When and How Much to Imagine: Adaptive Test-Time Scaling with World Models for Visual Spatial Reasoning

Shoubin Yu[*1] **Yue Zhang**[*1] **Zun Wang**[1]
Jaehong Yoon[2] **Huaxiu Yao**[1] **Mingyu Ding**[1] **Mohit Bansal**[1]

[1]University of North Carolina, Chapel Hill [2]Nanyang Technological University

https://adaptive-visual-tts.github.io

Abstract

Despite rapid progress in Multimodal Large Language Models (MLLMs), visual spatial reasoning remains unreliable when correct answers depend on how a scene would appear under unseen or alternative viewpoints. Recent work addresses this by augmenting reasoning with world models for visual imagination, but questions such as when imagination is actually necessary, how much of it is beneficial, and when it becomes harmful, remain poorly understood. In practice, indiscriminate imagination can increase computation and even degrade performance by introducing misleading evidence. In this work, we present an in-depth analysis of test-time visual imagination as a controllable resource for spatial reasoning. We first study when static visual evidence is sufficient, when imagination improves reasoning, and how unnecessary imagination affects accuracy and efficiency. To support this analysis, we then introduce AVIC, an adaptive test-time framework with world models that explicitly reasons about the sufficiency of current visual evidence before selectively invoking and scaling visual imagination. Finally, to further learn this gating and planning behavior without any annotation of when and how much to imagine, we introduce AVIC-R, which trains the policy end-to-end via GRPO from QA-correctness rewards and penalties by imagination cost. Across spatial reasoning benchmarks (SAT, MMSI) and an embodied navigation benchmark (R2R), our results reveal clear scenarios where imagination is critical, marginal, or detrimental, and show that selective control can match or outperform fixed imagination strategies with substantially fewer world-model calls. Our AVIC-R surpasses strong proprietary baselines including GPT-4o and GPT-4.1 while invoking the world model less often. Overall, our findings highlight the importance of analyzing and controlling test-time imagination for efficient and reliable spatial reasoning.

1 Introduction

Recent advances in multimodal large language models (MLLMs) [19, 22] have led to impressive progress in visual understanding and reasoning across various tasks. These models can follow natural language instructions, perceive visual scenes, and reason over multimodal input to support decision making. Despite the progress, *visual spatial reasoning* remains a persistent challenge [41, 7, 30, 33], particularly for questions whose answer depends on unseen regions, viewpoint changes, or transformations that cannot be reliably inferred from a single static observation.

A natural way to address this challenge, mirroring how humans operate, is through *visual imagination* [18]: when the observed visual evidence is insufficient, people mentally simulate how a scene

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: Different cases in always-on visual imagination. Imagined views are generated independently for different beam-searched actions (shown by multiple arrows). Case 1 (Helpful): Visual imagination reveals previously unseen viewpoints, enabling helpful spatial reasoning. Case 2 (Misleading imagined evidence): Imagination fails to preserve a task-relevant object (the white table in the red box). Case 3 (Unnecessary): The required information is already clearly observable in the original view (e.g., the bathtub in the blue box), making additional imagined views redundant.

would appear from alternative viewpoints or after potential movements, leveraging strong world priors learned from years of physical interaction and visual experience. Inspired by this intuition, recent work [43, 5, 28] has begun to integrate MLLMs with visual world models that can generate controlled novel views conditioned on hypothetical action at inference time. However, existing approaches often invoke visual imagination using fixed and exhaustive strategies (see Figure 1), without first reasoning about whether additional imagination is necessary and helpful. This lack of deliberation can lead to problematic imagination, producing misleading (Figure 1 (b)) or redundant (Figure 1 (c)) views that not only incur substantial computational overhead but can also distract downstream reasoning and result in worse performance than relying on the original observation alone. Through a systematic analysis of always-on imagination (Section 3), we show that such strategies are both inefficient and unreliable, motivating the need for more adaptive use of world models.

Based on these observations, we aim to answer two fundamental questions for visual spatial reasoning with world model imagination: *when* should a model invoke visual imagination, and *how much* imagined visual evidence is necessary if imagination is required. Rather than treating visual imagination as an always-on operation, we seek to make it a controllable, self-adaptive component during inference time. In this paper, we introduce **A**daptive **V**isual **I**magination **C**ontrol (**AVIC**), a framework that gates and plans world-model usage with an explicit policy model. Given an observation and a question, the policy first reasons about the sufficiency of the available visual evidence and conditionally decides whether to invoke the world model. If it decides not to, it answers directly from the observed view; otherwise, it generates a dynamic-length action plan that specifies how the imagination should move or reorient to acquire informative viewpoints, which are rendered by the visual world model and consumed by a downstream reasoner. A trajectory-level verifier then selects the most informative imagined trajectory among multiple policy samples, in contrast to prior beam-search approaches that score isolated keyframes. It enables instance-dependent test-time scaling, allowing us to move beyond fixed or exhaustive imagination strategies and to study visual spatial reasoning systematically.

While AVIC can be instantiated in a training-free manner via prompting and self-consistency sampling from a strong MLLM, training a stronger policy is non-trivial: no ground-truth supervision exists for what the optimal imagination trajectory should be, ruling out standard supervised fine-tuning or behavior cloning. To overcome this challenge and learn this behavior end-to-end, we further propose **AVIC-R**, which trains the gating policy on top of small open-source model (Qwen2.5-VL [4]) via reinforcement learning, using a composite reward built around QA correctness that requires no human supervision on *when* imagination is necessary. Specifically, we adopt Group-Relative Policy Optimization (GRPO) [32] to compute group-normalized advantages over rollouts that share the same prompt; the reward augments QA correctness with an action-count cost that discourages over-imagination, and a wrong-skip penalty that prevents the policy from collapsing to always-skip WM calling. These designs enable the policy model to discover better gating and planning behaviors from the QA accuracy and WM cost.

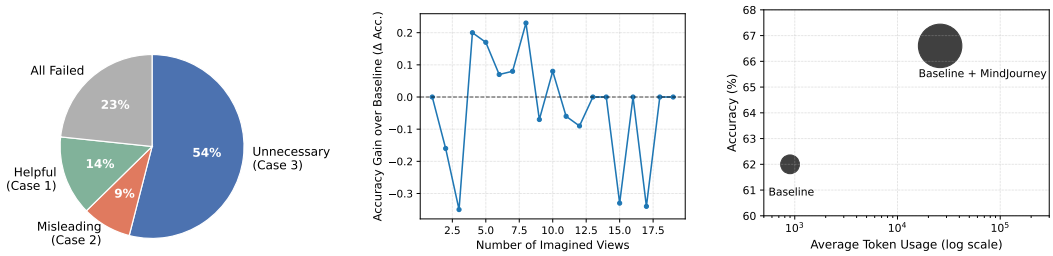
We evaluate the proposed framework on challenging spatial reasoning benchmarks (SAT [30], MMSI [42]), and the navigation benchmark R2R [2]. Across these settings, adaptive test-time scaling achieves SoTA or competitive performance while requiring substantially fewer extra language tokens and world-model calls compared to fixed imagination strategies. Notably, with only a small training set and LoRA updates, AVIC-R boosts a 7B open-source policy enough that the resulting pipeline outperforms variants using GPT-4o or GPT-4.1 as the policy model. Overall, beyond improved performance, our results reveal that the benefits of visual imagination are highly instance-dependent and structured by the nature of the spatial reasoning query. In particular, we find that world models are most beneficial for action-conditioned spatial reasoning, where answers depend on how a scene would evolve under specific movements or viewpoint changes, while offering limited gains for queries that can be resolved from existing observations. At the same time, our analysis shows that effective visual spatial reasoning typically requires only targeted imagination, and excessive or indiscriminate simulation can introduce noise and degrade performance. Together, these findings indicate that visual imagination is a selective, query-dependent test-time resource, requiring adaptive, uncertainty-aware allocation of world-model computation.

2 Related Work

Visual Spatial Reasoning with MLLMs. The rapid evolution of Multimodal Large Language Models (MLLMs) has made significant progress in various downstream tasks [23, 29, 25, 12, 54, 48, 49, 47, 10, 46, 8]. In particular, spatial reasoning has attracted considerable attention due to its critical role in bridging visual perception with embodied tasks [56, 38, 55, 57]. However, recent comprehensive evaluations indicate that current MLLMs still struggle with robust spatial reasoning [41, 7, 30, 33]. While recent efforts aim to enhance spatial capabilities through scaling training data [6, 14, 21] or chain-of-thought prompting [52, 17], they fundamentally process visual information as static 2D snapshots. In contrast, robust spatial reasoning requires an active and dynamic process where the agent can selectively acquire new visual evidence, similar to human mental simulation.

World Models and Visual Imagination. Recent advances in video generation have demonstrated the potential of serving as world models, enabling agents to imagine future frames or outcomes for improved decision-making [60, 9, 20, 15, 28]. This capability is further boosted by the emergence of controllable video generation, which allows for action-conditioned simulation [11, 3, 50, 37]. Notably, recent works such as MindJourney [43] have pioneered the use of world models to enhance visual spatial reasoning by synthesizing novel viewpoints. However, their model blindly generates a set number of views regardless of the question's difficulty or necessity. In contrast, we show that the utility of visual imagination is highly query-dependent, motivating selective rather than uniform use of world models at test time.

