Mementos: A Comprehensive Benchmark for Multimodal Large Language Model Reasoning over Image Sequences

Xiyao Wang[1†] **Yuhang Zhou**[1] **Xiaoyu Liu**[1]
Hongjin Lu[1] **Yuancheng Xu**[1] **Feihong He**[2] **Jaehong Yoon**[2] **Taixi Lu**[2] **Fuxiao Liu**[1]
Gedas Bertasius[2] **Mohit Bansal**[2] **Huaxiu Yao**[2‡] **Furong Huang**[1‡]
[1]University of Maryland, College Park
[2]UNC-Chapel Hill, Chapel Hill
[†]xywang@umd.edu [‡] Equal advising

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated proficiency in handling a variety of visual-language tasks. However, current MLLM benchmarks are predominantly designed to evaluate reasoning based on static information about a single image, and the ability of modern MLLMs to extrapolate from image sequences, which is essential for understanding our ever-changing world, has been less investigated. To address this challenge, this paper introduces Mementos, a new benchmark designed to assess MLLMs' sequential image reasoning abilities. Mementos features 4,761 diverse image sequences with varying lengths. We also employ a GPT-4 assisted method to evaluate MLLM reasoning performance. Through a careful evaluation of nine recent MLLMs on Mementos, including GPT-4V and Gemini, we find that they struggle to accurately describe dynamic information about given image sequences, often leading to hallucinations/misrepresentations of objects and their corresponding behaviors. Our quantitative analysis and case studies identify three key factors impacting MLLMs' sequential image reasoning: the correlation between object and behavioral hallucinations, the influence of co-occurring behaviors, and the compounding impact of behavioral hallucinations.

1 Introduction

The recent emergence of Multimodal Large Language Models (MLLMs) such as GPT-4V (OpenAI, 2023b) and Gemini (Team, 2023) has shown strong visual-language understanding and generation capabilities in many areas, like image captioning and visual question answering. Despite the notable performance of existing MLLMs, they often suffer from hallucination (a phenomenon where MLLMs produce inaccurate descriptions of the given images) due to insufficient reasoning capabilities, generating inaccurate responses in visual inference (Liu et al., 2023a; Yue et al., 2023). Thus, monitoring the reasoning capability is of great importance in understanding the ability and the limitations of MLLMs and applying MLLMs in the real world. Previous benchmarks, such as Liu et al. (2023a) and Yue et al. (2023), have primarily addressed evaluating reasoning in each individual image, relying on static and object-centric knowledge. However, they are insufficient to comprehensively assess the reasoning capabilities of MLLMs due to a lack of time-varying object behaviors or events.

To investigate the capabilities of Multi-Modal Language Models (MLLMs) in dynamic reasoning across image sequences, we present a new benchmark, *Mementos*. This benchmark focuses on the complex task of monitoring and deciphering the *positional changes* of objects within an image sequence, followed by the inference of behavioral patterns and logical connections among them. Such an endeavor requires the interpretation of the overarching context based on time-variant visual elements, posing a greater challenge than the analysis of static scenes. Concretely, Mementos consists of 4,761 image sequences with varying episode lengths, encompassing diverse scenarios from everyday life, robotics tasks, and comic-style storyboards. An episode refers to a specific event or series of events depicted in the image sequence. Each sequence is paired with a human-annotated description of the key objects and their behaviors.

To assess the reasoning capability of MLLMs on Mementos, we employ a *GPT-4*-assisted evaluation procedure: after an MLLM produces a description for an image sequence, we extract behavior and object keywords from both AI-generated and human-annotated descriptions using GPT-4. We then use keyword matching to assess the degree of behavioral and object hallucinations. To refine the correctness of this evaluation, we have developed behavior and object synonym graphs for each domain. These graphs facilitate more precise key-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 416–442

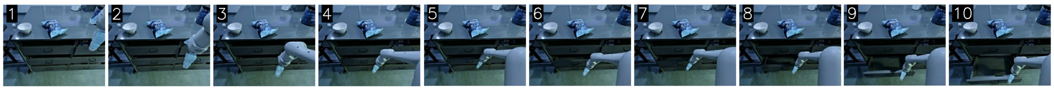
[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

Daily-life domain

[GPT-4V]: *In the sequence of images provided, we observe two men engaged in what appears to be a playful and humorous interaction on a set of outdoor stairs leading to a building entrance. In the first image, one man is seen holding a long, flexible object, possibly a strip of metal or rubber, and appears to be playfully threatening to strike the other man, who is ascending the stairs and holding a similar object. The second image captures a moment of anticipation, with the first man poised to 'attack' and the second man looking back with a smile, suggesting a friendly and joking atmosphere. The third image shows the first man in motion, as if he has just attempted to playfully hit the second man, who is now halfway through the doorway, still participating in the jest. In the final image, the second man has turned around completely, facing the first man with a grin, indicating that the mock confrontation has ended amicably, and both men seem to be enjoying the interaction. The overall mood is light-hearted, and the episode seems to be a moment of camaraderie and fun between the two individuals.*

[Human annotation]: *The sequence of images depicts three men interacting on an outdoor staircase. The man in brown is ascending the stairs slowly, holding onto the railings with both hands, while engaging in conversation with the man in black, who is attentively watching him and holding onto the railing as well. During the ascent, the man in brown turns back to continue his dialogue with the man at the bottom. Meanwhile, the man at the door observes the interaction between the two.*

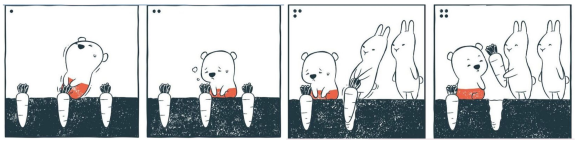
Robotics domain



[GPT-4V]: *The sequence of images depicts a robotic arm interacting with a drawer in a piece of furniture. Initially, the drawer is closed, and the robotic arm is positioned near it. As the sequence progresses, the arm extends towards the drawer and appears to grasp the handle. It then pulls the drawer open, revealing the contents inside. The images capture the motion of the robotic arm as it performs the task of opening the drawer, showcasing the arm's range of motion and its ability to interact with objects in its environment.*

[Human annotation]: *In this picture, there is a table with an empty bowl, a packet of snacks, and a blue can on it. The table has many closed drawers. A robotic arm has located the drawer in the first column and second row and has pulled it open.*

Comics domain



[GPT-4V]: *The image is a four-panel comic featuring a bear and rabbits. In the first panel, the bear is gleefully eating a large red apple while sitting in a hole with three carrots. In the second panel, the bear appears sad and contemplative, staring at the apple core, having finished the apple. In the third panel, three rabbits appear next to the bear, only their heads visible above the ground, mimicking the look of the carrots. In the final panel, the bear is happily eating a carrot, seemingly content with the new option, while the rabbits, now fully visible and standing in the hole, look on with one rabbit patting the bear's back in a comforting gesture. The comic conveys a message of companionship and finding contentment in what is available.*

[Human annotation]: *A little bear wearing red pants is sitting on the ground, trying to pull a carrot out of a hole, but he fails and sits on the ground, sighing in dismay. At this moment, two rabbits come over, and one of the rabbits pulls a carrot out of the hole and gives it to the little bear. The little bear looks very happy.*

Figure 1: Examples of hallucinations by GPT-4V in three domains on Mementos. The red box shows the description generated by GPT-4V based on the given prompt, and the human-annotated descriptions are in the blue box. Texts highlighted in yellow are hallucination parts generated by GPT-4V. This illustrates that even GPT-4V experiences severe hallucinations when reasoning from image sequences.

word matching, ensuring a thorough and nuanced analysis of the MLLMs' reasoning abilities. Besides, we also provide the comparison with human evaluation to demonstrates that the GPT-4-assisted evaluation procedure is very reliable.

We evaluated the reasoning proficiency of *nine leading-edge MLLMs* on Mementos, encompassing both black-box and open-source models. Our findings indicate that Mementos poses a considerable challenge to these current MLLMs. For instance, as depicted in Figure 1, GPT-4V exhibits notable behavioral and object hallucinations in various domains during image sequence reasoning. Behavioral hallucinations are defined as the MLLMs' erroneous interpretations or predictions of entity actions, while object hallucinations pertain to the inaccurate identification or creation of objects. Notably, behavioral hallucinations were more frequent than object hallucinations, highlighting a significant deficiency in MLLMs' capability to deduce events from image sequences.

Furthermore, our research pinpoints three principal factors that lead to the reasoning failures of MLLMs: (1) the interconnectedness of object and behavioral hallucinations, (2) the impact of co-occurring behaviors, and (3) the cumulative effect of behavioral hallucinations. The objective of our proposed benchmark and analyses is to shed light on innovative approaches to augment the reasoning abilities of MLLMs and to reduce hallucinations in their subsequent advancements.

2 Mementos

In this section, we introduce Mementos, a novel and challenging benchmark designed to test the reasoning capability of Multimodal Large Language

Model (MLLM) under sequential image input. Initially, we detail the data gathering and annotation methodology for Mementos, alongside an overview of data distribution. Subsequently, we outline the procedure and the metric employed to evaluate the reasoning capabilities of MLLMs on Mementos.

2.1 Mementos Benchmark

2.1.1 Dataset Composition

Mementos comprises 4,761 image sequences of varying lengths, predominantly sourced from Daily-life, Robotics, and Comics domains. Detailed statistics are provided in Table 1. This diverse collection is instrumental in evaluating the comprehensive time-varying reasoning abilities of MLLMs. Specifically, the robotics data, closely associated with embodied AI or real-world contexts, and the comic-style storyboard data, rich in stylistic and episodic diversity in image sequences, significantly enhance the benchmark's relevance and robustness.

Table 1: The number of image sequences in different categories within Mementos.

	Total	Train Set	Val set
Daily-life	3505	3055	450
Robotics	1101	902	199
Comics	155	105	50

Daily-life The Daily-life image sequences in Mementos are derived from video clips in the Next-QA dataset, as cited in Xiao et al. (2021). These sequences represent a range of everyday life scenarios. We have selectively extracted videos from the Next-QA Training set, specifically those with frame counts ranging from 400 to 2,500. To balance the challenge of testing MLLMs' reasoning capabilities against the risk of losing critical information, our methodology involves retaining the first frame of each video. Subsequently, we sample one image every 100 frames. The collected images from this sampling process then form an image sequence that corresponds to the original video. This approach ensures a rigorous yet feasible evaluation of MLLMs' reasoning abilities in dynamically evolving everyday scenarios.

