theorem 4) suggests that this modality-dependent distribution shift is closely related to the post-training sample efficiency of the corresponding world model. This highlights the importance of acquiring broad prior knowledge across modalities during pre-training, and of leveraging the proper modality whose priors are best aligned with the downstream task, in order to achieve stronger generalization at inference time and higher sample efficiency during post-training.

## D EXPERIMENT DETAILS

### D.1 VISWORLD-EVAL AND TRAINING DATA

To construct VisWorld-Eval, we consider the following tasks that primarily require simulating world dynamics over time: (1) *Paper folding*, (2) *Multi-hop manipulation*, and (3) *Ball tracking*. In addition, we include (4) *Maze* Ivanitskiy et al. (2023) and (5) *Sokoban* Tong et al. (2025a), as these two grid-world tasks are commonly used in prior work of studying visual generation for reasoning Xu et al. (2025); Li et al. (2025b). We also evaluate tasks that emphasize reconstructing underlying world structure from partial observations: (6) *Cube 3-view projection* and (7) *Real-world spatial reasoning*.

We summarize the training and test sample counts for each task in VisWorld-Eval, along with the corresponding original or referenced benchmarks, in Table 1.

We elaborate on the construction of training and test data for each task in VisWorld-Eval.

**Paper folding.** This task involves folding a paper grid with varying grid sizes (3–8) and folding steps (1–4). After folding, holes of different shapes—circles, triangles, stars, diamonds, and squares—are

Table 1: Overview of VisWorld-Eval and corresponding training data: features, statistics, and references.

Task	Capability	Domain	Training Samples	Test Samples	Source/Reference
Paper folding	Simulation	Synthetic	2,357	480	SpatialViz (2025b)
Multi-hop manipulation	Simulation	Synthetic	2,000	480	ZebraCoT (2025a), CLEVR (2017)
Ball tracking	Simulation	Synthetic	2,254	1,024	RBench-V (2025c)
Maze	Simulation	Synthetic	8,448	480	maze-dataset (2023)
Sokoban	Simulation	Synthetic	7,715	480	GameRL (2025a)
Cube 3-view projection	Reconstruction	Synthetic	2,500	480	SpatialViz (2025b)
Real-world spatial reasoning	Reconstruction	Real-world	10,661	522	MMSI-Bench (2025b)

punched into the paper. The model is then asked to predict the distribution of holes after the paper is completely unfolded, including queries such as the total number of holes, the number of holes of a specific shape, or the difference in counts between shapes. All test prompts are constructed at the highest difficulty level (grid size 8 with 4 folding steps). For SFT, we generate chain-of-thoughts using rule-based templates that follow a fixed procedure: unfold the paper step-by-step and then count the resulting holes by shape. These CoTs are then rewritten with Gemini 2.5 Pro to improve clarity and logical coherence. Under visual world modeling, we interleave reasoning steps with images of partially unfolded paper states. Under verbal world modeling, we represent intermediate states using two matrices encoding grid coverage status and hole shape at each position. Under implicit world modeling, we directly skip the explicit tracking of states from original CoTs.

**Multi-hop manipulation.** This task begins with an initial arrangement of several geometric objects (cubes, spheres, and cylinders) in various colors, rendered by Blender<sup>3</sup>. A sequence of text-based instructions is then provided, describing operations such as changing or swapping objects’ color or shape, adding new objects, or removing existing ones. To ensure these commands can be interpreted unambiguously in a 3D space, the instructions consistently use relative spatial references, with each object uniquely identified by its combined color and shape attributes—for example: “Place a purple cylinder between the black sphere and the yellow cube.” The model is asked to infer the resulting spatial layout. Queries may include the total number of objects of a specific shape, the directional relationship between two objects, or which object lies in a given direction relative to a reference object. Test prompts are constructed by varying both the number of initial objects (between 3 and 6) and the frequency (between 1 and 5) of different operation types. For SFT, chain-of-thought reasoning is generated using rule-based templates that simulate the stepwise execution of instructions before answering the final query, and these CoTs are subsequently refined with Gemini 2.5 Pro.

**Ball tracking.** This task features a red point-mass ball that moves at constant speed, reflects elastically off solid walls, and travels in the initial direction indicated by a green arrow. The model is asked to predict which numbered hole at the top of the image the ball will enter first. We generate input images with randomized resolution, initial ball position and direction, and a random number of holes (4–8). For test prompts, we select cases in which the ball trajectory reflects off at least one wall before entering a hole. For SFT, CoTs are generated by Seed 1.6, which is asked to explain the ball dynamics between adjacent frames.

**Sokoban.** Sokoban is a classic grid-based puzzle game. We generate instances with grid sizes ranging from 6 to 10, containing a single box and a target position. Test prompts are sampled from the same distribution as the training data. To construct CoTs, we use a search algorithm to compute an optimal solution path. To avoid excessively long trajectories, we render only key intermediate steps, including: (i) the player moving toward the box, (ii) pushing the box in a direction, and (iii) changing the pushing direction. To encourage reflective behavior, we additionally augment trajectories with randomized detours that involve walking into walls, reflecting, and backtracking to rejoin the optimal path. CoTs are generated by Seed 1.6, which explains the dynamics between adjacent frames. For visual world modeling, the rendered intermediate steps are interleaved with verbal CoTs. For pure verbal world modeling, these intermediate renderings are removed. For implicit world modeling, we additionally mask all explicit coordinates during CoTs with special tokens [masked].

<sup>3</sup><https://www.blender.org/>

**Maze.** Maze is a classic grid-based puzzle task. We generate both training and test samples with a fixed grid size of  $5 \times 5$ . To construct CoTs, we use rule-based templates followed by rewriting for improved naturalness. Under visual world modeling, rendered intermediate steps through points and lines are interleaved with verbal CoTs. The settings for verbal and implicit world modeling follow the same protocol as in Sokoban, with masking special tokens as `<point>[masked]</point>`.

**Cube 3-view projection.** This task considers stacks of colored cubes arranged on grids of varying sizes (3–5), with two cube colors. The input consists of one isometric view (either front-left or front-right) and two orthographic views of the stack. The question asks for the number of cubes of a specified color visible from another orthogonal view. Both the questions and answer choices account for ambiguity caused by occlusions, leading to uncertainty in the cube count. All test prompts are constructed using uniformly random grid sizes between 3 and 5. We generate CoTs using rule-based templates: the model first constructs the queried view, marks potentially occluded cubes using a third (auxiliary) color, and then counts cubes by color. These CoTs are subsequently rewritten by Gemini 2.5 Pro for improved naturalness. Under visual world modeling, we interleave reasoning steps with an image of the queried view. Under verbal world modeling, we represent intermediate views using character matrices, where different colors are encoded by different symbols.

**Real-world spatial reasoning.** For this real-world task, we directly adopt test samples from MMSI-Bench, focusing on positional relationship questions. We construct training prompts following a pipeline similar to Yang et al. (2025a). To obtain training CoTs, we run a visual-CoT model, which uses an SFT-trained BAGEL model for novel view synthesis as a tool. The resulting visual CoTs are subsequently filtered and rewritten by Gemini 2.5 Pro.

