

Peer Review

Review of: "Benchmark Evaluations, Applications, and Challenges of Large Vision Language Models: A Survey"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

This is a review article on vision-language models covering 2019 to the end of 2024. The abstract sets out five deliverables: model information for the major VLMs of the past five years; their architectures and training methods; a summary and categorization of popular benchmarks and evaluation metrics; applications, named as embodied agents, robotics, and video generation; and challenges, named as hallucination, fairness, and safety. The evidence base is three tables and three figures. Table 1 lists 23 models with year, architecture family, training data, parameter count, vision encoder, and pretrained backbone. Table 2 maps benchmark categories to datasets. Table 3 gives 40 rows of benchmark, evaluation format, category, annotation source, and size in thousands. A companion repository is given in the abstract (github.com/zli12321/Awesome-VLM-Papers-And-Models.git). Section 2 announces "three principal research directions" as the organizing frame: vision-language correlation, benchmarks and evaluation, and applications.

The argumentative core is Section 4. There, the authors claim that "most benchmarks prioritize data diversity and quantity while often overlooking improvements in evaluation quality, which hinders the effective assessment of VLMs," and Section 4.2 develops that into a specific mechanism: automatic metrics are cheap, cheap metrics push benchmark builders toward short-form and multiple-choice questions, and those formats can be gamed by the language component alone. Section 1 and the abstract also make a framing claim about the state of the field: "There are already several reviews on single-modality models[23][24] while the multi-modality one is still missing."

Some of this is good, and I want to be concrete about which parts. Section 5.3 on robotics is the strongest writing here. It is specific, it names systems rather than gesturing at trends, and Section 5.3.5 closes it with three genuine tensions (generalization versus flexibility, intelligence versus safety, embodiment versus effectiveness) that I have not seen framed that cleanly elsewhere. Section 4.1 is also a real contribution: sorting benchmark construction into fully human-annotated, synthetic-with-partial-validation, synthetic-with-full-validation, and simulator interaction gives a reader a way to reason about a benchmark they have never used. My problem is that the survey does not do the counting work its own summary claims require, and that the two benchmark tables disagree with the prose, with each other, and in several cells with the sources they describe.

2. Whether the conclusion holds on the data

2.1 The framing claim is refuted by the paper's own bibliography. Section 1 says the multimodal review "is still missing," and the abstract says "a comprehensive survey of existing studies on vlms is notably lacking." But this paper cites, as reference [14], Bordes et al., "An introduction to vision-language modeling"; as [15], Hartsock and Rasool, "Vision-language models for medical report generation and visual question answering: A review"; as [18], Lymperaiou and Stamou, "A survey on knowledge-enhanced multimodal learning"; as [80], Zhang et al., "Vision-language models for vision tasks: A survey"; and, twice, as [22] and again as [25], Liu et al., "A Survey on Hallucination in Large Vision-Language Models." Five multimodal surveys in the reference list of a paper whose stated motivation is that no multimodal survey exists. This does not sink the article. It does mean the positioning has to be rewritten into a claim about what this survey adds that those five do not, and that claim has to be argued rather than asserted.

2.2 The counts in the prose do not match the tables. Section 4 states that the authors "summarize and categorize existing 38 vision-language benchmarks". I counted Table 3 row by row and got 40 rows. TextVQA appears twice, so 39 distinct names. Section 4 also says the benchmarks are classified "to ten primary categories (Table 2)", and the caption of Table 2 repeats "categorized to 10 categories". Table 2 has 13 category rows: the ten the prose presumably means, plus "Robot simulator, web agent simulator", "Robotic benchmarks", and "Generative model, world model". Section 6 opens with "This section focuses on efforts on 3 challenging areas in vlm evaluation: hallucination, safety, and fairness" and then delivers six subsections, 6.1 through 6.6, the last three being alignment, efficient training, and dataset scarcity.

Three counts, three mismatches with what follows. For a survey, these are not cosmetic: the counts *are* the result.