Robotics For the Robotics data, we utilized videos from various sub-datasets within Open X-Embodiment (Collaboration et al., 2023). Open X-Embodiment aggregates video datasets from multiple university laboratories, showcasing a variety of tasks performed by different robotic systems. We meticulously selected sub-datasets from Open X-Embodiment that offer video resolutions exceeding 128x128 and exhibit a high degree of task diversity. From these chosen sub-datasets, a total of 1,101 videos were sampled. The precise number of videos sourced from each sub-dataset is detailed in Appendix A. For video sampling, our approach varied based on the length of the videos. Videos exceeding 100 frames were processed by sampling one image every $n/20$ frames, where n represents the total frame count. Conversely, for videos with frame counts ranging from 20 to 100, we sampled one image every 5 frames. This ensures the formation of comprehensive and representative image sequences for each video, catering to the evaluation of MLLMs in diverse and complex robotic contexts.

Comics The Comics data is composed of wordless multi-panel comics of diverse styles, curated from online sources. Unlike Daily-life and Robotics sections, where image sequences are uniformly extracted from video frames, the comics represent intentionally selected key moments within a narrative, manually illustrated by artists. This distinction sets our dataset apart from conventional video datasets. In addition to traditional comics, this category also incorporates 20 storyboards from movies reimagined in comic style. We have further deconstructed these comics into individual image sequences by taking screenshots. This approach enables a unique exploration of sequential visual reasoning, enhancing the diversity and complexity of the dataset for evaluating MLLMs.

2.1.2 Dataset Annotation

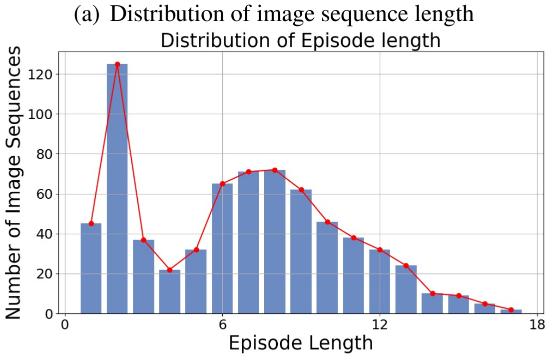
For each image sequence in Mementos, we have meticulously annotated a ground truth description that captures the unfolding events. These descriptions focus on the primary objects and their respective behaviors, where *behavior* refers to a verb or verb phrase associated with the object in question.

For the Daily-life data, we initially employed GPT-4V(ision) (OpenAI, 2023a), to amalgamate and reformulate the questions and answers from the Next-QA videos into single paragraph descriptions. This method significantly expedited the manual annotation process. Following this, we conducted a thorough manual review of these automated descriptions, making necessary adjustments. This included rectifying inaccuracies, removing non-existent episodes, and adding missing details to ensure alignment with the actual image sequences. To ensure reliability, we implemented a cross-validation step, where a separate set of annotators performed a secondary review. For the Robotics and Comics categories, the annotation process was entirely manual, conducted by human annotators. These annotations were then subjected to a verification process by the authors which ensures the accuracy and consistency of the descriptions across all categories.

2.1.3 Dataset Statistics

In showcasing the diversity of Mementos, we present a detailed overview of the data distribution within the Mementos validation set. Our analysis focuses on two key dimensions: the length of the image sequence and the length of the episode. The length of an image sequence is defined by the number of frames it contains, while the episode length is determined by the total number of events depicted in the sequence. A longer image sequence necessitates the MLLM to process a larger number of images, thereby challenging the model's capacity to manage sequences spanning broader timeframes. A greater episode length signifies that the image sequence encompasses more intricate scenarios.

(a) Distribution of image sequence length



(b) Distribution of episode length

Figure 2: Data distribution in Mementos Val set.

Image sequence length For the image sequence length, we count the number of frames in each image sequence. As shown in Figure 2(a), the majority of image sequences are between 4 and 14 frames in length. 67.38% of image sequences contain 4 to 14 frames, yet 31.90% of sequences are composed of longer frames - more than 15 frames.

Episode length To quantify the episode length within each image sequence of Mementos, we employed GPT-4 for extracting behavior keywords, specifically verbs associated with objects, from the human-annotated descriptions. This extraction was facilitated using a pre-defined manual prompt, details of which can be found in Appendix D. Following the extraction, we calculated the length of the behavior list for each image sequence. A lengthier behavior list signifies a more extended episode within the image sequence, which inherently poses a greater challenge for the MLLM in comprehending the entire image sequence. As illustrated in Figure 2(b), a significant portion of the image sequences, particularly those from the robotics data, feature episode lengths ranging between 1 and 3. This is mainly attributed to the dominance of two-action episodes like 'pick up and place', 'move and pull open', 'locate and push'. Meanwhile, the remaining data exhibits a normal distribution for episode lengths spanning 4 to 17.

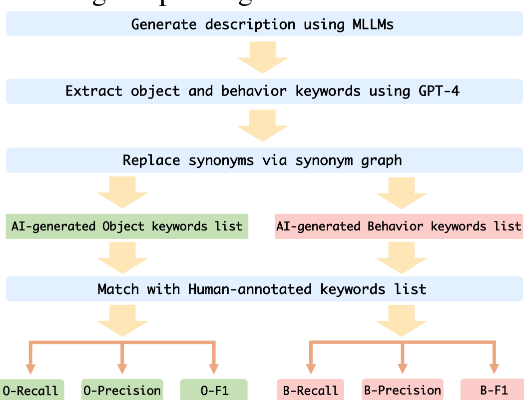


Figure 3: GPT-4-assisted evaluation procedure. We use "O-" for objects and "B-" for behaviors.

2.2 Evaluation Procedure and Metrics

In this section, we illustrate how to evaluate the descriptions generated by MLLMs, including the evaluation procedure and metrics.

Procedure As shown in Figure 3, we use an image sequence and a pre-designed prompt together as the input for MLLMs, and generate the description aligned with the corresponding image sequence. Next, we ask GPT-4 to extract object and behavior keywords in the AI-generated description. We then match the obtained keywords with the *synonym graph* we built, replacing the matched keywords with the root word from the synonym graph. Finally, we obtain two lists of keywords: AI-generated object list and AI-generated behavior list. We note that the proposed keyword extraction leveraging GPT-4 is surprisingly reliable and accurate, which is competitive with human extraction. Please refer to Appendix C for more details.

Synonym graph The synonym graph is an unilateral digraph where each edge connects two nodes representing words or phrases. For instance, given a synonym pair (pick up, lift up), an edge is directed from 'lift up' to 'pick up'. In each synonym pair, the first word, originating from the human-annotated keyword list, is referred to as the root word, while the second word is from the AI-generated keyword. To construct this synonym graph, we use GPT-4 to extract object and behavior keywords from all human-annotated descriptions in the Val set, forming a human-annotated keyword list. Then, we generate descriptions using GPT-4V, LLAVA, and Gemini and use GPT-4 to extract object and behavior keywords. After that, we manually match these words with the human-annotated keyword list to identify all synonym pairs and add them as edges to the synonym graph. Given a word or phrase, this synonym graph can quickly match the corresponding root word if a synonym exists in the human-annotated keyword list, completing the keyword replacement. For convenience in evaluation, we maintain separate synonym graphs for objects and behaviors of different categories. We make all constructed synonym graphs publicly available as open-source resources.

Metrics After obtaining the AI-generated object list and behavior list, we utilize the corresponding human-annotated object list and human-annotated behavior list as the ground truth to calculate 'Recall,' 'Precision,' and 'F1 metrics' at both the object and behavior levels. These metrics are used to measure the understanding capabilities regarding the image sequence episode. 'Recall' reflects the accuracy of an MLLM's reasoning about episodes in an image sequence, while 'precision' focuses on assessing the severity of hallucinations that occur when understanding the image sequence.

3 Experiments

In our experimental section, we delve into two key questions: (a) We examine the reasoning **capabilities** of current MLLMs on Mementos. Specifically, we assess the **severity** of object and behavioral hallucinations in these models. (b) We investigate the underlying **causes** of reasoning failures in MLLMs when interpreting image sequences.

3.1 Baseline evaluation

3.1.1 Models

We establish our baseline using 9 popular MLLMs. The black-box MLLMs include GPT-4V (OpenAI, 2023a) and Gemini (Team, 2023), and the open-source MLLMs are Video-LLaMA-2 (Zhang et al., 2023a), Chat-UniVi (Jin et al., 2023), LLaVA-1.5 (Liu et al., 2023c), MiniGPT4 (Zhu et al., 2023), MiniGPT5 (Zheng et al., 2023), mPLUG_Owl-v2 (Ye et al., 2023), and InstructBLIP (Dai et al., 2023). Considering that only a few open-source MLLMs are designed to process sequential images (Video-LLaMA-2 and Chat-UniVi), we adapt input for other models by combining all frames from an image sequence into one composite image, referred to as the combined-input (c-input) setting. For black-box MLLMs and Chat-UniVi, we conduct evaluations using both the c-input and an alternative approach where frames from the image sequence are input sequentially, termed the sequential-input (s-input) setting. For Video-LLaMA-2, we only test in s-input setting.

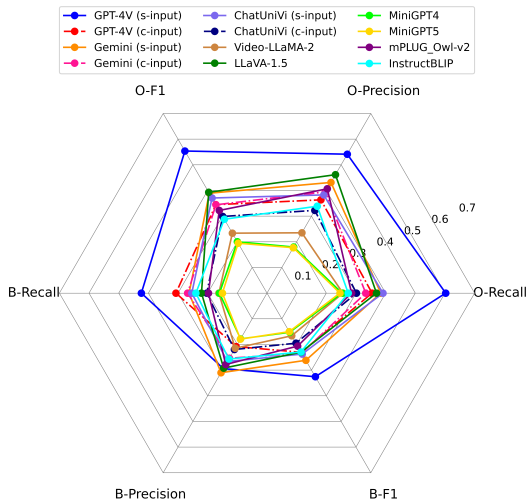


Figure 4: Comparison of metrics for different MLLMs.

3.1.2 Evaluation results

We evaluate all MLLMs on Mementos and report the results in Figure 4. Besides, we provide the performance of each baseline method in three different domains (Daily-life, Robotics, and Comics) in Table 2. We summarize our findings as follows: **GPT-4V (s-input) and LLaVA-1.5 are the best-performing models among black-box and open-source MLLMs, respectively.** As shown in Figure 4, except for being on par with Gemini (s-input) and LLaVA-1.5 in behavior precision, GPT-4V with s-input demonstrates the best reasoning

Table 2: Evaluation of different MLLMs on Mementos.