Examples of training CoTs are presented in Figure 7, 8, 9, 10, and 11.

## D.2 MODEL TRAINING

We adopt BAGEL (Deng et al., 2025), a state-of-the-art open-source unified multimodal model, as our base model. Most experiments are conducted by supervised fine-tuning (SFT) on task-specific datasets, where verbal and visual generation in both chain-of-thought reasoning and final answers are optimized using cross-entropy and flow-matching loss. Specifically, the loss for reasoning with visual world modeling is as follows:

$$\mathcal{L}_\theta(Q, I, R, A) = - \sum_{i=1}^{H+1} \sum_{j=1}^{|r_i|} \log p_\theta(r_{i,j} | r_{i,<j}, R_i) + \sum_{i=1}^H \mathbb{E}_{t,\epsilon} \left\| v_\theta(o_i^t, t | \tilde{R}_i) - (\epsilon - o_i) \right\|_2^2, \quad (12)$$

where  $o_i^t = t o_i + (1 - t)\epsilon$  are noisy observations. We emphasize that in our formulation,  $r_i$  refers to a verbal reasoning step, instead of a reward. We also perform reinforcement learning from verifiable rewards (RLVR) following SFT. During RL, only the verbal generation component is optimized by GRPO (Guo et al., 2025a), while visual generation is regularized via the KL-divergence with respect to the SFT-trained reference model:

$$\mathcal{J}_\theta(Q, I) = \mathbb{E}_{o,r \sim p_{\theta_{\text{old}}}} \left[ \sum_{i=1}^{H+1} \sum_{j=1}^{|r_i|} \left( \min \left( \frac{p_\theta(r_{i,j} | r_{i,<j}, R_i)}{p_{\theta_{\text{old}}}(r_{i,j} | r_{i,<j}, R_i)} A, \right. \right. \right. \\ \left. \left. \left. \text{clip} \left( \frac{p_\theta(r_{i,j} | r_{i,<j}, R_i)}{p_{\theta_{\text{old}}}(r_{i,j} | r_{i,<j}, R_i)}, 1 - \epsilon, 1 + \epsilon \right) A \right) \right) - \sum_{i=1}^H \mathbb{E}_{t,\epsilon} \left\| v_\theta(o_i^t, t | \tilde{R}_i) - v_{\theta_{\text{ref}}}(o_i^t, t | \tilde{R}_i) \right\|_2^2 \right]. \quad (13)$$

**Hyperparameters.** We perform supervised fine-tuning (SFT) of BAGEL based on its official repository<sup>4</sup>, using 8 GPUs, and conduct reinforcement learning from verifiable rewards (RLVR) using `verl`<sup>5</sup> on 64 GPUs. Hyperparameters for SFT and RLVR are reported in Table 2 and Table 3, respectively.

<sup>4</sup><https://github.com/ByteDance-Seed/Bagel>

<sup>5</sup><https://github.com/volcengine/verl>

Table 2: Hyperparameters for supervised fine-tuning UMMs.

Hyperparameter	Value
Learning rate	$3 \times 10^{-5}$
LR Schedule	Constant
Optimizer	AdamW
Loss weight (CE:MSE)	1:10
Warm-up steps	200
Training steps	4000
Gen. resolution	(256, 1024) for paper folding, cube 3-view (240, 1024) for multi-hop manipulation (256, 512) otherwise
Und. resolution	(224, 980)
Sequence length per rank	32K
Num. ranks	8

Table 3: Hyperparameters for reinforcement learning UMMs.

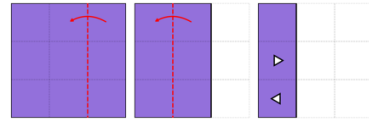
Hyperparameter	Value
Learning rate	$1 \times 10^{-5}$
Batch size	128
GRPO mini batch size	32
Group size	16
KL loss coefficient for visual gen.	0.1
KL loss coefficient for verbal gen.	0.0

The Qwen-VL baselines are trained using `LLaMA-Factory`<sup>6</sup> for supervised fine-tuning (SFT) and `verl` for reinforcement learning from verifiable rewards (RLVR).

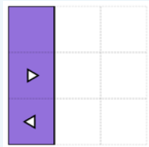
<sup>6</sup><https://github.com/hiyouga/LLaMA-Factory>

### Paper Folding

**Question:** An image shows a sheet of paper that has been folded, with some shapes cut out in its final folded state. Unfold the paper step by step to determine the final pattern when fully opened. Then, answer the question: how many triangle\_left cutouts are there in the unfolded paper?



**Response:** <think> Let's first enlarge the top-rightmost image that shows the cutting pattern on the folded paper.



```

[[0, -1, -1],
 [1, -1, -1],
 [1, -1, -1]]

[['', '', ''],
 ['triangle_right', '', ''],
 ['triangle_left', '', '']]

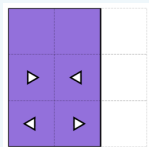
```

Let's systematically, step by step, to reveal the final appearance of the completely unfolded paper. And then answer the question accordingly. Let's trace the process of unfolding the paper step by step, beginning with the most recent fold.

First, we will reverse the second fold. This was a vertical fold where the right side of the paper was folded over the left. To undo this, we unfold the paper from left to right, using the vertical line between column 0 and column 1 as the axis of reflection. Any shapes on the top folded layer (column 0) will be mirrored into the newly revealed column 1. Let's trace this reflection:

- The 'triangle\_right' at position (1, 0) is reflected across the vertical axis. This creates a mirrored shape at (1, 1). The horizontal flip inherent in this reflection transforms the 'triangle\_right' into a 'triangle\_left'.
- Likewise, the 'triangle\_left' at (2, 0) is reflected to position (2, 1). The horizontal flip transforms this 'triangle\_left' into a 'triangle\_right'.

After this step, the original shapes remain in column 0, and their mirror images now occupy column 1.



```

[[0, 0, -1],
 [1, 1, -1],
 [1, 1, -1]]

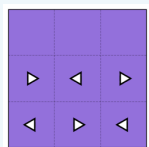
[['', '', ''],
 ['triangle_right', 'triangle_left', ''],
 ['triangle_left', 'triangle_right', '']]

```

Next, we reverse the first fold. This fold was made by placing the right third of the paper over the middle third. Unfolding this section to the right reveals the final pattern, using the fold line between the middle and right columns as the reflection axis.

- The shapes in the leftmost column were not part of this fold, so they remain unchanged.
- The shapes in the middle column were punched through a double layer. When unfolded, this single punch reveals two mirrored shapes. For each 'triangle\_left' visible on the folded paper's middle column, the result is:
  - A 'triangle\_right' on the stationary layer, which becomes the new middle column.
  - A 'triangle\_left' on the unfolding layer, which becomes the new right column. This is the mirror image of the 'triangle\_right' in the middle column.

