Can Video LLMs See Through Illusions? Video-Illusion QA Benchmark Dataset

Souto Ohira, Tosho Hirasawa, Mamoru Komachi
Hitotsubashi University
souto@scl.sds.hit-u.ac.jp, tosho@scl.sds.hit-u.ac.jp, mamoru.komachi@r.hit-u.ac.jp

Abstract

Recent advances in multimodal learning have sparked growing interest in understanding how large vision-language models interpret optical illusions. While the behavior of image LLMs—which handle one image and text but not video input—on visual illusion images has been actively explored, research on their video counterparts remains limited. Video LLMs, which process sequential frames, are gaining prominence in areas such as robotics and autonomous driving. Understanding how they handle visual illusions over time is crucial for safety and may also reveal their potential as computational models of human cognition. To address this gap, we present the Video-Illusion QA Benchmark (VILQA), a novel video question answering (QA) benchmark mainly composed of carefully curated illusion videos that exhibit temporally driven perceptual phenomena. To the best of our knowledge, VILQA is the largest and most comprehensive benchmark for temporally-driven visual illusions. We evaluate several video LLMs on this benchmark from multiple perspectives. Some models were able to perceive visual illusions in a way similar to the general human experience and demonstrated an ability to resist illusions even more effectively than humans. The constructed dataset is available at https://github.com/SDS-NLP/VILQA.

Keywords: Video LLMs, Optical Illusion, Human Comparison, First-Person Perspective Video

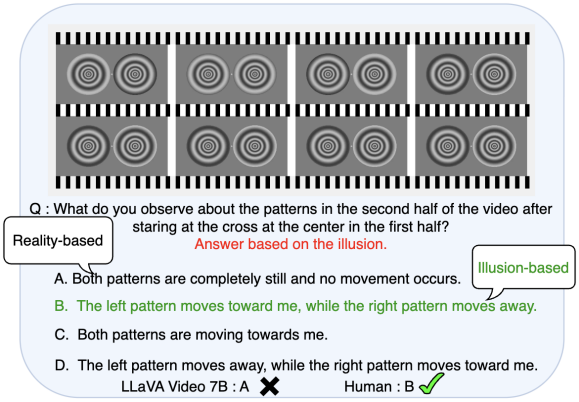

Figure 1: VILQA example illustrating an illusion-based question where prolonged viewing of motion induces a perceived motion in the opposite direction after the stimulus stops, highlighting the gap between illusion- and reality-based reasoning.[2]

1. Introduction

Visual illusion is a phenomenon in which there is a discrepancy between the physical properties of a stimulus and the perception it evokes (Todorović, 2020). This discrepancy provides valuable insights into how the brain interprets visual information. In this context, illusions help researchers understand human visual information processing by disentangling basic sensory input from higher-level cognitive interpretation (Day, 1984). In particular, numerous visual illusions have been reported in static images, arising from visual attributes such as brightness, color, shape, and depth. A well-known example is the Müller-Lyer illusion, in which adding arrowheads to the ends of a line segment alters the perceived length of the segment.

In contrast, videos involve visual information that changes over time, which can give rise to dynamic illusions that depend on temporal changes that cannot be reproduced with static images. One example of such a video-specific illusion is the *motion aftereffect* (Figure 1). In this illusion, prolonged observation of motion in a single direction leads to neural adaptation, whereby neurons tuned to that direction gradually reduce their activity. When the motion ceases, neurons responsive to the opposite direction become relatively more active, resulting in the illusion of motion in the opposite direction (Mather et al., 1998). It represents a perceptual phenomenon fundamentally different from static image illusions, in that temporal change is essential to its emergence.

Visual illusions, which reflect perceptual mismatches with physical reality, are a hallmark of biological perception and serve as valuable probes for comparing human cognition with that of computational models. If video LLMs, which process both visual and linguistic input, exhibit similar perceptual tendencies, such behavior may provide supporting evidence for their potential as computational models of human cognition. Moreover, prior works show that PredNet trained on first-person videos can replicate motion illusions (Watanabe et al., 2018;

[2]All illusion figures used in this study were sourced from the Illusion Forum (https://illusion-forum.ilab.ntt.co.jp/)

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9291–9300
11–16 May 2026.

Kobayashi et al., 2022). Examining their internal dynamics and behavioral patterns could yield insights into the underlying mechanisms of human perception and cognition.