Domain	Input type	Model	Object Recall	Object Precision	Object F1	Behavior Recall	Behavior Precision	Behavior F1
Daily-life	Sequential	GPT-4V	**59.80%**	**50.96%**	**53.51%**	**36.71%**	32.97%	**33.59%**
		Gemini	35.92%	42.06%	37.10%	18.80%	29.42%	21.64%
		Video-LLaMA-2	31.59%	30.01%	29.37%	17.05%	28.19%	20.12%
		Chat-UniVi	40.74%	40.78%	39.13%	22.30%	31.10%	24.90%
	Combined	GPT-4V	39.45%	39.64%	38.04%	26.43%	23.59%	23.98%
		Gemini	31.17%	37.39%	32.38%	17.71%	25.65%	19.74%
		Chat-UniVi	36.19%	38.88%	36.02%	21.80%	28.52%	23.73%
		LLaVa-1.5	37.72%	47.01%	40.18%	22.17%	**37.33%**	26.65%
		MiniGPT4	32.25%	23.14%	25.75%	18.09%	24.16%	19.45%
		MiniGPT5	31.39%	22.62%	24.91%	18.42%	24.56%	19.85%
		mPLUG_Owl-v2	32.59%	47.17%	37.04%	17.96%	33.57%	22.13%
		InstructBLIP	31.82%	41,14%	34.28%	22.40%	30.30%	24.55%
Robotics	Sequential	GPT-4V	**63.94%**	**65.42%**	**62.99%**	**60.72%**	24.43%	33.95%
		Gemini	43.80%	46.26%	43.15%	46.43%	**38.13%**	**39.38%**
		Video-LLaMA-2	13.41%	10.33%	11.15%	17.04%	8.96%	11.23%
		Chat-UniVi	35.40%	32.57%	32.39%	32.24%	16.69%	21.14%
	Combined	GPT-4V	27.87%	31.86%	28.58%	44.72%	16.54%	23.58%
		Gemini	34.78%	41.66%	36.16%	47.29%	29.59%	34.17%
		Chat-UniVi	17.74%	18.32%	17.07%	19.81%	10.01%	12.54%
		LLaVa-1.5	36.88%	46.62%	39.31%	25.27%	14.80%	17.95%
		MiniGPT4	10.97%	7.28%	8.16%	13.40%	5.88%	7.76%
		MiniGPT5	9.75%	6.52%	7.16%	8.96%	4.53%	5.43%
		mPLUG_Owl-v2	19.75%	26.70%	21.99%	26.46%	16.59%	19.51%
		InstructBLIP	17.96%	18.65%	17.29%	31.41%	19.08%	22.69%
Comics	Sequential	GPT-4V	**49.53%**	37.57%	**41.71%**	**19.97%**	17.29%	**18.11%**
		Gemini	38.57%	40.64%	38.53%	15.23%	19.11%	16.30%
		Video-LLaMA-2	20.26%	17.59%	18.09%	5.45%	11.07%	6.81%
		Chat-UniVi	28.04%	31.61%	28.13%	10.42%	15.74%	11.97%
	Combined	GPT-4V	29.23%	24.64%	25.90%	13.19%	13.09%	12.90%
		Gemini	41.25%	**45.07%**	41.18%	15.37%	**20.55%**	16.42%
		Chat-UniVi	25.12%	28.08%	25.51%	8.85%	10.67%	9.31%
		LLaVa-1.5	29.44%	35.61%	30.97%	8.63%	13.56%	10.27%
		MiniGPT4	20.50%	13.94%	15.74%	7.95%	8.64%	7.98%
		MiniGPT5	22.94%	18.11%	19.42%	8.88%	11.92%	9.94%
		mPLUG_Owl-v2	26.82%	37.74%	29.49%	8.70%	20.85%	11.74%
		InstructBLIP	25.02%	29.15%	25.10%	8.25%	10.48%	8.97%

capability compared with all other MLLMs in understanding image sequences. Among open-source models, LLaVA1.5 performs the best, nearly matching or even surpassing the black-box model Gemini in object comprehension, but its ability to infer behaviors from image sequences is weaker compared to Gemini and GPT-4V. Although Video-LLaMA-2 and Chat-UniVi are designed for video understanding, they do not show an advantage over LLaVA-1.5, especially Video-LLaMA-2, which performs notably worse compared to LLaVA-1.5. The weakest models in understanding image sequences are MiniGPT4 and MiniGPT5, with a significant gap in every metric compared to the other baselines. It's noteworthy that under c-input setting, the performance of black-box MLLMs does not significantly differ from that of open-source MLLMs. LLaVA-1.5 and mPLUG_Owl-v2 meet or even exceed the black-box MLLMs on many metrics.

MLLMs possess a much stronger ability on reasoning objects in image sequences than they do on reasoning behaviors. We find that all MLLM methods perform significantly better on the three metrics for objects than those for behaviors. Taking the best-performing GPT-4V as an example, it achieves over 50% on all three object metrics, with recall even reaching 60%, indicating it can effectively recognize the main objects in an image sequence. However, for behaviors, GPT-4V scores only around 30%, with the best recall metric barely exceeding 40%. Despite this, GPT-4V is still the best-performing MLLM in reasoning behaviors. This suggests that current MLLMs do not possess strong abilities to autonomously infer the behaviors from given sequential images, indicating the importance of our benchmark in highlighting the limitations in the reasoning abilities of MLLMs.

Reasoning capability of MLLMs varies across different domains. From Table 2, we find that black-box models perform best in the robotics domain across the three domains, while open-source models show relatively better performance in the daily-life domain. Analyzing each domain specifically, it is evident that in the daily-life domain, the performance of all methods, except for GPT-4V (s-input), does not vary significantly. The main reason for the performance gap between open-source MLLMs and black-box MLLMs is the noticeably lower metrics of open-source models compared to black-box models in the robotics and comics domains. The recall, precision, and F1 of both object and behavior for black-box MLLMs are almost more than double those of open-source models. We speculate that one reason for this phenomenon is the distribution shift between Mementos and the training data of open-source MLLMs. The limitations of the training data lead to weaker reasoning capability of open-source MLLMs.

3.2 Analysis of Failure Reasoning

In this section, we will provide reasons for failure reasoning results in current MLLMs, combining specific quantitative analyses and case studies. Since behavioral hallucination is a unique phenomenon in image sequence reasoning, and the causes of object hallucination are not significantly different from those in single image reasoning, we only present the reasons leading to behavioral hallucination in this paper. Due to space limitations, please refer to the Appendix E for specific case studies. The following are our main findings:

Interplay between object and behavioral hallucinations in MLLMs. A key hypothesis underpinning behavioral hallucination is that incorrect object identification leads to subsequent inaccuracies in behavior identification. To test this, we evaluated the correlation coefficients between object and behavioral hallucinations across different domains for various MLLMs, as detailed in Appendix B Table 4. Our findings reveal that, for most MLLMs, the correlation coefficients in the three domains fluctuate between 0.1 and 0.4, suggesting a weak yet present correlation. This outcome supports the hypothesis that object hallucination contributes to behavioral hallucination to some extent. Case studies further reveal that after an object hallucination occurs, MLLMs tend to describe behaviors related to the hallucinated object, even if these behaviors do not exist in the image sequence. As shown in Figure 5, after recognizing a scene as a tennis court, a MLLM might describe a person playing tennis. Interestingly, in the robotics domain, there is a negligible correlation between object and behavioral hallucinations in black-box MLLMs. This divergence is likely because behaviors in robotics are predominantly linked to robotic arms, which these MLLMs generally identify correctly.

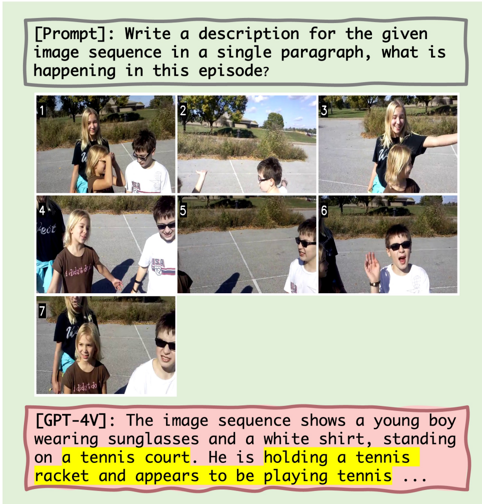
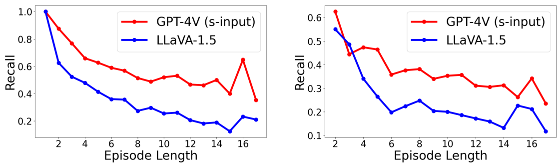


Figure 5: A sample of failure reasoning case in Daily-life domain. The failure reason is object hallucination, correlation between object hallucination and behavioral hallucination, and co-occurrence behavior. Following the object hallucination of ***tennis court***, the LVLM subsequently exhibits behavioral hallucinations of ***holding a tennis racket*** (correlation between object hallucination and behavioral hallucination) and ***appears to be playing tennis*** (co-occurrence behavior).

The impact of co-occurrence on behavioral hallucination. In line with object hallucination phenomena, as noted in Li et al. (2023c) and Zhou et al. (2023a), MLLMs demonstrate a tendency to generate behaviors that are commonly paired together. This proclivity exacerbates the problem of behavioral hallucination, especially in the field of robotics. Consider the case in Figure 1 where a robotic arm is tasked with opening a drawer by ***grabbing its side***. MLLMs might erroneously depict the sequence as the arm ***grabbing the handle*** first, followed by pulling the drawer open, since ***grabbing the handle*** is a more co-occurring behavior with 'pull open'. Despite the final outcome being accurately described, such errors in key details are unacceptable in robotics. This issue is of particular concern given the growing inclination to utilize MLLMs as reward functions in robotic training (Ma et al., 2023; Sontakke et al., 2023; Rocamonde et al., 2023; Baumli et al., 2023). Such behavioral hallucinations can critically affect the quality of the reward function, leading to potential mislearning of behaviors in robotic systems. Detailed case studies are shown in Appendix E.

The Snowball effect in behavioral hallucinations. The Snowball effect is a well-documented phenomenon in machine learning, referring to the progressive accumulation or intensification of errors in a system, as discussed in Asadi et al. (2019); Zhang et al. (2023b); Wang et al. (2023c); Liu et al. (2023d). Zhang et al. (2023b) notably highlight this phenomenon in Large Language Models. Experiments on Mementos reveal that the snowball effect in both behavioral and object hallucinations becomes markedly pronounced when reasoning through image sequences. The temporal nature of image sequences, consisting of a series of frames rather than a solitary image, demands that MLLMs sequentially infer the narrative. This process makes models susceptible to exacerbating hallucinations if errors occur early in the sequence. We specifically examined the trend of object and behavioral hallucination in GPT-4V and LLaVA-1.5 within the daily-life domain, correlating it with the episode length. As shown in Figure 6, there is a noticeable decrease in object and behavior recall for both MLLMs as the episode length extends. This trend suggests a heightened susceptibility to hallucinations and a pronounced snowball effect in MLLMs when processing image sequences with a greater array of objects and behaviors. Detailed case studies can be found in Appendix E.



(a) Object (b) Behavior

Figure 6: The trend of changes in object and behavior recall for GPT-4V and LLaVA-1.5 in the Daily-life domain as the episode length increases.

