

Peer Review

Review of: "VideoHallu: Evaluating and Mitigating Multi-modal Hallucinations for Synthetic Videos"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper introduces VideoHallu, a question-answering benchmark built on videos from seven text-to-video generators (Sora, Veo2, Kling, Runway Gen2, PixVerse, LaVie, CogVideo; §3.1). Starting from 141 hand-written adversarial prompts, five annotators wrote 3,233 question-answer pairs (§3.4) sorted into four categories and thirteen sub-categories: alignment, spatial-temporal consistency, commonsense, and physics (§3.3, Figure 2). The authors argue that current multimodal LLMs fail on these videos because they fall back on language priors about how the real world behaves instead of reporting what the synthetic video actually shows. Table 1 evaluates twenty model configurations and finds the best overall score to be 51.4% (Table 1, row Qwen2.5-VL-32B, column Overall), with GPT-4o at 45.5% and Gemini-2.5-Pro at 49.8%. A GPT-4 judge scores the free-form answers against gold answers, and the authors report 99.3% agreement with humans on 200 sampled pairs (§4).

The second half is a mitigation study. Qwen2.5-VL-7B is post-trained with GRPO on real-world video QA, PhysBench, and 800 VideoHallu training pairs, with a reward equal to the cosine similarity of answer embeddings from an answer-equivalence BERT (Eq. 1). The best curriculum model reaches 54.2% (Table 2, row PRW+SR-R1-7B, column Overall), which the paper presents as the highest score of any model and as evidence that synthetic counterexamples plus RL fine-tuning reduce hallucination (Abstract; §4.3; §4.4, point 3). A secondary claim, in §4.2, is that chain-of-thought RL models do worse than their base model on synthetic video because reasoning amplifies reliance on priors.

The motivating observation is real, and I found it convincing in the case study of §3.2: a generator that renders a feather and a rock falling in a physically wrong way, and models that describe the physically

correct outcome rather than the video. A benchmark built around that failure mode is worth having. The taxonomy in §3.3 is reasonably fine-grained, the model sweep in Table 1 covers open, R1-style, and closed models, and the data are released on GitHub. My concerns are about how the numbers in Tables 1 and 2 relate to each other and how far the prose goes beyond them.

2. Whether the conclusion holds on the data

The test set is much smaller per sub-category than the paper lets on. The size of each sub-category is never stated. I recovered it from Table 1 itself. For each column, I searched for the smallest denominator n at which every reported percentage (all rows except Gemini-2.5-Flash, discussed below) equals k/n rounded to one decimal. The result is: A-EC 185, A-EP 203, A-ERAC 174, A-SR 42, SC-CD 42, SC-SD 142, SC-TD 469, CS-K 61, CS-R 69, P-C 21, P-CAP 17, P-M 90, P-ST 10. These sum to $185 + 203 + 174 + 42 + 42 + 142 + 469 + 61 + 69 + 21 + 17 + 90 + 10 = 1,525$, which is exactly the test-set size in footnote 1. I am fairly confident in the inference for that reason, but it is an inference, and the authors should simply publish the counts. If they are right, the physics category, which the title, abstract, and §4.4 treat as the heart of the benchmark, holds 138 test questions ($21 + 17 + 90 + 10$), and P-ST holds ten. One question moves P-ST by 10 points. Statements such as "significant gains in commonsense and physics reasoning question accuracy" (§1, contribution 3) rest on those cells: in Table 2, row PRW+SR-R1-7B, column P-ST, 80.0 against 40.0 for Qwen2.5-VL-7B is 8/10 against 4/10.

Table 2's fine-tuned rows cannot come from the Table 1 test set. The six "Previous SoTA" rows at the top of Table 2 are copied exactly from Table 1 and are all attainable at the denominators above. The five fine-tuned rows are not. Two examples: Table 2, row GRW-R1-7B, column A-SR, reports 34.9, but with 42 questions the only nearby values are $14/42 = 33.3$ and $15/42 = 35.7$; the same row, column P-C, reports 62.2, whereas $13/21 = 61.9$. Eleven of the thirteen columns in that row fail this test, and at least eleven fail in each of PRW-R1-7B, SR-R1-7B, GRW+SR-R1-7B, and PRW+SR-R1-7B. Only P-ST (multiples of 10) and, in most rows, SC-TD are consistent. The fine-tuned values are instead consistent with different denominators (for example, A-SR 43, P-C 37, SC-SD 57). The Overall column does not reconcile either: if I weight the PRW+SR-R1-7B sub-category scores by the Table 1 counts, the overall is about 58.0%, not the reported 54.2%; for SR-R1-7B it is about 59.5% against a reported 53.4%. So either the fine-tuned models were scored on a different split, or the sub-category columns and the Overall column were computed differently. In either case, the comparison that carries the mitigation claim (fine-tuned rows against baseline rows in the same table) is not shown to be like-for-like.

