SCOPE: Symbolic Constraint-Guided Active Perception for Multimodal Reasoning

Wentao Cao[*1] **Po-Hsun Chang**[*1] **Sicheng Chen**[*1] **Xuanming Liu**[1] **Haotian lu**[2]

Abstract

Current multimodal reasoning systems often allocate extra inference-time computation to sampling additional textual rationales or tool-use trajectories. In visually ambiguous settings, however, the more consequential limitation is often the lack of an explicit intermediate state that records what the system currently believes about the scene, which alternatives remain plausible, which observations support those alternatives, and which new observations would be most informative for separating them. We introduce **SCOPE**, a symbolic active-perception framework that performs posterior sharpening over uncertainty-aware world hypotheses built from typed predicates, provenance, and explicit contradiction structure. SCOPE alternates between symbolic state induction, consistency-based world pruning, and conflict-driven evidence acquisition, so that additional computation is spent on refining explicit scene hypotheses rather than merely enlarging an unstructured pool of candidate outputs. This shift turns multimodal inference into a process of ambiguity management: the model preserves multiple plausible scene explanations early, prunes inconsistent ones as evidence accumulates, and focuses later perception on the few local conflicts that still influence the answer posterior. Across high-resolution and general multimodal reasoning benchmarks, SCOPE yields strong improvements over competitive baselines while also providing a more interpretable account of how extra computation changes the model's internal decision state.

1. Introduction

Building multimodal systems that can form explicit beliefs about a visual scene is a fundamental goal in vision-language research. For such systems to be useful on difficult reasoning problems, we argue that their inference procedure should satisfy three requirements.

> **Question.** *What should an inference procedure preserve, expose, and resolve when several plausible scene explanations coexist?*

First, it should preserve ***uncertainty*** over multiple plausible scene interpretations rather than collapsing too early to a single reading. Second, it should make those interpretations ***explicit*** enough to compare, revise, and relate to candidate answers. Third, it should support ***active evidence acquisition***, so that additional perceptual effort can be directed toward the distinctions that matter most for the final decision.

Recent multimodal systems have taken an important step toward the third requirement by interacting with images at inference time through operations such as zooming, cropping, and localized inspection (Achiam et al., 2023; Zheng et al., 2025; Zhang et al., 2025b). Yet, despite this progress, existing pipelines often fall short on one or more of the three requirements above. Methods centered on free-form trajectories can preserve some diversity, but they usually leave competing scene interpretations implicit in natural-language context. Other systems can retrieve more local evidence, but they often lack an explicit state that records which claims about the scene are currently supported, which alternatives remain unresolved, and which new observation would be most informative.

> **Question.** *Why does additional compute often help less than expected in visual reasoning?*

The central difficulty is that challenging visual questions rarely involve a single axis of uncertainty. A tiny OCR token, a subtle spatial relation, or a small region that changes the preferred answer can induce several plausible scene explanations at once. In such cases, the key issue is not merely how to generate more reasoning, but how to maintain and update an epistemic state over these alternatives. For decades, researchers have approached related problems from multiple angles, including structured reasoning, active perception, and inference-time computation. Inspiring as

[1]Jiangning Campus of the High School Affiliated to Nanjing Normal University [2]Southeast University. Correspondence to: Xuanming Liu <xuanmingliu0120@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

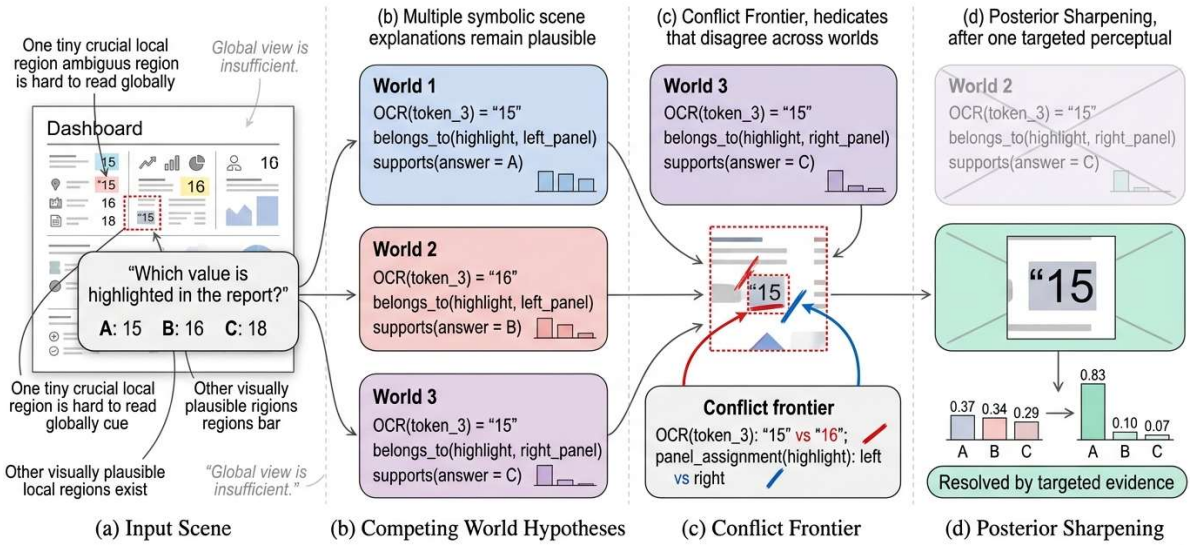

Figure 1. **Symbolic scene ambiguity** under partial visual evidence.

they are, these directions are typically missing one of the ingredients we need here: explicit symbolic alternatives, explicit contradiction structure, or a mechanism that uses those contradictions to guide subsequent perception.

This observation suggests a different view of multimodal inference. Instead of treating additional computation as a way to accumulate more outputs, we should treat it as a way to sharpen a posterior over candidate scene explanations. Under this view, the right intermediate object is a revisable symbolic state: one that can express typed claims about the scene, preserve incompatible alternatives, detect contradictions among them, track where each claim came from, and connect those contradictions to concrete perceptual actions. This perspective is especially important for multimodal reasoning because many failures originate from partially correct but mutually incompatible observations. A model may read one region correctly and another region incorrectly, or it may preserve a plausible but ultimately wrong relational interpretation because nothing in its intermediate state forces those observations to be reconciled. By elevating contradictions to first-class objects, SCOPE aims to turn these scattered local uncertainties into a structured inference problem.