However, despite the growing interest in video LLMs, no illusion video QA dataset has been developed to facilitate their analysis, and progress in this area remains limited. Based on the above, we propose the Video-Illusion QA Benchmark (VILQA), which encompasses not only illusion videos that induce perceptual mismatches with physical reality, but also videos that contain anomalous phenomena without any perceptual mismatch. Using the mismatch videos, we investigate whether video LLMs can accurately perceive physical reality without being misled by illusions—an essential capability for safety–critical applications such as robotics and autonomous driving—and whether they can recognize illusions in a human-like manner, providing insights into the alignment between human and model visual processing. Furthermore, by using the anomaly videos, we explore how closely the model's sense aligns with that of humans from the perspective of surprise. VILQA is a novel QA benchmark, illustrated in Figure 1, primarily composed of carefully curated illusion videos that exhibit temporally driven perceptual phenomena. It covers a wide range of visual illusion videos and, to the best of our knowledge, is the first QA benchmark to include such a diverse set of illusion videos.

Using VILQA, we conduct a comprehensive analysis of several video LLMs, including a fine-tuned model trained on first-person video datasets, to evaluate (1) their ability to perceive illusions in a human-like manner, (2) their capacity to resist illusions and accurately recognize the physical world, and (3) the extent to which their perceptual sense aligns with that of humans.

The main contributions of this study are:

1. We introduce the Video-Illusion QA Benchmark (VILQA), a novel VQA dataset primarily composed of systematically curated illusion videos that exhibit temporally driven perceptual phenomena.

2. We perform a detailed analysis of the behavior of existing video LLMs when exposed to illusion videos and compare their responses with those of humans.

3. We investigate whether fine-tuning video LLMs on first-person perspective videos, which reflect what humans observe in their daily lives, enhances their ability to perceive illusions. We show that such fine-tuning can partially align model behavior with human illusion-related responses.

2. Related Works

2.1. Visual Illusions and Image LLMs

In recent years, research on the effects of visual illusions on deep learning models has become increasingly active. Previous studies have reported that models such as Convolutional Neural Networks (CNNs), which are capable of recognizing visual patterns, exhibit illusion-like responses similar to those of humans when exposed to various types of visual illusions, including motion illusions, brightness and color illusions, and completion phenomena (Watanabe et al., 2018; Kobayashi et al., 2022; Sun and Dekel, 2021; Gómez-Villa et al., 2019; Gomez-Villa et al., 2020).

Building on these findings, recent studies have rapidly advanced in evaluating the impact of illusion images on image LLMs. QA (Question Answering) benchmarks that cover multiple categories of visual illusions have been developed, enabling detailed analyses of how the state-of-the-art image LLMs respond to illusion stimuli. Several studies have claimed that certain image LLMs can be deceived by visual illusions in a human-like manner under certain conditions (Guan et al., 2024; Rostamkhani et al., 2025; Shahgir et al., 2024).

2.2. Video LLMs

Spurred by the rapid advancement of image LLMs, research on video LLMs has also progressed rapidly (Maaz et al., 2024; Zhang et al., 2025; Wang et al., 2024). Video LLMs integrate video and natural language. By leveraging temporal context that static images cannot capture, they enable the processing of more complex information and open the door to more advanced applications, such as robotics and autonomous driving.

However, several challenges remain to be addressed for real-world deployment. For example, in the context of autonomous driving, numerous visual illusions are known to occur on the road, some of which have been identified as contributing factors in fatal accidents (Ekroll et al., 2021; Clark et al., 2013; Dong et al., 2021; Redelmeier and Raza, 2018). Accordingly, video LLMs deployed in such scenarios may encounter unexpected visual illusions. Given that existing deep neural networks (DNNs) have been shown to be vulnerable to such illusions, it is essential to verify that video LLMs exhibit strong robustness against these perceptual challenges.

Furthermore, from the perspective of the type of information being processed, video LLMs can be regarded as handling information structures that are more similar to those used by humans than image LLMs. This is because human visual perception is not based on single static images, but rather

Illusory contours	Motion-induced blindness
Dynamic luminance	Motion contrast
Color afterimage	Motion assimilation
Scene afterimage	Apparent motion (subjective square)
Troxler fading	Position shift (moving pattern)
Blur adaptation	Cycloid illusion
Reverse perspective	Visual jitter
Motion aftereffect	Deformation lamps
Lilac chaser	Hybrid image
MacKay rays	Drop-shadow illusion
Illusory motion (static pattern)	Shading illusion
Enigma illusion	Crater illusion
Pinna illusion	Hollow-mask illusion
Breathing square	

Table 1: VILQA includes 27 types of illusion videos sourced from the Illusion Forum.

on continuous streams of visual input over time, combined with linguistic information and other contextual cues. Consequently, video LLMs are not merely tools for extending practical applications; they also hold significant potential as computational models of human cognition.

2.3. Alignment of Video LLMs with Human Perception

Previous studies have reported that models trained on natural images tend to exhibit human-like misperceptions in response to visual illusions, whereas models trained on synthetic images, random patterns, or those that remain untrained show little to no susceptibility to such illusions (Kim et al., 2021). From the perspective of video data, first-person perspective videos, which are captured from a human's point of view, are considered to better capture human daily experiences and subjective perceptions, as opposed to third-person videos. In line with this, prior studies have demonstrated that training temporal prediction models such as PredNet on first-person perspective videos induces illusion-like responses similar to those observed in humans (Watanabe et al., 2018; Kobayashi et al., 2022).