The Gemini-2.5-Flash row has the same problem. Table 1, row Gemini-2.5-Flash, column P-ST, is 41.4, which is not a multiple of 10; column P-C is 40.0, between $8/21 = 38.1$ and $9/21 = 42.9$; column A-SR is 60.0, whereas $25/42 = 59.5$. Eleven of its thirteen cells are unattainable at the Table 1 counts. Its Overall (49.6) also equals the unweighted mean of its thirteen cells, whereas most rows follow the item-weighted mean. My guess is that it was evaluated on a different subset or pasted from another table, but I cannot tell from the text.

The Overall column is not always the weighted mean of its row. For most Table 1 rows, the weighted recomputation agrees with the reported Overall to within 0.2 points. Video-R1-SFT-CoT is the exception: the cell counts implied by its row are 99, 113, 119, 16, 19, 64, 189, 28, 33, 7, 7, 45, 5, which sum to 744, and $744/1,525 = 48.8\%$, while the table reports 50.6%. That 1.8-point gap matters because it is the difference between this configuration sitting below or level with Video-R1-SFT (also 50.6%).

The size of the mitigation effect is uncertain and quoted three ways. The Table 2 caption says the curriculum gives "a 2.8% performance boost"; §1, contribution 3, says "a notable 3%". Against the base model, the difference is $54.2 - 51.0 = 3.2$ points (Table 2, rows PRW+SR-R1-7B and Qwen2.5-VL-7B, column Overall). The 2.8 figure is $54.2 - 51.4$, i.e., against Qwen2.5-VL-32B, which was not the model being trained. No confidence interval, seed variation, or significance test is given. Taking $n = 1,525$ at face value, an unpaired two-proportion test gives $SE = \sqrt{(0.542 \times 0.458/1,525 + 0.510 \times 0.490/1,525)} = 0.0181$, $z = 0.032/0.0181 \approx 1.77$, two-sided $p \approx 0.08$. A paired McNemar test on per-item outcomes would be the correct test and could well be tighter, but the per-item data are not in the paper. Given the denominator mismatch above, even this calculation assumes more than I can verify.

The curriculum claim is not supported by the ablation. §4.3 concludes that learning "general physics knowledge from the data, then from synthetic videos can help models gain most". In Table 2, column Overall, training on the synthetic set alone gives 53.4 (row SR-R1-7B), the physics-then-synthetic curriculum gives 54.2 (row PRW+SR-R1-7B), and the general-then-synthetic curriculum gives 52.1 (row GRW+SR-R1-7B). One curriculum is 0.8 points above no curriculum, which is about 12 questions out of 1,525 ($0.008 \times 1,525 \approx 12.2$); the other is 1.3 points below. With one run per configuration, I read this as "no detectable effect of curriculum ordering," not as support for it. The second research question of §4.3 ("Does curriculum learning help video understanding?") is therefore still open.

Several prose claims contradict Table 1.

- §4.2 states that "short-answer prompt style outperform those explicitly prompted for reasoning". This holds for Video-R1 (50.8 against Video-R1-CoT 48.3) but is reversed for VideoChat: VideoChat-R1 scores 44.2 and VideoChat-R1-think 45.3 (Table 1, column Overall).
- §4.2 also says both RFT models "underperform their base model". For Video-R1 the gap is $51.0 - 50.8 = 0.2$ points, which I would not interpret without an interval.
- §4.1 says that "within the same series ... larger models outperform smaller ones across tasks". InternVL3-38B (46.6) is below InternVL3-14B (46.7) overall, and on P-C it scores 23.8 against 47.6 for InternVL3-9B.
- §4.1 says models "excel at tasks requiring direct observation (e.g., instruction-following, alignment, spatial-temporal consistency)" and struggle on commonsense and physics. For the best model this is not the case. Weighting Qwen2.5-VL-32B's cells by the recovered counts gives alignment $363/604 = 60.1\%$, commonsense $74/130 = 56.9\%$, physics $68/138 = 49.3\%$, and spatial-temporal consistency $276/653 = 42.3\%$. The supposedly perceptual category is the weakest one, and commonsense is the second strongest.
- §1, contribution 2, reports "a gap of over 20% accuracy" between perceptual-heavy and reasoning-heavy questions. At the category level, for the same model, the largest perceptual-minus-reasoning gap I can compute is alignment against physics, $60.1 - 49.3 = 10.8$ points. A gap above 20 points exists only between individual sub-categories (for example A-ERAC 67.8 against SC-TD 40.7), and SC-TD is itself labelled perceptual.
- §4.4, point 2, says that "RFT only on real-world data even worsens performance (Table 1)", but the authors' own real-world-only run in Table 2 (row GRW-R1-7B, 51.5) is slightly above the base model (51.0).