2.3 The taxonomy is not disjoint, and the paper never says so. In Table 2, Habitat[154], Gibson[155], and iGibson[156] appear in the "Robot simulator, web agent simulator" row and then again, the same three, in the "Robotic benchmarks" row. The descriptions distinguish online rule-based simulation from offline interaction records, a reasonable distinction, but giving both rows the identical three datasets means it is doing no work. Elsewhere, MM-Vet[124] sits in three categories and MMMU[131] in two. Overlap is defensible in a benchmark taxonomy. Presenting it as a partition into "ten primary categories" without a sentence saying membership is non-exclusive is not.

2.4 The summary claim about multiple choice is checkable, and it is thinner than stated. Section 4.2 concludes: "As a result, most benchmarks rely on a multiple-choice format to assess vlms (Figure 2b)." I counted the Evaluation column of Table 3. Twenty-one of the 40 rows contain "Multiple Choice": MMTBench, MM-En/CN, VCR, MMMU, SEEDBench, RealWorld QA, MMMU-Pro, CMMLU, C-Eval, MathVista, MathVision, MMLU, MMStar, M3GIA, InternetAGIEval, EgoSchem, MVBench, MLVU, VideoMME, Perception-Test, NaturalBench. That is $21/40 = 52.5\%$, and five of those rows (MMMU, MathVista, MathVision, NaturalBench, and MM-En/CN by way of its dual category) list multiple choice jointly with another format, so the pure multiple-choice share is $16/40 = 40\%$. A bare majority is not "most benchmarks rely on". The number should be in the text with its denominator, not delegated to a figure.

2.5 Text-only benchmarks are counted as vision-language benchmarks. Table 3 lists MMLU (category "Multimodal general intelligence", size 15.9K), CMMLU, and C-Eval (both "Multilingual multi-modal understanding"). None of the three contains images. MMLU is also the example the closing summary reaches for, "with MMLU[138] alone covering 57 distinct tasks", which is a fact about a text benchmark. Table 3 also lists VQAScore (size 665K) as a benchmark; VQAScore is a metric, and Section 4.2 treats it as one. Three of the rows carry categories that appear nowhere in Table 2: "Vision-Language Alignment" (VQAScore), "Generative AI Evaluation" (GenAI-Bench), "Vision-Language Adversarial Testing" (NaturalBench). So the two tables use two different category vocabularies.

2.6 Individual cells are wrong. Table 3 assigns VQAv2 to "Chart graphic understanding"; Table 2 assigns VQAv2[49] to "Visual reasoning, understanding, recognition, and question answering". VQAv2 has nothing to do with charts, so Table 3 is the erroneous one, and the two tables contradict each other on the same dataset. MMMU gets the same treatment in Table 3 ("Chart graphic understanding") while Table 2

lists it in both chart understanding and multimodal general intelligence. The duplicate TextVQA rows disagree with each other: one says annotation CrowdSource, size 45; the other says Expert Human, size 28.6. Both cannot describe the same dataset. DocVQA and InfographicVQA are both given "Answer Matching" (the DocVQA cell reads "Answermatching"), but the official metric for both is ANLS, a normalized edit-distance score, and that is exactly the partial-credit case Section 4.2 ought to be discussing. MSCOCO-30K appears with "BLEU / Rouge / Similarity" under a text-to-image category; the standard metrics there are FID and CLIPScore, and FID is not mentioned anywhere in the paper.

2.7 Table 1 contains a model that is not a VLM. The last row is OLMo-2[78], with vision encoder / tokenizer listed as "GPT-NeoX-20B[79]". OLMo-2 is a text-only language model, and GPT-NeoX-20B is a tokenizer, not a vision encoder. The VLM built on that family is Molmo, which already occupies the row above. In the same table, CLIP[1] is labelled "Encoder-decoder", which contradicts the paper's own Section 3.2, where CLIP is described as using "separate encoders for visual and textual inputs", and Section 3.1, which states flatly that "CLIP[1] has no cross-attention". CLIP is a dual encoder. DeepSeek-VL2's parameter cell reads "4.5B x 74", which I could not interpret from anything in the text.