By combining the stationary shapes from the left column with the revealed and mirrored shapes in the middle and right columns, we arrive at the fully unfolded pattern.



```

[[0, 0, 0],
 [1, 1, 1],
 [1, 1, 1]]

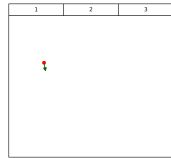
[['', '', ''],
 ['triangle_right', 'triangle_left', 'triangle_right'],
 ['triangle_left', 'triangle_right', 'triangle_left']]

```

Finally we can calculate that the number of triangle\_left cutoffs is 3. </think> 3

Figure 7: Examples of chain-of-thought SFT data for the paper folding task, under visual world modeling (left) and verbal world modeling (right).

### Ball Tracking



**Question:** In the given image:

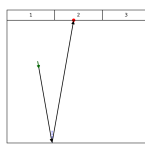
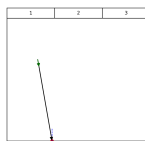
- There is a red ball and a green arrow; the arrow indicates the ball's initial direction of motion;
- At the top of the image, there are several rectangular holes arranged horizontally, numbered from left to right starting at 1;
- The black boundaries of the image represent solid walls. When the ball hits a wall, it undergoes ideal reflection;
- The red ball is considered a point mass-its size is negligible;
- The ball reflects according to the following rules:
  - The component of velocity perpendicular to the wall is reversed in direction;
  - The component of velocity parallel to the wall remains unchanged;
  - The ball moves at a constant speed in an ideal environment until it enters one of the holes.

Your task is:

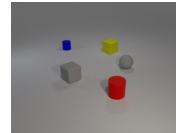
Estimate which hole the red ball will enter first during its motion. Please reason step by step, and put your final answer within `\boxed{}`.

#### Response:

- First, analyze the original image: The image shows a rectangular area with three horizontal holes at the top, numbered 1, 2, 3 from left to right. A red dot with a downward green arrow indicates the ball's initial position and direction of motion. The black lines are solid walls, and the ball undergoes ideal reflection (perpendicular velocity component reverses, parallel component stays the same).
  - Track the ball's initial motion: The ball moves downward in the direction of the green arrow. To determine where it will reflect and which hole it enters, we can follow its path and apply the reflection rule.
  - First, the ball moves downward until it hits the bottom wall. `<image>` (the image showing the ball at the bottom wall) demonstrates this first segment of motion. When hitting the bottom wall, the perpendicular (vertical) component of velocity reverses, so the ball now moves upward, while the horizontal component of velocity remains unchanged. The angle of incidence (angle between incoming path and perpendicular to the wall) is equal to the angle of reflection.
  - Next, track the upward motion after reflection: After reflecting off the bottom wall, the ball moves upward along a path that will take it toward the top wall where the holes are located. `<image>` (the image showing the ball moving toward hole 2) shows this upward path. Since the horizontal velocity component didn't change, and the vertical component was reversed, the ball's path after reflection is symmetric to its incoming path with respect to the perpendicular of the bottom wall.
  - Determine the target hole: The upward path of the ball, after reflecting off the bottom wall, leads directly to hole 2 at the top. There's no other reflection needed here because the ball's path, after the single reflection off the bottom wall, intersects with hole 2.
- So the final answer is `\boxed{2}`



### Multi-Hop Manipulation



**Question:** You are given an initial arrangement of objects viewed from an oblique front perspective. In this view, objects closer to the camera are considered 'front', and the 'left' and 'right' sides of the scene correspond to the left and right of the frame.

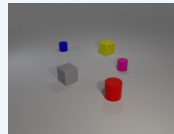
Perform the following operations in sequence:

- Change the object directly in front of the yellow cuboid into a rose cylinder.
- Place a gray cylinder behind and to the left of the object that is directly behind the rose cylinder.
- Place a gray sphere to the left of the rose cylinder.

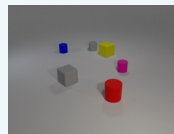
After completing all operations, what is the object to the right of the blue cylinder?

- A. gray cylinder, B. gray cuboid, C. red cylinder, D. blue cylinder.

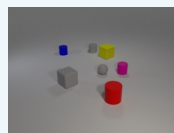
**Response:** `<think>` First, I will execute the operations step-by-step, starting from the initial scene which contains five objects: a blue cylinder, a gray sphere, a gray cuboid, a red cylinder, and a yellow cuboid. The first step is to change the object closest to the front of the yellow cuboid into a rose cylinder. In the initial image, this object is the gray sphere. Accordingly, I will change the gray sphere to a rose cylinder. `<image>`



Now that the gray sphere has become a rose cylinder, I'll proceed to the second step: place a gray cylinder to the left-back of the object closest to the back of the rose cylinder. The object directly behind the new rose cylinder is the yellow cuboid. Therefore, I will place a new gray cylinder to the left-back of the yellow cuboid. `<image>`



For the third and final operation, I need to place a gray sphere to the left of the rose cylinder. I will now add a new gray sphere into the scene, positioned to the left of the rose cylinder. `<image>`

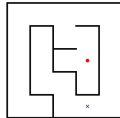


With all the operations complete, I can determine the answer to the question: 'What is the object to the right of the blue cylinder object?'. In the final scene, I can locate the blue cylinder. The object immediately to its right is the gray cylinder. Thus, the correct option is A. gray cylinder . `</think>` A

Figure 8: Examples of chain-of-thought SFT data for the ball tracking and multi-hop manipulation task.

**Maze**

**Question:** You are provided with an image of a maze. A red dot marks the starting position, and a blue X marks the goal. Your objective is to navigate the maze step-by-step from the start to the goal. Return your final path as a list of waypoints in the following format: '<point>x y</point><point>x y</point>...'



**Response:**

I will now solve the maze step by step. I can see the starting point, the red dot, is at <point>165 83</point>, located in the upper-left area of the maze. My destination, the blue 'X', is at <point>83 410</point>, located in the lower-left corner of the maze. [...]

Step 1: I begin at the starting point, <point>165 83</point>. Looking around, the path upwards is blocked by the top boundary of the maze, and the path to the left is blocked by a wall. This leaves me with two possible directions: right or down. After observing the maze layout, I've decided to move right. The path downwards leads into a small, U-shaped enclosure which is clearly a dead end. The path to the right, however, opens into a long corridor that seems to wind through a large portion of the maze, making it the only viable option for making significant progress. I will proceed right until I hit the wall at <point>246 83</point>.

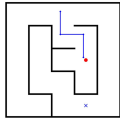
Since there are no new choices, I will continue along the only available path.

From my new position at <point>246 83</point>, the top is the maze boundary, the right is blocked by a wall, and the left is the direction I just came from. My only option is to go down. I will follow this path until I reach <point>246 246</point>, where another wall blocks further downward movement.

Again, there are no new choices, so I'll keep going. [...]

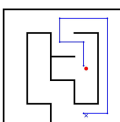
I will move left until I reach <point>83 328</point>. Here, the path to the left is blocked by the maze boundary, and I finally have a new decision to make.