Test-Time Scaling. Test-time scaling (TTS) improves performance by allocating additional inference computation without retraining. Prior work has explored various scaling strategies in language/visual-language models, including self-consistency [34], tree-based search [44, 36], verifier-guided method [24, 51], and (multimodal) CoT [40, 39, 35]. In the visual spatial reasoning, recent works [43, 5] realized through generating novel views and ensembling, but typically apply uniform computation across instances. Our method introduces adaptive visual test-time scaling, enabling targeted imagination only when necessary and improving computational efficiency.



(a) Statistics of Different Cases of Always-on Imagination

(b) Performance Gain v.s. Amount of Imagined Views

(c) Trade-off Among Accuracy, Token Numbers and Inference Time

Figure 2: **(a)**: In the majority of cases, visual imagination is unnecessary, while a smaller fraction is helpful or misleading, highlighting the need for selective invocation rather than uniform use. **(b)**: Accuracy gain over the baseline over the number of imagined views. Performance improvements are non-monotonic, indicating that additional imagination does not consistently translate to better reasoning and may even degrade accuracy when there are too many generated views. **(c)**: Accuracy v.s. average token usage. Bubble size indicates average running time. Fixed imagination strategies achieve higher accuracy at the cost of substantially increased computation, motivating adaptive test-time scaling that balances performance and efficiency.

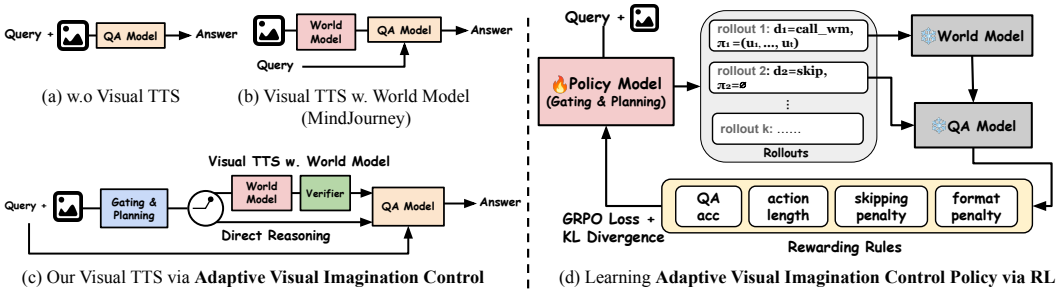
3 Analysis of Always-on World Model Calling

We consider a test-time setting where an MLLM is equipped with a visual world model that generates imagined observations from hypothetical viewpoints. Existing methods commonly invoke the world model in an always-on, which calling it on every instance and exhaustively exploring action branches, implicitly assuming that additional imagination is consistently beneficial. In practice, however, this incurs substantial computational cost and may yield ambiguous or noisy observations, making imagined views redundant when the answer is already evident from the initial observation and outright misleading when the world model produces noise.

To diagnose this strategy, we categorize each instance on SAT-Real [30] into three cases (Figure 1):

- **Case 1 (Imagination Helpful):** The model calls the world model and produces a correct answer, indicating that imagined views provide beneficial spatial information.
- **Case 2 (Misleading Imagined Evidence):** The model produces an incorrect answer after using imagined views. This outcome can reflect rendering errors, uninformative trajectories, or failure to use the imagined evidence.
- **Case 3 (Imagination Unnecessary):** The model produces a correct answer without calling the world model, suggesting that visual imagination is redundant.

We then examine three aspects of always-on behavior in Figure 2: case distribution, performance vs. amount of imagination, and computational cost. **(1) Case distribution.** As shown in Figure 2a, the majority of instances (54%) fall into Case 3, where the model already answers correctly without any world model invocation, while imagination is genuinely helpful in only 14% (Case 1), indicating that always-on imagination is unnecessary for most instances. **(2) Performance vs. amount of imagination.** Figure 2b shows that adding more imagined views does not consistently improve accuracy and even degrades it, suggesting that simply increasing imagination is not an ideal strategy. **(3) Cost–accuracy trade-off.** While always-on imagination yields only a 4.6% accuracy gain over the baseline, it requires nearly two orders of magnitude more tokens and about $30\times$ higher inference time (Figure 2c), a steep computational price for limited return. **(4) Selective imagination upper bound.** We further quantify the potential of selective WM usage by assuming imagination is applied only when it leads to a correct prediction. The baseline reaches **62.0%** on SAT-Real and always-on imagination only marginally improves to **66.6%**, whereas this selective upper bound jumps to **75.3%**, demonstrating that selective imagination policies are strongly motivated. The 75.3% selective upper bound applies to the fixed imagination procedure in this diagnostic, rather than to all possible trajectories or to AVIC-R; Appendix C.5 reports oracle interventions for AVIC-R. These results motivate selective, adaptive imagination.



(a) w.o Visual TTS (b) Visual TTS w. World Model (MindJourney)

(c) Our Visual TTS via **Adaptive Visual Imagination Control** (d) Learning **Adaptive Visual Imagination Control Policy via RL**

Figure 3: (a) Direct QA from the current observation. (b) Always-on world-model exploration. (c) Ours: a policy model decides *whether* and *how much* to imagine, selectively querying the world model only when warranted. (d) Our RL training loop for the gating policy: The policy model samples a set of rollouts that either query the world model (`call_wm`) or bypass it (`skip`). A frozen QA model answers them, and a four-component reward drives a GRPO update on the policy model.

4 Adaptive Visual Imagination Control

We now introduce **AVIC (ADAPTIVE VISUAL IMAGINATION CONTROL)**, an adaptive test-time framework that selectively invokes a world model only when additional visual evidence is likely to be useful (Figure 3c), in contrast to the always-on baseline analyzed in Sec. 3. We first formalize the problem (Sec. 4.1), then describe the framework (Sec. 4.2), and finally describe an RL procedure that trains the gating policy from QA Model and WM rewards (Sec. 4.3).

4.1 Problem Formulation

We consider a visual spatial reasoning task defined by an input tuple $\langle I, q, \mathcal{A} \rangle$, where I is the current egocentric observation (one or multiple views), q is a multiple-choice question, and $\mathcal{A} = \{a_1, \ldots, a_K\}$ is the answer set. The correct answer may depend on spatial relations that are ambiguous, occluded, or unobserved in I. The agent may optionally invoke a visual world model that renders novel imagined observations I_π from a sequence of egocentric actions π, with $I_\pi = \varnothing$ when no imagination is used. The predicted answer is $\hat{a} = \arg\max_{a \in \mathcal{A}} P\big(a \mid I, I_\pi, q\big)$.

4.2 Adaptive Imagination Framework

AVIC couples a **policy model** that gates and plans imagination with a **trajectory-level verifier** that picks a single targeted imagined trajectory for downstream reasoning.

Policy gating with test-time scaling. A policy θ maps $(I, q, \mathcal{A})$ to a decision $d \in \{\texttt{skip}, \texttt{call_wm}\}$ together with a short discrete action plan π drawn from a fixed low-level egocentric action space $\mathcal{U}$:

$$(d, \pi) \sim \theta(d, \pi \mid I, q, \mathcal{A}), \quad \pi = \begin{cases} \varnothing, & d = \texttt{skip}, \\ (u_1, \ldots, u_T),\ u_t \in \mathcal{U}, & d = \texttt{call_wm}. \end{cases} \tag{1}$$

Here $\mathcal{U}$ contains `turn-left` (9°), `turn-right` (9°), and `move-forward` (0.25 m), following MindJourney. A plan contains at most six atomic actions; repeated primitives compose larger motions, with cancelling or oscillatory sequences disallowed. Plans are converted to relative camera poses for world-model conditioning. We sample the policy M times ($M = 5$ by default) and apply majority voting only to the gating decision d. This vote determines whether to imagine but not select a plan.

Action execution and trajectory selection. When the majority decision is `call_wm`, the world model renders the non-empty candidate plans from the policy samples, producing $\mathcal{I}_{\pi^{(m)}} = W(I, \pi^{(m)})$. The verifier scores each complete rendered trajectory, and only the highest-scoring trajectory is passed to the reasoner; it is not applied only to one preselected plan. Different policy samples produce trajectories of varying usefulness; unlike prior beam-search approaches [43] that score isolated keyframes, we evaluate the *entire* trajectory as a coherent unit via a verifier V, preserving temporal and geometric consistency across sequential actions. Such that:

$$s^{(m)} = V(I, q, \mathcal{I}_{\pi^{(m)}}), \quad \pi^* = \arg\max_{\pi^{(m)}} s^{(m)}, \tag{2}$$

Final prediction. A vision-language reasoner ϕ predicts the answer using the original observation and the selected imagined views (or I alone when the gate selected `skip`, in which case $I_{\pi^*} = \varnothing$):

$$\hat{a} = \arg\max_{a \in \mathcal{A}} P_\phi(a \mid I, I_{\pi^*}, q)\,. \tag{3}$$

4.3 Learning When and How Much to Imagine via RL

Learning when and how much to imagine is non-trivial, as no ground-truth supervision exists for what the optimal imagination trajectory should be, ruling out supervised fine-tuning or behavior cloning. We therefore propose AVIC-R, which trains θ end-to-end with reinforcement learning, using QA correctness as the primary signal augmented by lightweight reward shaping (Figure 3 (d)).