2.8 Two citations point at the wrong papers. Section 2 says that older approaches "primarily train models from scratch (CLIP[82])". Reference [82] is Yao et al., "Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection". That is DetCLIP, a different model. CLIP is [1] in the same paper, and [26] as well. Table 1 gives InternVL's pretrained backbone as "QLLaMA[53]", and reference [53] is Cui et al., "Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca", which is not the QLLaMA component of InternVL. I checked both against the reference list as rendered on Qeios, and they are unambiguous.

2.9 Two statements about metrics are the wrong way round. Section 4.2 defines containment exact match as a check "whether the normalized predicted answer is contained in the normalized gold answer". The sentence immediately before it explains the motivation: generative models "often generate verbose but correct answers". If the prediction is the verbose one, the containment that fixes the problem is gold inside prediction, not prediction inside gold. The definition is inverted relative to its own rationale. The same paragraph then says "exact match tends to have high recall, which often fails to account for semantic equivalence between the gold and predicted answers, frequently misjudging human-acceptable correct answers as incorrect". Marking correct answers as incorrect is a false-negative problem, that is, low recall. As written, the clause states the opposite of what the rest of the sentence argues.

2.10 Several evaluative claims arrive without numbers. Section 4.2 asserts that "llm evaluations are shown to have the highest correlations to human evaluation" [177][178][179], with no coefficient, no task, and no comparison baseline anywhere in the paper. The blind-LLM result, "llms have demonstrated an unusual ability to select correct answers even without access to the actual questions[180]", is the load-bearing evidence for the survey's central criticism of multiple choice, and it too arrives with no number: how often, on which benchmark, against what chance rate. A survey is allowed to compress. But these are the claims the argument rests on, and each is a place where two or three reported figures would cost a sentence and settle the point. Section 6.2 describes "MM-Safetybench" with no citation at all, which I take to be an oversight rather than a choice.

2.11 The announced organizing frame is dropped. Section 2 introduces "three principal research directions" and maps them to Sections 3, 4, and 5. Section 6, Challenges, is a fourth pillar the frame does not accommodate, and after Section 2, the three directions never reappear. Sections 5.1 and 5.4.1 also overlap without acknowledgement: ScreenAgent is described in both, cited as [190] in one and [260] in the other, and the boundary between device-screen agents and web agents is never drawn.

3. Reproducibility

For a primary study, reproducibility means code, seeds, and model snapshots. For a survey, it means the selection protocol, and that is what I went looking for. I could not find it.

No search strategy is reported. There is no statement of which databases or preprint servers were searched, no query strings, no date of the search, no screening counts, no inclusion or exclusion criteria. Nothing resembling a PRISMA-style flow. So none of the three tables can be reconstructed by another group, and the word "comprehensive" in the abstract cannot be assessed at all.

The sampling frame is named but not identified. The caption of Table 2 says the categories "are surveyed from 15 SoTA vlm papers." Those 15 papers are never listed. That single sentence is the entire methodological basis of the taxonomy, and it is unverifiable as it stands. Similarly, Table 1 is titled as the state-of-the-art models, but the criterion for "state of the art" is never given. The window is described as "the past five years (2019-2024)," which is six calendar years, and 2020 has zero entries.

Table 3's size column has no defined unit. The header reads "Size (K)" and the values range from 0.2 (MM-Vet) to 22,000 (GQA), five orders of magnitude, with no statement of whether the count is questions, images, image-text pairs, or annotations, and no statement of which split is counted. The duplicate

TextVQA rows, 45 against 28.6, show exactly this ambiguity producing two incompatible numbers for one dataset.