To better visualize my progress and analyze the new options, I will now generate the current path to draw the path I've taken so far. <image>



Step 2: Looking at the path I've drawn, I am now at the intersection <point>83 328</point>. The path to the left is the maze boundary, and the path to the right is the one I just came from. I can either go up or down. I will choose to go down. The destination, the blue 'X', is in the lower-left section of the maze. My current position is directly above it. Going down is the most logical and direct move to get closer to the end point. Going up would take me back towards the area I started in, moving me further away from the goal. I will proceed down to <point>83 410</point>, which appears to be right at the location of the blue 'X' destination.

To confirm if I have reached the end, I will generate the current path to record this final move. <image>



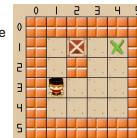
Step 3: In the end, observing the path I have traced with the blue lines, I can see that my current position at <point>83 410</point> coincides with the destination, the blue 'X' at <point>83 410</point>. Therefore, the exploration has been successfully completed. The path I took is: <point>165 83</point><point>246 83</point><point>246 246</point><point>165 246</point><point>165 328</point><point>83 328</point><point>83 410</point>

**Sokoban**

**Question:** This is a Sokoban task.

A player is in a small room with walls, an empty floor, one box, and one goal with green marks. Your objective is to guide the player to push the box onto the goal position.

The player can move 'up', 'down', 'left', or 'right' and can only push the box, not pull it. The player cannot walk through walls or the box.



Present your final movement sequence inside \bbox{ }.

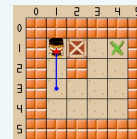
**Response:**

Let's begin by analyzing the initial setup. In the starting scene, I (the player character) am positioned at (3,1) on the grid. The box is located at (1,2), and the goal (marked with the green X) is at position (1,4). The room contains walls (the orange brick tiles) and floor tiles (the lighter, dotted ones). My objective is to push the box from (1,2) onto the goal at (1,4).

To accomplish this, I first need to position myself so I can push the box towards the goal. The box is currently in row 1, column 2, and the goal is located two columns to the right at (1,4). Since the goal is to the right of the box, I need to approach the box from its left side in order to push it right. Currently, I am at position (3,1). To get to the left side of the box (position (1,1)), I need to move up twice: first to (2,1) and then to (1,1), which places me adjacent to the box on its left side.

**Step 1,**

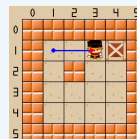
First, I check the positions: I am at (3,1), the box is at (1,2), and the goal is at (1,4). To push the box right towards the goal, I must first reach the left side of the box, which is at (1,1). The correct move here is to move up twice: first from (3,1) to (2,1), then from (2,1) to (1,1). After these moves, I will visualize the current game status: <image>



In this new status, I am at position (1,1), directly to the left of the box at (1,2). The goal remains at (1,4), two spaces to the right of the box.

**Step 2,**

Next, I check the updated positions: I am at (1,1), the box is at (1,2), and the goal is at (1,4). Since I am now to the left of the box, I can push it to the right. To move the box onto the goal, which is two spaces to the right of its current position, I need to push the box right twice. With the first push, I move to (1,2) and the box moves to (1,3). With the second push, I move to (1,3) and the box moves to (1,4), landing on the goal position. So, the correct move here is to push the box right twice, resulting in my movement being 'right, right'. After executing these moves, I will visualize the current game status: <image>



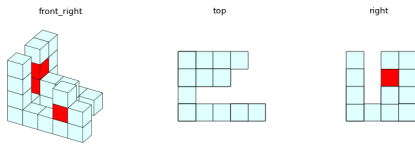
In this final status, the box is now at position (1,4), which is precisely on the goal, completing the task.

To summarize my reasoning: I first moved up twice to reach the left side of the box, as the goal was to the right of the box. From that position, I then pushed the box to the right twice, successfully placing it on the goal.

The final solution is: \bbox{up, up, right, right}.

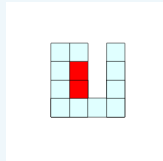
Figure 9: Examples of chain-of-thought SFT data for the maze and sokoban task.

**Cube 3-View Projection**

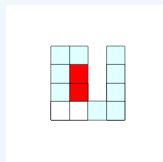


**Question:** A cube stack is made of equal-sized small cubes, mostly in lightcyan with a few red ones. The image shows its isometric view from front-right, top view, and right view from left to right. How many cubes in red can possibly be seen from the left view? Please answer from options: A. 4. B. 6. C. 5. D. 0.

**Response:** <think> First, let's construct the left view based on the cubes that are visible in the provided front-right, top, and right views. This initial construction shows us the cubes whose colors are known for certain.



However, this view is incomplete as it doesn't account for cubes that might be hidden from all given perspectives. To represent this uncertainty, let's regenerate the left view, marking the positions of these potentially hidden cubes in white. This will help us determine the full range of possible red cubes.



This complete left view shows 2 cubes that are confirmed to be red. It also identifies 2 positions (marked in white) where cubes could exist but whose colors are unknown. Since these hidden cubes could also be red, the total number of red cubes visible from the left could range from a minimum of 2 to a maximum of 4. Therefore, the highest possible number of red cubes is 4.</think> A

**Response:** <think> To determine the number of red cubes visible from the left, we must first construct the left view of the cube stack.

An initial attempt to represent this view, based on the cubes visible in the provided images, might look like the following character matrix. In this diagram, 'O' represents a red cube and 'X' represents a lightcyan cube.

```

+---+---+ +---+
| X | X | | X |
+---+---+ +---+
| X | O | | X |
+---+---+ +---+
| X | O | | X |
+---+---+ +---+
| X | X | X | X |
+---+---+ +---+

```

However, this representation is incomplete because some cubes that are visible from the left are occluded in the given views, making their color uncertain. A more accurate diagram should account for these possibilities. Let's refine the matrix, using 'O' for red cubes, 'X' for lightcyan cubes, and '\*' for cubes of an uncertain color.

```

+---+---+ +---+
| X | X | | X |
+---+---+ +---+
| X | O | | X |
+---+---+ +---+
| X | O | | X |
+---+---+ +---+
| * | * | X | X |
+---+---+ +---+

```

This corrected 4x4 view reveals the full structure as seen from the left. Based on this diagram, we can see 2 cubes that are definitely red (marked with 'O'). In addition, there are 2 cubes whose color is uncertain (marked with '\*'). Since each of these uncertain cubes could be red, the total number of red cubes visible from the left could range from a minimum of 2 (if both uncertain cubes are lightcyan) to a maximum of 4 (if both uncertain cubes are red).

Therefore, the number of red cubes that can possibly be seen from the left view is between 2 and 4, inclusive. Since 4 is a possible value and is available in the options, it is a correct answer.</think> A

Figure 10: Examples of chain-of-thought SFT data for the cube 3-view projection task, under visual world modeling (left) and verbal world modeling (right).

**Real-World Spatial Reasoning**

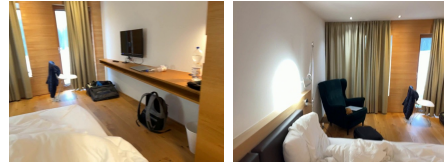


**Question:** If the round clock mounted on the dark wall is to the north of the heater below the windows, and you stand at the heater, in which direction will you find the white door in the room's background?  
Options: A. north B. southwest C. northwest D. east.