Training loop. For each question, the policy θ samples K rollouts $\{(d^{(i)}, \pi^{(i)})\}$. Each rollout is executed by frozen environment modules: `call_wm` rollouts query the world model W to render imagined views, while `skip` rollouts bypass it; both are answered by a frozen QA model ϕ. The resulting reward signals are aggregated into a GRPO update on the policy, while W and ϕ remain frozen throughout training. The verifier is used only at inference. QA correctness is computed by comparing the frozen QA model's predicted answer with the dataset ground-truth label. We also discussed dependence on the training reasoner in Appendix C.4.

Reward design. We assemble a composite reward with 4 components, as illustrated in Figure 3 (d):

$$r \;=\; \underbrace{\mathbf{1}_{\text{correct}}}_{\text{(i) QA correctness}} \;-\; \underbrace{c\,|\pi|}_{\text{(ii) action length}} \;-\; \underbrace{\beta_s\,\mathbf{1}_{\text{wrong-skip}}}_{\text{(iii) wrong-skip}} \;-\; \underbrace{\beta_p\,\mathbf{1}_{\text{parse-fail}}}_{\text{(iv) format}} \tag{4}$$

where $\mathbf{1}_{\text{correct}}$, $\mathbf{1}_{\text{wrong-skip}}$, and $\mathbf{1}_{\text{parse-fail}}$ indicate, respectively, that $\hat{a} = a^*$, that $d{=}\texttt{skip} \wedge \hat{a} {\neq} a^*$, and that the output schema cannot be parsed; $|\pi|$ counts atomic actions; and $c, \beta_s, \beta_p > 0$. Each term targets a specific failure mode:

- **QA correctness** is the only positive signal; all other terms are deductions.
- **Action-length cost** discourages over-imagination, a longer correct chain still scores below a shorter correct one, pushing the policy toward concise plans.
- **Wrong-skip penalty** is essential. Without it, an incorrect `skip` pays nothing while an incorrect `call_wm` still pays $c|\pi|$, biasing the policy toward skipping under uncertainty.
- **Format penalty** handles unrecoverable schema errors. We deliberately keep $\beta_{\text{p}} \approx \beta_{\text{s}}$ rather than larger: a heavier format penalty would incentivize the policy to collapse onto trivial outputs (e.g., always `skip` with empty π) merely to avoid parse failures.

We set $c = 0.1$ and $\beta_{\text{s}} = \beta_{\text{p}} = 0.5$, yielding a clean qualitative ordering: a correct `skip` scores $+1.0$, a correct `call_wm` with n actions scores $1 - 0.1n$, and a wrong `skip` (-0.5) is strictly worse than a short wrong `call_wm` (-0.1 to -0.3). The goal of this asymmetry is to prevent the policy from collapsing to *always-skip* under uncertainty: when in doubt, calling WM with a short plan is the safer bet than skipping. The action cost $c|\pi|$ then discourages over-long imagination chains.

GRPO objective. Without an external value function, we adopt Group-Relative Policy Optimization [32], which estimates advantages directly from the K rollouts sharing a prompt:

$$A_i = \frac{r_i - \mu_q}{\sigma_q + \epsilon}, \quad \mu_q = \tfrac{1}{K}\sum_j r_j, \quad \sigma_q^2 = \tfrac{1}{K}\sum_j (r_j - \mu_q)^2. \tag{5}$$

Group normalization absorbs per-question difficulty, where uniformly easy or uniformly hard groups contribute zero gradient. We optimize a token-level PPO-clipped objective [31] regularized by KL to a frozen reference θ_{ref}:

$$\mathcal{L} = -\mathbb{E}_{(q,i,t)}\big[\min\big(\rho_{i,t}A_i,\ \text{clip}(\rho_{i,t}, 1{-}\epsilon_c, 1{+}\epsilon_c)A_i\big)\big] + \beta_{\text{KL}}\,\widehat{\text{KL}}(\theta \| \theta_{\text{ref}})\,, \tag{6}$$

with per-token importance ratio $\rho_{i,t} = \theta(y_{i,t} \mid q, y_{i,<t}) / \theta_{\text{old}}(y_{i,t} \mid q, y_{i,<t})$. Only LoRA adapters in θ are updated; the KL anchor preserves the base VLM's general capabilities while shaping only gating and planning behaviors. Training hyperparameters, data curation, output parsing, and training-compute accounting are detailed in Appendix A.

Table 1: Comparison between TTS methods on SAT-Real. The best results are denoted by **bold**, and the second-best are underlined. **Avg. WM**: average world model calling times over the dataset.

Method	Policy Model	EgoM	ObjM	EgoAct	Goal	Pers	Avg.	# Token (K)	Avg. WM
InternVL3-14B [61]	–	56.5	69.5	54.0	73.5	45.4	59.3	0.2	0
+ MindJourney	–	69.6	60.9	**78.4**	79.4	42.4	66.7	2.5	12.34
+ AVIC	InternVL3-14B	**95.6**	**73.9**	62.1	76.4	42.4	68.0	2.0	0.64
+ AVIC	Qwen2.5VL-7B	73.9	47.8	67.5	73.5	42.4	61.3	4.4	1.81
+ AVIC-R	Qwen2.5VL-7B	82.6	52.1	70.2	**85.2**	**54.5**	**69.3**	4.8	3.03
GPT-4o [27]	–	56.5	**85.0**	50.0	64.0	45.0	60.3	0.9	0
+ MindJourney	–	78.3	60.9	78.4	70.6	57.5	69.3	26.0	12.34
+ AVIC	GPT-4o	**86.9**	60.9	64.8	82.3	48.4	69.3	9.5	0.72
+ AVIC	Qwen2.5VL-7B	65.2	73.9	64.8	**91.1**	**60.6**	71.3	5.0	1.81
+ AVIC-R	Qwen2.5VL-7B	82.6	82.6	**81.0**	**91.1**	51.2	**77.3**	5.4	3.03
GPT-4.1 [26]	–	95.7	73.9	78.3	**88.2**	39.4	74.0	0.7	0
+ MindJourney	–	**100.0**	82.6	**86.5**	79.4	45.4	77.3	67.1	12.34
+ AVIC	GPT-4.1	**100.0**	78.2	83.7	85.2	54.5	79.3	7.6	0.73
+ AVIC	Qwen2.5VL-7B	82.6	**86.9**	75.6	**88.2**	36.3	72.6	4.8	1.81
+ AVIC-R	Qwen2.5VL-7B	91.3	**86.9**	83.7	85.2	**57.5**	**80.0**	5.2	3.03
o1 [16]	–	78.3	82.6	73.0	73.5	**69.7**	74.6	1.4	0
+ MindJourney	–	**100.0**	65.2	78.4	82.4	63.7	77.3	39.4	12.34
+ AVIC	o1	**100.0**	**86.9**	**86.4**	91.1	66.6	**85.3**	14.6	1.28
+ AVIC	Qwen2.5VL-7B	86.9	65.2	78.3	**94.1**	69.6	79.3	5.7	1.81
+ AVIC-R	Qwen2.5VL-7B	86.9	65.2	81.0	**94.1**	69.6	80.0	6.1	3.03

Table 2: Results on MMSI.

Method	Accuracy
GPT-4o [27]	30.3
GPT-4o + AVIC	**32.3**
GPT-4.1 [26]	30.9
GPT-4.1 + AVIC	**33.8**

Table 3: Results on R2R embodied navigation dataset.

Methods	LLMs	NE↓	OSR↑	SR↑	SPL↑
NavGPT [58]	GPT-3.5	8.02	26.4	16.7	13.0
MapGPT [53]	GPT-4	5.80	61.6	41.2	25.4
MapGPT	GPT-4o	6.04	41.6	36.0	30.8
MapGPT + AVIC	GPT-4o	**5.97**	**45.3**	**37.5**	**31.9**

5 Experiments

5.1 Experiment Setup

Datasets and Benchmarks. We validate our proposed framework on both visual spatial reasoning benchmarks and embodied navigation tasks, covering a range of spatial ambiguities and interaction requirements. For visual spatial reasoning, we evaluate on SAT [30] and MMSI [42], two benchmarks designed to test visual spatial reasoning with single/multiple images. We also evaluate on the Room-to-Room (R2R) [2] for the embodied navigation task.

Implementation Details. Our framework is implemented on top of a vision-language model and a pretrained visual world model, Stable Virtual Camera (SVC) [59]. In the default same-backbone AVIC setting, the policy, verifier, and final QA model use the same base MLLM with different prompts. All decisions by AVIC are made at test time without additional fine-tuning, while we train AVIC-R with LoRA (More details in Appendix A).

5.2 Main Results

AVIC beats the always-on baseline in both efficiency and effectiveness across all backbones. Table 1 compares our framework against baselines on SAT-Real [30] across five categories. Across all open-source and proprietary backbones, AVIC consistently improves over the base MLLM and matches or surpasses the always-on baseline (MindJourney) while using $\sim$10% of the tokens and far fewer world-model calls. With GPT-4.1, accuracy rises from 74.0% to 79.3%; with o1, it reaches **85.3%** (+10.7% over base). Gains are most pronounced on Egocentric Movement, Action Consequence, and Perspective tasks, categories requiring action-conditioned spatial reasoning where selective imagination is most beneficial. Table 2 confirms these gains transfer to MMSI-Bench [42].