4 Related work

4.1 Benchmarking in MLLMs

The advent of MLLMs has prompted a reassessment of traditional benchmarks (Lin et al., 2014; Marino et al., 2019; Hudson and Manning, 2019). These benchmarks fail to sufficiently expose the hallucination issues in MLLMs. Consequently, there is a growing impetus to devise more challenging benchmarks. This trend spans various domains, from question and answering (QA) reasoning (Liu et al., 2023a; Yue et al., 2023), to optical character recognition (OCR) (Liu et al., 2023f), and extends to the study of hallucinations (Wang et al., 2023a), with benchmarks such as POPE (Li et al., 2023c) and Bingo (Cui et al., 2023). Additionally, comprehensive analyses of MLLMs, such as Mm-bench (Liu et al., 2023e), Mm-vet(Yu et al., 2023b), LVLM-eHub(Xu et al., 2023), SEED(Li et al., 2023a), GAVIE(Liu et al., 2023b), and LAMM (Yin et al., 2023), are emerging.

Our paper presents a novel benchmark using sequences from videos or comics to study behavioral hallucinations, diverging from single-image analysis. Unlike Chen et al. (2023a)'s vision QA tasks from uniformly sampled video frames, our benchmark challenges MLLMs to describe sequences without question guidance, offering a finer evaluation of hallucinations and reasoning in MLLMs.

4.2 Hallucination in MLLMs

Hallucinations in MLLMs, akin to those in Large Language Models (LLMs) (Zhang et al., 2023c; Li et al., 2023b; Zhou et al., 2024; Chen et al., 2023b), represent a significant challenge. In MLLMs, hallucinations are characterized by inconsistencies between the model's output and the visual content (Rohrbach et al., 2018; Wang et al., 2023a). Recent studies have explored various aspects of hallucination in MLLMs, covering topics such as object hallucination (Li et al., 2023c), hallucination assessment in GPT-4V (Cui et al., 2023), and knowledge hallucination (Liu et al., 2023a).

While there are methods proposed for mitigating hallucinations (Zhou et al., 2023a; Wang et al., 2023b; Leng et al., 2023; Zhou et al., 2023b; Chen et al., 2023c; Jiang et al., 2023; Huang et al., 2023; Yu et al., 2023a; Zhao et al., 2023), there is a noticeable gap in the literature regarding the study of behavioral hallucination. Moreover, the existing work does not offer a dedicated metric for evaluating behavioral hallucinations.

5 Conclusion

In this paper, we present Mementos, a novel and challenging benchmark designed to assess the reasoning abilities of Multimodal Large Language Models (MLLMs) in interpreting image sequences. We conduct evaluations on nine most recent MLLMs using GPT-4-assisted evaluation procedure. Our findings indicate that all tested MLLMs struggle with significant behavioral and object hallucinations in generating descriptions for image sequences. Through a mix of quantitative analysis and case studies, we identify three primary factors contributing to these reasoning failures.

Limitations

Domain courage Mementos is consisted of 4,761 image sequences from three domains: Daily life, Robotics, and Comics. It would be interesting to include a broader variety of data types. This expansion could include first-person navigation experiences, sequential medical CT scans, and interactive gaming data. MLLMs could behave different types of hallucinations in image sequences from other domains

Evaluation Process Our evaluation process focuses on the match of keywords to measure the reasoning ability of MLLMs. However, it would be possible that the MLLM generation is the same as human annotations in semantics but obtains low performance, since the generated tokens are not covered by our synonym graph. Future work could extend the evaluation method to semantic understanding rather than relying predominantly on keyword matching.

Hallucination Mitigation Our work identifies two kinds of hallucination: object and behavioral hallucinations and explore the failure reason of MLLMs. We have not yet proposed a mitigation method to reduce behavioral hallucinations. Future work could utilize the three causes of reasoning failures to bolster the reasoning faculties of MLLMs, making them more adept at accurately interpreting and describing complex image sequences.

Acknowledgement

Wang, Zhou, Liu, Xu, and Huang are supported by National Science Foundation NSF-IIS FAI program, DOD-ONR-Office of Naval Research under award number N00014-22-1-2335, DOD-AFOSR-Air Force Office of Scientific Research under award number FA9550-23-1-0048, Capital One and JP Morgan faculty fellowships. Yao is supported by the Cisco Faculty Research Award. He also thanks Center for AI Safety and Google Cloud Research Credits program for supporting our computing needs. Bansal is supported by DARPA ECOLE Program No. HR00112390060 and ONR Grant N00014-23-1-2356.

References

Kavosh Asadi, Dipendra Misra, Seungchan Kim, and Michel L. Littman. 2019. Combating the compounding-error problem with a multi-step model.

Kate Baumli, Satinder Baveja, Feryal Behbahani, Harris Chan, Gheorghe Comanici, Sebastian Flennerhag, Maxime Gazeau, Kristian Holsheimer, Dan Horgan, Michael Laskin, et al. 2023. Vision-language models as a source of rewards. *arXiv preprint arXiv:2312.09187*.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. 2023a. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *arXiv preprint arXiv:2311.14906*.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023b. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 245–255.

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023c. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*.

Open X-Embodiment Collaboration, Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Hao Su, Hao-Shu Fang, Haochen Shi, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jaehyung Kim, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jiajun Wu, Jialin Wu, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jitendra Malik, Jonathan Tompson, Jonathan Yang, Joseph J. Lim, João Silvério, Junhyek Han, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Zhang, Keyvan Majd, Krishan Rana, Krishnan Srinivasan, Lawrence Yunliang Chen, Lerrel Pinto, Liam Tan, Lionel Ott, Lisa Lee, Masayoshi Tomizuka, Maximilian Du, Michael Ahn, Mingtong Zhang, Mingyu Ding, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Pannag R Sanketi, Paul Wohlhart, Peng Xu, Pierre Sermanet, Priya Sundaresan, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel

Bustamante, Sean Kirmani, Sergey Levine, Sherry Moore, Shikhar Bahl, Shivin Dass, Shuran Song, Sichun Xu, Siddhant Haldar, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Sudeep Dasari, Suneel Belkhale, Takayuki Osa, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xuanlin Li, Yao Lu, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yueh hua Wu, Yujin Tang, Yuke Zhu, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zhuo Xu, and Zichen Jeff Cui. 2023. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2023. Hallucination augmented contrastive learning for multimodal large language model. *arXiv preprint arXiv:2312.06968*.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. 2023. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023c. Improved baselines with visual instruction tuning.

Xiaoyu Liu, Jiaxin Yuan, Bang An, Yuancheng Xu, Yifan Yang, and Furong Huang. 2023d. C-disentanglement: Discovering causally-independent generative factors under an inductive bias of confounder. *arXiv preprint arXiv:2310.17325*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023e. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023f. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Yecheng Jason Ma, William Liang, Vaidehi Som, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. 2023. Liv: Language-image representations and rewards for robotic control. *arXiv preprint arXiv:2306.00958*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Gpt-4v(ision) system card.

Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. 2023. Vision-language models are zero-shot reward models for reinforcement learning. *arXiv preprint arXiv:2310.12921*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Sumedh Anand Sontakke, Jesse Zhang, Séb Arnold, Karl Pertsch, Erdem Biyik, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. 2023. Roboclip: One demonstration is enough to learn robot policies. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023a. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2023b. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. *arXiv preprint arXiv:2312.01701*.

Xiyao Wang, Ruijie Zheng, Yanchao Sun, Ruonan Jia, Wichayaporn Wongkamjan, Huazhe Xu, and Furong Huang. 2023c. Coplanner: Plan to roll out conservatively but to explore optimistically for model-based rl.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. 2023. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2023a. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *arXiv preprint arXiv:2311.13614*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023b. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023b. How language model hallucinations can snowball.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023c. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023. Minigpt-5: Interleaved vision-and-language generation via generative vokens.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023a. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Yuhang Zhou, Suraj Maharjan, and Beiye Liu. 2023b. Scalable prompt generation for semi-supervised learning with language models. *arXiv preprint arXiv:2302.09236*.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2024. Explore spurious correlations at the concept level in language models for text classification.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Details of Open X-Embodiment Data Selection

In this section, we provide the names of all subsets selected from Open X-Embodiment dataset and the corresponding sampling video numbers. For detailed information, please refer to Table 3.

B Correlation Coefficients between Object and Behavioral Hallucinations

In this section, we provide detailed correlation coefficients between object and behavioral hallucinations in Table 4.

C Human Evaluation

In this section, to verify the reliability of the GPT-4 assisted evaluation procedure, we compare the results of GPT-4 assisted evaluation with those of human evaluation. We randomly select 200 image sequences from the entire Val set and manually extract object and behavior keyword lists for each image sequence's AI-generated description and human-annotated description. Then, we calculate six metrics and compare them with the metrics obtained using keyword lists extracted by GPT-4. We choose the four MLLMs that performed best in reasoning on Mementos as representatives: GPT-4V (s-input), Gemini (s-input), Chat-UniVi (s-input), and LLaVA-1.5. The evaluation results are shown in Table 5.

After comparison, we find that there is not a significant gap between the results of GPT-4 assisted evaluation and human evaluation, with the absolute value of the difference mostly ranging between 1% to 4%. For most metrics, the GPT-4 assisted evaluation tends to overestimate the performance of MLLMs, meaning the evaluation results are higher than those of human evaluation. However, the relative ranking among different MLLMs remains essentially unchanged. Overall, the GPT-4 assisted evaluation is quite reliable.

D Prompt Details

In this section, we provide all the prompts used in our paper, including those used to merge questions and answers from Daily-life videos into a single description, prompts for MLLMs to generate descriptions corresponding to image sequences, and prompts for extracting object and behavior keywords from both human-annotated and AI-generated descriptions. The detailed prompts are showm in Table 6.

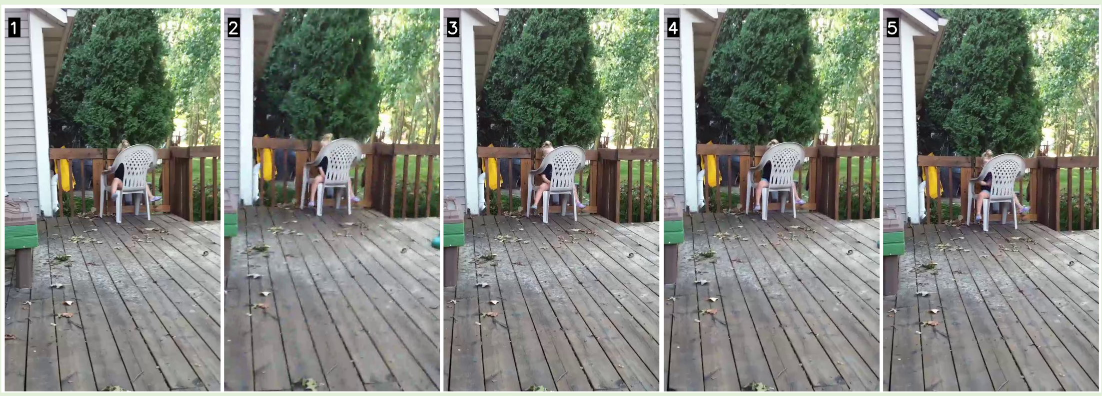
E Case Study

In this section, we present failure reasoning cases of different domains (Figure 7-22), with specific reasons for failure detailed in the captions of each figure.