Based on this idea, we develop **SCOPE** (**S**ymbolic **C**onstraint-guided **O**pen-world **P**erception **E**ngine). SCOPE keeps a frontier of uncertainty-aware world hypotheses, where each hypothesis is a typed symbolic scene explanation with explicit provenance and contradiction structure. The method alternates between inducing symbolic observations, consolidating them into candidate worlds, pruning implausible worlds using consistency constraints, and dispatching new visual actions to settle the most consequential disagreements. In this formulation, extra inference-time computation is spent on refining a symbolic belief state rather than simply multiplying raw trajectories.

Our formulation leads to a different role for intermediate computation. The relevant question is not which individual sample appears most trustworthy, but which symbolic distinctions still prevent commitment and which evidence would collapse them. Once the world posterior becomes sufficiently sharp, answer selection follows from marginalization over coherent scene explanations.

Our contributions are threefold. First, we frame scene-level ambiguity in multimodal reasoning as an inference-time state-estimation problem. Second, we introduce SCOPE, a symbolic active-perception framework built around hypothesis induction, constraint-based pruning, and conflict-driven acquisition. Third, we show empirically that this reframing improves performance and yields a more interpretable account of how inference-time computation is used.

2. Related Work

2.1. Inference-Time Computation in Foundation Models

A broad literature studies how to improve model behavior by reallocating effort from training time to inference time (Wang et al., 2022; Li et al., 2024; Yan et al., 2025; Zhang et al., 2025a; Muennighoff et al., 2025; Kang et al., 2025a). Representative strategies include self-consistency, answer reranking, verifier-based selection, iterative refinement, and search over candidate derivations (Aggarwal et al., 2023; Weller et al., 2025; Chen et al., 2024; Madaan et al., 2023; Zhuang et al., 2026). These methods are effective when uncertainty mainly lies in the reasoning path over an already observed input. Recent multimodal extensions follow a similar philosophy by sampling or refining visual-language trajectories (Li et al., 2026; Wang et al., 2025a).

Our setting differs in an important way: the uncertainty that matters is not only over reasoning traces, but over the latent state of the scene itself. SCOPE is therefore closer in spirit to posterior inference over competing scene explanations than to majority voting over outputs.

Another useful contrast is with multimodal tree-search and slow-thinking approaches, which also attempt to spend additional compute more deliberately. For example, recent work explores look-ahead search over interleaved visual-textual reasoning steps and demonstrates strong gains on geometry or spatial tasks (Yan et al., 2025). We share the emphasis on structured inference-time computation, but differ in the object being searched. Rather than searching directly over free-form multimodal trajectories, SCOPE searches over a compact frontier of symbolic worlds with explicit contradiction structure.

2.2. Interactive Visual Reasoning and Active Perception

A growing body of work augments multimodal models with localized visual operations such as cropping, zooming, panning, and iterative inspection (Zheng et al., 2025; Zhang et al., 2025b; Jiang et al., 2026; Wei et al., 2026). These systems show that active evidence acquisition can substantially improve performance on fine-grained and high-resolution tasks. Parallel efforts have also begun to evaluate this capability directly, highlighting a sizable gap between standard multimodal competence and genuinely active perception (Wang et al., 2025b). The main lesson from this line of work is that simply providing a model with access to visual tools is not enough; the system must also decide which perceptual action is worth taking next.

SCOPE builds on this insight but changes the control interface. Existing interactive systems often store intermediate results primarily as free-form text or short local observations. This can be sufficient for simple follow-up actions, but it makes it difficult to compare multiple plausible scene interpretations, detect incompatibilities among them, or quantify which unresolved distinction is most consequential for the answer. SCOPE introduces an explicit symbolic state so that perceptual control is conditioned on structured conflicts among candidate worlds rather than on unstructured textual context alone.

2.3. Structured and Symbolic Intermediate States

There is longstanding interest in introducing explicit structure into multimodal reasoning for compositionality, controllability, and interpretability (Dritsas et al., 2025; Xiong et al., 2025; Xu et al., 2025; Dong et al., 2025). Some approaches impose programmatic or modular decompositions, others encourage more grounded intermediate representations, and recent symbolic reasoning work in language has shown that explicit symbolic formats can improve faithfulness and verifiability. These efforts collectively suggest that natural-language chains are not always the best intermediate object when the task requires precise comparison among alternatives.

However, most existing structured formulations are used either as explanatory artifacts, as static outputs of a reasoning pipeline, or in narrowly constrained symbolic domains. Our use of symbolic structure is different in two respects. First, symbolic worlds are not end products attached after reasoning; they are the evolving state on which inference operates at every round. Second, contradiction structure is not merely diagnostic. It is operationalized to choose subsequent perceptual actions, making the symbolic state both representational and decision-theoretic.

2.4. Where SCOPE Sits

Taken together, prior work has established three ingredients: inference-time computation can improve reasoning, interactive perception can recover missing visual evidence, and structured intermediate states can improve interpretability and compositionality. What remains underdeveloped is a framework that combines all three around a single inference object. SCOPE occupies this gap by treating multimodal inference as posterior sharpening over explicit symbolic scene hypotheses, with contradiction-aware control over future perception.

3. Method

3.1. Problem Formulation

Given an image I, a question Q, and answer options $\mathcal{O}$, we view inference as posterior sharpening over a latent set of scene explanations. The central object in SCOPE is not a reasoning trace but a *symbolic world hypothesis*. At round n, the system maintains

$$\mathcal{W}_n = \{w_n^{(1)}, \ldots, w_n^{(M_n)}\}, \tag{1}$$

where each world hypothesis is a structured tuple

$$w = (\mathcal{P}, \Xi, \alpha, \rho, a). \tag{2}$$

Here $\mathcal{P}$ is a set of typed predicates, Ξ is a set of detected contradictions among predicates, α stores provenance over perceptual actions, ρ is the world score, and $a \in \mathcal{O}$ is the answer favored by the current world.