The repository is a bare link. No commit hash, no access date, no license, no description of its contents, no mapping from repository entries to the rows of Tables 1 to 3. The abstract presents it as where "detailed collections including papers and model repository links are listed," so it carries real weight and should be pinned to a release.

The figures could not be checked. As rendered on Qeios, Figures 1, 2, and 3 are not legible in the version I read; only their captions are present. This matters because Section 4.2 twice defers its quantitative claims to them, citing "Figure 2b" for the distribution of evaluation formats, while only a single Figure 2 caption exists and no sub-panel is labelled in the text. Whatever counts underlie Figure 2b should appear as numbers in the prose or as a column in Table 3, so a reader can check them without the image.

Model provenance. The "Undisclosed" cells for GPT-4V, Gemini, and Claude 3 in Table 1 are honest, and I have no complaint about them. What is missing is the version question: Sections 5 and 6 discuss GPT-4V and Gemini results without recording which snapshot the cited study used, although Section 4.2 itself warns that LLM-based evaluation struggles "with internal model updates".

4. What specifically to fix

1. Rewrite the last two sentences of the abstract and the closing sentences of Section 1. Drop "the multi-modality one is still missing". Name [14], [15], [18], [22], and [80] as existing multimodal surveys, state in two or three sentences what this one covers that they do not (my own reading is that the honest answer is the benchmark-construction taxonomy of Section 4.1 plus the robotics coverage of Section 5.3), and let the positioning rest on that.
2. Add a methods subsection, numbered 1.1 or 4.0, giving the search sources, the query terms, the cut-off date, the screening counts, and the inclusion criteria for Table 1 and Table 3 separately. Identify by citation the 15 papers named in the Table 2 caption.
3. Section 4, paragraph beginning "The number of VLM benchmarks": the count "38 vision-language benchmarks" does not match Table 3, which has 40 rows and 39 distinct names. Fix one or the other, and after removing MMLU, CMMLU, C-Eval, and VQAScore per item 5 below, restate the count.
4. Table 2 caption and the sentence in Section 4 that says "ten primary categories": the table has 13 rows. Either fold the three embodied rows into one and say so, or change the number to 13 everywhere. Then add one sentence stating that category membership is not exclusive, and

reconcile the duplicated Habitat / Gibson / iGibson entries between the "Robot simulator" and "Robotic benchmarks" rows.

5. Remove MMLU, CMMLU, and C-Eval from Table 3, or add a column marking them as text-only and explaining why a VLM benchmark survey lists them. Move VQAScore out of Table 3 into the metrics discussion in Section 4.2, where the paper already treats it as a metric. Bring the categories of GenAI-Bench and NaturalBench into the Table 2 vocabulary.
6. Delete one of the two TextVQA rows in Table 3 and say which split the surviving size refers to. While there, define the unit of the "Size (K)" column in the caption.
7. Table 3, VQAv2 row and MMMU row, Category column: VQAv2 is not chart understanding, and Table 2 already places it correctly. Correct both cells so the two tables agree.
8. Table 3, DocVQA and InfographicVQA rows: the metric is ANLS. Add it, fix the typo "Answermatching", and use the ANLS case in Section 4.2 as an example of partial-credit scoring, which the paper currently omits from a discussion where it belongs. Likewise, the MSCOCO-30K row needs FID and CLIPScore rather than BLEU and ROUGE, and FID should be defined somewhere in Section 4.2 given that Table 2 has a text-to-image category.
9. Table 1: drop the OLMo-2 row, or, if the authors mean to include it for the OLMo family lineage, move it into a footnote and state that it is a text-only backbone. The cell listing GPT-NeoX-20B as a vision encoder is wrong either way. In the same table, change CLIP's architecture from "Encoder-decoder" to dual encoder, which is what Sections 3.1 and 3.2 already say, and explain the "4.5B x 74" cell for DeepSeek-VL2.
10. Two mis-citations: Section 2, "train models from scratch (CLIP[82])" points to DetCLIP and should point to [1]; Table 1, InternVL backbone "QLLaMA[53]" points to the Chinese LLaMA and Alpaca paper. While auditing the bibliography, consolidate the duplicates I noticed: [1] with [26], [22] with [25], [27] with [58] and [104], [86] with [89], [190] with [260].
11. Section 4.2, "Answer matching" paragraph: reverse the containment definition so that the gold answer is checked for containment in the prediction, and change "exact match tends to have high recall" to describe the false-negative behavior the rest of that sentence correctly identifies.
12. Section 4.2, final paragraph: replace "most benchmarks rely on a multiple-choice format" with the count and its denominator. On Table 3 as it stands, I get 21 of 40 rows containing multiple choice, of which 5 are joint with another format, so 16/40 are pure multiple choice. Put those numbers in the text rather than in Figure 2b, and add the underlying counts for Figure 3 as well.

