

Peer Review

# Review of: "Large Vision-Language Model Alignment and Misalignment: A Survey Through the Lens of Explainability"

Evgenii Arsentev<sup>1</sup>

1. Independent researcher

## 1. What the work claims

This survey sets out to explain how large vision-language models connect what they see with what they say, and why that connection sometimes fails. It defines alignment in two aspects, representational alignment (how close the visual and textual representations are in the shared embedding space) and behavioral alignment (accurate, factual, and consistent text answers about an image), in Section 2.1. It then describes three training stages (contrastive training of the visual encoder as in CLIP, adapter fine-tuning, and end-to-end fine-tuning; Section 2.2, Figure 1), gives a theoretical view (the Platonic Representation Hypothesis) and an algorithmic view of why alignment is possible (Section 2.3), and covers measurement at the representation level and the behavioral level (Section 2.4).

The main claims are as follows. The authors propose to say "misalignment" instead of "hallucination" and split it into three semantic levels: object, attribute, and relation misalignment (Section 3.1). Causes are grouped at three levels: dataset (Section 3.2.1), model (Section 3.2.2), and inference, where task discrepancy is treated as an out-of-distribution problem (Section 3.2.3). Mitigation methods are divided into parameter-tuning (training scheme and model architecture; Section 4.1) and parameter-frozen (augment-based, inference-based, decoding-based, and post-decoding; Section 4.2), with a taxonomy in Figure 3. For future work, the survey points to standardized benchmarks covering all three semantic levels (Section 5.1), explainability-based diagnosis through internal knowledge decoding and attribution methods (Section 5.2), and architectural innovations (Section 5.3).

## *2. Whether the conclusion holds on the data*

I want to start with what holds up well on the data. Section 4.2 sorts the parameter-frozen methods by where they intervene in the pipeline: augment-based, inference-based, decoding-based, and post-decoding. That split is easy to use, and it gives a reader a clear place to put a new method. I also appreciate that the Decoding-based Methods part states openly that OPERA [91] treats over-reliance on summary tokens as the cause, while SGD [44] uses summarization as the cure. The text says that subtle differences in root-cause understanding lead to contradicted methodologies. Naming a contradiction in the field like this is useful for readers. The green-apple example in Section 1 also makes the three semantic levels easy to grasp.

For a survey, the data is the literature, so I recounted the reference list as it is rendered on Qeios. By my count, it has 110 entries, and every entry is cited at least once. The caption of Figure 3 lists 66 references, from [27] to [106]. Of these, 27 appear only in the caption and are not discussed in the body text: [51], [52], [54], [55], [59] to [66], [71], [85], [92] to [101], and [104] to [106]. So ten works, [92] to [101], are listed but not discussed in Section 4.2, among them four that name decoding in the title ([94], [95], [100], [101]). I would either add a sentence on each of them or mark in the caption which works are only listed.

By my count, 71 of the 110 reference entries carry the year 2024. It would help to say something about this concentration. The survey also does not report how the literature was selected. I found no search terms, databases, date cutoff, or inclusion criteria in the text. A short paragraph on this would let readers judge coverage.

Some claims are stronger than the cited evidence. "Due to its coarse-grained nature, object misalignment is relatively straightforward to detect and mitigate" (Section 3.1) comes without a citation or figure. "While RLHF guarantees high-quality training data" (Section 4.1) is also a strong statement, and I would soften it or support it. "This makes inference-based methods particularly effective at reducing misalignment" (Section 4.2, Inference-based Methods) comes with no reported comparison. More generally, no benchmark numbers from the cited works are compared anywhere. A reader cannot see which family of methods works better on POPE or CHAIR, although both are named in Section 2.4. A small table with a few reported results per family would address this.

There is also a link between sections that I could not follow. Section 4.2, Augment-based Methods, opens with "As analyzed in Section 3, insufficient input of image information is one of the primary causes of

misalignment". But the causes in Section 3.2 are listed under other names, and this cause is not among them. I would either add it to Section 3.2 or reword the opening.

Finally, the explainability lens in the title. Methods of explainability (internal knowledge decoding, probing, attribution) appear mainly as future work in Section 5.2. That section cites three works, [108], [109], and [110], and all three share authors with this survey (Zhao H, Du M). The body, Sections 3 and 4, does not organize causes or methods by an explainability method. I would suggest either showing in Sections 3 and 4 where such methods already explain a cause or a fix, or adjusting the title so that it matches what the body delivers.

### *3. Reproducibility*

For a survey, reproducibility is a different question from the one an experimental paper raises. This paper has no experiments, code, or data release, so there is nothing to rerun. What I can ask is whether another team, starting from the same research question, could rebuild the same corpus of papers and arrive at the same taxonomy of mitigation methods. From the text alone, I could not tell.