Based on these, we hypothesize that fine-tuning a video LLM on first-person natural videos that closely mirror everyday human experience induces human-like perceptual traits and strong illusion recognition. We test this by fine-tuning on a first-person QA dataset and comparing model performance before and after.

3. Benchmark Dataset

3.1. Video Collection

Our primary selection criteria focused on videos that exhibit a discrepancy between physical properties and perceptual appearances, as well as those that present anomalous, unexpected or rare phenomena. These videos were curated to cover a wide range of illusion types and temporal dynamics, ensuring the benchmark captures diverse perceptual phenomena.

In this study, we adopted web-sourced videos rather than synthetic ones in order to collect a diverse range of illusions and investigate what types of videos tend to mislead perception, while also better reflecting real-world applications. To collect a wide range of illusion videos involving temporal change, we first searched YouTube using both English and Japanese queries, such as "illusion," "visual illusion," and "trick art," along with their Japanese equivalents. In addition, we included videos from two major channels known for high-quality illusion content: `@brusspup`[3] and `@TheIllusionContest`[4]. As YouTube videos often include extraneous content such as subtitles or verbal explanations of the illusion, we applied temporal cropping based on timestamps and spatial cropping using predefined coordinates and dimensions. Furthermore, to comprehensively cover temporally dependent illusions that only arise in video, we included 27 illusion types published on the Illusion Forum in our dataset, as summarized in Table 1.

As a result, we collected a total of 275 illusion videos. The statistics are shown in Table 2. This scale surpasses or is at least comparable to existing illusion image datasets including those with 100 images from 16 root stimuli and 374 images (Zhang et al., 2023; Shahgir et al., 2024).

3.2. QA Annotation

This QA task is designed to evaluate whether video LLMs demonstrate (1) *illusion recognition*, i.e., the ability to recognize visual illusions in a human-like manner; (2) *illusion robustness*, i.e., correctly perceiving illusions without being misled; and (3) *anomaly sensitivity*, i.e., the capacity to react to unexpected or unusual events in a manner similar to human observers.

To assess (1) and (2), we annotated QA pairs for videos that exhibit a discrepancy between physical properties and perceptual appearances (the "Discrepancy" category), as illustrated in Figure 1. For (3), we targeted videos that do not involve such a

[3] https://www.youtube.com/@brusspup

[4] https://www.youtube.com/@TheIllusionContest

Source	# video	Discrepancy	Anomaly
YouTube (En)	172	99	73
YouTube (Ja)	59	40	19
Illusion Forum	44	38	6
Total	275	177	98

Table 2: Statistics of video data sources.

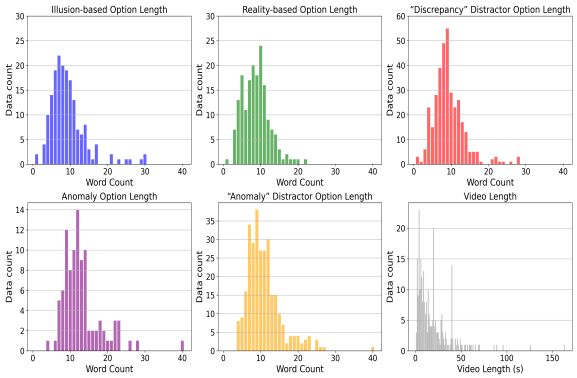

Figure 2: Distributions of option lengths and video durations in the VILQA dataset.

discrepancy, but instead present events that are rare, unexpected, or uncomfortable, an event likely to surprise a human observer (the "Anomaly" category).

To ensure diversity in questions and corresponding answer choices, we hired 31 native English-speaking annotators using Prolific. For each video, annotators first determined whether it belonged to the "Discrepancy" or "Anomaly" category. Based on this classification, they then followed one of two distinct annotation procedures: Discrepancy Annotation or Anomaly Annotation, each based on the category they judged by themselves. Finally, we manually reviewed the dataset for quality and rewrote if necessary.

3.2.1. Discrepancy Annotation

If the video was categorized as "Discrepancy," they were asked to create one question related to the illusion presented in the video, along with the following four answer choices:

1. **(I1) Illusion-based perception**: The perceptual interpretation typically experienced by most people under the illusion.
2. **(I2) Reality-based perception**: The perceptual interpretation when the illusion does not influence perception.
3. **(I3) Distractor 1**: A plausible but incorrect interpretation.
4. **(I3′) Distractor 2**: Another plausible but incorrect interpretation.