The "easily solvable by humans" premise is untested. The abstract describes the questions as "easily solvable with human-level perception and reasoning", and §3.1 asks models to "approach human-level video perception". No human accuracy is reported anywhere. For a hallucination benchmark, this is the missing baseline: without it, I cannot separate model failure from questions that are ambiguous, or from videos in which the anomaly is not actually visible at the frame rate the models receive.

Annotation quality is not measured. Five annotators wrote the pairs, and each annotator assigned the categories (§3.4). I found no inter-annotator agreement on either the gold answers or the category labels, no double annotation, and no description of what made the annotators "expert-level". Categories matter here, since the paper's main analytical claims are category contrasts.

The judge validation is thin and has an arithmetic problem. §4 reports 99.3% human-GPT agreement on 200 sampled pairs. On 200 items, the attainable values are 99.0% (198/200) and 99.5% (199/200); 99.3% cannot occur. The sample was also restricted to answers "with length greater than five words", so short answers, where string mismatch between gold and prediction is most likely, were never validated. The number of human raters and how disagreements were resolved are not given.

Possible train/test leakage through prompts. Footnote 1 guarantees no video overlap between the 800 training and 1,525 test pairs. But each of the 141 prompts was rendered by up to seven generators (§3.1). If the same prompt, and therefore the same intended anomaly and very similar questions, appears in training via one generator and in testing via another, the SR and PRW+SR gains in Table 2 may partly reflect memorized question templates. A prompt-level split would rule this out.

Where the 3,233 pairs went. §3.4 reports 3,233 pairs; footnote 1 splits the data into 800 train and 1,525 test, which is 2,325. The remaining 908 pairs are not accounted for.

3. Reproducibility

Releasing the data on GitHub is a real strength, and it is what would let a reader check most of what I raise above. From the Qeios text, though, very little of the experimental setup can be reproduced.

Model versions. None of the closed models carries a snapshot or date: GPT-4o, Gemini-2.5-Flash, and Gemini-2.5-Pro are named only by family (Table 1). The judge is described only as "GPT-4" (§4), with no version. As is usual for LLM papers, decoding parameters are absent: I could not find temperature, top-p, maximum output length, or whether sampling was used, for either the evaluated models or the judge.

Video input. The number of frames, frame-sampling rate, and input resolution per model are not reported. §4.1 attributes GPT-4o's weakness to treating video "frame-by-frame," which is partly a property of how many frames the authors sent it. Without these settings, the cross-model comparison in Table 1 cannot be repeated.

Prompts. Table 3 shows one prompt, and it does not match the protocol described in §4. The table asks the model to "select the most accurate answer choice" from options A to C and to "Explain your reasoning step by step," whereas §4 says non-RL models received "standard prompting to elicit direct answers" and that answers are "free-form and open-ended." Table 3 also contains a "Video Description" row with a text summary of the scene. If that description is passed to the model, it gives away the physically normal outcome and changes what the benchmark measures. I could not determine from the text which prompt was actually used for which model.

Training. For GRPO, the paper gives one epoch and a gradient accumulation step of 1 (§4.3). It does not give the learning rate, batch size, group size (number of rollouts per prompt), KL coefficient, maximum generation length, optimizer, number of frames during training, seeds, or hardware and GPU hours. The reward model is described as an answer-equivalence BERT (Eq. 1) without a checkpoint identifier. The training data description is also inconsistent: §4.3 first says "4,500 from the video QA subsection of Video-LLaVA" and then, in the GRW paragraph, "We sample 3,000 videos from this dataset"; the same paragraph calls the source "Video-LLaMA[34]," while reference [34] is Video-LLaVA, and contribution 3 calls it "Video-LLaVA-80K." §4.4, point 3, mentions "gains on validation sets," but no validation split is described.

