TRIAD: Benchmarking Omni-Modal Ambiguity in Multimodal Large Language Models

Zhaolu Kang[1,2*], Yidi Wang[2*], Yile Li, Siqi Zeng, Yingjie He, Richeng Xuan[1†], Zhichao Hu[1†]

[1]Hunyuan, Tencent [2]Peking University *Equal contribution †Corresponding author

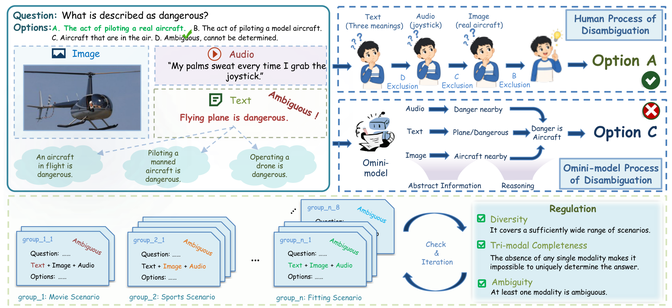
Paper
Abstract & paper

Email
zlkang25@stu.pku.edu.cn

TL;DR **TRIAD asks whether omni-modal LLMs can resolve ambiguity spread across text, image and audio. The best of 13 models trails humans by 20.4 points in group accuracy and drops to 4.6 without options.**

1 Motivation

- Real communication is **ambiguous**: a sentence, image or sound can stay underspecified until the other channels are considered.
- Existing multimodal benchmarks test perception or reasoning over **already determinate** inputs.
- Can omni-modal LLMs resolve ambiguity **distributed across text, vision and audio**?



Humans combine all three cues to exclude options and pick A; the model fuses shallow cues and picks C.

2 Benchmark design

- Each item: a **text–image–audio triplet**, a question and 2 to 8 answer options.
- **Well-posed**: the full triplet fixes one answer; every two-modality view leaves several options open.
- **Scene groups** of 2 to 8 sister queries perturb one modality, so the answer follows the altered cue.

$$|\mathcal{Y}_{\mathcal{M}}(x)| = 1, \qquad |\mathcal{Y}_{S}(x)| > 1 \quad \forall S \subset \mathcal{M},\ |S| = 2$$

Checked by five annotators; about 40% of candidate items are admitted.

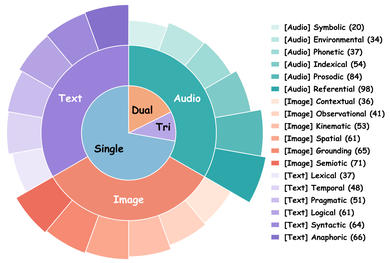
- **Three seeding routes**: start from an ambiguous sentence, image or sound, then author the other channels around it.
- **Natural media**: mostly licensed real photographs; AI images only when no photo fits, screened by two annotators. Natural or voice-acted audio; text-to-speech only with human-edited prosody.
- **Review**: five annotators (three NLP, two linguistics). Each query is written by one and reviewed by four, who list at least two consistent options for every modality subset. κ = 0.84 on answers, 0.79 on ambiguity tags.
- **Six categories per modality.** Text: anaphoric, syntactic, logical, pragmatic, temporal, lexical. Image: semiotic, grounding, spatial, kinematic, observational, contextual. Audio: referential, prosodic, indexical, phonetic, environmental, symbolic.
- **Limitations**: a diagnostic set, not a training corpus; English and Chinese only; open answers are graded by judge models.

3 Benchmark at a glance

327 hand-written queries

86 scene groups

18 ambiguity categories



Inner: ambiguity tier. Outer: the 18 leaf categories.

Ambiguous	Items	%
1 modality	230	70.3
2 modalities	59	18.0
all 3	38	11.6

- 184 English and 143 Chinese queries.
- Groups of 2, 3, 4 or 8 queries.
- 2 to 8 options per question.

4 Results: 13 omni-modal models

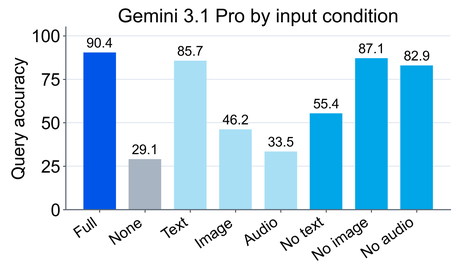
- **Setup**: 5 Gemini and 8 open-source omni models with native tri-modal input; Standard, CoT and open-ended prompts at temperature 0. **Group accuracy**, the primary metric, counts a scene group only if every sister query is right.