	Question	Option
Discrepancy	22.6%	59.9%
Anomaly	72.4%	78.6%
Total	40.4%	66.5%

Table 3: Percentage of VILQA questions and options undergoing major semantic revisions during manual review.

3.2.2. Anomaly Annotation

If the annotators categorize the video as belonging to the "Anomaly" category, the question was standardized as follows: "What is unusual about this video from the perspective of an optical illusion?" The annotators were then asked to provide a correct perceptual interpretation that describes the observed anomaly in general, along with three plausible but incorrect alternatives.

3.2.3. Quality Control

Every QA pair was manually reviewed by the first author after annotation. If any item did not conform to the annotation guidelines, or if issues such as excessively long answer choices or typographical errors were observed, the authors made corrections accordingly. In particular, illusion-based answer choices were revised, mainly following the explanations on the original video source sites, when they conflicted with typical human perception.

These modifications were carried out with careful consideration to preserve the overall diversity of the QA content.

As shown in Table 3, a substantial portion of the 275 QA items underwent major semantic revisions to ensure dataset quality. The table reports the proportion of items where at least one of the four answer choices was significantly modified, as well as those where the question itself required substantial changes. The values in Table 3 may appear large at first glance, as they also include category changes. For example, 63 items initially labeled as "Discrepancy" were reclassified as "Anomaly" after strict revision. Individual differences in susceptibility to visual illusions are likely a contributing factor (Schwarzkopf et al., 2011; Cretenoud et al., 2021).

Figure 2 summarizes the distributions of option lengths and video durations of the final dataset.

3.3. Benchmark Usage ("Discrepancy" Category)

As Shinozaki et al. (2025) pointed out, conventional analyses leave certain ambiguities unresolved. In particular, the answering policy is often unspecified, making it unclear whether responses should

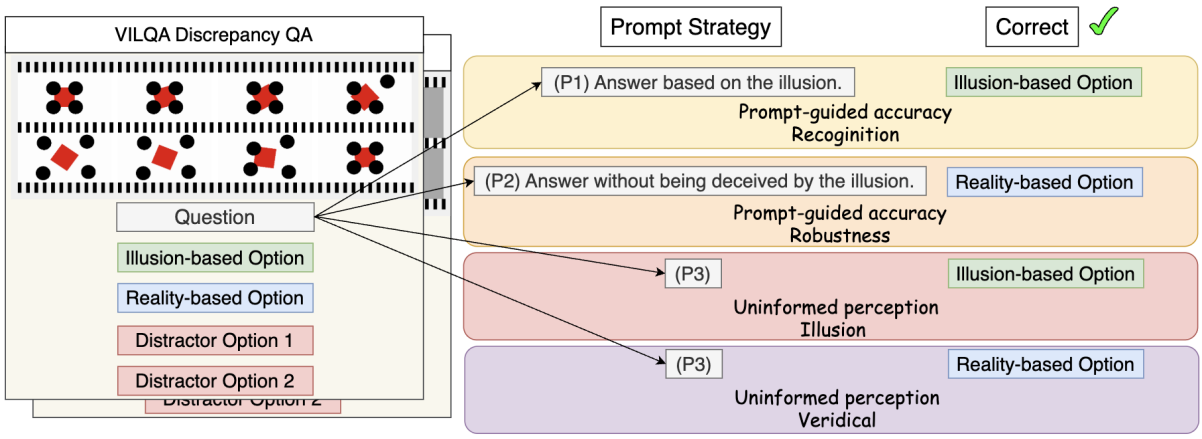

Figure 3: Overview of prompt-guided and uninformed perception evaluations in the VILQA "Discrepancy" category.

Prompt Type (Evaluation Objective)	Prompt Template
P1 (Illusion recognition)	{Question} Answer based on the illusion.
P2 (Illusion robustness)	{Question} Answer without being deceived by the illusion.
P3 (Ilusion perception or Veridical perception)	{Question}

Table 4: Prompt types and templates used for evaluating the "Discrepancy" category in VILQA.

be grounded in visual illusions or in physical reality. Such ambiguity compromises the consistency and interpretability of model evaluations. They attempted to address this issue by introducing two distinct prompts, employing terms such as "actual" and "apparent." However, these terms alone do not fully resolve the ambiguity regarding whether answers should be based on perceptual illusions or objective physical properties.

To resolve this issue, we introduce three types of simple instruction prompts that explicitly specify the intended answering policy and eliminate ambiguity. The following prompts are attached to the actual questions and used during evaluation, as shown in the red text in Figure 1, Table 4 and Figure 3.

- **(P1) Answer based on the illusion:** Assesses whether the model exhibits illusion-based perception similar to that of humans, providing insight into the alignment of the model's perceptual tendencies with human cognition.
- **(P2) Answer without being deceived by the illusion:** Evaluates the model's ability to accurately perceive reality, which is critical for real-world applications where avoiding misrecognition is essential.
- **(P3) No instruction:** Observes the model's natural response tendencies in ambiguous or unconstrained scenarios, offering a baseline for understanding its inherent behavior when confronted with illusions.