**Response:** <think> My goal is to determine the direction of the white door from the heater.  
First, let's establish the reference frame. The problem states that the round clock on the dark wall is north of the heater. This means if I stand at the heater and look towards the clock, I am facing North.  
Let's examine the provided images.  
Image 1 shows the white door, a bookshelf, and a TV on a media console.  
Image 2 shows the heater (a radiator under the window on the right), the clock on a dark gray wall section, a dining table, and the same TV and media console.  
The heater, my starting point, is visible in Image 2. The clock is also visible in Image 2, allowing me to establish the North direction from my position at the heater. The white door, however, is not in this view. It is visible in Image 1.  
The TV and media console are common elements in both images, indicating that the scenes are in the same room. From Image 2, the TV area is to the left of the heater. In Image 1, the door is to the left of the TV area. This suggests the door is further to the left from the perspective of Image 2.  
To find the door from my position at the heater (viewpoint of Image 2), I need to turn my view to the left. I'll simulate a 45-degree turn to the left (a "front left" action) from the viewpoint of the second image to see if the door comes into view.  
<image>



The new image, generated by turning 45 degrees left from the second image, now clearly shows the white door on the left side of the frame.  
Let's put it all together:  
1. I am standing at the heater.  
2. Facing the clock is facing North. From Image 2, the clock is roughly in front of me from the heater's position.  
3. To see the white door, I had to turn 45 degrees to my left.  
4. A 45-degree turn to the left from a North-facing direction is Northwest.  
Therefore, standing at the heater, the white door is in the northwest direction. This corresponds to option C. </think> C



**Question:** When table lamp (in frame 2) is south of pillow (in frame 2), what is the position of tv (in frame 1) in relation to pillow (in frame 2)?  
Options: A. northeast, B. southwest, C. northwest, D. west

**Response:** <think> To solve this spatial reasoning problem, I first need to establish a clear view of all the relevant objects: the table lamp and pillow from the second image, and the TV from the first image. The initial frames show these objects from different perspectives. Frame 1 shows the TV, while Frame 2 shows the pillow and the table lamp. To get a single, comprehensive view, a good strategy would be to move backward from the position of the second image. This should widen the field of view and hopefully bring the TV into sight. Let's move backward from the second image. <image>



With this new perspective providing a combined view, we can now proceed with the reasoning.  
1. Identify the pillow in frame 2: It is a white, fluffy object on the bed, near the dark headboard (left side of the bed in frame 2).  
2. Identify the table lamp in frame 2: It is a white wall-mounted lamp on the left wall (from frame 2's view) near the headboard, emitting light.  
3. The question states the table lamp is south of the pillow, so the pillow is north of the lamp. This defines the north-south axis: the pillow is in the northern direction relative to the lamp.  
4. Identify the TV in frame 1: It is a black flat-screen TV mounted on a white wall above a wooden shelf, to the right of the wooden door with curtains (from frame 1's view).  
5. Relate the TV's position to the pillow's position: The bed (with the pillow) is a common reference. In frame 1, the bed is in the foreground (lower left), and the TV is on a wall that is south and west of the pillow's location (near the bed's headboard).  
6. Confirm the relative position: Based on the north-south and east-west axes, the TV is southwest of the pillow.  
The position of the TV (in frame 1) in relation to the pillow (in frame 2) is `B` (southwest). </think> B

Figure 11: Examples of chain-of-thought SFT data for the real-world spatial reasoning task.

D.3 ANALYTIC EXPERIMENTS

**Sample efficiency.** For Figure 5a, we randomly subsample either 500 or 1000 training examples. The resulting models are evaluated under two settings: (i) a hard setting with the maximum difficulty (grid size 8 and 4 folding steps, default in VisWorld-Eval), and (ii) an in-distribution setting (denoted as Normal in the figure) with randomly sampled grid sizes (3–8) and folding steps (1–4).

**Task difficulties and world model fidelity.** For Figure 5b, we generate test samples with varying cube-stack sizes (3–6), where size 6 is out-of-distribution relative to the training data. To assess world-model fidelity, we compare the generated views with the ground-truth views: for verbal world modeling, we use string pattern matching; for visual world modeling, we use Gemini 3 Pro to compare images. Since accurately inferring colors becomes particularly challenging at larger stack sizes, we evaluate only the shapes of the views and ignore color information. We also find that overall accuracy can be bottlenecked by verbal subskills (e.g., counting holes) after SFT, thus we report the accuracy of RL-trained models in Figure 5b. In contrast, RL can distract verbal world modeling capabilities, leading to invalid formats of generated symbolic matrices, thus we report world-model fidelity of SFT-trained models.

**Implicit world modeling.** For Figure 5c, we supervised fine-tune (SFT) BAGEL on CoTs with implicit world modeling, in which all explicit point coordinates are replaced by the placeholder token sequence <point>masked<point>. After training, we extract the hidden representations at the position of the token masked from each transformer layer. We then split the extracted representations from different CoTs into training and validation sets with an 8:2 ratio and train a two-layer MLP (hidden size 4096) to predict the ground-truth point coordinates. Since all samples are 5 × 5 mazes, we formulate coordinate prediction as two 5-way classification tasks (for x and y, respectively). We compute classification accuracy for each coordinate and report the average of the two.



Figure 12: Probing implicit world models, by training a set of probes, i.e., MLPs which infer the masked point coordinates during reasoning from internal representations.

E EXTENDED EXPERIMENTAL RESULTS

E.1 EVALUATION OF ADVANCED VLMS ON VISWORLD-EVAL

Table 4 reports the zero-shot performance of advanced VLMS on VisWorld-Eval. Overall, these models perform suboptimally, highlighting limitations of current multimodal AI systems. Among them, Gemini 3 Flash and Gemini 3 Pro remarkably outperform the other models; however, their performance remains far from satisfactory on challenging tasks like paper folding, ball tracking, cube 3-view projection, and real-world spatial reasoning.

Table 4: Zero-shot evaluation of advanced VLMS on VisWorld-Eval. We report the average accuracy over five tasks (excluding Maze and Sokoban) and over all seven tasks. †For BAGEL as a UMM, we follow the official visual understanding evaluation protocol, which generates verbal reasoning only.

Models	Paper Folding	Multi-Hop Manip.	Ball Tracking	Cube 3-View	MMSI (Pos. Rel.)	Maze	Sokoban	Overall (5 tasks)	Overall (7 tasks)
<i>Proprietary Models</i>									
Gemini 3 Flash	25.6	75.4	55.3	52.7	41.3	73.9	99.3	50.0	60.5
Gemini 3 Pro	27.0	74.5	44.7	53.3	49.6	33.5	90.2	49.8	53.2
Seed 1.8	10.6	75.2	24.4	42.5	38.8	83.9	68.3	38.3	49.1
GPT 5.1	6.4	73.9	34.8	44.5	44.8	0.6	62.8	40.8	38.2
o3	13.5	68.1	24.7	37.7	44.4	0.0	36.0	37.6	32.0
<i>Open-Source Models</i>									
Qwen3-VL-8B-Thinking (2025a)	11.0	49.3	17.8	21.2	27.7	0.0	5.8	25.4	18.9
BAGEL-7B† (2025)	11.2	31.6	19.4	26.8	27.2	0.0	0.2	23.2	16.6

E.2 COMPARISON BETWEEN SFT-TRAINED VLMS AND UMMs

We present the performance of SFT-trained VLMS compared with UMMs across three tasks in Figure 13, supplementing Section 4.4.