F Status of Exemption from Institutional Review Board

Before starting any segments of the study involving human evaluation, the research team completed and submitted a “Human Subjects Research Determination" form to the appropriate Institutional Review Board (IRB). We obtained a determination letter from the IRB before any human study activities commenced, indicating that our project proposal had been granted ‘Exempt’ status. This classification implies that the proposed research was deemed ‘Not Human Subjects Research’.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?



[GPT-4V]: The sequence of images depicts a person sitting on a white plastic chair on a wooden deck, with trees and a railing in the background. In the first image, the person is seated normally. As the sequence progresses, the chair's back legs begin to give way, and the person starts to tilt backward. By the third image, the chair has collapsed entirely, and the person is falling backward with their legs in the air. In the fourth image, the person has almost completely fallen to the deck, and the chair is visibly broken. The final image shows the person lying on the deck, with the chair collapsed underneath them, suggesting an unexpected and possibly humorous mishap.

[Human annotation]: The sequence of images captures a moment on a wooden deck surrounded by trees, which explains the presence of dead leaves on the ground. A young girl is seated in a white plastic chair, swinging her legs.

Figure 7: A sample of failure reasoning case in Daily-life domain, we highlight the hallucination parts in yellow. Failure reason: co-occurrence behavior and Snowball.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

[GPT-4V]: The sequence of images depicts two individuals engaged in a winter activity, likely sledding down a snow-covered hill. In the first image, they are seen sitting at the top of the hill, preparing to slide down. They are both wearing winter clothing and appear to be in a cheerful mood, ready to enjoy the ride. As the sequence progresses, we see them in motion, sliding down the hill with increasing speed. The person in the front is holding onto the sled's handles, while the person behind is holding onto the front person's waist, ensuring they stay together. By the last image, they have gained momentum and are further down the hill, with the person in the back leaning to the side, possibly for balance or to steer. The snowy environment and their attire suggest it's a cold day, but the activity they're engaged in is a common and enjoyable way to experience winter weather.

[Human annotation]: In the sequence of images, we see a man and a girl engaging in a winter sledding activity. The man is seated on a large plastic disc to slide down the snowy slope. He is wearing sunglasses on his nose, likely to protect his eyes from the glare of the sun reflecting off the snow. The girl, wearing a beanie for warmth during the sledding adventure, is holding onto the man's shoulder to avoid falling off as they descend. She hugs his neck, and as they begin to slide, her reaction is to hug him even tighter, possibly for security and enjoyment. The man is seen wiggling his legs against the snow, an effort to propel themselves forward down the hill.

Figure 8: A sample of failure reasoning case in Daily-life domain, we highlight the hallucination parts in yellow. Failure reason: object hallucination, and correlation between object hallucination and behavioral hallucination.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

[GPT-4V]: In this sequence of images, we see a person engaging in a playful attempt to mimic the action of pushing a large aircraft. The aircraft, a fixed-wing propeller plane, is stationary on the tarmac. The person, dressed in a red shirt and shorts, starts by facing the aircraft with their hands extended as if they are about to exert force. They then proceed to push against the nose of the aircraft, maintaining the playful illusion that they are trying to move the massive plane with their own strength. Throughout the sequence, the person's actions are clearly for fun, as the size and weight of the aircraft make it impossible for a single individual to move it in such a manner. The background shows a clear day with other helicopters and planes visible, indicating that the setting is likely an airfield or aviation museum.

[Human annotation]: The sequence of images captures a man in red on an airport apron, performing the role of directing a plane to park using hand gestures. He is seen standing in front of the aircraft, guiding the pilot with his movements. As the plane approaches, the man raises both hands to signal the aircraft to stop moving. Once the plane has come to a complete stop, the man signals a cross with his arms, indicating that the plane should cease all movement. Having finished his job, the man then walks away from the track.

Figure 9: A sample of failure reasoning case in Daily-life domain, we highlight the hallucination parts in yellow. Failure reason: lack of common sense and Snowball.

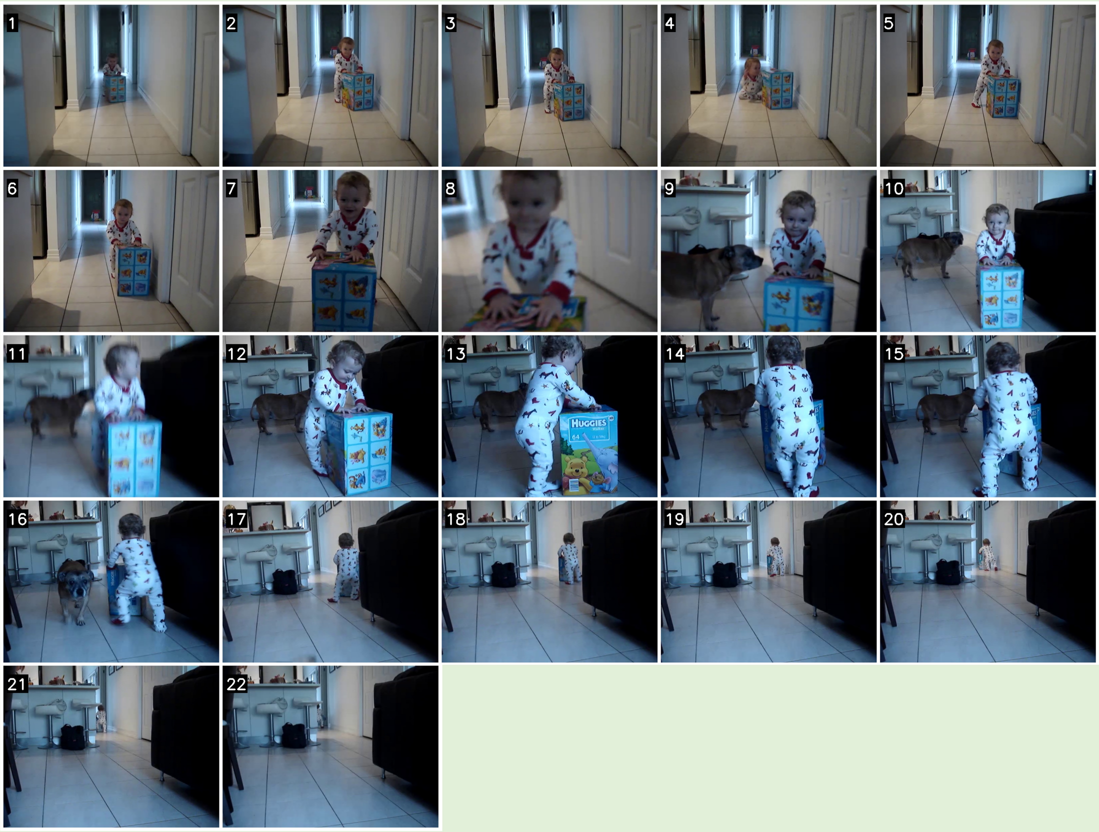
[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

[GPT-4V]: The image sequence shows a young boy wearing sunglasses and a white shirt, standing on a tennis court. He is holding a tennis racket and appears to be playing tennis. The scene is captured in multiple shots, showing the boy in various positions and angles, emphasizing his actions and the environment around him.

[Human annotation]: The collage of images captures a sequence of moments featuring three children, interacting outdoors by a roadside on a sunny day. The boy is wearing sunglasses to protect his eyes from the sun, while the girl in brown initially shields her eyes with her hands. Throughout the sequence, the children are seen pointing in various directions, possibly as part of a conversation or game they are engaged in. Towards the end of the sequence, the boy is seen waving his hand in front of the camera.

Figure 10: A sample of failure reasoning case in Daily-life domain, we highlight the hallucination parts in yellow. Failure reason: object hallucination, correlation between object hallucination and behavioral hallucination, and co-occurrence behavior.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

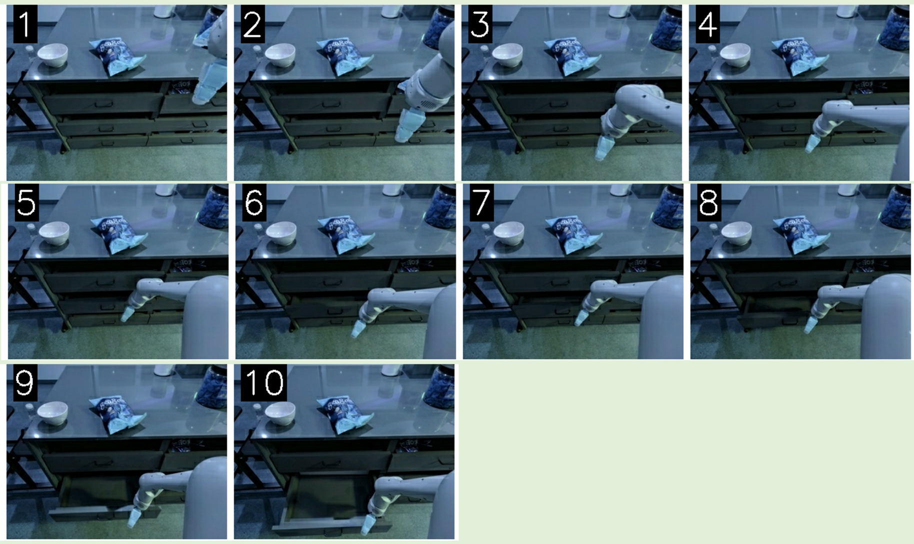


[GPT-4V]: In this sequence of images, we see a young child in pajamas with a playful design, pushing a large box of Huggies diapers along a tiled hallway. The child starts at one end of the hallway and makes their way towards the camera, showing determination and focus on the task. As the child progresses, a dog enters the scene, observing the child's actions with interest. The child continues to maneuver the box around, at times looking directly at the camera, possibly interacting with someone off-camera. Eventually, the child successfully navigates the box past the living area, where we can see an adult in the background, and heads towards a doorway at the end of the hallway, demonstrating persistence and mobility. Throughout the sequence, the child appears to be engaged in a playful activity, possibly turning a chore into a game, while the dog provides quiet companionship.

[Human annotation]: In the sequence of images, we see a baby in a living room pushing a box across the floor. As the baby walks along the corridor, he eventually sits down on the floor. After reaching the sofa, the baby turns around, presumably to walk back to where he started. Throughout this activity, a dog is present in the room, initially looking at the baby as he walks past and later walking towards the camera after the baby turns around. The baby resumes pushing the box back to the starting point.

