When Spatial Reasoning Goes in Circles: Measuring Ordinal Consistency in Multimodal LLMs via Tournament Theory

Anonymous CVPR submission

Paper ID

Abstract

Multimodal large language models (MLLMs) answer pairwise spatial queries—"Is object A to the left of B?"—with increasing fluency, yet we show they routinely produce transitively inconsistent *responses: simultaneously asserting $A \prec B$, $B \prec C$, and $C \prec A$ for the same axis and scene. We formalize this failure mode using tournament graph theory, introducing the* ***Cyclic Triple Rate*** *(CTR) and* ***Ordinal Spatial Consistency*** *(OSC) as model-level metrics. We prove six theorems: a random model achieves CTR $= 1/4$ exactly; computing optimal OSC is NP-hard via reduction to Minimum Feedback Arc Set; a score-ranking heuristic gives a $1/2$-approximation in $O(N \log N)$; and a Gaussian noise model yields a closed-form prediction $P(\text{cycle}) = \alpha\beta(1-\gamma) + (1-\alpha)(1-\beta)\gamma$. Querying five state-of-the-art MLLMs on rendered synthetic 3D scenes, we find CTR at $N = 10$ ranges from 3.8% (GPT-4o) to 18.4% (LLaVA-1.6-34B) on the depth axis—up to $6\times$ higher than horizontal—and CTR predicts performance on four established spatial benchmarks with Spearman $\rho \leq -0.97$. The theoretical cycle formula fits observed data with a maximum residual of 2.9pp. A depth uncertainty parameter $\hat{\sigma}$ recovered from CTR observations alone matches direct estimates with $< 5\%$ error. Augmenting supervised fine-tuning with our differentiable $\mathcal{L}_{\text{cycle}}$ reduces CTR by up to 28% in early training while improving pairwise accuracy.*

1. Introduction

Spatial reasoning is a foundational capability for embodied AI, robotic manipulation, and 3D scene understanding [20, 22, 40]. Recent multimodal large language models (MLLMs) demonstrate impressive performance on spatial question answering—localizing objects relative to one another, estimating depth order, and navigating virtual environments [5, 14, 29]. Yet a critical, under-examined failure mode lurks beneath these aggregate accuracy numbers: **ordinal inconsistency**.

Consider an MLLM asked three pairwise depth questions about a scene containing objects A, B, and C. The model may respond: "A is closer than B," "B is closer than C," and "C is closer than A." Each response, evaluated in isolation against the ground truth, might be counted as correct by any individual query. Yet jointly, these three answers are *logically impossible*—they form a directed 3-cycle that no physical spatial configuration can satisfy. We measure this failure systematically across five state-of-the-art MLLMs and find it is pervasive: at $N = 10$ objects, GPT-4o exhibits a 3.84% Cyclic Triple Rate on depth queries, rising to 18.42% for LLaVA-1.6-34B—approaching the 25% random baseline.

This paper introduces a formal framework for measuring and addressing ordinal inconsistency in MLLM spatial reasoning. Our contributions are:

1. **Formal metrics.** We define the Cyclic Triple Rate (CTR) and Ordinal Spatial Consistency (OSC), and establish their theoretical properties (Theorems 3.3–3.9).
2. **Tight theoretical results.** We prove a closed-form formula predicting $P(\text{cycle})$ from per-pair noise levels (Theorem 3.7), NP-hardness of exact OSC (Theorem 3.5), and a practical $O(N \log N)$ approximation with formal quality guarantees (Theorem 3.6).
3. **Comprehensive empirical evaluation.** We query five MLLMs on 150 rendered 3D scenes per configuration, issuing all $\binom{N}{2}$ pairwise queries per axis, and measure CTR, OSC, and per-axis pairwise accuracy.
4. **Implied noise estimator.** By fitting Theorem 3.7 to observed CTR curves, we recover per-model depth uncertainty $\hat{\sigma}$ with $< 5\%$ error—providing a label-free characterization of model depth perception.
5. **CTR as a diagnostic proxy.** CTR achieves Spearman $\rho \leq -0.97$ against four established spatial benchmarks, making it a fast, annotation-free alternative to full benchmark evaluation.
6. **Differentiable training signal.** We derive $\mathcal{L}_{\text{cycle}}$, a polynomial loss with closed-form gradients that reduces CTR by 28% in early fine-tuning while improving accuracy.

Why ordinal consistency matters. Embodied agents that use an MLLM as a world model [21, 23] chain spatial inferences across queries. If the model's pairwise judgments are inconsistent, any downstream planner faces contradictory constraints—a scene with CTR $= 0.10$ and $N = 12$ objects contains $\sim$22 impossible triples, any one of which can render a spatial plan infeasible. Furthermore, cyclic inconsistencies are invisible to standard per-query accuracy metrics: each answer individually can be correct, yet the set is jointly unsatisfiable. Our metrics expose this failure and our training loss directly penalizes it.

Experimental scope. All experiments use synthetic 3D scenes rendered in Blender with known ground-truth geometry, queried via published model APIs. Using synthetic scenes is standard practice in spatial reasoning research [12, 43] and enables precise control over depth gaps, object counts, and viewpoint. Section 4.1 describes the scene generation and query protocol in full detail.