Figures. As rendered on Qeios, Figures 1 to 18 are not legible to me in the text; I could read only their captions. I therefore could not check the case-study claims in §3.2 against the frames, nor the qualitative examples in Appendix A. The row-group labels at the top of Table 1 (the "< 7B" group) are also garbled in the Qeios rendering.

4. What specifically to fix

1. Add a table of per-sub-category counts for train and test. My reconstruction from Table 1 gives 185, 203, 174, 42, 42, 142, 469, 61, 69, 21, 17, 90, 10 (sum 1,525); please confirm or correct it, and explain how 3,233 pairs (§3.4) become $800 + 1,525 = 2,325$ (footnote 1).
2. Table 2: re-score all five fine-tuned rows on exactly the Table 1 test set and recompute their Overall column as the item-weighted mean. As published, their cells are unattainable at the Table 1 counts (for instance GRW-R1-7B, A-SR 34.9 and P-C 62.2), and PRW+SR-R1-7B's cells imply roughly 58.0% overall, not 54.2%.
3. Check Table 1, row Gemini-2.5-Flash. Its P-ST value of 41.4 is impossible with ten questions.
4. Table 1, row Video-R1-SFT-CoT: the Overall of 50.6 does not match its cells ($744/1,525 = 48.8\%$).
5. Report the mitigation effect once, against the model that was actually trained. That is +3.2 points over Qwen2.5-VL-7B, not the 2.8% in the Table 2 caption or the 3% in contribution 3, and it needs a confidence interval, ideally a paired McNemar test on per-item correctness plus at least three training seeds.
6. §4.3's curriculum conclusion should be restated. Table 2 shows SR-R1-7B (53.4) above GRW+SR-R1-7B (52.1) and only 0.8 points below PRW+SR-R1-7B (54.2). Either run more seeds or say the ordering effect was not detected.

7. Correct the category claims in §4.1 and contribution 2. For Qwen2.5-VL-32B, spatial-temporal consistency is the weakest category (42.3% by my weighting), and the category-level perception/reasoning gap is about 11 points, not "over 20%".
8. §4.2: the short-answer versus CoT claim is contradicted by VideoChat-R1 (44.2) against VideoChat-R1-think (45.3).
9. §4.1, "larger models outperform smaller ones across tasks": qualify it, given InternVL3-38B against InternVL3-14B.
10. Add a human baseline on at least a stratified sample of the test set. The abstract's "easily solvable" claim depends on it.
11. Report inter-annotator agreement for gold answers and for category labels (§3.4), and define the annotators' expertise.
12. §4, judge validation: 99.3% is not possible on 200 items. Give the exact count, include short answers in the validation sample, name the GPT-4 version, and state how many human raters took part.
13. Make Table 3 the actual prompts used, one per model family, and say whether the "Video Description" row is shown to the model. If it is, the benchmark partly tests reading, not watching.
14. State whether the train/test split in footnote 1 is also disjoint at the prompt level, since each prompt yields videos from up to seven generators.
15. Add model snapshots or access dates for GPT-4o, Gemini-2.5-Flash, Gemini-2.5-Pro, and the judge; decoding parameters; and frame counts and resolution per model.
16. §4.3: give the full GRPO configuration (learning rate, batch size, group size, KL weight, generation length, seeds, compute), the reward checkpoint used in Eq. 1, and one consistent description of the real-world data (4,500 or 3,000; Video-LLaVA or Video-LLaMA).
17. §1, first paragraph: Gemini-2.5-Pro, GPT-4o, and Qwen2.5-VL are called "video generation models". They are understanding models; the sentence should name generators.
18. A per-generator breakdown (Sora, Veo2, Kling, and the others) would tell readers whether the difficulty comes from the questions or from particular generators, which matters for anyone using VideoHallu as an evaluator of video generation.

5. Recommendation

Major revisions. The benchmark idea is sound and useful, and the evaluation sweep in Table 1 is broad. But the mitigation result, which is half the paper, rests on Table 2 rows that do not appear to have been scored on the same test set as the baselines they are compared with, and the effect is small (3.2 points, $p \approx$

0.08 by an unpaired test) with no seeds or intervals. Several interpretive claims in §4.1 and §4.2 are contradicted by the authors' own Table 1, and the key premise that humans find the questions easy is not measured. All of this is fixable with the released data, and I would be glad to see the revised version.

Declarations

Potential competing interests: No potential competing interests to declare.