Figure 11: A sample of failure reasoning case in Daily-life domain, we highlight the hallucination parts in yellow. Failure reason: Snowball. In this case, we observe that in addition to the significant behavioral hallucinations caused by Snowball effect mentioned in Section 3.2, another result of Snowball is that LVLMs may not fully describe all episodes in an image sequence. That is, after a behavioral hallucination occurs, the LVLM might assume the episode has ended and stop describing. For instance, in this case, the LVLM stopped describing after mentioning the child reaching the living room and the adult leaving, without continuing to describe the child pushing the box back along the hallway.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

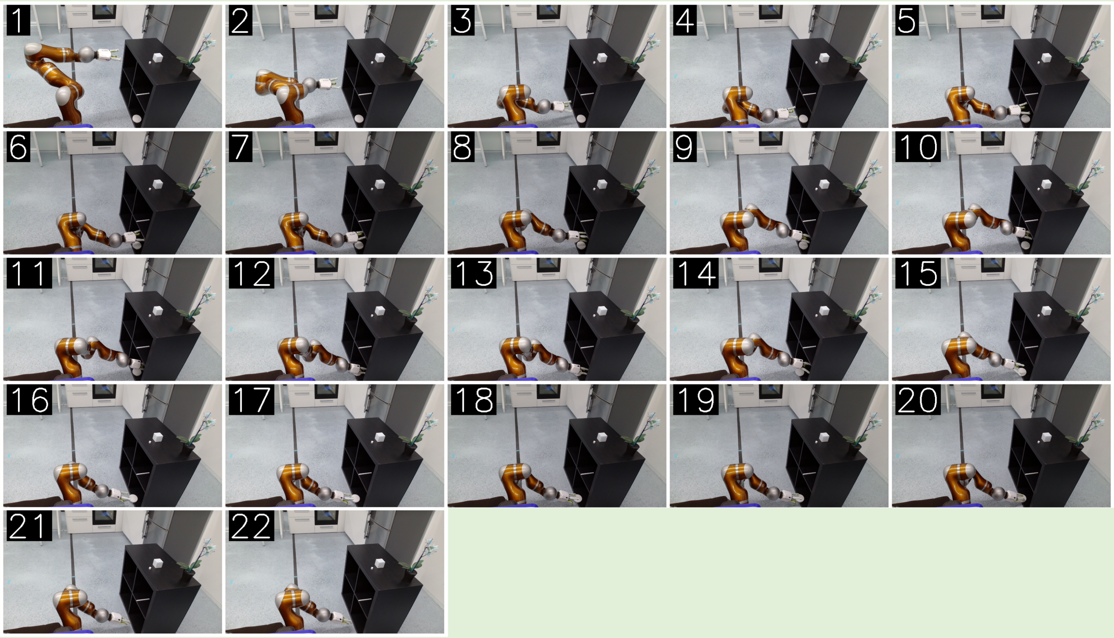


[GPT-4V]: The sequence of images depicts a robotic arm interacting with a drawer in a piece of furniture. Initially, the drawer is closed, and the robotic arm is positioned near it. As the sequence progresses, the arm extends towards the drawer and appears to grasp the handle. It then pulls the drawer open, revealing the contents inside. The images capture the motion of the robotic arm as it performs the task of opening the drawer, showcasing the arm's range of motion and its ability to interact with objects in its environment.

[Human annotation]: In this picture, there is a table with an empty bowl, a packet of snacks, and a blue can on it. The table has many closed drawers. A robotic arm has located the drawer in the first column and second row and has pulled it open.

Figure 12: A sample of failure reasoning case in Robotics domain, we highlight the hallucination parts in yellow. Failure reason: co-occurrence behavior.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

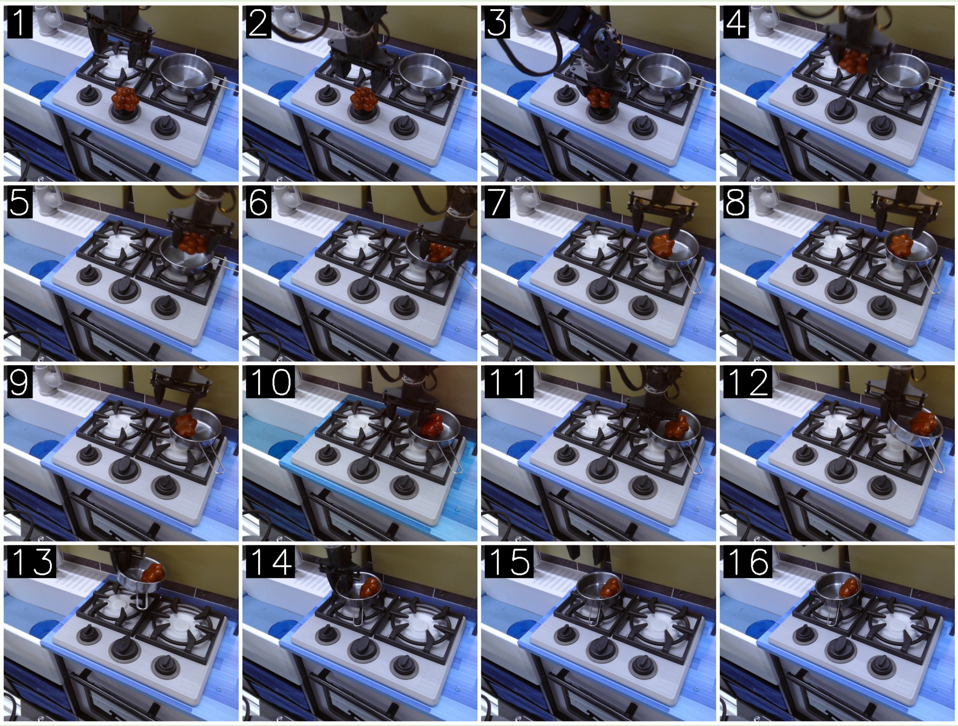


[GPT-4V]: In this sequence of images, we observe a robotic arm performing a task in a laboratory or testing environment. The arm, which has a brown and metallic appearance with a white end effector, is positioned near a black shelving unit. The sequence captures the robot in motion as it reaches towards a small white cube placed on top of the shelf. The images show the robot arm extending towards the cube, adjusting its position, and attempting to interact with the object. The sequence likely demonstrates the robot's ability to perform precise movements and interact with objects, which could be part of a test or demonstration of its capabilities in manipulation and object handling.

[Human annotation]: In this scene, in a room, there is a black cabinet with a flower pot, a cube, and a small ball on top of it. The cabinet has four compartments, and in the bottom right compartment, there is a flat, disc-shaped object. A robotic arm is picking up this disc-shaped object from the compartment in the bottom right corner of the cabinet.

Figure 13: A sample of failure reasoning case in Robotics domain, we highlight the hallucination parts in yellow. Failure reason: Snowball. This case effectively demonstrates the lack of LVLM's reasoning ability in image sequence comprehension. In the first image, the robotic arm indeed appears to be moving towards the cube, but from the second image, the arm lowers and moves towards the disc-shaped object. The LVLM failed to infer this behavior from the first two images and based its subsequent description solely on the understanding in the first image, leading to a Snowball effect.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

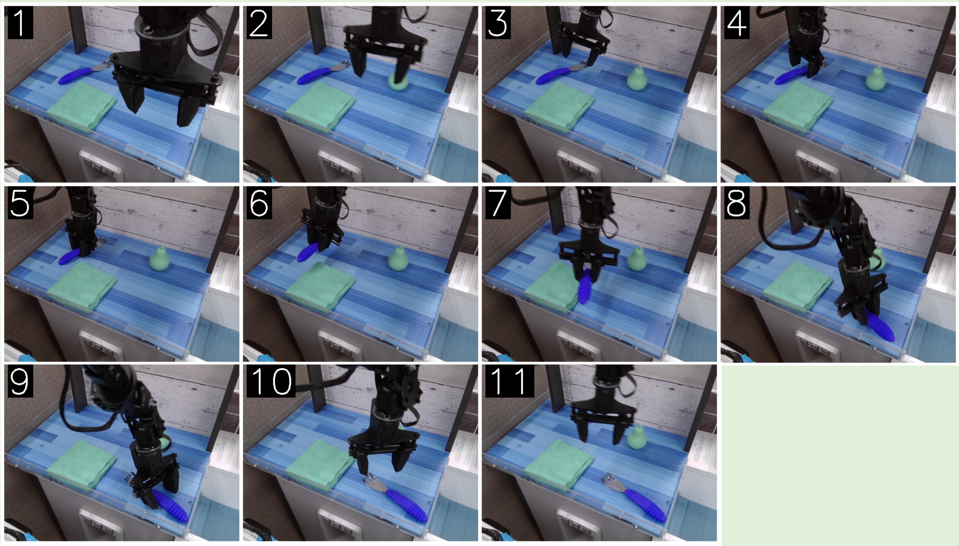


[GPT-4V]: In this sequence of images, we observe a robotic arm performing the task of cooking on a stove. Initially, the robotic arm is positioned above a pot on the stove, holding a cluster of brown objects that resemble sausages or some type of food. The arm then proceeds to lower the food into the pot. As the sequence progresses, the robotic arm releases the food into the pot and retracts, leaving the food to presumably cook. The sequence ends with the food items inside the pot, and the robotic arm retracted to its initial position, having successfully placed the items into the pot for cooking. The images depict a demonstration of automation in a kitchen setting, showcasing how robotics can be used to handle cooking tasks.

[Human annotation]: In this scene, there are two stoves and three switches. On the right stove, there is a silver pot. On the middle switch, there is a brown object. A robotic arm moves from the left stove to the middle switch, picks up the object, and places it into the silver pot on the right stove. Afterwards, the arm picks up the pot with the object inside and places it on the left stove.