13. Section 4.2: attach numbers to the two claims the argument depends on. For "llm evaluations are shown to have the highest correlations to human evaluation," give at least one correlation value and the task it was measured on. For the blind-LLM result [180], give the accuracy reached without the question and the chance rate on the same benchmark. Both are short additions, and both would strengthen the paper's best argument considerably.
14. Section 6, opening sentence: "3 challenging areas" should read six, matching subsections 6.1 to 6.6, or the last three subsections should move elsewhere. The cross-reference to subsection 5.4.4 in the same sentence points at Social Goodness and appears to be a broken pointer. Section 6.2 needs a citation for MM-Safetybench, and the sentence beginning "Yinbg et al.[283] SafeBench" is missing both a verb and the correct spelling of the author name; "Rorhrbach et el.[271]" in Section 6.1 has the same problem.
15. Gaps I would want covered in Section 4, in order of how much they would help: contamination and leakage, including the fact that MMStar, already cited as [139], exists to isolate vision-independent samples; prompt and option-order sensitivity in multiple-choice VLM evaluation, which bears directly on the survey's own thesis; the open evaluation toolkits practitioners actually run, since the conclusion addresses practitioners; and inference-side efficiency (resolution tiling, visual token compression), which appears only in passing under ShowUI[259].
16. Two smaller things. Section 5.2 reads as promotional: "visually stunning video effects" and the Supermeme.ai[197] description belong on a product page. Section 6.6 ends on an uncited quotation attributed to Ilya Sutskever; cite the talk with a date or drop it.

5. Recommendation

Major revisions.

I want to separate two judgments. The scope and the ambition are sound, and parts of this paper are genuinely useful: Section 4.1 gives practitioners a way to think about where benchmark questions come from, Section 5.3 is a well-organized entry point into VLM robotics, and Section 4.2 identifies a real and under-discussed failure mode in how the field measures progress. A reader new to the area would come away better oriented than they started.

But a survey is a measuring instrument, and this one is not yet calibrated. Three separate counts in the prose disagree with the tables they point at (38 against 40 benchmarks, ten against 13 categories, three against six challenge areas). Two tables assign the same dataset to different categories. One table row is a

text-only language model presented as a vision language model, with a tokenizer in the vision encoder column. Two citations point at different papers than the text claims. The definition of containment exact match is inverted, and "high recall" states the opposite of the sentence containing it. Above all, there is no selection protocol at all, which means the word "comprehensive" in the abstract is not something a reader can check, and the framing claim that no multimodal survey exists is contradicted by five entries in this paper's own reference list.

None of this requires new research. It requires an audit pass over the three tables against their primary sources, a methods subsection of perhaps 300 words, the counts moved out of the figures into the text, and a rewritten positioning paragraph. Weeks, not months, and I would read the revision with interest. One honest limit on my side: I could not inspect Figures 1 to 3 or the linked repository in the version I reviewed, so if the figures carry the quantitative backing Section 4.2 defers to them, part of item 12 may already be half-answered.

Declarations

Potential competing interests: No potential competing interests to declare.