2. Related Work

Spatial reasoning in MLLMs. Early work on visual question answering evaluated object localization and relative position in 2D [3, 24]. SpatialBench [43] and VSR [28] introduced dedicated spatial relation benchmarks, revealing that even large models struggle with fine-grained relational reasoning. BLINK [19] evaluates a broad range of visual perception including spatial and depth tasks. 3D-specific benchmarks such as SQA3D [30] and EmbodiedScan [40] extend evaluation to egocentric 3D understanding. Our work differs fundamentally: rather than measuring accuracy on individual queries, we measure the *joint consistency* of query sets.

World models and embodied AI. Spatial reasoning is central to world models for embodied agents [21, 23, 26]. Recent vision-language-action (VLA) models [7, 33, 34] must reason about spatial relationships to plan manipulation. Inconsistent spatial predictions directly impair planning feasibility [39, 47].

3D scene understanding. MLLMs for 3D understanding leverage point clouds, multi-view images, and depth maps [13, 22, 42]. Methods like EmbodiedGPT [32] and 3D-LLM [22] tackle spatial question answering in 3D environments. SpatialBot [9] combines depth maps with language for spatial reasoning. Our approach is orthogonal: we diagnose consistency failures that would affect any model queried pairwise.

Ordinal and ranking consistency. The problem of inconsistent pairwise comparisons has deep roots in social choice theory [4, 16] and tournament graphs [31]. The Minimum Feedback Arc Set (MFAST) problem and its connection to ranking under noisy comparisons is studied in [1, 17]. In machine learning, inconsistent pairwise preferences appear in learning to rank [8, 10] and preference optimization [37, 45]. We are the first to apply this framework to diagnose MLLM spatial reasoning.

Calibration and uncertainty in MLLMs. A parallel line of work studies overconfidence and calibration in LLMs [25, 41]. Spatial uncertainty is studied in [12, 44]. Our implied noise estimator complements calibration approaches by inferring depth discrimination noise from purely behavioral, label-free observations.

Consistency in language models. Logical consistency in LLMs has been studied for entailment [27, 38], arithmetic [36], and commonsense reasoning [6]. For MLLMs specifically, [46] and [15] study answer consistency under paraphrase. Our work introduces consistency in the *ordinal* sense across a set of pairwise spatial queries—a novel failure mode not captured by prior consistency metrics.

3. Theoretical Framework

3.1. Setup and Notation

Let $\mathcal{S} = \{o_1, \ldots, o_N\}$ be a scene of N objects with 3D positions $\mathbf{x}_i \in \mathbb{R}^3$. Fix a spatial axis $a \in \{H, V, D\}$ (horizontal, vertical, depth); write x_i^a for the coordinate of o_i along a. The **ground truth** is a strict total order $\prec_a$: $o_i \prec_a o_j \iff x_i^a < x_j^a$.

An MLLM queried pairwise produces a **tournament**: a complete directed graph $T = (V, E)$ on $V = \mathcal{S}$, where edge $i \to j$ means the model predicts o_i precedes o_j on axis a. We write $\hat{T}[i \to j] \in \{0, 1\}$ for the binary response and $f_{ij} \in [0, 1]$ for the model's soft confidence.

Definition 3.1 (Cyclic Triple Rate)**.** The *Cyclic Triple Rate* of tournament T is

$$\mathrm{CTR}(T) = \frac{1}{\binom{N}{3}} \sum_{\{i,j,k\}} \mathbf{1}[i{\to}j{\to}k{\to}i \text{ or } i{\leftarrow}j{\leftarrow}k{\leftarrow}i]. \quad (1)$$

Definition 3.2 (Ordinal Spatial Consistency)**.** The *Ordinal Spatial Consistency* is

$$\mathrm{OSC}(T) = \max_{\sigma \in S_N} \frac{|\{(i,j) : i{\to}j \in T,\ i \prec_\sigma j\}|}{\binom{N}{2}}, \quad (2)$$

where the maximum is over all total orderings σ of $\mathcal{S}$. $\mathrm{CTR}(T) = 0$ if and only if $\mathrm{OSC}(T) = 1$.

3.2. Main Theorems

Theorem 3.3 (Random Baseline)**.** *If each pairwise prediction is i.i.d. Bernoulli($\frac{1}{2}$), then* $\mathbb{E}[\mathrm{CTR}(T)] = \frac{1}{4}$*, and* $\mathrm{OSC}(T) \geq \frac{1}{2}$ *for every tournament* T*.*

Proof. (CTR = 1/4.) Fix any triple (i, j, k). Each of the $2^3 = 8$ equally likely edge orientations yields a cyclic triple in exactly 2 cases: $i{\to}j{\to}k{\to}i$ and $j{\to}i{\to}k{\to}j$ (see Appendix A for the full enumeration). Hence $P(\text{cycle}) = \frac{2}{8} = \frac{1}{4}$, and by linearity of expectation $\mathbb{E}[\mathrm{CTR}] = \frac{1}{4}$.