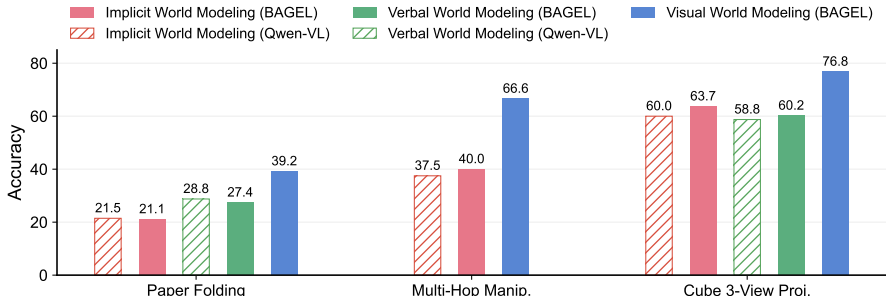


Figure 13: Performance of SFT-trained VLMS compared with UMMs across three tasks.

E.3 FULL RESULTS OF SFT-TRAINED UMMs ON MMSI-BENCH

We report all scores on positional relationship tasks of MMSI-Bench in Table 5.

Table 5: Full results of SFT-trained UMMs on MMSI-Bench positional relationship tasks.

Models	MMSI-Bench (Positional Relationship)						Overall
	Cam.-Cam.	Obj.-Obj.	Reg.-Reg.	Cam.-Obj.	Obj.-Reg.	Cam.-Reg.	
Implicit WM	33.1	31.2	31.8	46.5	29.1	37.3	34.8
Visual WM	29.6	29.5	31.6	60.9	25.8	54.4	38.4

E.4 ADDITIONAL QUALITATIVE EVALUATION

We provide additional qualitative evaluation of trained UMMs’ reasoning, particularly failure cases.

**Real-world spatial reasoning.** Figure 14 (left) provides a showcase of correct interleaved reasoning for real-world spatial reasoning. For failure analysis, as shown in Figure 15a, reasoning with implicit world modeling is prone to hallucinations. In contrast, visual generation (Figure 15b) yields more faithful world models, but still suffers from insufficient quality, including blurring and corrupted details. Moreover, we find that current VLMS and UMMs continue to exhibit limited understanding of positions and directions across different viewpoints. We expect that stronger base models and better-curated post-training data will enable more effective use of visual world models for spatial reasoning in future work.

**Paper folding.** Figure 14 (right) provides a showcase of correct interleaved reasoning for paper folding. For failure analysis, as illustrated in Figure 16, verbal reasoning about geometric symmetry is prone to hallucinations, leading to inaccurate verbal world modeling. In contrast, visual world models, benefiting from stronger prior knowledge, generate correct intermediate unfolding steps even in the presence of erroneous verbal reasoning.

**Cube 3-view projection.** As shown in Figure 17, visual world models are able to approximately generate novel views of cube stacks even in the challenging out-of-distribution setting with an unseen stack size of 6, indicating strong prior knowledge of spatial transformations. Nevertheless, overall task performance remains limited by subtle shape-generation errors (Figure 17b,d) and inaccurate color inference (Figure 17c). We expect these issues to be alleviated through improved post-training and stronger base models.

**Real-World Spatial Reasoning**




**Question:** The pictures are taken continuously from a first-person perspective, with two pictures overlapping at the frame. At the moment of the last picture, in which direction is the potted plant relative to you?  
Options: A: Front right, B: Back left, C: Back right, D: Front left

**Model Output:** <think> My goal is to determine the direction of the potted plant relative to the camera at the end of the second image. First, I need to understand the spatial relationship between the two images. The problem states that the two pictures are taken from a first-person perspective, and at the end, the camera is facing the scene in the second image. This means the second image represents the camera's "front" view at the final moment.


The first image shows a potted plant on a small table, next to a window with curtains and a chair. The second image shows a wall with framed pictures, a teapot cabinet, and a lamp. These are two different views of the same room.

To figure out how these two views relate, I need to determine the camera's movement between the two shots. I'll start from the view in the second image and try to find the view from the first image by rotating the camera. Let's see what happens if I turn the camera "front left" (a 45-degree turn to the left) from its position in the second image. <image>



This new view shows a different part of the room, with a framed picture and the teapot cabinet, but it does not show the potted plant from the first image. This means the plant is not in the "front left" direction.

Now, let's try turning in the opposite direction. I will rotate the camera "front right" (a 45-degree turn to the right) from its original position in the second image. <image>

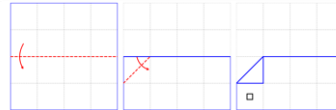


This view is a match! It shows the same framed pictures, lamp, and chair, and importantly, it shows the window with the curtains and the chair that are next to the potted plant in the first image.

This confirms the spatial layout. The camera's "front" view is the scene in the second image. To see the potted plant, I had to turn the camera to the "front right". This means that from the camera's final position (facing the second image), the potted plant is located in the front-right direction.


Therefore, the correct option is A: Front right. </think> A

**Paper Folding**



**Question:** Analyze the image showing a folded paper with cutouts. Mentally reverse the folding process to reconstruct the final unfolded design, then provide your answer to: how many cutouts are there in the unfolded paper?

**Model Output:** <think> Let's start by enlarging the rightmost image in the upper area, the one displaying the folded paper's cutting pattern. <image>




Let's systematically, step by step, to reveal the final appearance of the completely unfolded paper. And then answer the question accordingly. Let's unfold the paper step-by-step to see how the final pattern of holes is created.

First, we'll reverse the second and final fold. This was a diagonal fold where the top-left corner was folded down over the main body of the paper. The fold line is the diagonal connecting the top-left and bottom-left corners. To reverse this, we unfold the small triangular flap back up and to the left.

The key principle here is that unfolding creates a mirror image of any holes located on the moving flap, using the fold line as the axis of symmetry.

In this step, the square hole is located in the bottom-left section of the paper. This area was the stationary part of the paper during the fold; it was not the flap that was folded over.

Therefore, since the hole is not on the moving flap, it does not create a mirror image when we unfold the paper. The action simply restores the paper to its previous rectangular shape, leaving the single square hole unaffected in its original position. <image>




Now, we will reverse the first fold. This was a horizontal fold where the top half of the paper was folded down over the bottom half. The fold line is the horizontal centerline of the paper. Unfolding the top half back up will mirror the existing hole across this centerline.