AVIC-R turns a small policy model much stronger, even better than proprietary ones. AVIC-R adds two further findings on top of AVIC. First, a small RL-trained policy can drive a much larger backbone: with only Qwen2.5VL-7B as the gating policy, AVIC-R outperforms AVIC variants that

Table 4: Ablation over action scaling, gating, and world model. Based on GPT-4.1.

Action Scaling	Gating	WM	Avg. WM	Acc. (%)
–	–	–	0	74.0
–	–	✓	12.34	77.3
–	✓	✓	0.51	73.3
✓	✓	✓	0.73	79.3

Table 5: Effect of policy and QA model choice on SAT-Real.

Policy	QA	Acc. (%)
GPT-4o	GPT-4o	68.0
GPT-4o	o1	80.0
o1	o1	81.3
o1	GPT-4o	68.6

Table 6: **Reward ablation: the *skip-wrong* penalty.** Removing the -0.5 penalty for an incorrect `skip` causes excessive skipping, dropping overall accuracy by 14.66 points and degrading every question type. Values are test accuracy (%).

	EgoM	ObjM	Goal	EgoAct	Pers	**Avg.**
AVIC-R *w/o* skip-wrong	65.22	65.22	79.41	62.16	42.42	62.67
AVIC-R (full)	**82.61**	**82.61**	**91.18**	**81.08**	**51.52**	**77.33**
Δ	+17.39	+17.39	+11.77	+18.92	+9.10	+14.66

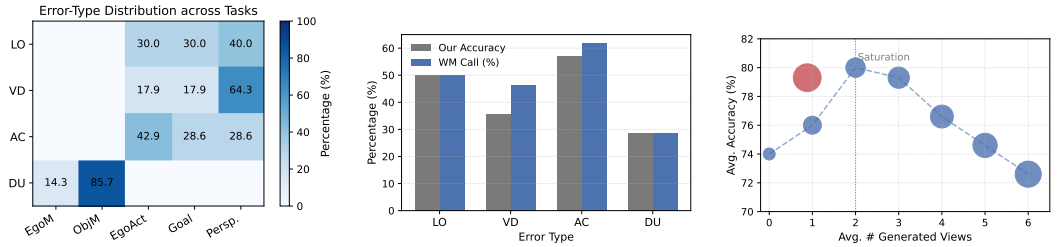
use the proprietary backbone itself as the policy on three of four backbones. As shown in Tab. 1, with GPT-4o used as the QA model, AVIC-R with Qwen2.5VL-7B outperforms GPT-4o used as the policy model by 8.0% in QA accuracy, suggesting that AVIC-R enables better gating and planning of the world model. Second, RL is essential to making the small policy work: prompting Qwen2.5VL-7B alone as the AVIC policy often underperforms even the always-on baseline (e.g., 61.3% vs. MindJourney's 66.7% on InternVL3-14B; 72.6% vs. 77.3% on GPT-4.1), as the 7B model lacks the in-context reasoning to gate reliably. Our lightweight RL training fixes this, lifting the same Qwen2.5VL-7B by 6–8 points across InternVL3/GPT-4o/GPT-4.1 and turning a policy that previously underperformed the always-on baseline into one that drives the full pipeline to top results.

AVIC's selective imagination also transfers to embodied navigation. Table 3 applies AVIC to embodied navigation, integrated into MapGPT's [53] step-wise framework on the 72-scene R2R [2] evaluation. At each step, our policy decides whether to augment currently observed panoramas with imagined views before the next-action prediction (more details in Appendix C.7). Compared to MapGPT with GPT-4o, AVIC achieves higher OSR/SR/SPL and lower navigation error (NE), indicating more reliable goal reaching with shorter, less redundant trajectories, transferring the benefits of selective imagination from static spatial reasoning to embodied tasks.

5.3 Ablation Studies

Effect of selective gating and action-level scaling. As listed in Tab. 4, we analyze the contributions of world-model (WM) imagination, gating, and action-level test-time scaling. Vanilla baseline achieves 74.0% w.o any scaling. Always-on invoking the WM with spatial beam search and without gating or action scaling improves performance to 77.3%, but at the cost of excessive computation, requiring an average of 12.34 WM calls. Introducing a gating mechanism via a policy model alone drastically reduces WM usage (0.51 calls) but also hurts accuracy (73.3%), indicating that binary WM invocation without action-level control is insufficient and can suppress necessary imagination. In contrast, our full method that combines gating with action scaling achieves the best performance (79.3%) while keeping WM usage low (0.73 calls). This demonstrates that *when* to invoke the WM and *how* to use it are both critical: selective gating must be paired with targeted action planning to ensure that limited imagination is informative rather than restrictive. Overall, it highlights that effective visual TTS requires control over both WM invocation and action planning.

Effect of policy and QA model choice. As listed in Tab. 5, we compare different combinations of policy models and QA models on SAT-Real. We find upgrading the QA model yields substantial improvements regardless of the policy model used (68.0% $\rightarrow$ 80.0%). These comparisons show that the choice of QA model affects accuracy, but they do not uniquely identify the system bottleneck. The separate selection and oracle studies in Appendices C.2 and C.5 show additional headroom from gating and trajectory selection with the reasoner held fixed.



(a) Error types vary across task categories.

(b) World-model usage and accuracy gains depend on error type.

(c) Accuracy saturates as more imagined views are generated.

Figure 4: Analysis of when and how much to invoke world-model imagination.

The skip-wrong penalty is essential to RL training. We isolate the contribution of the *skip-wrong* reward term, a -0.5 penalty applied whenever the policy chose `skip` (bypassing the world model) and answered incorrectly. Without this term, a wrong `skip` costs nothing while a wrong `call_wm` costs at least the per-action cost (0.30 for a typical 3-step plan with action_cost$=0.1$), which favors skipping under uncertainty. Reintroducing the penalty makes a wrong skip worse than a short wrong WM call (e.g., -0.5 versus -0.3 for three actions), restoring an incentive to query the world model. Table 6 reports the result. The skip-wrong term lifts overall test accuracy from 62.67% to 77.33% ($+14.66$ points), with the largest gains on *action-consequence* ($+18.92$), *ego-movement* ($+17.39$), and *obj-movement* ($+17.39$); without it, the policy over-skips and its accuracy approaches the no-WM baseline. The broader penalty sweep is given in Appendix C.3.

5.4 When and How Much a World Model is Needed for Visual Spatial Reasoning?

To diagnose *when* world-model imagination is necessary, we manually classify MLLM failures on SAT-Real into four error types: **(1) Limited Observability** (LO, 15.2%): required information occluded or out of view; **(2) Viewpoint Dependence** (VD, 42.4%): answer depends on transforming between egocentric and object-centric frames; **(3) Action-Conditioned Reasoning** (AC, 31.8%): answer depends on the scene state after a hypothetical action; and **Dynamics Understanding** ((4) DU, 10.6%): temporal reasoning about camera or object motion. Figure 4a shows that SAT task categories do not map one-to-one with error types but exhibit *compositional* patterns: EgoAct is dominated by AC errors (post-action viewpoints), Pers. by VD errors (reference-frame transformation), and ObjM by DU errors (temporal dynamics). LO errors appear across multiple categories, indicating that occlusion and limited field-of-view are general failure sources. This decomposition lets us study WM utility through error structure rather than surface task labels.

RQ1: When to call WM?

WM should be used selectively, primarily when reasoning requires predicting future states under hypothetical actions rather than reinterpreting existing visual evidence.

World models are most needed for action-conditioned reasoning. Figure 4b shows that WM imagination yields the largest gain (**+57.1%**) on AC errors, where the answer depends on the post-action scene state, e.g., counterfactual queries like "what if I turn left by 90°?". By contrast, DU errors require only reference-frame transformation over the current view and benefit much less (+28.5%). LO and VD errors fall between these extremes: rendering can reveal occluded content (LO) or visualize the scene from a different vantage (VD), but adds little when the transformation can be inferred symbolically from the original observation. These findings indicate that WM utility is highly instance-dependent. It is most useful when the question references a future or counterfactual scene, often unnecessary for static reinterpretation of what is already observed.

RQ2: How much imagination is needed?

Visual spatial reasoning benefits from *targeted* rather than extensive WM imagination.