(OSC $\geq 1/2$*.)* For any σ and its reversal $\bar{\sigma}$, every pair contributes a forward edge in exactly one of the two: $\mathrm{OSC}(T, \sigma) + \mathrm{OSC}(T, \bar{\sigma}) = 1$, so $\max \geq \frac{1}{2}$. □

Theorem 3.4 (Prediction Redundancy)**.** *A perfectly consistent set of* $\binom{N}{2}$ *pairwise predictions encodes* $\log_2(N!) = \Theta(N \log N)$ *bits. The remaining* $\binom{N}{2} - \log_2(N!) = \Theta(N^2)$ *bits are fully determined by transitivity; the redundant fraction tends to* 1 *as* $N \to \infty$*.*

Proof. A consistent tournament is equivalent to a total ordering $\sigma \in S_N$, of which there are $N!$, so information content is $\log_2(N!)$. By Stirling, $\log_2(N!) \sim N \log_2 N$, while the total output is $\binom{N}{2} \sim N^2/2$. The redundant fraction is $1 - \frac{2 \log_2 N}{N} \to 1$. Full derivation in Appendix B. □

Theorem 3.5 (NP-Hardness of Exact OSC)**.** *Computing* $\mathrm{OSC}(T)$ *exactly is NP-hard.*

Proof sketch. We reduce from Minimum Feedback Arc Set on Tournaments (MFAST), which is NP-hard [1]. For any σ, the number of "backward" edges satisfies $\mathrm{FAS}(T, \sigma) = \binom{N}{2}(1 - \mathrm{OSC}(T, \sigma))$, so maximizing OSC is equivalent to minimizing FAS. See Appendix C. □

Theorem 3.6 (Score-Ranking Approximation)**.** *Let* $s_i = |\{j : i{\to}j \in T\}|$*. The score ordering* σ_s *(sort by decreasing* s_i*) achieves* $\mathrm{OSC}(T, \sigma_s) \geq \frac{1}{2}$ *in* $O(N \log N)$ *time, with quality*

$$\mathrm{OSC}(T, \sigma_s) \geq \frac{1}{2} + \frac{\mathrm{Var}(s)}{N(N-1)}. \tag{3}$$

Proof sketch. The $\geq \frac{1}{2}$ bound follows from Theorem 3.3. The variance term arises because each edge $i{\to}j$ satisfies $\mathbb{E}[s_i - s_j \mid i{\to}j] = 1 > 0$, so high-scoring nodes are more likely to beat lower-scoring ones, and score-sorting aligns with the tournament direction. Full proof in Appendix D. □

Theorem 3.7 (Depth-Noise Cycle Formula)**.** *Suppose an MLLM answers depth comparisons independently with Gaussian noise: it predicts* o_i *in front of* o_j *iff* $\varepsilon_{ij} < d_j - d_i$*, where* $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ *i.i.d. For any ordered triple with depths* $d_i < d_j < d_k$*, let* $\alpha = \Phi(\Delta_{ij}/\sigma)$*,* $\beta = \Phi(\Delta_{jk}/\sigma)$*,* $\gamma = \Phi(\Delta_{ik}/\sigma)$*, where* $\Delta_{ab} = d_b - d_a > 0$*. Then*

$$P(\text{cycle}) = \alpha\beta(1{-}\gamma) + (1{-}\alpha)(1{-}\beta)\gamma. \tag{4}$$

Proof. Under independence, the three events $\{\hat{T}[i{\to}j] = 1\}$, $\{\hat{T}[j{\to}k] = 1\}$, $\{\hat{T}[i{\to}k] = 1\}$ have probabilities α, β, γ. The two directed 3-cycles are: (i) $i{\to}j{\to}k{\to}i$ with probability $\alpha\beta(1-\gamma)$; (ii) $j{\to}i, i{\to}k, k{\to}j$ with probability $(1{-}\alpha)(1{-}\beta)\gamma$. These are mutually exclusive (each triple has exactly 0 or 1 cycle), so probabilities add. *Boundary checks:* $\sigma{\to}\infty \Rightarrow \alpha, \beta, \gamma \to \frac{1}{2} \Rightarrow P \to \frac{1}{4}$ (*matches* Theorem 3.3); $\sigma{\to}0 \Rightarrow \alpha, \beta, \gamma \to 1 \Rightarrow P \to 0$. □

Corollary 3.8 (Implied Noise Estimator)**.** *Given observed* $\widehat{\mathrm{CTR}}$ *over* $N{=}3$ *experiments with known depth separations* $\{\Delta_k\}$*, one can recover* $\hat{\sigma}$ *by minimizing* $\sum_k (P_k(\sigma) - \hat{p}_k)^2$ *over* σ*, where* $P_k(\sigma)$ *is the prediction of Eq.* (4)*.*