Let's examine the hole's position. The single square hole is located in the bottom-left quadrant of the folded paper. When we unfold the paper along the horizontal centerline, this hole will be mirrored onto the top-left quadrant.

This mirroring process creates a new, symmetrical hole. It is positioned directly above the original hole, equidistant from the horizontal fold line. Because the original hole is in the bottom-left, its reflection appears in the top-left, creating a vertical pair of holes.

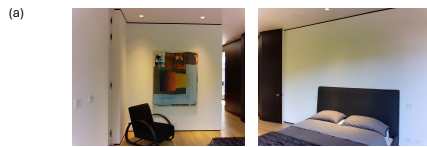
Consequently, after this final unfolding step, we have two square holes: the original one in the bottom-left and its new, mirrored counterpart in the top-left. <image>



Finally we can calculate that the number of cutouts is 2. </think> 2

Figure 14: Showcases of interleaved verbal-visual chain-of-thought reasoning, generated by post-trained UMMs, where visual generation serves as world models. <image> denotes a placeholder indicating the position of a generated image.

**Real-World Spatial Reasoning**



**Question:** The images are taken continuously from a first-person perspective. At the moment of the last image, in which direction is the black chair relative to you?  
Options: A: Front right, B: Back left, C: Front left, D: Back right

**Model Output (Implicit World Modeling):** <think> To solve this, I need to determine the direction of the black chair relative to the camera's position in the last image.

First, I'll analyze the two initial images to build a mental map of the room.

Image 1 shows a black chair in front of a white wall with a large painting. To the right of the painting, there's a dark hallway.

Image 2 shows a bed with a black headboard against a white wall. To the left of the bed, there's a dark door.

The hallway seen on the right in Image 1 appears to be the same hallway that starts on the right side of the bed in Image 2. This suggests the two images are taken from different positions in the same room, looking in roughly opposite directions.

The question asks for the chair's direction relative to the camera at the moment of the last image. The last image is Image 2, which shows the bed. This means the camera's starting position is near the bed, looking towards it.

To find the chair, I need to explore the surroundings from this starting position. I'll simulate turning the camera to see what's around the bed. I'll start by turning left from the perspective of Image 2.

This new view shows a desk and a window. The black chair is not here. This means the chair is not to the left of the bed.

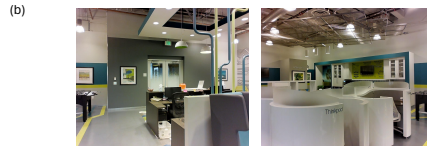
Now, I'll return to the original perspective of Image 2 (facing the bed) and try turning in the opposite direction. I will turn right.

This view reveals the dark hallway seen in the first image, along with the painting on the wall. The black chair is located in this area, as seen in the first image.

So, to summarize the findings:

1. My starting position is at the camera for Image 2, facing the bed.
2. To see the black chair, I had to turn around 180 degrees to face the opposite direction (towards the painting).
3. A direction that is 180 degrees opposite to "front" is "back".

Therefore, the black chair is in the "back" direction relative to my position in the last image. Looking at the options, "Back left" (B) and "Back right" (D) are possibilities. Since turning left from my starting position did not reveal the chair, the chair must be on the right side of my original viewpoint. Combining these facts, the chair is located in the back-right direction. </think> D



**Question:** When you enter the room through the white door, which way should you go to reach the TV area?  
Options: A: Go straight, B: Go left, C: Cannot be determined, D: Go right

**Model Output (Visual World Modeling):** <think> My goal is to determine the direction from the white door to the TV area. The first image shows a white door in the background, and the second image shows a TV area. To understand their relative positions, I need to explore the space. Starting from the first image, I will move forward to get a better view of the room.



After moving forward, the view is wider. The white door is still visible, and the TV area seems to be to the left. To confirm this, I will turn slightly to the left from this new position.



After turning left, the TV area is now more clearly visible. To get a final, comprehensive view that includes both the white door and the TV area, I will move backward from this current position.



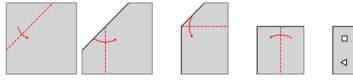
Now that I have a clear overview of the room's layout, I can determine the direction.

1. **Identify the white door:** In the first image, the white door (with glass panels) is visible in the background, slightly to the right of the center.
2. **Locate the TV area:** The second image shows the TV area, which includes a TV on a yellow wall and the "thinkpad" sign. This area is consistent with the left side of the first image (via matching blue/green wall art and foosball table).
3. **Determine direction:** When facing the white door (from the entrance), the TV area is positioned to the left. This is confirmed by the spatial relationship: moving left from the door's position aligns with the TV area's location.

The correct option is `B` (go left). </think> B


Figure 15: Showcases of reasoning generated by post-trained UMMs in the real-world spatial reasoning task. We highlight hallucinations or incorrect reasoning steps in red.

Paper Folding



**Question:** Analyze the image showing a folded paper with cutouts. Mentally reverse the folding process to reconstruct the final unfolded design, then provide your answer to: calculate the number of triangle\_left minus the number of triangle\_right.


**Model Output (Visual World Modeling):** <think> First, zoom into the upper section's rightmost image showing the cutting pattern on the folded paper.



Let's systematically, step by step, to reveal the final appearance of the completely unfolded paper. And then answer the question accordingly. Let's begin by unfolding the paper step-by-step, starting with the final fold made.

First, we will reverse the fourth and final fold. This was a vertical fold where the left side of the paper was folded over the right. To undo this, we unfold the left side back out, using the vertical crease as a line of symmetry. The principle of unfolding is that any cuts on the moving flap are mirrored across the fold line.


In this case, the existing cutouts—the square and the triangle—are located on the right-hand portion of the paper, which remains stationary. The flap being unfolded is blank. Therefore, as we unfold the left side, **no new shapes are created by mirroring. The original square and triangle simply remain in place**, leaving the newly revealed left portion of the paper as a solid, uncut surface.



Next, we reverse the third fold. This was a horizontal fold where the top half was folded down over the bottom half. We now unfold the paper upwards, using the horizontal crease as an axis of reflection. All existing cutouts on the bottom half will be mirrored onto the newly revealed top half. Let's trace the reflection of each shape:

- The square on the bottom half is reflected across the fold line, creating an identical square directly above it.
- Similarly, the left-pointing triangle on the bottom half is reflected, creating a new, identical left-pointing triangle directly above the original.
- Finally, the right-pointing triangle on the bottom half is also reflected, creating a new, identical right-pointing triangle in the top half.


By combining the original cutouts with their newly created reflections, we arrive at the pattern shown in the image.



Continuing the process, we now undo the second fold. This was a vertical fold where the left section was folded over the right. We will unfold this flap back to the left, with the vertical crease acting as our line of reflection. The shapes on this moving flap will be mirrored onto the newly exposed area.

The key change here involves the small, diagonal cutout at the top of the folded flap. This cutout was punched through both layers of paper. When we unfold the flap, the original cutout on the flap and the corresponding cutout on the layer beneath it combine to form a new, complete square.

The other existing cutouts—the full squares and the two triangles—were not on this moving flap, so they remain unaffected by this step. By completing the square while leaving the other shapes untouched, we arrive at the configuration shown.