These prompts establish a clear framework that removes ambiguity and enables precise, multi-perspective analysis of model perception.

3.4. Contamination Filtering

Given web-scale training of video LLMs, VILQA may include contamination. Following (Shahgir et al., 2024), we pseudo-classified videos with Gemini 2.5 Pro (prompt: "Please describe this video.") and flagged those with detailed, accurate identifications.

As a result, out of 275 illusion videos, 161 were judged to be free from the risk of contamination. We release all 275 videos and the corresponding QA pairs; however, the following main accuracy result is reported on this filtered subset (On the full dataset, we observed increases in prompt-guided and anomaly accuracy across models).

4. Experimental Settings

4.1. Evaluation Prompt Strategy

As explained in the previous section and summarized in Table 4, we evaluate video LLMs using three distinct instruction prompt strategies. The metrics are defined as follows:

4.1.1. Prompt-guided Evaluation

For this evaluation, we use only videos from the "Discrepancy" category of VILQA.

- Illusion recognition: This is the accuracy rate when the prompt (P1: "Answer based on the illusion") is given. It is the percentage of responses interpreted as I1: "The perceptual interpretation typically experienced by most people under the illusion."

- Illusion robustness: This is the accuracy rate when the prompt (P2: "Answer without being deceived by the illusion") is given. It is the percentage of responses interpreted as I2: "The perceptual interpretation when the illusion does not influence perception."

4.1.2. Uninformed Perception Evaluation

For this evaluation, we also use only videos from the "Discrepancy" category of VILQA, testing the baseline condition without any additional prompt (P3).

- Illusion perception: The percentage of responses interpreted as I1: "Perceptual interpretation influenced by the illusion."

- Veridical perception: The percentage of responses interpreted as I2: "Perceptual interpretation not influenced by the illusion."

4.1.3. Anomaly Evaluation

For this evaluation, we use only videos from the "Anomaly" category of VILQA. Unlike the "Discrepancy" category, we directly use a fixed question—"What is unusual about this video from the perspective of an optical illusion?"—without providing any additional instruction prompts.

- Anomaly detection accuracy: This refers to the accuracy rate determined by selecting the correct answer (the anomalous, rare or surprising event shown in the video) along with three plausible but incorrect alternatives.

4.2. Human Performance Evaluation

We recruited nine independent annotators (not involved in the original QA) and collected three responses per item on a random sample of 100 VILQA instances (66 "Discrepancy", 34 "Anomaly"). For each question, the human label was the majority vote of the three.

4.3. Models

4.3.1. Video LLMs

In this study, we evaluated a total of eight video LLMs, including LLaVA-Video-Qwen2 (Zhang et al., 2025), Qwen2.5-VL-Instruct (Bai et al., 2025), Video-R1 (Feng et al., 2025), and Gemini 2.5 Pro (Google, 2025), among others.

Video-R1 is a variant of Qwen2.5-VL-Instruct fine-tuned using a reinforcement learning method called GRPO, allowing us to compare how additional training influence the perception. For inference, we adopted the official default generation parameters for each model.

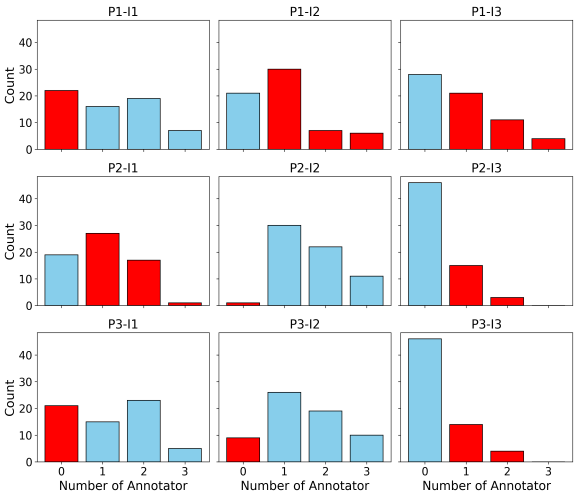

Figure 4: The number of annotators selecting each option under three instruction prompt settings ("Discrepancy").

4.3.2. Fine-tuning

For first-person perspective fine-tuning, we used 23,734 QA pairs from the EgoTaskQA[5] training set. Because Ego4D is widely used, we opted for EgoTaskQA from LEMMA, which broadly covers everyday human tasks and supports activity grounding.

We fine-tuned only the ViT and MLP components, which are primarily responsible for the visual processing in the video LLMs. To reduce the risk of overfitting, we fine-tuned each model for a single epoch. Qwen was trained with a batch size of 126, whereas LLaVA was trained with a batch size of 14.