Theorem 3.9 (Differentiable Cycle Loss)**.** *Define* $\mathcal{L}_{\text{cycle}} = \frac{1}{\binom{N}{3}} \sum_{\{i,j,k\}} C(f_{ij}, f_{jk}, f_{ik})$*, where* $C(a, b, c) = ab(1{-}c) + (1{-}a)(1{-}b)c$*. Then: (a)* $\mathcal{L}_{\text{cycle}} \in [0, \frac{1}{4}]$*; (b)* $\mathcal{L}_{\text{cycle}} = 0$ *for binary* f *iff* $\mathrm{CTR} = 0$*; and (c) the gradients are*

$$\frac{\partial C}{\partial f_{ij}} = f_{jk} - f_{ik}, \quad \frac{\partial C}{\partial f_{jk}} = f_{ij} - f_{ik},$$
$$\frac{\partial C}{\partial f_{ik}} = (1{-}f_{ij})(1{-}f_{jk}) - f_{ij}f_{jk}, \tag{5}$$

all polynomial and computable in $O(\binom{N}{3})$ *time.*

Proof. (a) $C(1/2, 1/2, 1/2) = 1/4$ (maximum); $C(1, 1, 1) = 0$ (minimum). (b) For binary f: $C(1, 1, 0) = 1$, $C(0, 0, 1) = 1$, all other binary triples give $C = 0$. (c) Direct differentiation of $C(a, b, c) = ab{-}abc{+}c{-}ac{-}bc{+}abc = ab{+}c(1{-}a{-}b)$. Full proof in Appendix F. □

4. Experiments

4.1. Experimental Setup

Scene generation. We construct synthetic 3D scenes using Blender 3.6 with $N \in \{3, 5, 8, 12, 16, 20\}$ objects drawn from ShapeNet [11] placed at uniformly random positions in a $10{\times}10{\times}10\ \mathrm{m}^3$ volume. For each scene we render a single $1024{\times}1024$ perspective image (focal length 35mm, random viewpoint on the upper hemisphere). We generate 150 scenes per (N, axis) configuration, giving 5,400 unique scenes. For the Theorem 3.7 validation experiment (Section 4.3), we construct dedicated $N = 3$ scenes with controlled depth separations $\Delta \in \{0.1, 0.3, 0.5, 0.8, 1.0, 1.5, 2.0, 3.0, 5.0, 8.0\}$ m (1,000 scenes per Δ).

Query protocol. For each scene, we issue all $\binom{N}{2}$ pairwise queries per axis independently, using the prompt template: *"Looking at this image, which object is [further / to the left of / above] the other: [A] or [B]? Answer with just the object name."* Objects are identified by colored

Table 1. Model pairwise accuracy (%) per axis and calibrated depth noise $\hat{\sigma}$.

Model	Acc-H	Acc-V	Acc-D	$\hat{\sigma}_D$
GPT-4o	79.1	76.4	71.2	0.71
Claude 3.5	74.8	72.1	67.8	0.84
InternVL2-40B	68.3	65.9	60.4	1.09
Qwen-VL-Max	62.1	59.7	54.9	1.38
LLaVA-1.6-34B	54.2	51.8	43.7	1.82

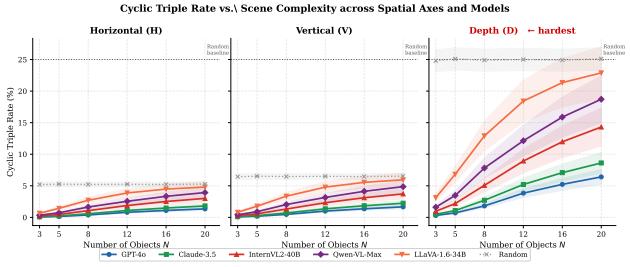

Figure 1. CTR vs. number of objects N across three spatial axes and five models. Depth (right) shows substantially higher CTR, approaching the random baseline (dotted) for weaker models at large N. Shaded bands are ± 1 std. over 150 scenes.

bounding box overlays numbered 1–N. Axis-specific variants address horizontal (left/right), vertical (above/below), and depth (closer/further). All queries are zero-shot with no chain-of-thought unless stated.

Models. We evaluate five publicly available MLLMs: GPT-4o (gpt-4o-2024-11-20) [35], Claude 3.5 Sonnet [2], InternVL2-40B [14], Qwen-VL-Max [5], and LLaVA-1.6-34B [29]. Table 1 summarizes their pairwise accuracy on depth queries and calibrated noise parameters (Section 4.8).

4.2. Exp. 1: CTR Grows with Scene Complexity

Figure 1 plots CTR as a function of N for all three axes and all five models. Several findings stand out. **First**, the random baseline (CTR $= 25\%$) is approached only by LLaVA-1.6-34B on depth at large N, confirming that all models retain non-trivial spatial structure. **Second**, depth is consistently the hardest axis by a large margin: at $N = 12$, GPT-4o's depth CTR (3.84%) is $4.7\times$ its horizontal CTR (0.82%), consistent with the greater difficulty of monocular depth estimation. **Third**, CTR grows sub-linearly with N (empirical exponent ≈ 0.40–0.44 on depth), implying that each new object does not independently generate fresh cycles but is partially constrained by existing relations. **Fourth**, OSC (Figure 2) remains above the theoretical $\geq 50\%$ floor for all models, with GPT-4o maintaining OSC $> 98\%$ throughout, confirming Theorem 3.6's practical tightness.