Figure 14: A sample of failure reasoning case in Robotics domain, we highlight the hallucination parts in yellow. Failure reason: co-occurrence behavior and Snowball. This case also reflects another outcome of the Snowball effect that we mentioned in Figure 11. After assuming that the robotic arm is cooking, the LVLM do not continue to describe the behavior of the robotic arm moving the pot from the right stove to the left.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

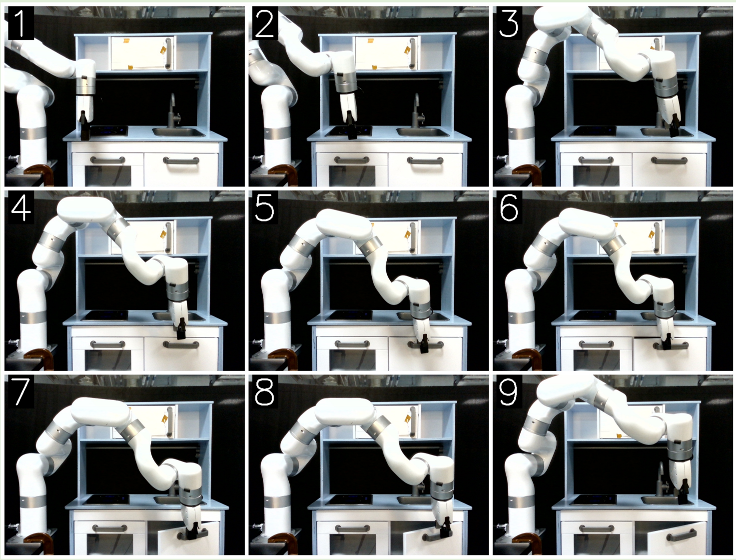


[GPT-4V]: The sequence of images depicts a robotic arm performing a task of picking up a fork from a table. Initially, the table is set with a fork lying flat, a green cloth, and a green bulbous object. The robotic arm, equipped with a gripper, descends towards the fork. It adjusts its position and orientation to align with the fork. The gripper then closes around the fork's handle, securing it. Subsequently, the arm lifts the fork off the table, demonstrating a successful grasp. The sequence ends with the robotic arm holding the fork in the air, having completed the task of picking it up. Throughout the sequence, the green cloth and the green object remain undisturbed on the table, indicating that the robot's action was precise and targeted only at the fork.

[Human annotation]: In this scene, there is a blue fork, a green towel folded into a square, and a green pear toy on a blue table. A robotic arm picks up the fork from the table and places it to the right of the square towel.

Figure 15: A sample of failure reasoning case in Robotics domain, we highlight the hallucination parts in yellow. Failure reason: Snowball.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

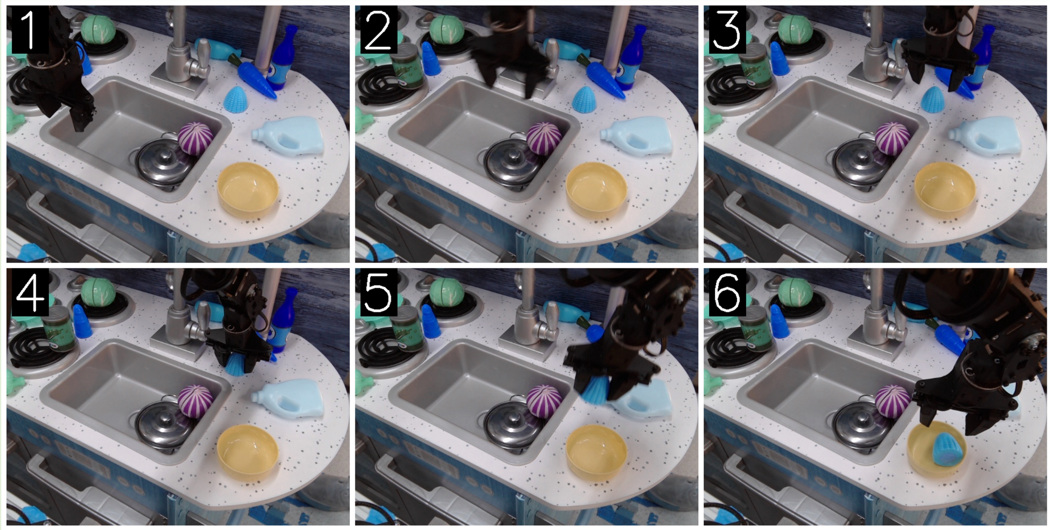


[GPT-4V]: The sequence of images depicts a robotic arm interacting with a kitchen setup. Initially, the robot arm is positioned above a smartphone resting on the counter. As the sequence progresses, the arm descends towards the phone, makes contact, and appears to be manipulating or pressing on the phone's screen. The robot's movements suggest that it is programmed to perform a task involving the phone, possibly simulating a human interaction such as selecting an app or entering a command. Throughout the sequence, the robot's precise and controlled movements are consistent with the capabilities of modern robotic technology used for automation and interaction with various devices.

[Human annotation]: In a room, there is a counter with a black induction stove on the left and a sink on the right. Below them, there is a cabinet with two doors. A robotic arm moves from the left side of the counter to the right and opens the door of the cabinet on the right side.

Figure 16: A sample of failure reasoning case in Robotics domain, we highlight the hallucination parts in yellow. Failure reason: object hallucination, correlation between object hallucination and behavioral hallucination, and Snowball.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

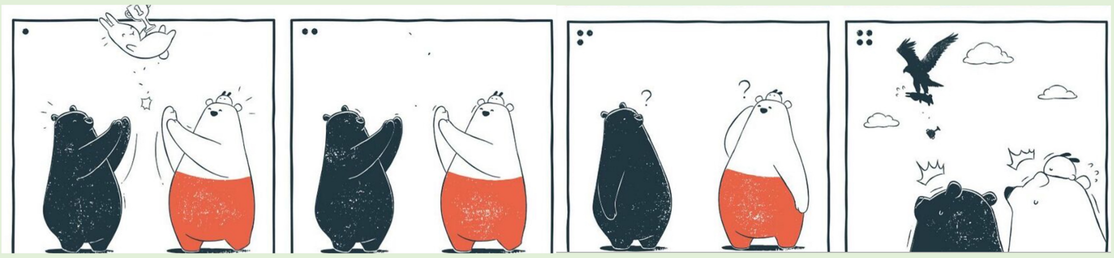


[GPT-4V]: The image sequence depicts a series of six frames showing the process of washing a dish. In the first frame, we see a dirty plate with food residue on it placed next to a sink that contains a purple sponge and a dishwashing brush. There are also bottles of dish soap and a hand soap dispenser nearby. In the second frame, the plate is being rinsed with water from the faucet, and the food residue is being washed away. By the third frame, the plate appears to be clean and is being rinsed off thoroughly. In the fourth frame, the clean plate is placed on the drying mat to the left of the sink. The fifth frame shows the sink area with the plate now absent, indicating it has been moved away after cleaning. Finally, in the sixth frame, we see a person's hands as they turn off the faucet, concluding the dishwashing process. Throughout the sequence, the position of the cleaning tools and soap bottles changes slightly, suggesting activity and interaction with these items during the cleaning process.

[Human annotation]: On a countertop with a sink, on the left, there are two stoves: one with a green cabbage model and the other with a jar. To the right of the sink, there are models of a blue carrot, half a blue corn, a blue bottle, a horizontally placed dishwashing liquid model, and a yellow plastic bowl. Inside the sink, there is a purple and white onion model and a metal basin with a lid. A robotic arm moves to the half blue corn model, picks it up, and places it inside the yellow plastic bowl in the bottom right corner.

Figure 17: A sample of failure reasoning case in Robotics domain, we highlight the hallucination parts in yellow. Failure reason: object hallucination, correlation between object hallucination and behavioral hallucination, and Snowball.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

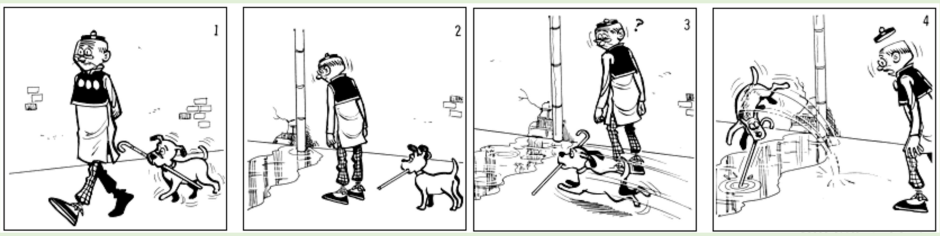


[GPT-4V]: The sequence of images depicts a charming and humorous interaction between two bears. In the first image, we see a black bear and a white bear with red shorts looking up as a small bird is falling from the sky. The white bear is preparing to catch the bird with its hands cupped together. In the second image, the white bear successfully catches the bird, and both bears seem pleased with the catch, as indicated by the little star symbolizing the successful action. In the third image, confusion sets in; both bears are looking at their empty hands with question marks above their heads, wondering where the bird has gone. The final image reveals the twist: a large eagle is flying away with the small bird in its talons, leaving the bears looking up in surprise as they realize what has happened. The sequence tells a story of an unexpected turn of events in a lighthearted and whimsical manner.

[Human annotation]: A black bear and a white bear together toss a small rabbit into the air to celebrate its first-place win. After being tossed into the air, the rabbit doesn't come down, leaving the two bears puzzled on the ground. However, when they look up, they are surprised to find a black eagle in mid-air snatching away the rabbit.

Figure 18: A sample of failure reasoning case in Comics domain, we highlight the hallucination parts in yellow. Failure reason: object hallucination, correlation between object hallucination and behavioral hallucination, and Snowball.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

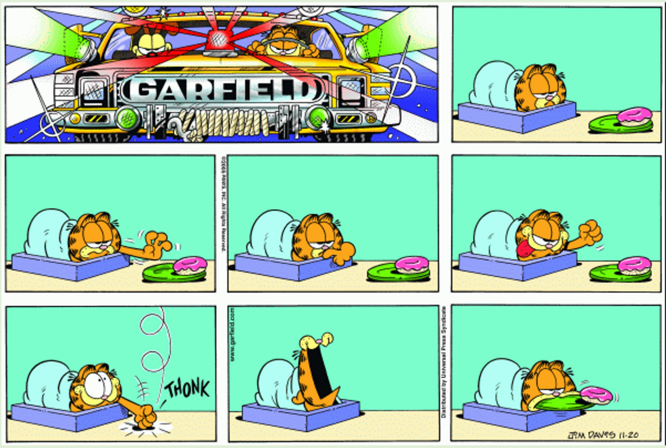


[GPT-4V]: In the sequence of images, we observe a humorous situation involving a person and their dog. In the first image, the person is walking confidently with their dog on a leash. In the second image, the person bends down to inspect something at the base of a pole while the dog waits patiently. By the third image, the person looks puzzled, scratching their head, as they notice water beginning to pool around the pole's base. The final image reveals the comedic twist: the dog is secretly turning a valve at the base of the pole, causing water to spray out, much to the person's surprise and confusion. The person's hat is even popping off in astonishment at the unexpected fountain created by the mischievous dog.

[Human annotation]: A man is walking down the street, followed by a dog with a cane in its mouth. Suddenly he came across a puddle and stopped, not knowing how to get over. At this time, the dog behind him rushed over, used the crutch in his mouth to support the ground and jumped up over the puddle. The man expressed great shock.

Figure 19: A sample of failure reasoning case in Comics domain, we highlight the hallucination parts in yellow. Failure reason: object hallucination, correlation between object hallucination and behavioral hallucination, and Snowball.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?