[.]</think> -1

**Model Output (Verbal World Modeling):** <think> Begin by zooming in on the rightmost picture in the upper section that illustrates the cutting pattern on the folded paper.

```

...
[[-1, -1, -1],
 [-1, 1, -1],
 [-1, 1, -1]]

[['', '', ''],
 ['', 'diamond', ''],
 ['', 'triangle_left', '']]
...

```

Let's systematically, step by step, to reveal the final appearance of the completely unfolded paper. And then answer the question accordingly. Let's unfold the paper step by step, reversing each fold to reveal the final pattern.

First, we will reverse the fourth and final fold. This was a vertical fold where the right half of the paper was folded onto the left half. To unfold it, we open the paper from left to right. The vertical crease acts as a line of symmetry, meaning any cuts on the folded portion will be mirrored onto the newly revealed section.

Let's analyze the shapes on the folded paper:

- The square cutout is on the stationary left side, so it remains in its original position.
- The left-pointing triangle is on the folded-over section. As we unfold, this cutout is mirrored across the vertical crease. A reflection across a vertical line flips a shape horizontally, so the left-pointing triangle creates a new, right-pointing triangle on the right side of the paper.

Therefore, after this first unfolding step, the original square and left-pointing triangle remain, and a new, symmetrically placed right-pointing triangle is added.

```

...
[[-1, -1, -1],
 [-1, 1, 1],
 [-1, 1, 1]]

[['', '', ''],
 ['', 'diamond', 'diamond'],
 ['', 'triangle_left', 'triangle_right']]
...

```

Now, let's reverse the third fold. This was a horizontal fold where the top half was folded down. Unfolding it involves lifting that section back up. The horizontal crease again acts as a line of symmetry. Any cuts on the folded portion will be mirrored onto the newly revealed top section.

Let's trace the existing shapes from the previous step:

- The square and the left-pointing triangle are on the stationary bottom half, so they remain in place.
- The right-pointing triangle is on the folded top section. As we unfold this flap, the cutout is mirrored across the horizontal crease. A reflection across a horizontal line creates a vertical mirror image, so the right-pointing triangle generates a new, downward-pointing triangle on the top half of the paper.

After this step, the existing shapes on the bottom half remain, and the newly created downward-pointing triangle appears on the top half.

```

...
[[-1, 1, 1],
 [-1, 1, 1],
 [-1, 1, 1]]

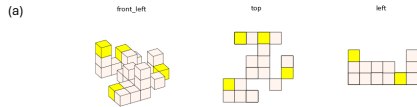
[['', 'triangle_down', 'triangle_right'],
 ['', 'diamond', 'diamond'],
 ['', 'triangle_left', 'triangle_right']]
...

```

[.]</think> -1

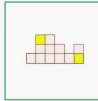
Figure 16: Showcases of reasoning generated by post-trained UMMs in the paper folding task. We highlight hallucinations or incorrect reasoning steps in red, but also mark correctly generated visual unfolding intermediate steps with green borders.

**Cube 3-View Projection**

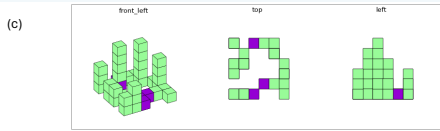


**Question:** A cube stack is made of equal-sized small cubes, mostly in seashell with a few yellow ones. The image shows its isometric view from front-left, top view, and left view from left to right. How many cubes in yellow can possibly be seen from the front view?  
Please answer from options: A. 2. B. All three other options are possible. C. 4. D. 1.

**Model Output:** <think> To determine the number of yellow cubes visible from the front, we first need to construct the front view of the stack. By synthesizing the provided isometric, top, and left views, we can generate the required front view.

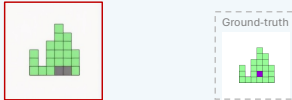


Examining the generated front view, we can directly count the number of visible yellow cubes. There are 2 yellow cubes visible. Therefore, the correct answer is 2, which corresponds to option A. </think> A

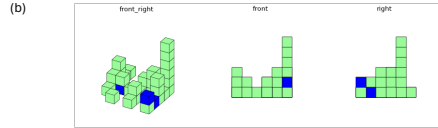


**Question:** A cube stack is made of equal-sized small cubes, mostly in palegreen with a few darkviolet ones. The image shows its isometric view from front-left, top view, and left view from left to right. How many cubes in darkviolet can possibly be seen from the right view?  
Please answer from options: A. 3. B. All three other options are possible. C. 1. D. 2.

**Model Output:** <think> To determine the number of darkviolet cubes visible from the right, we first need to construct the right view of the stack. The colors of some cubes visible from the right might be unknown or 'white' because they are occluded in the given front-left, top, and left views. We will mark these cubes in gray to represent this uncertainty.

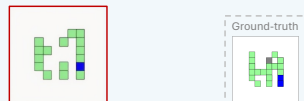


The resulting right view is shown above. We can see 0 cubes that are confirmed to be darkviolet. However, there are 2 cubes (shown in gray) whose color is unknown. Since these two unknown cubes could potentially be darkviolet, the total number of darkviolet cubes visible from the right could be 0, 1, or 2. Therefore, all three options (1, 2, and 3) are possible. </think> B

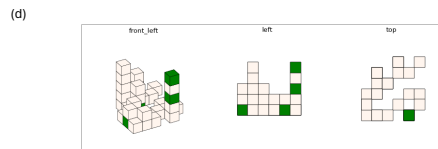


**Question:** A cube stack is made of equal-sized small cubes, mostly in palegreen with a few blue ones. The image shows its isometric view from front-right, front view, and right view from left to right. How many cubes in blue can possibly be seen from the top view?  
Please answer from options: A. 4. B. 5. C. 3. D. 0.

**Model Output:** <think> To determine the number of blue cubes visible from the top, we first need to construct the top view of the stack. By synthesizing the information from the given front and right views, we can generate the top view as shown below.



Examining the generated top view, we can count the number of blue cubes. There is exactly 1 blue cube visible. However, we must also consider the possibility of occluded cubes whose colors might be unknown from the provided front and right views. [...] </think> C



**Question:** A cube stack is made of equal-sized small cubes, mostly in seashell with a few green ones. The image shows its isometric view from front-left, left view, and top view from left to right. How many cubes in green can possibly be seen from the front view?  
Please answer from options: A. All three other options are possible. B. 0. C. 4. D. 2.

**Model Output:** <think> To determine the number of green cubes visible from the front, we first need to construct the front view of the stack. By synthesizing the provided isometric, left, and top views, we can generate the required front view.



Examining the generated front view, we can observe the visible green cubes. There are 2 green cubes visible. Therefore, the correct answer is 2, which corresponds to option D. </think> D

Figure 17: Showcases of reasoning generated by post-trained UMMs in the paper folding task. We mark correct and incorrect generated cube views with green and red borders, respectively. For incorrect generations, the corresponding ground-truth views are provided for reference (note that these are shown only for readers and are never provided to the models during reasoning).