A predicate is not an arbitrary text span or generic textual note. It is a typed symbolic record designed to support comparison, conflict detection, and answer conditioning. We use a typed schema

$$p = \big(\tau, u, v, r, q, s\big). \tag{3}$$

where $\tau \in \mathcal{T}$ is the predicate type, u is the subject anchor, v is the value or target, r is the supporting region,

$q \in [0, 1]$ is confidence, and s is the supporting evidence source. In our implementation, $\mathcal{T}$ contains four families: unary attribute predicates, counting predicates, relational predicates, and option-conditioned predicates. Examples include $\texttt{ocr}(r_3) =$ "15", $\texttt{count}(\texttt{red_dot}) = 6$, $\texttt{left_of}(o_i, o_j)$, and $\texttt{supports}(p, a_k)$.

This typed state lets us define explicit contradiction structure. We consider four contradiction classes: *value conflicts* (same slot, different value), *cardinality conflicts* (incompatible counts), *spatial conflicts* (geometrically incompatible relations), and *answer conflicts* (predicates that jointly support incompatible options). These categories are useful because they correspond to different failure modes and different follow-up actions. Value conflicts often call for higher-resolution inspection, cardinality conflicts typically require wider or repeated counting evidence, spatial conflicts require inspecting the geometry of a local arrangement, and answer conflicts often arise when several locally plausible claims imply different global conclusions. Rather than collapsing these discrepancies into generic uncertainty, SCOPE keeps them visible and uses them to decide where subsequent perception should focus.

The resulting inference problem is to transform a diffuse posterior over worlds into a concentrated posterior over a small number of mutually coherent explanations. This is the central move of SCOPE: extra compute is allocated to ***collapsing symbolic ambiguity***, not to proliferating independent textual trajectories.

3.2. Framework Overview

SCOPE alternates among three operations: **Induce**, **Prune**, and **Disambiguate**. Induction converts local observations into typed predicates. Pruning expands and scores candidate worlds, then removes worlds that are weakly supported or internally inconsistent. Disambiguation chooses the next perceptual action by asking which unresolved conflict among high-posterior worlds is most consequential for the answer posterior.

3.3. Induce: Uncertainty-Aware Predicate Construction

Each perceptual action returns local evidence, such as a crop, a zoomed patch, OCR output, or an inspected region with associated model commentary. SCOPE maps this evidence into typed predicates rather than appending it directly to a textual scratchpad. The key design choice is that induction is *multi-valued*: when a region admits several plausible readings, the system emits several candidate predicates rather than prematurely collapsing them.

Let $\widetilde{\mathcal{P}}_n$ denote the induced predicates at round n. For a region r, the inducer may produce a local ambiguity set

$$\mathcal{B}(r) = \{p_1^{(r)}, \ldots, p_{m_r}^{(r)}\}, \tag{4}$$

with normalized local support

$$\mu(p \mid r) = \frac{\exp(\beta q(p))}{\sum_{p' \in \mathcal{B}(r)} \exp(\beta q(p'))}. \tag{5}$$

These ambiguity sets are the basic atoms from which worlds are built. The point is not simply to record uncertain observations, but to preserve the *alternative symbolic commitments* that will later compete during pruning.

3.4. Prune: Consistency-Structured World Scoring

World expansion combines prior worlds with newly induced ambiguity sets. A candidate world is produced by selecting at most one predicate from each ambiguity set and merging it with a parent world. This gives a controlled combinatorial space in which each world corresponds to one coherent interpretation of the currently observed evidence. Importantly, the point is not to enumerate every possible symbolic completion of the scene. The pruning stage is designed to keep only a compact frontier of plausible worlds, so the representation remains expressive enough to preserve ambiguity while staying small enough to support iterative refinement. This balance between representational coverage and computational tractability is central to the method: if too few worlds are preserved, the model commits too early; if too many are preserved, later perception loses focus and contradiction structure becomes noisy.

The strength of a world is measured by four quantities. First, evidence support aggregates predicate confidence,

$$U(w) = \sum_{p \in \mathcal{P}(w)} \omega_{\tau(p)} \log(\varepsilon + q(p)). \tag{6}$$

Second, answer alignment measures how strongly the world favors its attached answer,

$$A(w) = \sum_{p \in \mathcal{P}(w)} \psi(p, a(w)). \tag{7}$$

Third, contradiction cost penalizes pairwise and higher-order incompatibilities,

$$C(w) = \sum_{\xi \in \Xi(w)} \eta(\xi). \tag{8}$$

Finally, redundancy regularization discourages worlds that repeatedly rely on near-duplicate evidence,

$$R(w) = \sum_{r} \max\{0, n_r(w) - 1\}, \tag{9}$$

where $n_r(w)$ counts how many predicates in w are supported by the same local region.