[GPT-4V]: In this sequence of images, we see a humorous cartoon strip featuring a cat lying on a bed next to a plate with a donut. The cat, initially resting, is tempted by the donut and reaches out to grab it with its paw. As the cat tries to pull the donut closer, it accidentally flings the donut into the air. The cat then watches in dismay as the donut lands on its face, sticking to it due to the donut's icing. In the final image, the cat appears resigned to its fate, with the donut still covering its face, as it lies back down on the bed. The sequence is a playful depiction of the cat's failed attempt to sneak a treat and ending up in a comical predicament.

[Human annotation]: Lazy Garfield is nestled in his blanket with a plate of doughnuts in front of him. The doughnuts are on one side of the plate, just out of Garfield's reach. He withdraws his hand to think of a solution, then has a bright idea. He makes a fist and forcefully knocks on one side of the plate, causing both the doughnuts and the plate to fly into the air. In his plan, the thrown doughnuts will fall into his open mouth, so he eagerly opens wide to catch them. Unfortunately, the plate drops into his mouth before the doughnuts, and the doughnuts end up teetering on the edge of the plate.

Figure 20: A sample of failure reasoning case in Comics domain, we highlight the hallucination parts in yellow. Failure reason: Snowball.

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?

[GPT-4V]: In this sequence of three black-and-white cartoon images, we observe a humorous episode at a bar. In the first image, a man is holding a beer mug with a shining, seemingly magical beer that is attracting his gaze with its radiance. In the second image, the bartender, who is behind the bar, is revealed to be holding a flashlight under the mug, creating the illusion of the glowing beer. The third image shows the man's realization of the trick as he looks through the now empty mug, with beer spilled all over the bar and floor, and the flashlight's beam visible through the bottom of the mug. The man's expression of surprise and the bartender's absence suggest the man has been duped by a simple yet effective prank.

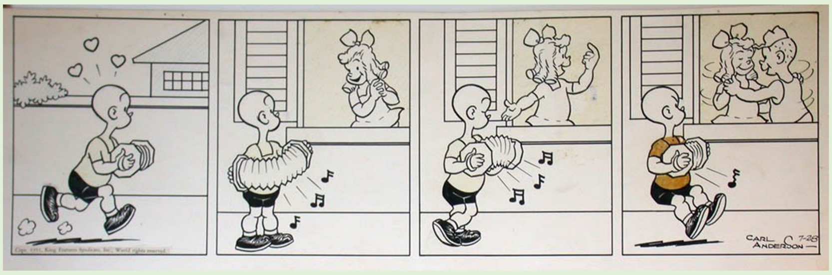
[Human annotation]: A man roughly smashes his glass onto the bar counter, hoping the loud noise will prompt the bartender to refill his drink. Hearing the sound, the bartender comes over with a bottle to pour more drink into the man's glass. However, when the man picks it up, he finds that the bottom of the glass has a big hole, and the drink spills all over the table. It turns out that his earlier action of smashing the glass was so forceful that he accidentally broke the bottom of the glass, causing the drink to leak out as soon as he lifted it.

Figure 21: A sample of failure reasoning case in Comics domain, we highlight the hallucination parts in yellow. Failure reason: object hallucination, correlation between object hallucination and behavioral hallucination, and Snowball.

Table 3: Number of videos selected from each sub-dataset of Open X-Embodiment.

Sub-dataset name	Number of videos selected
fractal20220817_data	400
kuka	50
bridge	300
jaco_play	50
berkeley_autolab_ur5	50
toto	10
columbia_cairlab_pusht_real	5
stanford_hydra_dataset_converted_externally_to_rlds	5
ucsd_kitchen_dataset_converted_externally_to_rlds	50
bc_z	50
utokyo_pr2_opening_fridge_converted_externally_to_rlds	5
utokyo_pr2_tabletop_manipulation_converted_externally_to_rlds	10
utokyo_xarm_pick_and_place_converted_externally_to_rlds	1
utokyo_xarm_bimanual_converted_externally_to_rlds	5
dlr_sara_pour_converted_externally_to_rlds	5
dlr_edan_shared_control_converted_externally_to_rlds	100
asu_table_top_converted_externally_to_rlds	20
utaustin_mutex	30
berkeley_fanuc_manipulation	30

[Prompt]: Write a description for the given image sequence in a single paragraph, what is happening in this episode?



[GPT-4V]: In this sequence of images, we see a story unfold in four parts. In the first image, a bald boy is happily running with a box of chocolates and hearts floating above his head, indicating he is in love or has affectionate feelings. In the second image, he has arrived at a house and is presenting the chocolates to a girl who is leaning out of an upstairs window, while musical notes suggest he is serenading her. In the third image, the boy continues to play the accordion, serenading the girl with music, still hopeful and eager to impress her. However, in the final image, the boy's romantic gesture takes an unexpected turn as another boy appears at the window, and the girl joyfully shares the chocolates with him, leaving the first boy shocked and dismayed as he continues to play the accordion, now with a look of surprise and disappointment on his face. The story captures a classic comedic twist on unrequited love and the pitfalls of romantic gestures.

[Human annotation]: A boy, holding his accordion, runs to the house of the girl he adores. He plays music under her window, and the girl listens happily. As she listens, she gestures with her finger, calling someone from inside the house. To the boy's surprise, it's her boyfriend who appears. The girl and her boyfriend start dancing to the accordion music. This upsets the boy, who has a crush on the girl, and he stops playing.

Figure 22: A sample of failure reasoning case in Comics domain, we highlight the hallucination parts in yellow. Failure reason: object hallucination, correlation between object hallucination and behavioral hallucination, and Snowball.

Table 4: Correlation coefficient between behavioral hallucination and object hallucination of different MLLMs on Mementos.

Domain	**Input type**	**Model**	Recall	Precision	F1
Daily-life	Sequential	GPT-4V	0.120	0.188	0.132
		Gemini	0.165	0.179	0.146
		Video-LLaMA-2	0.197	0.067	0.125
		Chat-UniVi	0.138	0.178	0.137
	Combined	GPT-4V	0.242	0.182	0.199
		Gemini	0.158	0.179	0.152
		Chat-UniVi	0.127	0.184	0.172
		LLaVa-1.5	0.112	0.134	0.106
		MiniGPT4	0.135	0.145	0.115
		MiniGPT5	0.126	0.188	0.146
		mPLUG_Owl-v2	0.106	0.113	0.069
		InstructBLIP	0.133	0.125	0.127
Robotics	Sequential	GPT-4V	-0.012	0.022	0.011
		Gemini	0.027	0.144	0.101
		Video-LLaMA-2	0.107	0.107	0.109
		Chat-UniVi	0.038	0.121	0.089
	Combined	GPT-4V	0.041	-0.022	0.008
		Gemini	-0.049	-0.086	-0.106
		Chat-UniVi	0.189	0.242	0.207
		LLaVa-1.5	0.135	0.123	0.157
		MiniGPT4	0.186	0.316	0.233
		MiniGPT5	0.056	0.027	0.045
		mPLUG_Owl-v2	0.244	0.163	0.231
		InstructBLIP	0.227	0.235	0.253
Comics	Sequential	GPT-4V	0.045	0.225	0.158
		Gemini	0.176	0.081	0.144
		Video-LLaMA-2	0.261	0.280	0.299
		Chat-UniVi	0.239	0.331	0.221
	Combined	GPT-4V	0.343	0.539	0.471
		Gemini	0.187	0.121	0.167
		Chat-UniVi	0.293	0.113	0.279
		LLaVa-1.5	0.062	0.101	0.088
		MiniGPT4	0.199	0.134	0.213
		MiniGPT5	0.324	0.366	0.339
		mPLUG_Owl-v2	0.231	-0.043	0.157
		InstructBLIP	0.288	0.005	0.262

Table 5: Human evaluation.

Model	Eval type	Object			Behavior		
		Recall	Precision	F1	Recall	Precision	F1
GPT-4V (s-input)	GPT-4	60.91%	51.04%	54.13%	38.02%	33.05%	34.12%
	Human	57.69%	49.54%	52.01%	35.26%	31.60%	32.67%
Gemini (s-input)	GPT-4	37.54%	39.43%	36.88%	23.38%	34.19%	24.02%
	Human	35.82%	38.11%	37.09%	20.46%	33.72%	22.99%
ChatUnivi (s-input)	GPT-4	40.32%	42.04%	39.52%	24.95%	28.06%	27.15%
	Human	37.65%	38.59%	36.46%	25.73%	27.40%	26.64%
LLaVA-1.5 (c-input)	GPT-4	35.77%	44.18%	38.09%	24.47%	38.79%	28.59%
	Human	36.84%	41.37%	39.77%	22.95%	39.82%	29.18%

Table 6: All prompts used in our paper.

Prompt
Task: Rewrite questions and answers into a single paragraph
Image: <Image sequence> Text: <Write a description for this image based on the following questions and answers in one paragraph. Please remember that some objects or actions in the following questions and answers may not be included in the images. Please do not include the excluded items in your description. Here are the questions and answers: Question: {**Question 1**} Answer: {**Answer 1**} Question: {**Question 2**} Answer: {**Answer 2**} ... Question: {**Question n**} Answer: {**Answer n**}>
Task: Generate description for the given image sequence
Image: <Image sequence> Text: <Write a description for the given image sequence in a single paragraph, what is happening in this episode?>
Task: Extract object and behavior keywords
Text: <I will provide you two paragraphs. The first paragraph is human-composed and the second paragraph is generated by AI models. I want to evaluate the hallucination in the second paragraph. Please extract the object and action words or phrases from the following text. The objects should have a tangible meaning and consist of no more than two words; non-tangible objects should not be extracted. The action words or phrases should only relate to the extracted objects. Also, you must convert the corresponding actions to their complete root form. Then, for the final answer, please examine 4 lists and must transfer the synonyms in 4 lists into the same word. Please directly output the final object and action lists in two paragraphs, respectively as in the form in the example below without any justifications or intermediate steps. Here is an example: 1. The sequence of images captures a dog's cautious interaction with a metal toy inside a house. The dog appears wary and maintains a distance from the unfamiliar object, barking to express its disapproval and possibly intimidation. As the toy moves, the dog's reaction is to bark and lean backward, showing a clear sign of being unsettled by the toy's motion. When the toy momentarily ceases movement, the dog also stops, remaining alert and attentive. At the end of the image, when the toy comes to a halt, the dog looks up, still processing the strange encounter with the inanimate object. 2. The image is a collage of multiple pictures featuring two dogs playing with a toy alligator. The dogs are in various positions, with some of them standing on the toy alligator, while others are interacting with it in different ways. The collage captures the dogs' playfulness and excitement as they engage with the toy alligator. The lists are Object list 1: [dog, toy, house] Action list 1: [interaction, bark, express intimidation, move, lean backward, stop, look up] Object list 2: [dog, toy] Action list 2: [play, stand, interaction] Here is the paragraphs: 1. {**Human-annotated description**} 2. {**AI-generated description**} The lists are:>