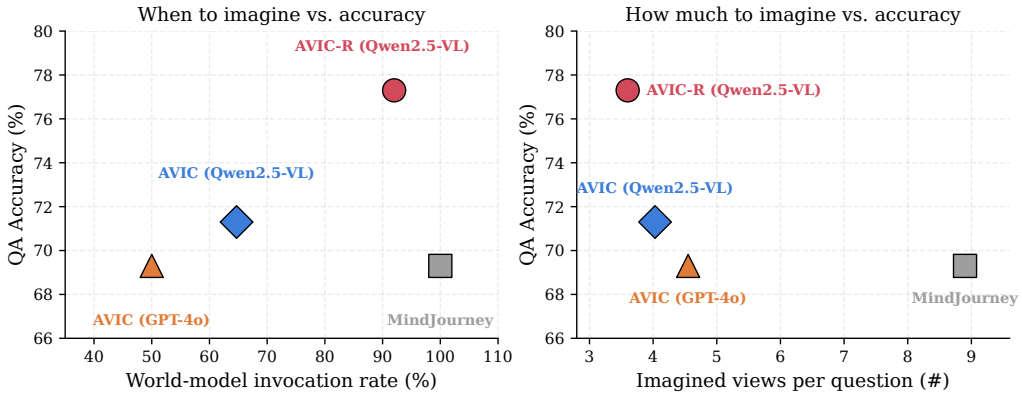


Figure 5: **Adaptive imagination achieves higher accuracy at lower cost on SAT-Real.** *Left:* world-model invocation rate (*when* to imagine) vs. QA accuracy. *Right:* imagined views per question (*how much*) vs. QA accuracy. AVIC-R achieves the highest accuracy (**77.3%**) using the fewest imagined views (**3.60** vs. 8.90 for always-on).

Spatial reasoning requires limited imagination. Figure 4c sweeps fixed-budget baselines with a predetermined number of imagined views. Even a single targeted view raises accuracy by 2% over the no-imagination baseline (74.0% $\rightarrow$ 76%), and a second view captures most of the remaining headroom (76% $\rightarrow$ 80%). Beyond two views, additional rollouts bring no further gains and eventually degrade performance, as accumulated rendering artifacts and redundant content begin to confuse the downstream reasoner. Targeted, low-budget imagination is what spatial reasoning needs, while exhaustive scaling is wasteful at best and harmful at worst.

AVIC-R learns better when and how much to imagine. The findings above point to a simple rule for visual test time scaling with WM: we should call WM for visual spatial reasoning mainly on action-conditioned questions, and use only targeted views per call. Figure 5 (and Appendix Table 8) shows that AVIC-R learns both parts, while alternative policies do not. On the *when* axis, AVIC-R calls WM on 92% of questions on average, with strong differences across categories: 100% on EgoAct (dominated by AC errors) versus 78.8% on Pers.(dominated by VD errors). This category-aware behavior emerges from QA-correctness and WM-cost rewards alone, and no per-category labels are provided. Other alternatives are far less stable: AVIC with GPT-4o calls 100% on EgoM but 0% on ObjM, and zero-shot Qwen2.5VL over-skips at 64.7%, ending 6 points behind AVIC-R (71.3% vs. 77.3%). Calling WM more often is not the point, but calling it on the right questions is. On the *how-much* axis, AVIC-R uses 3.60 views per question, fewer than every selective baseline (zero-shot Qwen 4.03, GPT-4o 4.55) and less than half of MindJourney's 8.90, while still reaching the best accuracy. In short, both behaviors emerge from our lightweight RL scheme: the policy learns when to imagine and how much to imagine, just from QA correctness and a WM cost penalty.

6 Conclusion

In this paper, we study visual spatial reasoning with world models through the lens of adaptive test-time scaling, finding that always-on imagination is often unnecessary and even misleading. We introduce AVIC, which selectively decides *when* and *how much* to imagine at inference time, and AVIC-R, which trains this gating policy end-to-end via lightweight RL from QA-correctness and WM-cost signals. Across spatial reasoning and embodied navigation benchmarks, our framework achieves competitive or state-of-the-art results while substantially reducing world-model calls, tokens, and inference time; notably, AVIC-R with a small open-source policy outperforms pipelines that use proprietary backbones as the policy. Our analysis shows world-model imagination is most beneficial for action-conditioned reasoning and requires only limited, targeted views, highlighting the importance of instance-dependent TTS for efficient and reliable reasoning with world models.

Limitations & Broader Impacts. See Appendix for limitations and broader impacts discussion.

7 Acknowledgments

This work was supported by ONR Grant N00014-23-1-2356, ARO Award W911NF2110220, DARPA ECOLE Program No. HR00112390060, Microsoft Accelerate Foundation Models Research (AFMR) grant program, NSF AI Engage Institute DRL-2112635, and Cisco and Capital One Faculty Awards. The views contained in this article are those of the authors and not of the funding agency.

References

[1] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[3] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22875–22889, 2024.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report, 2025.

[5] Meng Cao, Xingyu Li, Xue Liu, Ian Reid, and Xiaodan Liang. Spatialdreamer: Incentivizing spatial reasoning via active mental imagery. *arXiv preprint arXiv:2512.07733*, 2025.

[6] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas J. Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, 2024.

[7] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. *ArXiv*, abs/2406.01584, 2024.

[8] Andong Deng, Tongjia Chen, Shoubin Yu, Taojiannan Yang, Lincoln Spencer, Yapeng Tian, Ajmal Saeed Mian, Mohit Bansal, and Chen Chen. Motion-grounded video reasoning: Understanding and perceiving motion at pixel level. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8625–8636, 2025.

[9] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and P. Abbeel. Learning universal policies via text-guided video generation. *ArXiv*, abs/2302.00111, 2023.

[10] Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 105–116, 2025.

[11] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *ArXiv*, abs/2404.02101, 2024.

[12] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *ArXiv*, abs/2307.12981, 2023.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[14] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *ArXiv*, abs/2311.12871, 2023.

[15] Yidong Huang, Zun Wang, Han Lin, Dong-Ki Kim, Shayegan Omidshafiei, Jaehong Yoon, Yue Zhang, and Mohit Bansal. Planning with sketch-guided verification for physics-aware video generation. *arXiv preprint arXiv:2511.17450*, 2025.

[16] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. OpenAI o1 system card, 2024.

[17] Binbin Ji, Siddharth Agrawal, Qiance Tang, and Yvonne Wu. Enhancing spatial reasoning in vision-language models via chain-of-thought prompting and reinforcement learning. *ArXiv*, abs/2507.13362, 2025.

[18] Stephen M Kosslyn, William L Thompson, and Giorgio Ganis. *The case for mental imagery*. Oxford University Press, 2006.

[19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[20] Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *ArXiv*, abs/2305.19195, 2023.

[21] Jialu Li, Jaemin Cho, Yi-lin Sung, Jaehong Yoon, and Mohit Bansal. Selma: Learning and merging skill-specific text-to-image experts with auto-generated data. In *Neural Information Processing Systems*, 2024.

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

[24] Shalev Lifshitz, Sheila A McIlraith, and Yilun Du. Multi-agent verification: Scaling test-time compute with multiple verifiers. *arXiv preprint arXiv:2502.20379*, 2025.

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023.

[26] OpenAI. GPT-4.1 technical overview, 2024.

[27] OpenAI. Hello GPT-4o. OpenAI Blog, 2024.

[28] Cheng Qian, Emre Can Acikgoz, Bingxuan Li, Xiusi Chen, Yuji Zhang, Bingxiang He, Qinyu Luo, Dilek Hakkani-Tür, Gokhan Tur, Yunzhu Li, et al. Current agents fail to leverage world model as tool for foresight. *arXiv preprint arXiv:2601.03905*, 2026.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[30] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, et al. SAT: Dynamic spatial aptitude training for multimodal language models, 2024.

[31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[32] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[33] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *ArXiv*, abs/2406.16860, 2024.

[34] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[35] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.

[36] Ziyang Wang, Jaehong Yoon, Shoubin Yu, Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Conference on Empirical Methods in Natural Language Processing*, 2025.

[37] Zun Wang, Jaemin Cho, Jialu Li, Han Lin, Jaehong Yoon, Yue Zhang, and Mohit Bansal. Epic: Efficient video camera control learning with precise anchor-video guidance. *ArXiv*, abs/2505.21876, 2025.

[38] Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023.

[39] Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot++: Test-time scaling with soft chain-of-thought reasoning. *arXiv preprint arXiv:2505.11484*, 2025.

[40] Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184, 2026.

[41] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10632–10643, 2024.

[42] 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*, 2025.

[43] Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. *Advances in Neural Information Processing Systems*, 38:109855–109885, 2026.

[44] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36:11809–11822, 2023.

[45] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views, 2025. Structural Priors for Vision Workshop at ICCV 2025.

[46] Jaehong Yoon, Shoubin Yu, and Mohit Bansal. Raccoon: A versatile instructional video editing framework with auto-generated narratives. *arXiv preprint arXiv:2405.18406*, 2024.

[47] Shoubin Yu, Difan Liu, Ziqiao Ma, Yicong Hong, Yang Zhou, Hao Tan, Joyce Chai, and Mohit Bansal. Veggie: Instructional editing and reasoning video concepts with grounded generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15147–15158, 2025.

[48] Shoubin Yu, Jaehong Yoon, and Mohit Bansal. Crema: Generalizable and efficient video-language reasoning via multimodal modular fusion. In *International Conference on Learning Representations*, 2025.

[49] Shoubin Yu, Yue Zhang, Ziyang Wang, Jaehong Yoon, and Mohit Bansal. Mexa: Towards general multimodal reasoning with dynamic multi-expert aggregation. *Findings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

[50] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2024.

[51] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.