The world score is then

$$\rho(w) = U(w) + \lambda_a A(w) - \lambda_c C(w) - \lambda_r R(w). \tag{10}$$

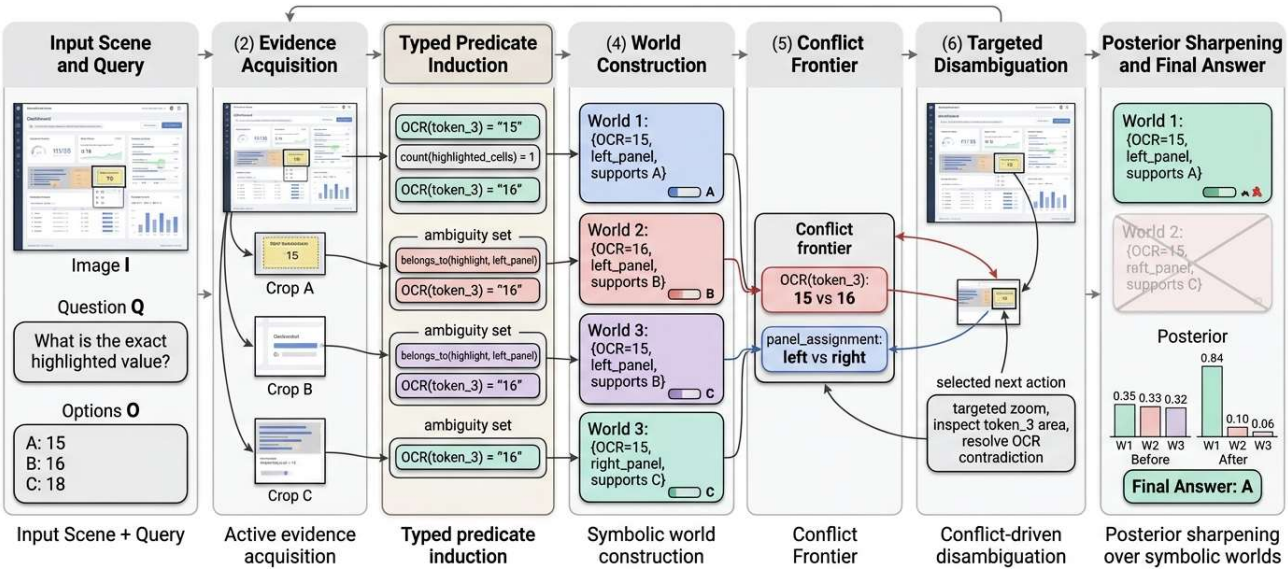

Figure 2. **Overview of SCOPE.** The system alternates between symbolic predicate induction, world construction under consistency constraints, and conflict-driven evidence acquisition. Instead of aggregating trajectories directly, SCOPE maintains a compact set of candidate worlds and updates the answer posterior through symbolic pruning and disambiguation.

This is a fundamentally different object from confidence over a single trajectory. A world may contain locally uncertain predicates but still achieve high global support if its claims are mutually compatible and answer-coherent; conversely, a fluent trajectory can correspond to a low-scoring world if its induced predicates disagree. The scoring function therefore plays two roles simultaneously. First, it aggregates local support into a global symbolic explanation. Second, it regularizes the hypothesis space so that later action selection is driven by a small set of serious alternatives rather than by a diffuse collection of weak candidates. In this sense, pruning is not just a computational optimization; it is part of the inference logic of the method.

SCOPE converts world scores into an answer posterior via

$$\pi_n(a) = \frac{\sum_{w \in \mathcal{W}_n :\, a(w)=a} \exp\big(\rho(w)/\tau\big)}{\sum_{w \in \mathcal{W}_n} \exp\big(\rho(w)/\tau\big)}. \tag{11}$$

The pruning operator keeps the top-M worlds by $\rho(w)$ after deduplicating worlds with identical symbolic content. This turns inference into posterior sharpening over a compact symbolic frontier.

3.5. Disambiguate: Conflict-Driven Action Selection

The most distinctive step in SCOPE is action selection. Not every uncertain predicate deserves another visual query, and not every unresolved local discrepancy is equally important for the answer. The right action is the one expected to distinguish among the *highest-posterior competing worlds*. This is where SCOPE differs most sharply from generic iterative inspection strategies: it does not revisit regions merely because they are hard to perceive, but because the surviving symbolic worlds disagree about them in ways that still influence the answer posterior. Let $\mathcal{W}_n^\star$ be the top-K worlds after pruning. We define the conflict frontier as

$$\Gamma_n = \big\{\xi : \xi \in \Xi(w),\ w \in \mathcal{W}_n^\star\big\}. \tag{12}$$

which collects contradictions that survive among plausible worlds.

For a candidate action a_c targeting region r_c, we define its utility as

$$\Delta(a_c) = \sum_{\xi \in \Gamma_n} m(\xi, r_c)\, \upsilon(\xi)\, \kappa(\xi). \tag{13}$$

where $m(\xi, r_c) \in \{0, 1\}$ indicates whether the action can inspect evidence relevant to contradiction ξ, $\upsilon(\xi)$ is the posterior mass currently entangled in that contradiction, and $\kappa(\xi)$ measures how much resolving ξ is expected to sharpen the answer posterior. Concretely,

$$\begin{aligned} \upsilon(\xi) &= \sum_{w \in \mathcal{W}_n^\star :\, \xi \in \Xi(w)} \pi_n\big(a(w)\big), \\ \kappa(\xi) &= \Big| \mathbb{E}_{w:\, \xi \in \Xi(w)}[a(w)] - \mathbb{E}_{w:\, \xi \notin \Xi(w)}[a(w)] \Big|. \end{aligned} \tag{14}$$

where the second term is implemented as option-posterior separation rather than literal numeric expectation. This criterion embodies the central principle of the method: spend new perceptual effort only where it can collapse disagreements among serious candidate worlds.

Once new evidence arrives, contradicted worlds lose score mass through $C(w)$, while compatible worlds gain relative

posterior weight. The answer prediction after the final round is

$$\hat{a} = \arg\max_{a \in \mathcal{O}} \pi_N(a). \tag{15}$$

This aggregation is not voting over outputs; it is marginalization over a structured set of scene explanations.

4. Experiments

4.1. Experimental Setup

Benchmarks. We evaluate SCOPE on both high-resolution and general multimodal reasoning benchmarks. The high-resolution suite includes V∗ Bench (Wu & Xie, 2024), HR-Bench at 4K and 8K (Wang et al., 2025b), and TreeBench (Wang et al., 2026). Together these datasets stress fine-detail perception, spatial discrimination, and reasoning over large visual scenes, making them particularly suitable for testing whether explicit symbolic ambiguity management is beneficial. We also report results on MME-RealWorld-Lite (Zhang et al., 2024) to assess whether the same inference procedure remains useful outside explicitly high-resolution settings. This combination of benchmarks allows us to evaluate not only whether SCOPE improves final accuracy, but also whether its design transfers across different forms of multimodal difficulty, from OCR-heavy questions to relational and scene-level reasoning.