5. Results and Discussion

5.1. Human performance

Figure 4 shows that, under the "Illusion recognition" condition (P1), a larger proportion of participants selected incorrect options I3. Although I2 selections decreased slightly relative to the other

[5]Because answers in EgoTaskQA are often short phrases rather than complete sentences, we used GPT-4.1 to convert them into natural sentences based on the corresponding questions and original answers.

Model	Prompt-guided accuracy Recognition	Robustness	Uninformed perception Illusion	Veridical	Anomaly Accuracy
Open Model					
LLaVA-Video-Qwen2-7B	0.328	0.647	0.277	0.529	0.548
LLaVA-Video-Qwen2-72B	0.555	0.605	0.412	0.454	0.643
Qwen2.5-VL-Instruct-7B	0.370	0.655	0.345	0.479	0.476
Qwen2.5-VL-Instruct-32B	0.420	**0.840**	0.336	0.504	0.595
Video-R1-7B	0.387	0.664	0.269	**0.563**	0.571
InternVL3-8B	0.403	0.597	0.328	0.437	0.429
InternVL3-14B	0.471	0.580	0.345	0.529	0.667
VideoLLaMA3-7B	0.370	0.555	0.319	0.496	0.524
GLM-4.1V-9B	0.420	0.437	0.370	0.429	0.476
First-person Perspective FT Model					
LLaVA-Video-Qwen2-7B (+ft)	0.395	0.622	0.303	0.496	0.524
Qwen2.5-VL-Instruct-7B (+ft)	0.387	0.605	0.378	0.462	0.476
Proprietary Model					
Gemini 2.5 Pro	**0.597**	0.781	**0.454**	0.412	**0.690**
Gemini 2.5 Flash	0.571	0.739	0.437	0.311	0.619
Human (Subset)	0.406	0.516	0.438	0.453	0.889

Table 5: Illusion recognition, illusion resistance and anomaly sensitivity performance of video LLMs and humans on the filtered VILQA.
Recognition: accuracy on "Discrepancy" with P1 (selecting I1); **Robustness**: accuracy on "Discrepancy" with P2 (selecting I2); **Illusion**/**Veridical**: percentage of responses interpreted as I1/I2 under P3 (no instruction); **Anomaly Accuracy**: accuracy on the "Anomaly" category.

prompt conditions, which suggests that the instruction had some effect, the effect appears limited and item-dependent, pointing to challenges in prompt comprehension, thereby hindering task understanding. Moreover, because visual illusions are not perceived uniformly across individuals (Schwarzkopf et al., 2011; Cretenoud et al., 2021) and only three participants completed the tasks, sampling variability may have amplified individual differences; in particular, these three may have exhibited lower sensitivity to the relevant illusions. These observations suggest two complementary remedies: refining the instruction prompt to reduce misinterpretation and, in future studies, expanding the participant pool to mitigate idiosyncratic effects and increase the proportion of respondents who arrive at the correct answer.

Under the "Illusion robustness" condition (P2-I2), at least one participant selected the correct answer in most cases. This indicates that the task is at least partially solvable by humans.

In the case of P3, the proportion of participants who selected I3 was low, and selections were almost evenly split between I1 and I2. This suggests that, when no specific instruction is given, humans do not consistently rely on either illusion-based or reality-based judgments.

5.2. Illusion recognition

Table 5 shows that Gemini 2.5 Pro achieved a relatively high illusion recognition accuracy, significantly outperforming random guessing (0.25). Compared to open models, Gemini 2.5 Pro likely benefits from substantially larger training data and model size, which may contribute to its enhanced performance. Moreover, it also surpassed the average score of human participants in this setting. This contrast suggests that the model selects the most representative illusion-based answers defined in the benchmark. Since these answers are constructed to reflect perceptual interpretations widely observed in the general population, the model's consistent performance could indicate alignment with a more normative human perception, rather than with individual variability.

On the other hand, the illusion recognition accuracies of the open models tested in this study were approximately on par with those of human participants, but noticeably lower than that of Gemini 2.5 Pro. Increasing the model size tended to improve performance. However, the comparison between Qwen2.5-VL-Instruct-7B and Video-R1-7B suggests that post-hoc reinforcement learning did not contribute significantly to score improvements.

Both humans and LLaVA are similarly deceived by certain illusions, such as color aftereffects and hybrid images. In contrast, some illusions appear

to affect only humans. For example, the breathing square, the blur aftereffect, the Pinna illusion, and the lilac chaser. Conversely, illusions that only the model recognized include motion-induced blindness and the Enigma illusion and etc. The model tended to recognize illusions in which static images appear to move more effectively than humans. Moreover, trick art is often easy to see through when looked at carefully, which can make humans confused about what the instruction prompt is actually trying to convey. As a result, they might miss the intended message and go with a reality-based answer instead.