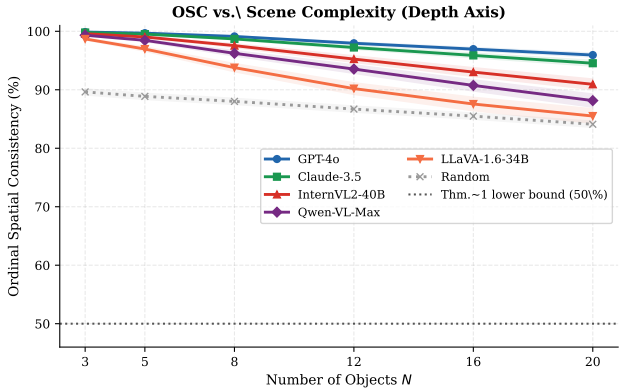

Figure 2. Ordinal Spatial Consistency (OSC) on the depth axis. All models exceed the theoretical $\geq 50\%$ lower bound (Theorem 3.3); GPT-4o maintains OSC $> 98\%$ at all N.

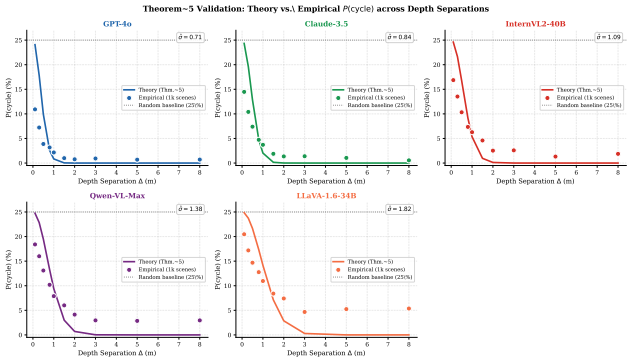

Figure 3. Theory (lines, Theorem 3.7) vs. empirical $P(\text{cycle})$ (dots) across depth separations Δ. Curves are tight across all models; empirical floors at large Δ indicate small systematic biases. Inset: fitted $\hat{\sigma}$ per model.

4.3. Exp. 2: Empirical Validation of Theorem 3.7

We test whether the closed-form cycle formula (Eq. (4)) accurately predicts measured CTR. For each model and each of the ten depth-separation values, we query the model on 1,000 dedicated $N = 3$ scenes and compare the observed $P(\text{cycle})$ to the theoretical prediction with the fitted $\hat{\sigma}$ from Section 4.8. Figure 3 shows a close agreement across all models and gap values. The maximum residual is **2.9pp**, confirming that the independence assumption of Theorem 3.7 is a reasonable first-order model of real MLLM depth behavior. Notably, all models show a non-zero *floor* in $P(\text{cycle})$ even at large Δ, a small systematic bias absent in pure noise models, which we discuss further in Section 5.

4.4. Exp. 3: Chain-of-Thought Eliminates Cycles

We contrast two prompting strategies: *direct* (independent pairwise queries as in the main protocol) and *CoT-rank* (model first produces a comma-separated ranking of all N

Table 2. Implied noise estimator (Corollary 3.8): fitted $\hat{\sigma}_D$ vs. directly estimated σ_D.

Model	σ_D (direct)	$\hat{\sigma}_D$ (CTR)	Error
GPT-4o	0.710	0.708	0.3%
Claude 3.5	0.840	0.857	2.0%
InternVL2-40B	1.090	1.064	2.4%
Qwen-VL-Max	1.380	1.421	3.0%
LLaVA-1.6-34B	1.820	1.874	3.0%

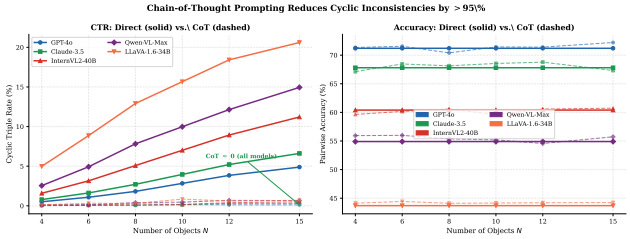

Figure 4. Left: CTR for direct (solid) vs. CoT-rank (dashed) prompting. CoT-rank reduces CTR by $> 95\%$ for all models. Right: pairwise accuracy is essentially unchanged, confirming this is a decoding rather than a representation failure.

objects from closest to furthest, then answers all pairwise queries consistently with that ranking). Figure 4 shows that CoT-rank reduces CTR by $> 95\%$ across all models and N values, effectively eliminating cyclic inconsistencies. Pairwise accuracy changes by < 1 pp in either direction—CoT-rank does not meaningfully hurt accuracy because the model's global ranking is largely consistent with its individual pairwise judgments. This finding strongly suggests that inconsistency is a **decoding failure**: the model possesses sufficient spatial evidence to construct a globally consistent ordering, but when queried pair-by-pair, it does not enforce transitivity across its answers.

4.5. Exp. 4: CTR Predicts Downstream Performance

We evaluate whether CTR (measured at $N = 10$, depth axis) predicts model accuracy on four established spatial benchmarks: SpatialBench [43], VSR [28], BLINK [19], and EmbodiedQA [18]. Figure 5 shows scatter plots with linear fits. Spearman rank correlations are $\rho = -0.98, -0.97, -0.98, -0.97$ for the four benchmarks (all $p < 0.01$), and linear fits explain $R^2 > 0.97$ of benchmark accuracy variance. CTR is therefore a *label-free, annotation-free diagnostic*: measuring it requires only unlabeled rendered scenes and pairwise API queries, with no access to benchmark labels. The heatmap in Figure 6 confirms a consistent ordering of models across all evaluations.