Baselines. Our comparisons cover both strong multimodal systems and representative inference-time methods. The former group includes GPT-4o (Achiam et al., 2023), Thyme (Zhang et al., 2025b), DeepEyes (Zheng et al., 2025), and Pixel-Reasoner (Wang et al., 2025a). The latter group includes self-consistency (Wang et al., 2022), CISC (Taubenfeld et al., 2025), iterative refinement variants (Madaan et al., 2023; Zhuang et al., 2026), confidence-oriented methods (Kang et al., 2025b; Fu et al., 2025), and RTWI (Li et al., 2026). This comparison set lets us evaluate SCOPE against both competitive base systems and diverse ways of spending additional inference-time compute.

Backbones and implementation. Unless stated otherwise, SCOPE is instantiated with Qwen3-VL 4B and 8B backbones (Bai et al., 2025). The method uses a bounded action budget per round, a capped set of surviving world hypotheses, and dedicated prompts for symbolic induction, consolidation, and action planning. No task-specific fine-tuning or external symbolic supervision is introduced; all structured state updates occur during inference. This choice is deliberate: we want the empirical gains to reflect the value of the inference procedure itself rather than the effects of additional training. In other words, the experiments isolate the contribution of symbolic state organization at test time.

4.2. Main Results

Tables 1 and 2 show that SCOPE improves consistently over both single-pass baselines and prior inference-time strategies. The gains are most visible on settings where answer quality depends on resolving small but consequential perceptual differences, including OCR-heavy examples, fine-grained spatial judgments, and large-scene reasoning. This pattern is consistent with the intended role of symbolic state management: when multiple local interpretations remain plausible, explicitly tracking and pruning them is more helpful than simply adding more loosely organized samples. Equally important, the improvements are not confined to a single benchmark family or a single backbone scale. This suggests that the proposed inference object—explicit symbolic worlds—captures a recurring source of multimodal difficulty rather than a narrow artifact of one dataset.

Relative to self-consistency-style baselines, SCOPE changes not just the amount but the organization of inference-time computation. Relative to RTWI, the main difference is that contradictions among candidate scene explanations are represented and acted upon directly. The especially strong results on TreeBench and HR-Bench suggest that this reorganization matters most when minor perceptual discrepancies can flip the final answer.

4.3. Ablation Study

The ablation results show that the symbolic components play distinct and complementary roles. In particular, the ablations help disentangle where the gains come from. If the method improved solely because it performs more iterative actions, then removing explicit symbolic structure should have limited impact. Instead, we observe that removing the symbolic control mechanisms produces substantial degradation, indicating that the gains come from how inference-time computation is organized rather than from simply increasing its quantity. Removing world-level scoring weakens the system's ability to suppress implausible scene explanations. Removing conflict-driven acquisition reduces the value of later rounds because new observations are no longer directed toward the ambiguities that matter most. Replacing structured predicate induction with a less explicit intermediate representation causes the largest drop, suggesting that the gain does not come merely from taking more actions, but from maintaining a structured state on which those actions can condition.

4.4. Scalability

Figure 3 highlights two complementary ways in which the method benefits from extra computation. The figure is particularly important because it shows that the improvement pattern is structural rather than accidental. Width and depth

Table 1. Performance on high-resolution reasoning benchmarks. **Red** numbers emphasize the strongest improvements, while blue-tinted headers separate benchmark groups for easier scanning.