I found no description of the literature search anywhere in the whole text. There are no databases, no search strings, no date cutoff, and no inclusion or exclusion criteria. I would add a short paragraph on this, so readers can see how the surveyed works were chosen.

The model versions are also not pinned. Section 1 names GPT-4V [1], but reference 1 is an OpenAI web page, "ChatGPT can now see, hear, and speak", accessed September 25, 2023. Section 1 also names GPT-3.5 [6], while reference 6 is Brown TB (2020), "Language models are few-shot learners", which is the GPT-3 paper. It would help to cite a source that matches the model named in the text, or to say which model the reference stands for.

To make the taxonomy checkable, I would suggest a supplementary table with one row per surveyed method. The columns could be the reference number, the semantic level addressed (object, attribute, relation), the cause level targeted (dataset, model, inference), whether the model is tuned or frozen, and the benchmarks used. Readers could then verify each classification, and the authors could update the table later.

The main artifact of this survey is Figure 3, the taxonomy of mitigation methods, together with the reference list. In the text version I worked from, Figure 3 comes through only as its caption and a list of 66 references, so I could not check which leaf each method sits in. The table above would solve this as well.

#### 4. *What specifically to fix*

1. In Section 4.2 and Figure 3, by my count, 27 of the 66 references in the Figure 3 caption are not discussed in the body. These include [92] to [101] (four of them, [94], [95], [100], and [101], are decoding methods by title) and the post-decoding works [104] to [106]. I would either give each a sentence or say in the text that Figure 3 lists more works than are discussed.
2. I would add a short methods paragraph with the databases, search strings, date cutoff, and inclusion criteria, since I could not find any of these in the text. By my count, 71 of the 110 references carry the year 2024, so the cutoff date matters for the reader.
3. It would help to add a comparison table of reported results on shared benchmarks for the main method families, taken from the original papers. POPE [24], CHAIR [25], and MME [26] are already named in Section 2.4. With such a table, claims like "particularly effective" in Section 4.2 (Inference-based Methods) would rest on numbers.
4. The explainability lens, which is the promise of the title, could be more visible in the body. For each cause in Section 3.2, I would say which explainability evidence (probing, attention analysis, attribution) supports it, instead of keeping explainability mostly in Section 5.2. Section 5.2 cites [108], [109], and [110], all by authors of this survey, so I would also add independent interpretability work on vision-language models, for example [23], which is already in the list.
5. In Section 4.2, Augment-based Methods, the text says "As analyzed in Section 3, insufficient input of image information". This is not one of the causes named in Section 3.2. I would either add it to Section 3.2 or point to the matching cause.
6. In Section 2.4, "GAVIR[28]" should be checked against reference 28 (Liu F et al.), because the evaluation method from that work is usually called GAVIE. Also, in Section 2.4, cosine similarity for alignment is cited to [15], a paper on protein representation (Shu D et al.), and a vision-language source would fit better. The sentence "High alignment corresponds to scores close to 1, while low alignment corresponds to scores closer to 0" should also mention that cosine similarity can be negative.
7. In Section 1, GPT-3.5 is cited to Brown TB (2020) (reference 6), which is the GPT-3 paper, and GPT-4V is cited to an OpenAI web page (reference 1). A source that matches each model would help the reader.
8. In Section 4.2, Augment-based Methods, references [72] to [75] are retrieval-augmented captioning models (video captioning, LMCap, Smallcap, retrieval-augmented image captioning) rather than

LVLMM methods. I would say so in the text or move them.

9. A few language points. In Section 4.2, Decoding-based Methods, the sentence "However, the imbalance attention between what still remain debated." needs rewriting. In Section 4.1, "This can often achieved" is missing a word. Also, in Section 4.1, "While RLHF guarantees high-quality training data" is too strong, and "aims to provide" would match the cited works better.

## *5. Recommendation*

**Minor revisions.** The survey gives a clear, readable map of LVLMM misalignment: the three-level taxonomy (object, attribute, relation; Section 3.1) crossed with three levels of causes (Section 3.2), and the split of training-free methods by where they intervene (Section 4.2), are both useful. I would add a short paragraph on how the literature was selected. It would also help to discuss, or at least explicitly note, the 27 references found only in the Figure 3 caption (my count). A results table on shared benchmarks would make the comparison easier to follow. Finally, I would let the explainability lens reach into Sections 3 and 4, rather than mainly Section 5.2. None of this changes the structure of the paper.

## **Declarations**

**Potential competing interests:** No potential competing interests to declare.