

Peer Review

Review of: "Are Vision-Language Models Truly Understanding Multi-vision Sensor?"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The authors argue that current vision-language models read thermal, depth, and X-ray images as if they were ordinary RGB photographs, and so miss what the sensor physically measures (heat emission rather than reflected sunlight, distance rather than haze). To show this, they build MS-PR, a multiple-choice benchmark of roughly 7k images and 10k questions drawn from 13 public datasets (§5.1.1, Appendix A), split into four perception dimensions (Existence, Count, Position, General Description) and two reasoning dimensions (Contextual Reasoning, Sensory Reasoning). Questions are generated by ChatGPT/GPT-4o from a sensor-knowledge prompt and then reviewed by human annotators (§3.2, Figure 4, Appendix B). Ten models, seven open and three closed, are scored in Table 1.

The second contribution is a training objective, Diverse Negative Attributes (DNA) optimization. As far as I can reconstruct Eq. 2 from the rendered text, it adds to the SFT loss a reference-free, length-normalized pairwise term of the form $\log \sigma(\beta \log \pi_{\theta}(y^+|x)/|y^+| - \beta \log \pi_{\theta}(y^-_j|x)/|y^-_j| - \gamma)$, averaged over k negatives. Trained on 600 held-out images (200 per sensor, 3,600 QA pairs), DNA is reported to lift multi-vision reasoning well above plain SFT for Phantom-7B, Qwen2-VL-7B, and InternVL2-8B (Table 2) without harming general benchmarks (Table 5).

2. Whether the conclusion holds on the data

I think the central observation is real and useful. Table 1 shows a pattern that is hard to explain away: perception averages sit in a fairly narrow band, while Sensory Reasoning on depth collapses for almost every model. Qwen2-VL-7B scores 19.3 (Table 1, row Qwen2-VL-7B, Depth, Sensory Reasoning), LLaVA-

1.5-7B 22.1, and even GPT-4o only 33.6, against 69.7 and 79.3 for the same model on thermal and X-ray. That contrast inside one model is more convincing than any cross-model ranking, and the authors deserve credit for surfacing it. I also checked the averaging: for BLIP-2, Thermal, $(59.2 + 32.2 + 57.8 + 65.3) / 4 = 53.625$, which matches the reported 53.6; for Claude-3.5-Sonnet, Thermal, $(75.3 + 46.2 + 64.1 + 67.8) / 4 = 63.35$, reported as 63.3. The table is internally consistent where I tested it.

The human study in Appendix C is a good addition. Twenty Prolific participants answered 45 reasoning questions (15 per sensor) with 95.1% accuracy, against 73.3 for GPT-4o on the same subset. The stated gap of 21.8 is correct as a difference in percentage points ($95.1 - 73.3 = 21.8$); it is not a 21.8% relative gap, and I would word it that way. With 45 items, though, the uncertainty is wide. A rough binomial standard error at $p = 0.73$ is $\sqrt{0.73 \times 0.27 / 45} \approx 0.066$, so a 95% interval on the GPT-4o subset score spans about ± 13 points. The claim that subset scores "align within the margin of error" with Table 1 is therefore true but weak; almost any value would align. It also validates only the reasoning half of the benchmark, and only 45 of about 10k questions.

My main questions concern the DNA claim.

How much of the gain is SFT, and how much is DNA? Table 2 shows that SFT alone already delivers most of the improvement in many cells. For Phantom-7B, Thermal, Multi-vision Reasoning goes from 64.0 (base) to 78.9 (+SFT) to 84.6 (+DNA): +14.9 from SFT, then +5.7 from DNA. For Phantom-7B, Depth, Sensory Reasoning, it goes from 29.6 to 65.3 to 86.1. So §5.2.2's sentence that SFT gives "slight improvements" does not match the table; a jump of 35.7 points is not slight. The DNA increment is still consistent and positive, which is the honest and sufficient claim.

The missing baseline. The loss in Eq. 2 is very close to SimPO (ref. [40]) with a joint SFT term and several negatives. The natural comparison is therefore SFT + DPO or SFT + SimPO with $k = 1$, not SFT alone. Table 3 partly speaks to this: with $k = 1$, Phantom-7B reaches 83.1 on Thermal Multi-vision Reasoning, versus 84.0 at $k = 2$ and 84.6 at $k = 3$. Most of the DNA gain over SFT (78.9 in Table 2) is thus already present with a single negative, which suggests the pairwise preference term, rather than the diversity of negatives, does most of the work. That is worth knowing, and it deserves to be stated.

In-distribution training. The 600 training images come from the same 13 source datasets as the benchmark (§5.1.1), and I assume the QA pairs come from the same GPT-4o pipeline and templates. A model can improve here by learning the question style and the distractor style, without learning more about sensor physics. A held-out source dataset per sensor (train on seven thermal sources, test on the eighth) would separate the two readings.

Chance level. I could not find how many options each question has. This matters. If there are four options, several Depth Sensory Reasoning scores in Table 1 are below the 25% chance line (Qwen2-VL-7B 19.3, LLaVA-1.5-7B 22.1), which would point to systematically attractive distractors rather than random guessing. That is an interesting finding in itself, but the reader cannot judge it without the option count.