	V* Bench			HR-Bench-4K			HR-Bench-8K			TreeBench		
Method	Attr.	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall	Perc.	Reas.	Overall
GPT-4o	72.2	60.5	67.5	66.8	63.3	65.0	60.8	58.5	59.6	61.7	38.3	46.9
Thyme	83.5	80.3	82.2	91.0	63.0	77.0	86.5	57.5	72.0	57.0	28.9	39.3
DeepEyes	91.3	88.2	90.1	91.3	59.0	75.1	86.8	58.5	72.6	57.7	29.3	39.8
Pixel-Reasoner	83.5	76.3	80.6	86.0	60.3	72.9	80.0	54.3	66.9	54.3	30.1	39.0
Qwen3-VL-4B-Instruct												
Base	91.3	89.5	90.6	90.0	70.0	80.0	89.8	64.0	76.9	60.4	36.3	45.2
SC	92.2 (-2.6)	90.8 (-1.3)	91.6 (-2.1)	95.0 (-2.3)	72.3 (-2.0)	83.7 (-2.1)	91.5 (-2.0)	73.5 (-1.0)	82.5 (-1.5)	64.4 (-6.1)	37.9 (-1.5)	47.7 (-3.2)
CISC	93.0 (-1.8)	90.8 (-1.3)	92.1 (-1.6)	94.8 (-2.5)	72.3 (-2.0)	83.6 (-2.2)	91.5 (-2.0)	73.3 (-1.2)	82.4 (-1.6)	63.8 (-6.7)	37.5 (-1.9)	47.2 (-3.7)
TRT	92.2 (-2.6)	90.8 (-1.3)	91.6 (-2.1)	95.0 (-2.3)	72.3 (-2.0)	83.7 (-2.1)	91.5 (-2.0)	73.3 (-1.2)	82.4 (-1.6)	63.8 (-6.7)	37.5 (-1.9)	47.2 (-3.7)
Self-Ref.	91.3 (-3.5)	89.5 (-2.6)	90.6 (-3.1)	95.3 (-2.0)	72.5 (-1.8)	83.9 (-1.9)	91.3 (-2.2)	73.5 (-1.0)	82.4 (-1.6)	64.4 (-6.1)	37.9 (-1.5)	47.7 (-3.2)
Self-Cer.	91.3 (-3.5)	89.5 (-2.6)	90.6 (-3.1)	93.0 (-4.3)	72.3 (-2.0)	82.7 (-3.1)	91.5 (-2.0)	73.5 (-1.0)	82.5 (-1.5)	63.7 (-6.8)	37.5 (-1.9)	47.2 (-3.7)
Deepconf	93.0 (-1.8)	**92.1** (-0.0)	92.6 (-1.1)	95.3 (-2.0)	72.5 (-1.8)	83.9 (-1.9)	92.8 (-0.7)	73.8 (-0.7)	83.3 (-0.7)	65.8 (-4.7)	37.9 (-1.5)	48.2 (-2.7)
RTWI	93.9 (-0.9)	90.8 (-1.3)	92.7 (-1.0)	95.5 (-1.8)	73.0 (-1.3)	84.3 (-1.5)	93.0 (-0.5)	73.8 (-0.7)	83.4 (-0.6)	66.4 (-4.1)	39.0 (-0.4)	49.1 (-1.8)
Ours	**95.1** (+1.2)	**92.6** (+0.5)	**93.8** (+1.1)	**97.6** (+2.1)	**74.5** (+1.5)	**86.1** (+1.8)	**94.2** (+1.2)	**74.9** (+1.1)	**84.6** (+1.2)	**71.2** (+4.8)	**40.7** (+1.7)	**51.1** (+2.0)
Qwen3-VL-8B-Instruct												
Base	92.2	89.5	91.1	93.0	69.0	81.0	89.8	71.3	80.6	67.1	34.4	46.4
SC	93.0 (-2.7)	90.8 (-2.6)	92.1 (-2.7)	95.8 (-1.0)	76.8 (-3.0)	86.3 (-2.0)	93.0 (-2.0)	74.8 (-3.0)	83.9 (-2.5)	69.8 (-3.4)	37.1 (-2.3)	49.1 (-2.7)
CISC	93.9 (-1.8)	92.1 (-1.3)	93.2 (-1.6)	95.5 (-1.3)	75.8 (-4.0)	85.7 (-2.6)	91.0 (-4.0)	75.8 (-2.0)	83.4 (-3.0)	67.8 (-5.4)	36.7 (-2.7)	48.2 (-3.6)
TRT	92.2 (-3.5)	90.8 (-2.6)	91.6 (-3.2)	95.3 (-1.5)	77.0 (-2.8)	86.2 (-2.1)	93.0 (-2.0)	74.8 (-3.0)	83.9 (-2.5)	69.1 (-4.1)	37.1 (-2.3)	48.9 (-2.9)
Self-Ref.	92.2 (-3.5)	90.8 (-2.6)	91.6 (-3.2)	93.8 (-3.0)	77.0 (-2.8)	85.4 (-2.9)	93.0 (-2.0)	74.8 (-3.0)	83.9 (-2.5)	69.1 (-4.1)	36.7 (-2.7)	48.6 (-3.2)
Self-Cer.	93.0 (-2.7)	89.5 (-3.9)	91.6 (-3.2)	95.3 (-1.5)	75.8 (-4.0)	85.6 (-2.7)	88.8 (-6.2)	75.3 (-2.5)	82.1 (-4.3)	67.8 (-5.4)	36.7 (-2.7)	48.2 (-3.6)
Deepconf	93.9 (-1.8)	92.1 (-1.3)	93.2 (-1.6)	96.0 (-0.8)	76.8 (-3.0)	86.4 (-1.9)	93.5 (-1.5)	75.5 (-2.3)	84.5 (-1.9)	70.4 (-2.8)	37.9 (-1.5)	49.9 (-1.9)
RTWI	94.8 (-0.9)	92.1 (-1.3)	93.7 (-1.1)	95.8 (-1.0)	77.0 (-2.8)	86.4 (-1.9)	94.5 (-0.5)	77.0 (-0.8)	85.8 (-0.6)	71.1 (-2.1)	37.9 (-1.5)	50.1 (-1.7)
Ours	**96.6** (+1.8)	**94.5** (+2.4)	**95.1** (+2.4)	**96.9** (+1.1)	**80.3** (+3.3)	**89.5** (+3.1)	**96.1** (+1.6)	**78.2** (+1.2)	**86.9** (+1.1)	**73.7** (+2.6)	**40.1** (+1.8)	**52.3** (+2.4)

Table 2. Performance on the general multimodal benchmark MME-RealWorld-Lite. We adopt a more saturated visual hierarchy to make **overall gains** immediately visible.

	Perception						Reasoning				
Method	OCR	RS	DT	MO	AD	Overall	OCR	DT	MO	AD	Overall
Pixel-Reasoner	89.6	52.0	86.0	38.9	30.9	53.1	71.0	72.0	46.0	32.5	45.6
DeepEyes	90.0	52.7	89.0	43.3	33.4	55.4	76.0	69.0	44.0	35.0	46.8
Qwen3-VL-4B-Instruct											
Base	89.6	50.7	85.0	37.3	34.3	53.4	73.0	75.0	42.7	31.0	44.8
SC	91.2 (-2.0)	53.3 (-2.7)	89.0 (-2.0)	38.6 (-1.2)	36.9 (-3.1)	55.5 (-2.2)	78.0 (-3.0)	78.0 (-5.0)	43.3 (-0.7)	32.8 (-1.5)	47.0 (-2.0)
CISC	91.2 (-2.0)	53.3 (-2.7)	89.0 (-2.0)	38.6 (-1.2)	37.4 (-2.6)	55.7 (-2.0)	78.0 (-3.0)	78.0 (-5.0)	42.7 (-1.3)	32.3 (-2.0)	46.6 (-2.4)
TRT	90.8 (-2.4)	53.3 (-2.7)	89.0 (-2.0)	38.6 (-1.2)	36.9 (-3.1)	55.5 (-2.2)	77.0 (-4.0)	79.0 (-4.0)	43.3 (-0.7)	32.8 (-1.5)	47.0 (-2.0)
Self-Ref.	91.2 (-2.0)	53.3 (-2.7)	89.0 (-2.0)	38.2 (-1.6)	36.9 (-3.1)	55.4 (-2.3)	78.0 (-3.0)	78.0 (-5.0)	42.7 (-1.3)	32.3 (-2.0)	46.6 (-2.4)
Self-Cer.	90.8 (-2.4)	54.0 (-2.0)	89.0 (-2.0)	38.6 (-1.2)	36.9 (-3.1)	55.5 (-2.2)	78.0 (-3.0)	78.0 (-5.0)	42.7 (-1.3)	32.8 (-1.5)	46.8 (-2.2)
Deepconf	91.6 (-1.6)	53.3 (-2.7)	89.0 (-2.0)	38.9 (-0.9)	37.4 (-2.6)	55.9 (-1.8)	78.0 (-3.0)	79.0 (-4.0)	43.3 (-0.7)	32.8 (-1.5)	47.1 (-1.9)
RTWI	92.4 (-0.8)	54.7 (-1.3)	89.0 (-2.0)	39.2 (-0.6)	38.3 (-1.7)	56.6 (-1.1)	79.0 (-2.0)	80.0 (-3.0)	43.3 (-0.7)	32.8 (-1.5)	47.2 (-1.8)
Ours	**93.8** (+1.4)	**56.8** (+2.1)	**91.2** (+2.2)	**40.5** (+1.3)	**40.9** (+2.6)	**58.4** (+1.8)	**81.9** (+2.9)	**83.5** (+3.5)	**45.1** (+1.8)	**35.1** (+2.3)	**49.6** (+2.4)
Qwen3-VL-8B-Instruct											
Base	89.2	54.7	86.0	34.8	33.1	52.9	74.0	80.0	34.0	35.3	46.2
SC	90.0	58.0	89.0	39.2	38.6	56.6	81.0	83.0	40.7	36.0	49.2
CISC	90.0	60.0	89.0	39.2	38.6	56.8	80.0	83.0	40.7	36.0	49.1
TRT	90.0	60.0	89.0	39.2	38.6	56.8	81.0	83.0	40.7	36.0	49.2
Self-Ref.	90.4	60.0	89.0	38.9	38.3	56.7	82.0	83.0	40.0	36.3	49.4
Self-Cer.	90.0	60.0	92.0	39.2	38.6	57.1	81.0	83.0	40.7	36.3	49.4
Deepconf	90.0	61.3	90.0	39.5	38.9	57.2	82.0	83.0	41.3	36.5	49.7
RTWI	90.4	62.7	92.0	39.8	38.9	57.8	83.0	84.0	42.0	36.5	50.1
Ours	**91.5** (+1.1)	**66.1** (+3.4)	**95.0** (+3.0)	**41.5** (+1.7)	**40.0** (+1.1)	**60.0** (+2.2)	**85.0** (+2.0)	**88.0** (+4.0)	**45.7** (+3.7)	**39.5** (+3.0)	**53.4** (+3.4)