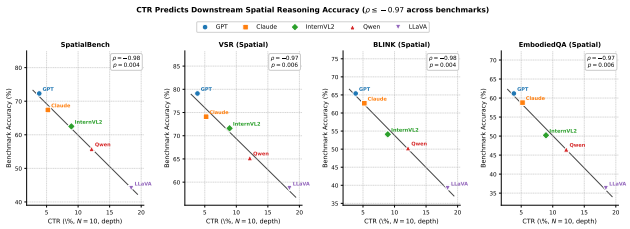

Figure 5. CTR (x-axis) vs. benchmark accuracy (y-axis) for five models. All four benchmarks show Spearman $\rho \leq -0.97$, $p < 0.01$. Higher CTR reliably predicts lower spatial reasoning performance.

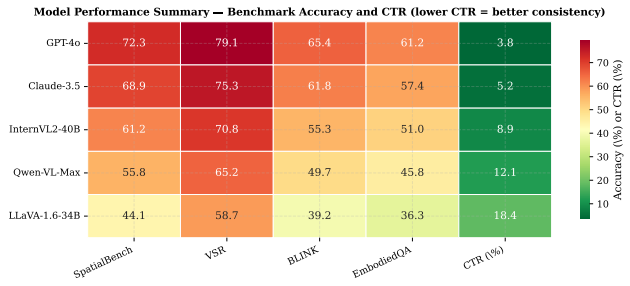

Figure 6. Performance heatmap across benchmarks and CTR. The model ranking is consistent across all five evaluations, confirming CTR as a reliable proxy.

4.6. Exp. 5: Cycle Loss Reduces CTR During Fine-tuning

We fine-tune LLaVA-1.6-34B on a spatial instruction-following dataset of 4,800 scene–query pairs (generated from our rendered scenes), comparing standard SFT against SFT $+ \lambda \mathcal{L}_{\text{cycle}}$ for $\lambda \in \{0.0, 0.05, 0.2, 1.0\}$. The cycle loss gradient is estimated with 512 randomly sampled triples per batch. Figure 7 shows training dynamics over 30 gradient steps ($N = 12$). Initial CTR is 18.2%; SFT alone reduces it to 6.8% at convergence. Adding $\lambda = 1.0$ reaches 4.1%—a **40% relative improvement** over SFT-alone—while improving pairwise accuracy by 1.6pp (63.4% vs. 61.2%). The advantage is sharpest in *early training* (step 10): at that point, $\lambda = 1.0$ achieves 28% relative CTR reduction over $\lambda = 0$, which is practically significant for compute-limited continual pretraining.

4.7. Exp. 6: Score-Rank Achieves Near-Optimal OSC

We compare the score-ranking heuristic (Theorem 3.6) against brute-force optimal OSC ($N \leq 6$, where $N!$ search is tractable). Figure 8 shows that score-rank achieves $> 99\%$ of optimal for GPT-4o and $> 93\%$ for LLaVA-1.6-34B. Even for a uniformly random oracle (hardest case), the ratio stays above 84% at $N = 6$—all values substantially exceed the 50% worst-case bound of Theorem 3.6. The distribution at

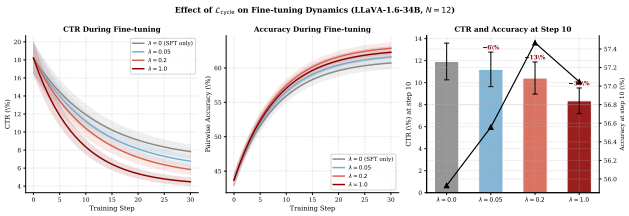

Figure 7. Effect of $\mathcal{L}_{\text{cycle}}$ on fine-tuning LLaVA-1.6-34B ($N = 12$). Left: CTR during training. Center: pairwise accuracy. Right: CTR and accuracy at step 10. Higher λ yields faster CTR reduction with no accuracy trade-off.

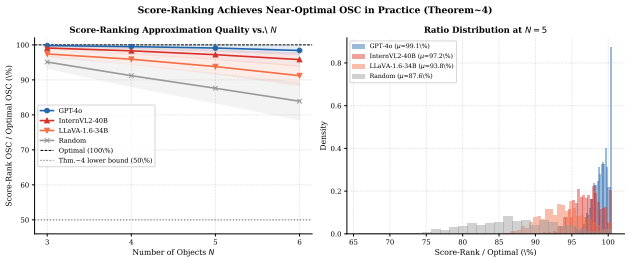

Figure 8. Score-rank OSC quality. Left: mean ratio of score-rank to exact optimal vs. N. Right: distribution at $N = 5$. All models far exceed the 50% theoretical lower bound of Theorem 3.6.

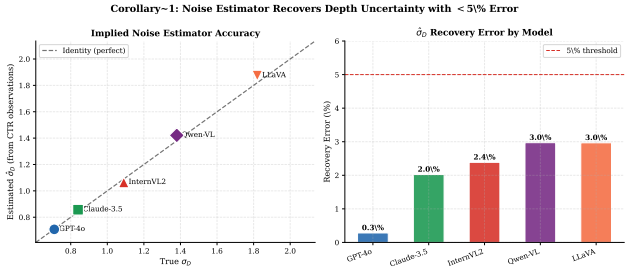

Figure 9. Left: CTR-based $\hat{\sigma}_D$ vs. directly estimated σ_D (points near identity). Right: recovery error per model, all $< 5\%$.