5.3. Illusion robustness

With regard to the illusion robustness accuracy, the models generally outperformed human participants. This may be because the task is inherently difficult for humans who are susceptible to illusions: even when they are consciously aware of the illusion, they may not be able to suppress its influence. In contrast, models may be able to disregard such perceptual bias more systematically. Therefore, such settings may be considered closer to general VQA scenarios for the model.

Model size comparisons show that larger capacity does not necessarily improve illusion robustness. The LLaVA-Video-Qwen2-7B vs. Video-R1-7B comparison suggests that reinforcement learning may improve the ability to perceive reality without being misled by illusions.

In particular, the 7B-scale models, such as LLaVA, tend to perform well when tested on fully synthetic illusion videos. In contrast, they generally underperform on illusions that involve real-world footage. This is likely because real-world illusion videos feature high contextual variability, dynamic backgrounds, and multiple interacting objects, all of which impose a greater cognitive load on the model than fully synthetic videos. As with general VQA tasks, this effect is particularly pronounced in smaller-scale models, which appear to be more sensitive to such complexity and often show reduced robustness in these settings.

5.4. Uninformed perception

Under the "uninformed perception" condition, illusion perception and veridical perception were nearly evenly distributed, with a slight bias toward the latter. This observation suggests that, in the absence of explicit instruction, video LLMs may produce outputs that reflect either illusion-driven interpretations or objective, reality-based judgments. Therefore, it is important to be aware that both tendencies can coexist in their behavior during open-ended interactions.

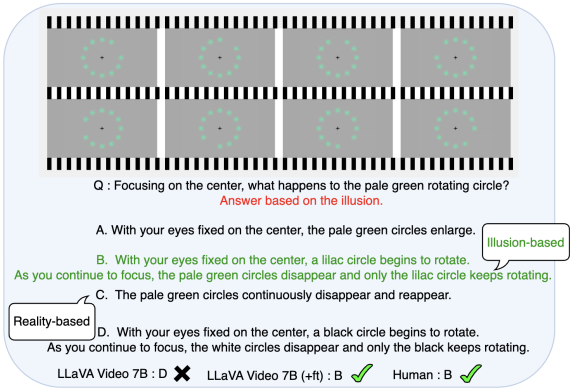

Figure 5: A example showing that first-person fine-tuning improves the alignment of the video LLM with human illusion-based perception in VILQA.

In addition, the performance differences between illusion perception and the illusion recognition accuracy, as well as between veridical perception and the illusion robustness accuracy, indicate that the instructional prompts used in this study were effective in guiding the models' behavior.

5.5. Anomaly sensitivity

In the evaluation of the "Anomaly" category, which captures perceptual violations often linked to emotional salience, the models performed worse than human participants. However, increasing model size tended to narrow the gap, with larger models approaching human-level performance. Although LLaVA struggled with these cases, humans successfully identified the anomalies, including videos where objects appear to differ in color or size but are identical, or the well-known dancer illusion that appears to be rotating in both directions.

5.6. Alignment with human perception

As mentioned earlier, our study primarily focuses on evaluating existing models. However, motivated by the future potential of video LLMs as computational models of human cognition, we additionally conducted first-person perspective fine-tuning to promote more human-aligned behavior. In this subsection, we report the results.

As shown in Table 5, illusion recognition scores improved under prompt-guided evaluation, and both prompt-guided and uninformed perception accuracies shifted toward human performance levels. These results support the hypothesis that first-person perspective natural videos, which closely mirror everyday human visual experience, may help the model acquire human-like perceptual characteristics.

Specific illusions that became recognizable after fine-tuning include the motion aftereffect, as shown

in Figure 1, as well as the lilac chaser illusion with a pale green color, as shown in Figure 5, both of which are illusions readily perceived by humans.

Overall, these findings suggest that first-person perspective videos are likely to be critical for developing video LLMs with more human-like perceptual capabilities.

6. Conclusion

We constructed a dataset to investigate how video LLMs perceive visual illusions. Our analysis indicates that some models appeared to recognize illusions in a human-like manner based on their output, and most models seemed to exhibit human-like emotional responses. Furthermore, we found that using instructional prompts enabled certain models to resist illusions even more effectively than humans. In addition, our results suggest that fine-tuning on first-person perspective videos may be an important ingredient for improving alignment between the model and human perception in the context of visual illusions.

7. Limitation

To more conclusively demonstrate the effectiveness of fine-tuning with first-person perspective videos, it would be ideal to conduct a comparative study using third-person perspective videos fine-tuned on an equivalent task. However, to the best of our knowledge, there is currently no third-person video dataset that matches the task setting of EgoTask closely enough to enable a fair comparison. We therefore leave this ablation as future work.