[52] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

[53] Yifan Zhang, Zhengting He, Jingxuan Li, Jianfeng Lin, Qingfeng Guan, and Wenhao Yu. Mapgpt: an autonomous framework for mapping by integrating large language model and cartographic tools. *Cartography and Geographic Information Science*, 51(6):717–743, 2024.

[54] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *European conference on computer vision*, pages 399–415. Springer, 2024.

[55] Yue Zhang and Parisa Kordjamshidi. Vln-trans: Translator for the vision and language navigation agent. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[56] Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *Trans. Mach. Learn. Res.*, 2024, 2024.

[57] Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. Spartun3d: Situated spatial understanding of 3d world in large language models. *arXiv preprint arXiv:2410.03878*, 2024.

[58] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7641–7649, 2024.

[59] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models, 2025.

[60] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification. *ACM Transactions on Graphics (TOG)*, 37:1 – 12, 2018.

[61] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

A Implementation Details

AVIC-R Training Details. We post-train Qwen2.5-VL-7B-Instruct with LoRA [13] (r=8, α=16, dropout 0.05, applied to the `q/k/v/o` projections; $\sim$5M trainable parameters out of 8.3B) using online GRPO [32]. For each question we sample K=16 rollouts from the policy with temperature 1.0, top-p=0.95, top-k=50, and a 512-token response budget; advantages are computed by per-question (group) reward normalization. The PPO surrogate uses a clip ratio of ϵ=0.2 and a K3 KL penalty to the frozen base policy with β=0.1, where the same backbone with the LoRA adapter *disabled* serves as the reference policy (no separate frozen copy). Each rollout's reward combines task correctness, computed by matching the frozen GPT-4o answer to the dataset ground-truth label, with an action cost of 0.1 per atomic step, a parse-failure penalty of -0.5, and an additional -0.5 *skip-wrong* penalty when the policy bypasses the world model and answers incorrectly; this last term is essential to prevent the policy from collapsing to "always skip". The action vocabulary consists of 9° rotations and 0.25 m forward steps, capped at 6 atomic actions per plan. We optimize with AdamW, lr=2×10^{-5}, weight decay 0, gradient clipping at 1.0, per-device batch size 1 and no gradient accumulation, yielding an effective batch of 8 questions per optimizer step on 8×A100-80 GB GPUs (DDP via `torchrun`). The world model is Stable Virtual Camera [59] run in `img2trajvid_s-prob` mode (CFG 4.0, 8 target views, trajectory prior, `interp` chunking, short side 576). Training data is a curated 30/70 mix of GPT-4o-prescored *easy-skip* and *needs-WM* questions; we train for up to 300 optimizer steps and select the checkpoint that achieves the best held-out accuracy. We train AVIC-R using GPT-4o as the frozen QA model and apply the resulting policy to other QA backbones without retraining (Table 1).

Training compute and accounting. Training uses eight A100-80 GB GPUs for up to 300 optimizer steps, with checkpoint selection based on held-out accuracy. Only the approximately 5M LoRA parameters are updated. World-model generation and frozen-QA queries supply the online rewards; the trajectory verifier is absent from training and is used only during inference. Parameter-efficient updates do not remove the cost of these online queries. Inference timing is reported separately in Table 7; that table does not isolate verifier latency as a separate column and should not be read as a training-cost measurement.

Balanced data curation. Random training sampling is inefficient: as shown in Sec. 3, most SAT questions are already answered correctly without imagination (Case 3), so every reasonable rollout produces $r \approx +1$, the within-group variance vanishes, and Equation (5) yields zero gradient. The questions that drive learning are those the base VLM fails on but a targeted imagined view would correct. We therefore pre-score a candidate pool with a strong reference QA model under the `skip` policy and split each question into *easy-skip* (base correct) or *needs-imagination* (base wrong). The final training set ($\tilde{3}$000 examples) retains every needs-imagination instance and sub-samples easy-skip ones at a 30% ratio, balancing learning signal against an anchor that prevents collapse onto always-call-WM.

Lenient schema parsing. High-temperature sampling occasionally yields JSON outputs that are syntactically broken but semantically recoverable: d set to an action verb (`turn-left`) rather than a meta-decision, or π written as free-form strings (`"turn-right 9 degrees"`). A strict parser would reject these as parse failures, conflating format errors with semantic confusion. We instead use a lenient parser with a salvage stage that infers d = `call_wm` whenever π is non-empty regardless of the literal d value, and parses natural-language action strings into structured records via regex. Truly unrecoverable outputs still incur r_{parse}. The finer-grained reward distinguishes "intent recovered, format mangled" (taught via the call-WM reward) from "no plan at all" (parse failure), supplying GRPO with a smoother gradient.

Metrics. We evaluate SAT/MMSI spatial reasoning benchmark with multiple-choice QA accuracy. In the R2R navigation setting, the agent must follow natural language instructions in indoor environments. We integrate our adaptive visual test-time scaling framework into the navigation pipeline and measure performance. We evaluate navigation performance using four standard metrics: Navigation Error (NE), Oracle Success Rate (OSR), Success Rate (SR), and Success weighted by Path Length (SPL). NE measures the geodesic distance between the agent's final position and the target, while SR reports the fraction of episodes where the final position is within a predefined success threshold. OSR measures whether the agent ever reaches within the success threshold at any point along its trajectory, reflecting exploration ability independent of stopping. SPL jointly evaluates success and efficiency

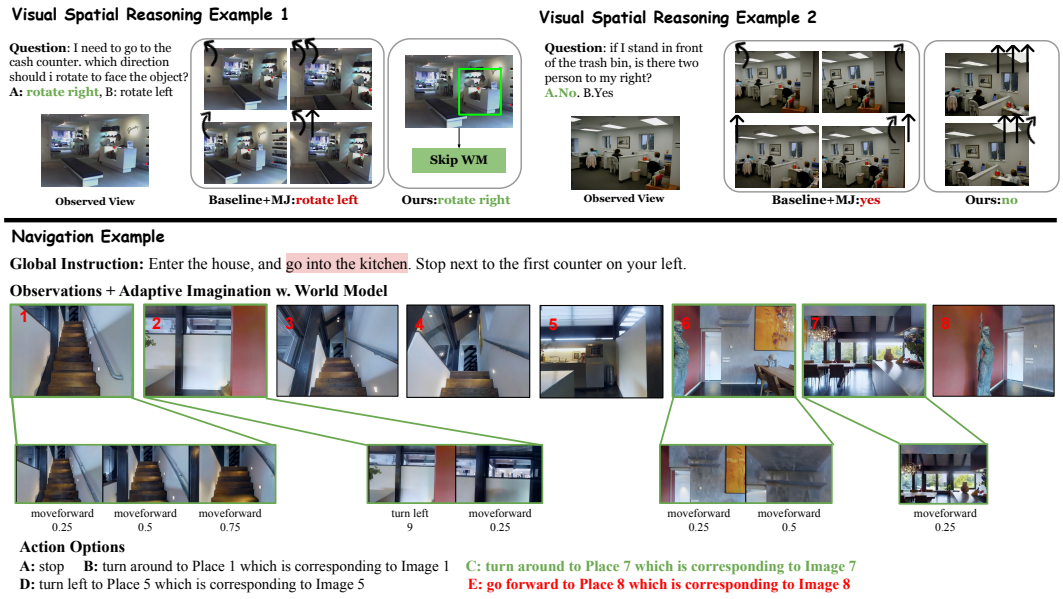


Figure 6: Qualitative examples on SAT of the always-on imagination method and our adaptive method, as well as the R2R navigation task. In the navigation example, the green option is selected by the model with adaptive imagination via our method, while the red one is without world model imagination.

by weighting successful episodes by the ratio between shortest-path length and the actual trajectory length.

Prompts. We provide extra technical details of our adaptive visual test time scaling framework. In Tab. 18, we provide verifier prompts that are used to score each generated trajectory, and in Tab. 19, we provide prompts for world model gating and action planning in our policy model.

B Extra Experiments

Qualitative Analysis. We also provide qualitative examples as illustrated at the top of Figure 6. We compare our adaptive visual TTS method with the always-on imagination method, MindJourney (MJ). In the first example, the target object (the *cash counter*) is already clearly visible in the observed view. Our method correctly identifies that additional visual imagination is unnecessary and directly skips world model. In contrast, MJ indiscriminately invokes the world model, generating multiple imagined views that introduce misleading evidence and ultimately lead to an incorrect prediction. In the second example, AVIC yields the correct answer by selectively imagining the state where the agent is in front of the *trash bin*. In contrast, MJ performs dense imagination and generates views that do not accurately reflect this critical spatial condition, leading to an incorrect prediction. Furthermore, we present a qualitative navigation example at the bottom of Figure 6. Our adaptive visual test-time scaling selectively augments informative indoor observations (e.g., zooming in or turning to explore nearby views), enabling the agent to better inspect the environment and align its actions with the global instruction (*"go to the kitchen"*). In contrast, the baseline without visual imagination lacks sufficient perceptual evidence and consequently chooses an incorrect direction.

Sensitivity to Errors from World Models. Our work is motivated by the observation that imperfect imagination can introduce misleading or noisy evidence that may degrade performance (Sec. 3, Figs. 1–2). To further evaluate robustness, we conduct additional experiments using **Cosmos** [1] in Tab. 10 as an alternative world model, which tends to produce less visually stable camera trajectories and noisier geometric structures compared to SVC.