Generalization claim. §5.3 says DNA "enhances its generalizability". Table 5 shows gains on MMMU (47.8 to 49.3), MMBench (79.8 to 80.2), and SEED (75.3 to 75.7), and drops on MME (2126 to 2113) and Q-Bench (69.9 to 68.2). "No meaningful loss" is supported; "enhances" is not. Only Phantom-7B is shown.

No result in Tables 2 to 5 carries a seed count, standard deviation, or confidence interval, and the Phantom-7B and Qwen2-VL-7B runs are a single epoch each (§5.1.2). Some ablation differences, such as 83.7 vs 82.7 on Depth Multi-vision Reasoning between $k = 3$ and $k = 2$ (Table 3), are small enough that I cannot tell them from run-to-run noise.

3. Reproducibility

The authors point to a public repository (github.com/top-yun/MS-PR) for code and data. I did not inspect it, so I cannot say what it contains. On the page itself, the source datasets are named and described in Appendix A, which makes the benchmark reconstructable in principle, and the generation prompt is said to appear in Figures 5 and 6.

Several items are missing or, as rendered on Qeios, not legible.

- §5.1.2 gives $\beta = 2$, $\gamma = 0.2$, and $k = 3$, QLoRA with AdamW, and a learning rate of $2e-5$ with one epoch for Phantom-7B (rank 256, alpha 256) and Qwen2-VL-7B (rank 64, alpha 64). The same details are missing for InternVL2-8B, although it appears in Table 2.
- Closed models are named without snapshots. "GPT-4o" and "Claude-3.5-Sonnet" have several dated versions; "Gemini-Pro" is cited to the Gemini 1.5 report, so I am not sure which model was queried. Evaluation dates, temperature, max tokens, and the answer-parsing rule are not stated, as is usual in this literature but still worth fixing.
- The generator is "ChatGPT/GPT-4o" (§3.2) with no version or sampling settings. How many human annotators reviewed the questions, what share was edited or discarded, and whether any agreement statistic was computed, are not reported.
- No seeds, batch size, GPU type, or training time.
- As rendered on Qeios, I could not read Figures 3 to 13, so I could not check the per-task sample counts in Figure 3 or the human-agreement plot in Figure 7.

4. What specifically to fix

1. Table 1 and Table 2 disagree on one baseline row. Qwen2-VL-7B, X-ray, reads Count 39.8 and Position 63.8 in Table 1, but 39.7 and 63.7 in Table 2. The Table 2 values reproduce the reported perception average exactly $((71.0 + 39.7 + 63.7 + 84.4) / 4 = 64.7)$, while the Table 1 values give 64.75. Please pick one and correct the other.
2. §4.1 says the reasoning problems are revealed "in Table 3". Table 3 is the k ablation; I believe Table 1 is meant.
3. I would add SFT + DPO and SFT + SimPO baselines to Table 2, trained on the same 3,600 pairs with one negative. Without them, the reader cannot tell what the multiple-negative design adds beyond a known preference loss.
4. §5.2.2 should mention InternVL2-8B, which is in Table 2 but absent from the text, and should drop "slight" for SFT given the Phantom-7B Depth Sensory Reasoning jump from 29.6 to 65.3.
5. State the number of answer options and the chance level in §3.2, and comment on below-chance cells in Table 1.
6. Table 4 stops at $n = 100$ images per sensor. The $n = 200$ point is the Phantom-7B + SFT and + DNA rows of Table 2; adding them to Table 4, or saying so in §5.5, would close the curve.
7. One out-of-source test (a thermal, depth, or X-ray dataset never used for training) would make the "sensor understanding" reading of Table 2 much stronger.
8. Please report mean and standard deviation over at least three seeds for Tables 2 and 3.
9. In §5.3, rephrase "enhances its generalizability" to match Table 5, where MME and Q-Bench decrease.
10. Because GPT-4o both wrote the questions and is the strongest closed model on Multi-vision Reasoning in Table 1, a short discussion of possible generator bias would help. Even a small set of questions written by a different model, scored for GPT-4o, would address it.
11. Give the InternVL2-8B training settings in §5.1.2, and give model snapshots, decoding settings, and evaluation dates for the three closed models.
12. Minor: LLaVA-1.5-7B is cited as [32] in Table 1 but as [20] in §5.1.2; the Appendix C difference of 21.8 should be labelled as percentage points.

5. Recommendation

Minor revisions. The benchmark addresses a gap that matters for driving, security screening, and medical imaging, and Table 1 makes the depth-reasoning failure visible in a way others can build on. The DNA gains are consistent across three open models and all three sensors. What I would ask for is mostly about attribution and wording: a preference-loss baseline, seeds, the option count, and a few corrected numbers and sentences. None of this changes the direction of the findings, and I think the revised paper will be a solid reference for sensor-aware VLM evaluation.

Declarations

Potential competing interests: No potential competing interests to declare.