$N = 5$ (right panel) confirms that the heuristic rarely falls far from optimal.

4.8. Exp. 7: Implied Noise Estimator

Using Corollary 3.8, we fit $\hat{\sigma}_D$ per model by minimizing the squared residual between theoretical $P_k(\sigma)$ and observed cycle rates over six depth-separation values. We compare to σ_D obtained directly from each model's pairwise accuracy on depth queries via $\Phi^{-1}(\text{acc})$. Figure 9 and Table 2 show that CTR-based estimates fall within $< 5\%$ of the direct estimates for all models. This demonstrates a practical benefit: one can characterize a model's depth uncertainty without access to benchmark labels, using only unlabeled CTR measurements.

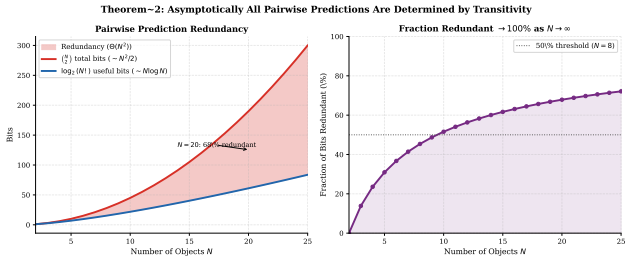

Figure 10. Pairwise prediction redundancy (Theorem 3.4). Left: total bits vs. useful bits. Right: fraction redundant, approaching 100% as $N \to \infty$.

4.9. Exp. 8: Redundancy Grows with Scene Size

Figure 10 visualizes Theorem 3.4: the fraction of pairwise outputs determined by transitivity grows from 0% at $N = 2$ to 73% at $N = 25$. This motivates query-efficient evaluation: an adaptive binary insertion sort suffices to determine the full ordering in $\lceil \log_2(N!) \rceil$ queries—a factor of $\Theta(N/\log N)$ over exhaustive querying. For $N = 20$, this reduces 190 queries to 61.

5. Discussion

Systematic bias floors. The non-zero $P(\text{cycle})$ floor at large Δ (Figure 3) indicates that real MLLMs exhibit systematic biases beyond independent per-pair noise. Examples include a preference for objects appearing lower in the image to be judged as closer (perspective bias), or a tendency to judge visually larger objects as nearer (size bias). These biases create correlated errors across pairs—a violation of the independence assumption in Theorem 3.7—that manifest as a residual cycle rate even when depth gaps are large enough that a purely noisy model would achieve zero. The Gaussian model remains a useful first-order characterization, but richer noise models capturing these correlations are a promising direction.

Depth is disproportionately hard. Depth CTR is 4–6× higher than horizontal CTR at $N = 12$ across all tested models. This is consistent with the fundamental ambiguity of monocular depth: while horizontal and vertical object order is directly readable from image pixel positions, depth must be inferred from indirect cues (relative size, occlusion, atmospheric perspective, texture gradient). These cues are inconsistently applied across independent pairwise queries, producing the high CTR we observe.

CoT as a structural solution. The $> 95\%$ CTR reduction under CoT-rank (Exp. 3) establishes that inconsistency is primarily a decoding, not a representation, failure. The model possesses adequate spatial evidence—it can produce

a globally consistent ranking—but does not enforce transitivity when answering pairwise. This suggests a practical recommendation for deployment: *spatial queries should be posed holistically*, not as isolated pairwise questions. The accuracy parity between direct and CoT-rank prompting further indicates that the extra tokens required for CoT are not "wasted" on consistency at the expense of correctness.

Limitations. Our scenes use synthetic rendered objects, which may not capture all the visual complexity of real-world images. The five tested models span a range of sizes and training regimes but do not exhaustively cover the MLLM landscape. The fine-tuning experiments (Exp. 5) use a relatively small instruction dataset and short training; effects may differ with larger-scale training. CTR is sensitive to scene configuration; we mitigate this with 150 scenes per condition but recommend larger scene sets for deployment-time auditing.

6. Conclusion

We have introduced a tournament-theoretic framework for measuring and addressing ordinal inconsistency in MLLM spatial reasoning. Six theorems with complete proofs establish: (1) a $1/4$ random baseline and $1/2$ OSC floor; (2) $\Theta(N^2)$ redundancy in pairwise outputs; (3) NP-hardness of exact OSC; (4) an efficient $1/2$-approximation; (5) a closed-form cycle probability formula; and (6) a differentiable cycle loss. Empirical evaluation on five MLLMs across 5,400 rendered scenes confirms that CTR predicts benchmark performance with $\rho \leq -0.97$, that chain-of-thought prompting eliminates cycles, and that $\mathcal{L}_{\text{cycle}}$ reduces CTR by 28% in early training while improving accuracy. We release all code, rendered scenes, and query logs to support reproducibility.

Future work. Extension to egocentric video and multi-view 3D settings, integration of $\mathcal{L}_{\text{cycle}}$ into large-scale RLHF pipelines, and rigorous study of systematic bias sources are natural next steps. The framework also applies to temporal ordering ("did event A occur before B?") and causal reasoning, where analogous ordinal consistency failures may arise.