Stage-wise Computation Cost. The decision module adds a lightweight inference step whose cost is small relative to world-model (WM) generation. Rather than always invoking expensive

imagination, it predicts *when* and *how* to use the WM, reducing unnecessary calls. The table shows that policy cost is minor compared to WM cost and is offset by large savings. Although AVIC introduces $\sim$14.9 s of policy overhead, it reduces WM time by $\sim$153.7 s (163.32 $\rightarrow$ 9.59), yielding a $\sim$**6**$\times$ **reduction in total time** (177.84 $\rightarrow$ 29.04) and $\sim$**20**$\times$ **fewer tokens** (162.6 $\rightarrow$ 7.6 k), while improving accuracy by 2.0 points over always-on. AVIC achieves both higher accuracy and substantially lower overall computation.

Table 7: **Inference cost and accuracy on SAT-Real.** AVIC adds modest policy-inference overhead but substantially reduces world-model time and token usage, achieving the highest accuracy at $\sim$6$\times$ lower total time and $\sim$20$\times$ fewer tokens than always-on imagination.

Method	Policy (s)	WM (s)	QA (s)	Total (s)	Tokens (k)	WM Calls	Acc. (%)
Baseline (no WM)	0	0	3.9	3.9	0.7	0	74.0
MindJourney (always-on)	0	163.32	14.52	177.84	162.6	12.42	77.3
AVIC	14.92	9.59	4.53	29.04	7.6	0.73	79.3

Robustness across runs. To assess the stability of AVIC-R, we repeat our main evaluation three times with independently sampled rollouts; all other hyperparameters are held fixed. The overall accuracy gave a mean of 69.33 with a sample standard deviation of 0.77 (standard error of the mean 0.44).

The Cosmos result remains above the o1 baseline (77.3% versus 74.6%), but the gain is smaller than with SVC (85.3%). Thus the method remains sensitive to world-model quality. Appendix C.4 extends this comparison to the RL-trained policy without retraining, and Appendix C.7 discusses why these results do not isolate the causes of misleading imagined evidence.

Evaluations on Additional Benchmark. We extend our evaluation to **MindCube** [45] in Tab. 9. AVIC improves performance on this fine-grained spatial reasoning benchmark, indicating that our method generalizes across tasks.

Framework Error Analysis. Table 11 reports outcome-based recall and precision for the zero-shot GPT-4o policy. These metrics depend on the generated plans, the rendered evidence, and the final reasoner as well as the gating decision. The controlled Qwen2.5-VL-7B comparison in Appendix C.3 distinguishes zero-shot from RL-trained behavior. Invocation rates and view budgets are reported separately in Table 8.

Table 8: **Analysis of *when* and *how much* to imagine on SAT-Real.** We compare four imagination policies along three axes: fraction of questions on which the world model is invoked (*top*), average number of imagined views per question (*middle*), and resulting QA accuracy (*bottom*).

Method	EgoM	ObjM	EgoAct	Goal	Pers	Avg.
when to call world model (call_wm %)						
MindJourney (always-on)	100.0	100.0	100.0	100.0	100.0	100.0
AVIC (GPT-4o)	100.0	0.0	62.2	32.4	54.5	50.0
AVIC (Qwen2.5VL)	47.8	60.9	78.4	73.5	54.5	64.7
AVIC-R (Qwen2.5VL)	91.3	91.3	100.0	97.1	78.8	92.0
how much to imagine (# img per question)						
MindJourney (always-on)	11.13	4.83	8.37	11.03	8.55	8.90
AVIC (GPT-4o)	3.30	0.00	5.30	4.18	5.39	4.55
AVIC (Qwen2.5VL)	4.09	3.07	4.90	3.28	4.39	4.03
AVIC-R (Qwen2.5VL)	2.95	3.00	4.59	2.79	4.23	3.60
task accuracy (%)						
MindJourney (always-on)	78.3	60.9	78.4	70.6	57.5	69.3
AVIC (GPT-4o)	**87.0**	69.6	64.9	82.4	48.5	69.3
AVIC (Qwen2.5VL)	65.2	73.9	64.9	**91.2**	**60.6**	71.3
AVIC-R (Qwen2.5VL)	82.6	**82.6**	**81.1**	**91.2**	51.5	**77.3**

Table 9: Results on MindCube-Tiny

Method	Acc. (%)
GPT-4o	36.5
GPT-4o + AVIC	**38.7** (+2.2)

Table 10: Performance across world models. Based on o1.

Method	Acc. (%)
o1	74.6
o1 + AVIC (SVC)	85.3
o1 + AVIC (Cosmos)	77.3

Table 11: **Gating recall and precision** (%) per category for the zero-shot GPT-4o policy in AVIC, rather than the RL-trained AVIC-R policy. *Recall*: fraction of WM-needed questions on which the policy calls WM. *Precision*: fraction of WM-called questions that actually benefit from WM.

Metric	EgoM	ObjM	EgoAct	Goal	Pers.	Avg.
Recall	100.0	33.3	55.6	33.3	52.6	43.9
Precision	5.6	50.0	19.2	22.2	62.5	27.1

C Additional Ablations and Analysis

We further examine the contributions of gating, planning, and trajectory selection; the wrong-skip penalty; transfer across world models and reasoners; and performance on SAT-Synthesized. We distinguish the configurations used in these experiments because the policy and reasoner need not be the same model.

C.1 Contributions of Gating, Planning, and Verification

Table 12 evaluates a Qwen2.5-VL-7B policy with a GPT-4o reasoner on SAT-Real. Gating and planning are produced jointly in one policy forward pass. Consequently, the gating-only condition retains the policy's gate but substitutes the always-on baseline's greedy search whenever imagination is requested. In the gating-and-planning conditions without the MLLM verifier, we answer each rollout separately and aggregate answers by majority vote. This *answer-level* self-consistency is distinct from majority voting over the policy's *gating decisions* in the full framework.

Table 12: Component ablation on SAT-Real with a Qwen2.5-VL-7B policy and GPT-4o reasoner. Calls and views are averages per question. "Plan" denotes policy-generated plans; the gating-only variants instead use baseline greedy search. For the gating-and-planning variants, removing the verifier substitutes answer-level self-consistency.

Method	Gate	Plan	Verify	WM calls	Views	Acc. (%)
No imagination	–	–	–	0	0	60.3
Always-on	–	–	–	12.34	8.90	69.3
AVIC	✓	–	–	7.40	5.34	68.0
AVIC	✓	✓	–	2.21	4.93	69.3
AVIC	✓	✓	✓	1.81	4.03	71.3
AVIC-R	✓	–	–	7.94	5.73	72.6
AVIC-R	✓	✓	–	3.37	4.49	75.3
AVIC-R	✓	✓	✓	3.03	3.60	**77.3**

For the zero-shot policy, gating alone reduces computation but does not retain all the accuracy of always-on imagination. Adding policy-generated plans raises zero-shot accuracy from 68.0% to 69.3% while reducing average calls from 7.40 to 2.21. Replacing answer voting with the verifier adds 2.0 percentage points for both zero-shot and RL-trained policies. RL adds 6.0 points in each gating-and-planning configuration. The full AVIC-R achieves 77.3%, compared with 69.3% for always-on imagination, with 3.03 rather than 12.34 calls. These are system-level comparisons: their computation budgets also differ, so they do not isolate a component's effect at matched compute.

C.2 Effect of the Trajectory Selection Rule

We additionally compare selection rules using GPT-4o as both policy and reasoner with $M = 5$ policy samples (Table 13). Random selection chooses a candidate trajectory; answer self-consistency votes

over the answers obtained from the candidates; and the MLLM verifier scores each complete rendered trajectory before selecting the highest-scoring one. The verifier prompt is provided in Table 18. The candidate-set oracle counts a question as successful if at least one of the five candidates yields the correct answer.

Table 13: Downstream QA accuracy under different selection rules on SAT-Real. Both the policy and reasoner are GPT-4o, with $M = 5$. These scores measure answer accuracy, not the verifier's trajectory-ranking accuracy.

Selection rule	Acc. (%)
Random	70.0
Answer self-consistency	75.3
MLLM verifier	77.3
Candidate-set oracle	88.0

The verifier improves accuracy by 7.3 points over random selection and 2.0 points over answer self-consistency, recovering approximately 41% of the random-to-oracle gap. Substantial selection headroom remains. This separate selection experiment uses a GPT-4o policy; its 77.3% result should not be conflated with the numerically equal AVIC-R result in Table 12.

C.3 Wrong-Skip Penalty and Outcome-Based Precision

Table 14 varies the wrong-skip penalty β_s using the same Qwen2.5-VL-7B policy backbone. With action cost $c = 0.1$, the setting $\beta_s = 0.3$ equalizes the penalties for a wrong skip and a wrong call with a three-action plan. It is not symmetric for every plan length, because the call cost is $c|\pi|$.

Table 14: Wrong-skip penalty sweep on SAT-Real with a Qwen2.5-VL-7B policy. Precision and accuracy are percentages. The zero-shot row is evaluated without RL; the remaining rows use RL with the specified penalty.

