VERDICT: Training-Free Step-Wise Verification of Multimodal Reasoning via Disagreement-Aware Consensus

Rohit Sinha* · Kunal Tilaganji* · Tanuja Ganu · Nagarajan Natarajan · Amit Sharma · Vineeth Balasubramanian
Indian Institute of Technology Hyderabad · Microsoft Research * equal contribution

ECCV 2026 | MALMÖ | SEPT 8–12

Microsoft Research

IIT Hyderabad
Indian Institute of Technology

The Problem

Reasoning chains fail quietly

Individual steps carry gaps that look locally plausible, then cascade into a wrong final answer.

Verification must happen at the step level.

Key Insight

Disagreement is the signal

A valid step makes judges converge. When they cannot agree, the step is unstable.

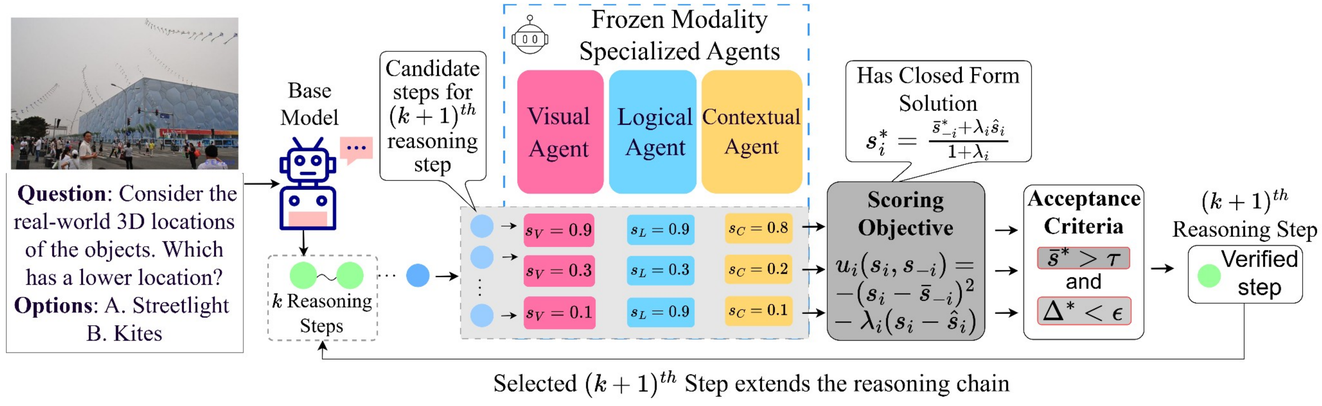
```
(0.7, 0.7, 0.7)  Δ* = 0        accepted
(0.9, 0.3, 0.9)  Δ* ≈ 0.11     rejected
```

Same mean - averaging cannot tell them apart.

Contributions

1. **Coupled scoring, not aggregation.**
   Frozen, modality-specialised judges. No training, no labels.
2. **Closed-form consensus.**
   The unique equilibrium of a coordination game.
3. **Provably not an average.**
   Dispersion is non-separable in raw scores (Prop. 1).

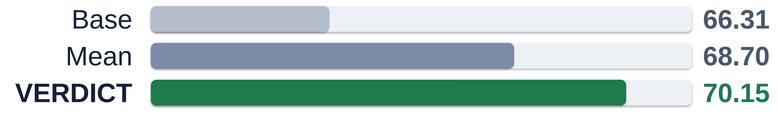
How VERDICT Works



Three frozen judges score every candidate step in isolation. A closed-form consensus redistributes those scores around a fixed mean. It can never inflate confidence, and the dual criterion decides which candidate extends the chain.

Results

BENCHMARK	Base	Mean	VERDICT	Δ
3DSRBench	56.12	58.34	**59.02**	+2.90
CV-Bench-3D	76.39	79.77	**82.34**	+5.95
CV-Bench-2D	74.27	77.57	**79.22**	+4.95
BLINK	48.31	50.17	**51.32**	+3.01
MMStar	61.25	64.14	**65.88**	+4.63
AI2D	81.52	82.18	**83.14**	+1.62
Average	**66.31**	**68.70**	**70.15**	**+3.84**



Accuracy (%) — no benchmark falls below the base model.

Ablations & Analysis

Ranking drives the gain
Dropping ranking costs −1.17 to −2.17; dropping filtering only −0.26 to −1.08.

Consensus, not averaging
Raw Average trails by 2.59-2.95 under the identical dual criterion.

Stubbornness is asymmetric
λ_V spans 2.90 pp, λ_C only 1.19. Swapping V ↔ C costs −1.76.

Weaker judges, bigger win
Margin over Mean grows +0.68 (7B) → +0.89 (4B) → +1.21 (2B).

Robust to both thresholds
ε plateaus over [0.1, 1.0]; τ = 0.6 is best on five of six benchmarks.

Base-model agnostic
+2.45 to +4.00 across four base-model families, judges held fixed.

Paper, Code & Project Page

Project page
kunaltilaganji.github.io/verdict

arXiv
arxiv.org/abs/2608.10665

Next: reward-model ensembles, multi-agent evaluation and compositional code verification.