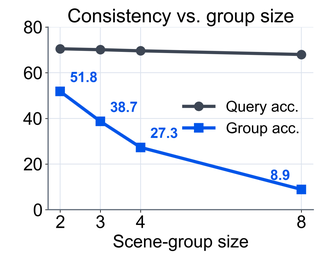
Model	Query	Group	CoT group	Open group
Human	**95.1**	**91.6**	–	**79.2**
Gemini 3.1 Pro	**90.4**	**71.2**	**72.5**	**4.6**
Gemini 3 Pro	89.2	68.9	70.4	3.1
Gemini 3 Flash	82.5	59.4	58.1	3.5
Gemini 2.5 Pro	80.7	55.3	58.0	3.3
Qwen3.5-Omni-Plus	78.0	52.8	51.4	3.1
Gemini 2.5 Flash	73.2	45.6	46.8	2.9
Qwen3-Omni-30B	72.5	41.7	40.9	2.3
MiniCPM-o-4.5	66.2	36.4	38.2	1.6
Qwen2.5-Omni-7B	61.8	30.2	32.5	1.2
OmniVinci-9B	58.7	26.9	25.4	1.0
Ola-7B	55.4	23.8	24.7	0.9
Qwen2.5-Omni-3B	54.1	22.1	20.6	0.7
Baichuan-Omni-1.5	52.9	20.4	19.2	0.6
Random	24.3	2.2	–	–

Accuracy (%), full text–image–audio input; humans on a 100-query stratified subset.

- **Query accuracy hides failures**: the human–model gap grows from **4.7** to **20.4** points at the group level.
- **Options are scaffolding**: without them the best model falls from 71.2 to **4.6** group accuracy; humans keep 79.2.
- **Chain-of-thought barely helps**: it changes group accuracy by at most 2.7 points.
- **Unstable answers**: Standard and CoT prompts agree on 69 to 88% of queries; all three prompts on only 6 to 17%.

5 Diagnostics





Left: shortcuts without image or audio. Right: group accuracy falls as groups grow.

Input	Human	Gemini 3.1 Pro	Qwen3.5-Omni
Text only	18.2	42.6	32.1
No image	70.5	57.2	43.6
No audio	60.7	51.8	39.7

Group accuracy with one modality alone or removed.

- **Text priors**: with text alone, models beat humans by 14 to 24 points.
- **Audio is underused**: without it Gemini 3.1 Pro still reaches 82.9 query accuracy (full: 90.4).
- **No counterfactual control**: per-query accuracy barely moves from groups of 2 to 8 (70.5 → 68.0), yet group accuracy falls from 51.8 to 8.9.

Models anchor on text and use image and audio to confirm, not revise.

6 Descriptions are not enough; why models fail

Model	Full	A+	I+	Both
Gemini 3 Pro	89.2	83.6	85.4	79.8
Gemini 3 Flash	82.5	77.1	79.3	74.0
Qwen3.5-Omni-Plus	78.0	71.4	74.2	68.1
MiniCPM-o-4.5	66.2	58.9	62.0	54.7
Human	95.1	88.7	90.5	84.1

Query accuracy; A+: audio as a transcript, I+: image as a description.

Error type	Share
Transcript-only listening	35.0%
Image-anchored commitment	25.5%
Option priors	20.0%
Cross-modal contradiction	19.5%

200 sampled errors, pooled across models and prompts.

- **Premature commitment**: models form a hypothesis from the most language-like cue, then use the other modalities to justify it instead of revising it.

7 Scale, open answers and limits

Answer format	Human	Gemini 3.1 Pro
Multiple choice	95.1	90.4
Open-ended	88.4	18.4
Open, strict	82.7	9.4

Query accuracy (%).

- **Scale helps but does not close the gap**: group accuracy rises from 22.1 to 52.8 across Qwen omni models and from 45.6 to 71.2 across Gemini, still 20 points below humans (91.6).