Policy / β_s	Precision	Acc.
Zero-shot	38.2	71.3
RL, 0	19.5	62.67
RL, 0.3	43.2	74.6
RL, 0.5	**50.5**	**77.3**
RL, 1.0	36.1	67.2

The default $\beta_s = 0.5$ achieves the highest precision and accuracy among the tested settings. Both removing the penalty and increasing it to 1.0 degrade performance; low precision therefore cannot be attributed to reward asymmetry alone. Here, precision is outcome-based: a call is counted as beneficial only when it improves the final answer. It consequently depends on planning, rendering, and evidence integration as well as the gate. The 27.1% precision in Table 11 belongs to the zero-shot GPT-4o policy, whereas the 38.2% and 50.5% values here belong to zero-shot and RL-trained Qwen policies, respectively.

C.4 Transfer across World Models and Reasoners

Table 15 extends the world-model comparison using o1 as the reasoner. The zero-shot AVIC policy has not been trained on SVC-generated views, so its SVC–Cosmos gap is not evidence of policy-training overfitting to SVC. For AVIC-R, we replace SVC with Cosmos at inference without retraining the policy originally optimized using SVC and the frozen GPT-4o QA model.

Both policies improve over the baseline with either world model, but gains are smaller with Cosmos. This supports transfer in the tested setting while showing that the benefit remains sensitive to the renderer; it does not establish invariance to world-model quality.

The correctness reward compares the frozen QA model's answer with the dataset label. The QA model is therefore an answer generator, not a judge that defines correctness. Its reasoning biases can nevertheless affect which trajectories receive reward. Table 1 evaluates the same trained AVIC-R

Table 15: Transfer across world models on SAT-Real with the o1 reasoner. The zero-shot AVIC policy is o1. The AVIC-R policy is Qwen2.5-VL-7B, trained using GPT-4o and SVC, then evaluated with either world model without retraining. Gains are percentage points over no imagination.

Policy	World model	Acc. (%)	Gain
No imagination	–	74.6	–
AVIC (zero-shot)	SVC	85.3	+10.7
AVIC (zero-shot)	Cosmos	77.3	+2.7
AVIC-R	SVC	80.0	+5.4
AVIC-R	Cosmos	76.0	+1.4

policy without retraining across four reasoners: gains over their respective baselines are 10.0 points for InternVL3-14B, 17.0 for GPT-4o, 6.0 for GPT-4.1, and 5.4 for o1. These results demonstrate transfer to the evaluated reasoners, while the largest gain remains on the reasoner used during training.

C.5 Conditional Oracle Analysis

We examine two privileged interventions on AVIC-R (Table 16). Oracle gating selects skip versus call using precomputed ground-truth information while retaining the policy's plans and trajectories. The second intervention additionally selects the best trajectory among the sampled candidates.

Table 16: Conditional oracle analysis for AVIC-R on SAT-Real. These interventions use ground-truth information unavailable at deployment and retain the sampled candidate set.

Setting	Acc. (%)
AVIC-R	77.3
With oracle gating	88.0
With oracle gating and trajectory selection	95.3

The successive gains of 10.7 and 7.3 points quantify conditional headroom under privileged decisions, not a unique causal decomposition or a model-independent upper bound. They do not establish reasoner reliability on arbitrary imagined evidence. The 75.3% oracle in Section 3 is restricted to MindJourney's fixed pipeline. The 88.0% selection oracle in Table 13 also uses a different policy configuration from the oracle-gating experiment here.

C.6 Additional Evaluation on SAT-Synthesized

Table 17 reports the additional SAT-Synthesized comparison. AVIC and AVIC-R improve accuracy over MindJourney by 1.5 and 4.8 points, respectively, while using fewer world-model calls. This extends the observed accuracy–computation trade-off beyond SAT-Real.

Table 17: Additional results on SAT-Synthesized. Average WM calls are reported per question.

Method	Acc. (%)	Avg. WM calls
GPT-4o, no imagination	61.0	0
MindJourney	72.3	12.34
AVIC	73.8	2.34
AVIC-R	**77.1**	3.92

C.7 Interpretation and Scope

Invocation frequency and imagination budget. The 54% "unnecessary" fraction in Section 3 diagnoses the original fixed-imagination setting; it is not a target rate for the learned AVIC-R gate, which invokes the WM on 92% of questions (Table 8). Its 50.5% outcome-based precision still leaves substantial room for improvement. Frequent invocation can nevertheless coexist with fewer calls and views per question when each call uses a shorter plan than exhaustive exploration.

Misleading imagined evidence. Case 2 is outcome-based. Rendering errors, plausible but uninformative trajectories, and failures to integrate imagined evidence can all contribute. The world-model comparison supports sensitivity to rendering quality but does not quantitatively separate these causes.

Imagination and navigation actions. In R2R, MapGPT's graph-defined navigation candidates remain unchanged. The 9° rotations and $0.25\,\mathrm{m}$ forward motions augment currently observed panoramas through virtual camera motion; they neither move the agent nor require hypothetical graph nodes. The discrete, six-action interface (Table 19) limits imagination resolution and horizon; continuous control and richer interactions remain outside this evaluation.

Statistical scope. The additional SAT-Real ablations use 150 questions and report point estimates without repeated-run variability or confidence intervals. Small differences should be interpreted cautiously; these results do not establish statistical significance or robustness across training seeds.

D Impact Statement

This work studies world-model-based visual imagination in visual spatial reasoning and highlights the limitations of existing always-on test-time imagination methods. Through systematic analysis, we show that indiscriminate visual imagination can be computationally inefficient and, in some cases, harmful due to misleading or redundant imagined views. Our findings emphasize the importance of adaptive test-time computation, demonstrating that effective spatial reasoning requires selectively invoking visual imagination only when necessary and scaling it appropriately. Beyond the specific benchmarks studied, our insights are broadly applicable to multimodal agents that rely on test-time simulation, including embodied AI and interactive systems.

E Limitations

Our work focuses on adaptive imagination control for visual spatial reasoning and short-horizon embodied navigation; extending the framework to longer-horizon decision-making, manipulation, and broader multi-modal tasks is a natural next direction. The framework also assumes access to a separate visual world model and a fixed discrete action space, leaving room for richer extensions such as continuous action spaces, multiple specialized world models, or joint training of the gating policy with the world model itself. Incorporating world-model uncertainty into the reward signal is another promising direction for improving robustness under noisier imagined rollouts. Appendix C.7 further discusses the outcome-based failure taxonomy, the distinction between call frequency and imagination budget, and the statistical scope of the additional experiments.

F License

We will make our code and models publicly accessible. We use standard licenses from the community and provide the following links to the licenses for the datasets, codes, and models that we used in this paper. For further information, please refer to the specific link.

QWen2.5VL [4]: Apache-2.0

SAT [30]: MIT

MMSI [42]: CC-BY-4.0

R2R [2]: MIT

Table 18: Verifier prompts for scoring imagined view plans.

Role. You are an *independent evaluator* for visual spatial reasoning.

Input. A multiple-choice question, answer options, the current observation image(s), and *one* candidate action plan. The plan includes imagined views rendered by a world model.

Task. Score how *useful* the imagined views are for answering the question.

Score Range. Integer from 0 (not helpful, irrelevant, or low quality) to 9 (highly helpful and informative).

Scoring Guidelines.

- Assign higher scores if the imagined views reveal missing evidence needed to answer the question (e.g., resolving occlusion or viewpoint ambiguity).
- Assign higher scores if the imagined views are sharp, coherent, and visually consistent.
- Assign lower scores if the views are redundant, uninformative, distorted, or unrelated to the question.
- If the original observations are already sufficient, most plans should receive low scores.

Rules.

- Do **not** answer the question.
- Output **only** a single integer between 0 and 9.
- Do not output any additional text.

Output Example.

5

Table 19: Policy model prompts for world model gating and action planning.

Role. You are a *policy model* for spatial reasoning in a 3D indoor environment. Your goal is to decide whether to invoke a world model (WM) and, if needed, plan actions that acquire the most informative imagined views.

Input. One or more images, a multiple-choice question, and answer options.

Tasks.

- Decide whether to `SKIP` or `CALL` the world model.
- If `CALL`, generate a short action plan (1–6 actions) to gather additional visual evidence.

Action Space (Discrete, Fixed).

- `move-forward` 0.25 meters
- `turn-left` 9 degrees
- `turn-right` 9 degrees

Action Composition Guidelines.

- Repeated turns approximate larger rotations (e.g., 2 turns $\approx 18°$, 5 turns $\approx 45°$, 10 turns $\approx 90°$).
- When a question specifies a larger angle, approximate it using repeated $9°$ turns.

When to Call the World Model.

- The answer is not directly observable from the current view.
- The question depends on perspective, facing direction, rotation, or left/right relations.
- The question requires reasoning about motion or state changes over time.

Constraints.

- Do not generate cancelling or oscillating actions (e.g., left then right).
- If turning, choose a single direction and turn monotonically.

Output Format (JSON only).

```
{
  "decision": "skip" | "call_wm",
  "reason": "<one sentence>",
  "actions": [
    {"type": "move-forward" | "turn-left" | "turn-right",
     "value": <number>}
  ]
}
```