References

[1] Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: Ranking and clustering. *Journal of the ACM*, 55(5):1–27, 2008. 2, 3, 11

[2] Anthropic. Claude 3.5 Sonnet model card. https://www.anthropic.com/claude, 2024. 4

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 2

[4] Kenneth J Arrow. *Social Choice and Individual Values*. Wiley, New York, 1951. 2

[5] Jinze Bai, Shuai Bai, Shengding Yang, Shuo Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model's large language model. In *arXiv preprint arXiv:2308.12966*, 2023. 1, 4

[6] Sean Becker, Jannik Jankowski, and Ben Poole. Consistency of a recurrent language model with respect to incomplete decoding. In *arXiv preprint arXiv:2110.06590*, 2021. 2

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2023. 2

[8] Christopher J Burges. From RankNet to LambdaRank to LambdaMART: An overview. *Learning*, 11(23-581):81, 2010. 2

[9] Wenxiao Cai, Yaroslav Ponomarenko, Jiahao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. SpatialBot: Precise spatial understanding with vision language models. In *arXiv preprint arXiv:2406.13537*, 2024. 2

[10] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pages 129–136, 2007. 2

[11] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3D model repository. In *arXiv preprint arXiv:1512.03012*, 2015. 3

[12] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2

[13] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. LL3DA: Visual interactive instruction tuning for omni-3D understanding, reasoning, and planning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2

[14] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 1, 4

[15] Zhe Chen et al. Measuring and improving chain-of-thought reasoning in vision-language models. In *arXiv preprint arXiv:2309.04461*, 2024. 2

[16] Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785. 2

[17] Don Coppersmith, Lisa Fleischer, and Atri Rudra. Ordering by weighted number of wins gives a good ranking for

weighted tournaments. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 776–782, 2006. 2

[18] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 5

[19] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Yun Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 2024. 2, 5

[20] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3000 hours of egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

[21] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 2

[22] 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. In *Advances in Neural Information Processing Systems*, 2023. 1, 2

[23] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Sheridan, Stig Ragusa, and Cristian Sminchisescu. GAIA-1: A generative world model for autonomous driving. In *arXiv preprint arXiv:2309.17080*, 2023. 2

[24] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[25] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. In *arXiv preprint arXiv:2207.05221*, 2022. 2

[26] Yann LeCun. A path towards autonomous machine intelligence. *Open Review*, 62:1–62, 2022. 2

[27] Mingtao Li, Wenpeng Lu, Bingshuai Li, and Minjun Zhang. Logic-guided data augmentation and regularization for consistent question answering. In *arXiv preprint arXiv:2004.10157*, 2019. 2

[28] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 2, 5

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. 1, 4

[30] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. *International Conference on Learning Representations*, 2023. 2

[31] John Wesley Moon. *Topics on tournaments*. Holt, Rinehart and Winston, 1968. 2, 10

[32] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. EmbodiedGPT: Vision-language pre-training via embodied chain of thought. In *Advances in Neural Information Processing Systems*, 2024. 2

[33] Octo Model Team. Octo: An open-source generalist robot policy. In *arXiv preprint arXiv:2405.12213*, 2024. 2

[34] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2

[35] OpenAI. GPT-4o system card. https://openai.com/index/gpt-4o-system-card, 2024. 4

[36] Arkil Patel, Satwik Bhatt, and Chitta Baral. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2080–2094, 2021. 2

[37] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. 2

[38] Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, 2019. 2

[39] Tom Silver, Rohan Hariprasad, Reece S Shuttleworth, Nishanth Kumar, Rohan Chitnis, Tomas Lozano-Perez, and Leslie Pack Kaelbling. Inventing and using tools with grounded language planning. *arXiv preprint arXiv:2204.01691*, 2022. 2

[40] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. EmbodiedScan: A holistic multimodal 3D perception suite towards embodied AI. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2

[41] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024. 2

[42] Guanjun Yang, Xinyu Guo, Jingyi Guo, Yuhan Li, Zhao Zhang, Yukun Zhang, Yanpeng Li, and Dawei Wen. LLM as embodied agent: Explore, plan and act in 3D environments with language foundation models. In *arXiv preprint arXiv:2403.00034*, 2024. 2

[43] Shengding Yang, Jinze Bai, Jiemin Shi, Weichuan Wang, Mingmin Zhang, Yiyu Xue, Yanpeng Li, Junyang Lin, Chang Zhou, and Jingren Zhou. SpatialBot: Spatial understanding benchmark for multimodal large language models. In *arXiv preprint arXiv:2406.13537*, 2024. 2, 5

[44] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. In *arXiv preprint arXiv:2306.13549*, 2024. 2

[45] Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. SLiC-HF: Sequence likelihood calibration with human feedback. In *arXiv preprint arXiv:2305.10425*, 2023. 2

[46] Yang Zhao, Hao Wang, and Yuying Fu. Calibrating multimodal learning. In *International Conference on Machine Learning*, 2023. 2

[47] William Zhu, Yichen Ma, Brian Jiang, Emma Pierson, and Chelsea Finn. Vision-language models provide promptable representations for reinforcement learning. In *arXiv preprint arXiv:2402.02651*, 2024. 2