Table 3. Ablation study of the key components of SCOPE. WS denotes symbolic world scoring, CD denotes conflict-driven acquisition, and SP denotes structured predicate induction. Colored rows and explicit +/- cues improve visual comparison.

	V* Bench		HR-Bench-4K		HR-Bench-8K	
Variants	Attr.	Spatial	FSP	FCP	FSP	FCP
Qwen3-VL-8B-Instruct						
SC	93.0 (-2.7)	90.8 (-2.6)	95.8 (-1.0)	76.8 (-3.0)	93.0 (-2.0)	74.8 (-3.0)
w/o RF	93.0 (-2.7)	**93.4** (-0.0)	96.3 (-0.5)	78.0 (-1.8)	**95.0** (-0.0)	77.5 (-0.3)
w/o SK	93.0 (-2.7)	90.8 (-2.6)	96.0 (-0.8)	77.0 (-2.8)	93.5 (-1.5)	75.8 (-2.0)
w/o WA	94.8 (-0.9)	**93.4** (-0.0)	96.3 (-0.5)	78.3 (-1.5)	94.5 (-0.5)	**77.8** (-0.0)
SCOPE	**96.6**	**94.5**	**96.9**	**80.3**	**96.1**	**78.2**

do not behave like redundant hyperparameters; instead, they correspond to two different resources in the inference process. Width governs how many plausible symbolic alternatives remain under consideration, while depth governs how many rounds of conflict resolution the method can perform. Increasing symbolic width allows SCOPE to keep more candidate scene explanations alive before pruning, which is useful when early observations are genuinely ambiguous. Increasing sequential depth gives the method more opportunities to revisit unresolved distinctions with targeted actions. The two axes therefore play different roles: one broadens

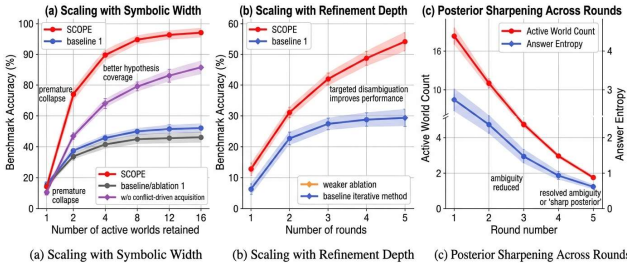

Figure 3. Scalability analysis of SCOPE along three dimensions: (a) model size, (b) perception width, and (c) perception depth.

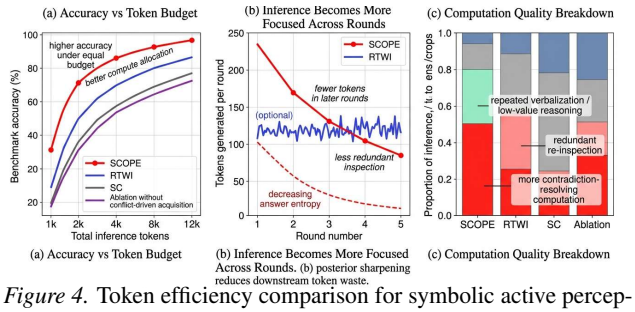

Figure 4. Token efficiency comparison for symbolic active perception.

representational coverage over possible worlds, while the other sharpens the posterior through iterative disambiguation.

4.5. Complexity and Efficiency

Computational profile. With an action budget of at most B per round and a world cap of M, the dominant cost of SCOPE lies in evidence acquisition together with symbolic world expansion and pruning. In practice, the structured bookkeeping remains lightweight relative to multimodal decoding because the maintained state stores compact predicate records rather than repeatedly re-verbalized observations. This distinction is important for understanding why SCOPE remains practical: the method adds structure to the inference loop, but the added structure is deliberately compact. It is therefore better viewed as a reallocation of test-time computation than as a qualitatively heavier inference regime.