Furthermore, regarding the human performance results in this study, only three participants completed each prompt condition. This sample size is insufficient to adequately reduce response variance. Increasing the number of participants will be pursued in future work.

For deeper analysis, it will be essential to determine whether models truly experience illusions by probing their internal representations, including where and how they allocate attention to visual information.

Acknowledgements

This work is partly supported by JST, PRESTO Grant Number JPMJPR2366, Japan. We also acknowledge the support of the Google Cloud credits provided through the Google Gemma Academic Program.

8. Bibliographical References

Helen E Clark, John A Perrone, and Robert B Isler. 2013. An illusory size–speed bias and railway crossing collisions. *Accident Analysis & Prevention*, 55:226–231.

Aline F Cretenoud, Lukasz Grzeczkowski, Marina Kunchulia, and Michael H Herzog. 2021. Individual differences in the perception of visual illusions are stable across eyes, time, and measurement methods. *Journal of vision*, 21(5):26–26.

R. H. Day. 1984. The nature of, perceptual illusions. *Interdisciplinary Science Reviews*, 9(1):47–58.

Bo Dong, Airui Chen, Yuting Zhang, Yangyang Zhang, Ming Zhang, and Tianyang Zhang. 2021. The foggy effect of egocentric distance in a nonverbal paradigm. *Scientific Reports*, 11(1):14398.

Vebjørn Ekroll, Mats Svalebjørg, Angelo Pirrone, Gisela Böhm, Sebastian Jentschke, Rob van Lier, Johan Wagemans, and Alena Høye. 2021. The illusion of absence: how a common feature of magic shows can explain a class of road accidents. *Cognitive research: principles and implications*, 6:1–16.

Alexander Gómez-Villa, Adrián Martín, Javier Vazquez-Corral, and Marcelo Bertalmío. 2019. Convolutional neural networks can be deceived by visual illusions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12309–12317. Computer Vision Foundation / IEEE.

Alexander Gomez-Villa, Adrian Martín, Javier Vazquez-Corral, Marcelo Bertalmío, and Jesús Malo. 2020. Color illusions also deceive cnns for low-level vision tasks: Analysis and implications. *Vision Research*, 176:156–174.

Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, and Michael C. Mozer. 2021. Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, 4(3):251–263.

Taisuke Kobayashi, Akiyoshi Kitaoka, Manabu Kosaka, Kenta Tanaka, and Eiji Watanabe. 2022. Motion illusion-like patterns extracted from photo and art images using predictive deep neural networks. *Scientific Reports*, 12(1):3893.

George Mather, Frans Verstraten, and SM Anstis. 1998. *The motion aftereffect: A modern perspective*. MIT Press.

Donald A Redelmeier and Sheharyar Raza. 2018. Optical illusions and life-threatening traffic crashes: A perspective on aerial perspective. *Medical hypotheses*, 114:23–27.

D Samuel Schwarzkopf, Chen Song, and Geraint Rees. 2011. The surface area of human v1 predicts the subjective experience of object size. *Nature neuroscience*, 14(1):28–30.

Taiga Shinozaki, Tomoki Doi, Amane Watahiki, Satoshi Nishida, and Hitomi Yanaka. 2025. Do large vision-language models distinguish between the actual and apparent features of illusions? In *Proceedings of the 47th Annual Meeting of the Cognitive Science Society, CogSci 2025*. Cognitive Science Society.

Eric D. Sun and Ron Dekel. 2021. ImageNet-trained deep neural networks exhibit illusion-like response to the Scintillating grid. *Journal of Vision*, 21(11):15.

Dejan Todorović. 2020. What are visual illusions? *Perception*, 49(11):1128–1199.

Eiji Watanabe, Akiyoshi Kitaoka, Kiwako Sakamoto, Masaki Yasugi, and Kenta Tanaka. 2018. Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in psychology*, 9:340023.

Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. 2023. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5718–5728. Association for Computational Linguistics.

9. Language Resource References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL technical report.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-R1: Reinforcing video reasoning in MLLMs.

Google. 2025. Gemini 2.5 pro. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro. Accessed: 2025-6-28.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14375–14385. IEEE.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12585–12602. Association for Computational Linguistics.

Mohammadmostafa Rostamkhani, Baktash Ansari, Hoorieh Sabzevari, Farzan Rahmani, and Sauleh Eetemadi. 2025. Illusory VQA: benchmarking and enhancing multimodal models on visual illusions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2025, Nashville, TN, USA, June 11-15, 2025*, pages 2995–3004. Computer Vision Foundation / IEEE.

Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. 2024. IllusionVQA: A challenging optical illusion dataset for vision language models. CONFERENCE ON LANGUAGE MODELING 2024, COLM 2024.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. 2024. InternVideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, volume 15143 of *Lecture Notes in Computer Science*, pages 396–416. Springer.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*.