Token usage. Figure 4 shows that SCOPE remains competitive in token consumption even though it maintains explicit symbolic state. A key reason is that recurring evidence is compressed into reusable predicates, which reduces the need to restate the same observations across rounds. The resulting tradeoff is favorable: the method gains more deliberate control over evidence gathering without a matching increase in language-generation overhead.

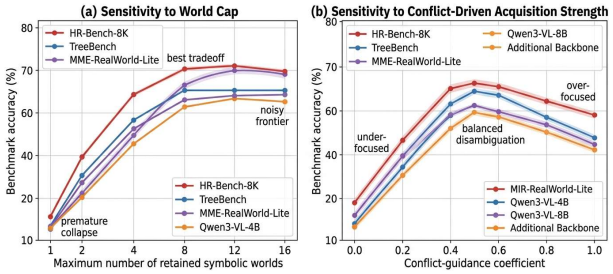

Figure 5. Hyperparameter sensitivity of SCOPE with respect to world cap and conflict-acquisition strength.

4.6. Hyperparameter Sensitivity

We examine two symbolic control parameters: the maximum number of worlds retained after pruning and the strength assigned to conflict-driven acquisition. Extremely small world caps force commitment too early, while overly large caps preserve too many weak explanations. Likewise, if conflict-driven acquisition is too weak, later rounds remain diffuse; if it is too strong, the policy may overfocus on one uncertainty source prematurely. The best performance appears in the middle regime, where the state remains expressive enough to encode alternatives but selective enough to support decisive updates.

5. Conclusion

We introduced SCOPE, a multimodal inference framework that represents intermediate uncertainty through explicit symbolic world hypotheses. Instead of treating extra test-time computation as unstructured repeated prompting, SCOPE organizes it around a persistent symbolic state that can be expanded, scored, and selectively refined.

Across diverse benchmarks, this state-centric formulation produces stronger results and offers a clearer picture of how additional inference-time computation changes the system's beliefs about the scene. The empirical pattern is consistent across backbone scales and task families: the gains are largest when success depends on resolving local ambiguities, comparing fine-grained alternatives, or coordinating evidence across multiple rounds.

Impact Statement

Our goal is to improve the reliability and interpretability of multimodal reasoning systems by making their intermediate decision state more explicit. The same capability could also strengthen automated processing in domains involving surveillance, document triage, or decision support. We therefore view symbolic transparency as necessary but not sufficient: deployment in sensitive settings still requires calibration, human oversight, and task-specific safeguards.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Aggarwal, P., Madaan, A., Yang, Y., et al. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12375–12396, 2023.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Dong, Y., Liu, Z., Sun, H.-L., Yang, J., Hu, W., Rao, Y., and Liu, Z. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9062–9072, 2025.

Dritsas, E., Trigka, M., Troussas, C., and Mylonas, P. Multimodal interaction, interfaces, and communication: a survey. *Multimodal Technologies and Interaction*, 9(1):6, 2025.

Fu, Y., Wang, X., Tian, Y., and Zhao, J. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025.

Jiang, Z., Guo, H., Fang, C., Xiao, C., Hu, X., Sun, L., and Xu, M. Medvr: Annotation-free medical visual reasoning via agentic reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Kang, Z., Zhao, X., and Song, D. Scalable best-of-n selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*, 2025a.

Kang, Z., Zhao, X., and Song, D. Scalable best-of-n selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*, 2025b.

Li, H., Yang, Y., Lin, Y., Xiang, D., Yang, M., and Peng, X. Reliable thinking with images. *arXiv preprint arXiv:2602.12916*, 2026.

Li, Y., Yuan, P., Feng, S., Pan, B., Wang, X., Sun, B., Wang, H., and Li, K. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. *arXiv preprint arXiv:2401.10480*, 2024.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., and Hashimoto, T. B. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Taubenfeld, A., Sheffer, T., Ofek, E., Feder, A., Goldstein, A., Gekhman, Z., and Yona, G. Confidence improves self-consistency in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, 2025.

Wang, H., Su, A., Ren, W., Lin, F., and Chen, W. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.

Wang, H., Li, X., Huang, Z., Wang, A., Wang, J., Zhang, T., Bai, S., Kang, Z., Feng, J., Zhuochen, W., et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *The Fourteenth International Conference on Learning Representations*, 2026.

Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Yu, W., and Tao, D. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025b.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Wei, L., He, L., Lan, J., Dong, L., Cai, Y., Li, S., Zhu, H., Wang, W., Kong, L., Wang, Y., et al. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Weller, O., Ricci, K., Yang, E., Yates, A., Lawrie, D., and Van Durme, B. Rank1: Test-time compute for reranking in information retrieval. *arXiv preprint arXiv:2502.18418*, 2025.

Wu, P. and Xie, S. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Xiong, T., Wang, X., Guo, D., Ye, Q., Fan, H., Gu, Q., Huang, H., and Li, C. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13618–13628, 2025.

Xu, W., Wang, J., Wang, W., Chen, Z., Zhou, W., Yang, A., Lu, L., Li, H., Wang, X., Zhu, X., et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025.

Yan, H., Xu, F., Xu, R., Li, Y., Zhang, J., Luo, H., Wu, X., Tuan, L. A., Zhao, H., Lin, Q., et al. Mur: Momentum uncertainty guided reasoning for large language models. *arXiv preprint arXiv:2507.14958*, 2025.

Zhang, Q., Lyu, F., Sun, Z., Wang, L., Zhang, W., Hua, W., Wu, H., Guo, Z., Wang, Y., Muennighoff, N., et al. A survey on test-time scaling in large language models: What, how, where, and how well? *arXiv preprint arXiv:2503.24235*, 2025a.

Zhang, Y.-F., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

Zhang, Y.-F., Lu, X., Yin, S., Fu, C., Chen, W., Hu, X., Wen, B., Jiang, K., Liu, C., Zhang, T., et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025b.

Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., and Yu, X. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Zhuang, Y., Singh, C., Liu, L., Shen, Y., Zhang, D., Shang, J., Gao, J., and Chen, W. Test-time recursive thinking: Self-improvement without external feedback. *arXiv preprint arXiv:2602.03094*, 2026.