

# Investigating Hallucination in Image Sequences using Mementos

Bruno Hime

University of Trento

State University of Campinas

b205652@dac.unicamp.br

Nella Zabrina Pramata

University of Trento

nellazabrina.pramata@unitn.it

## 1 Overall Problem

Multimodal Large Language Models (MMLMs) have exhibited the ability to perform visual and language tasks. However, the benchmarks available are commonly found in static data formats such as a single image. Mementos (Wang et al., 2024) is a new benchmark that offers data in a novel format, in the form of a sequence of images. Based on the results shown in the Mementos paper, most models exhibit below-chance performance which indicates that MMLMs struggle with object and behavioral hallucinations when generating descriptions from image sequences.

The MMLMs performance on hallucinations raises a question of whether the problem itself is due to the way the data is composed. Specifically if the way the sequence of images and the prompt is composed correlates with hallucinations. In addition, to understand more about hallucinations and possible ways to prevent them, some mitigation techniques will be taken into account. This also raises an interesting point since the existing techniques are usually effective for a single image, the effectiveness for image sequences remains uncertain.

## 2 Research Question

The objective of this study is to examine how data perturbations influence model predictions, with a particular focus on understanding which factors might lead to hallucinations and how much they can affect the model’s response. We further break it down into the following sub-questions: (1) How does incorrect text information influence the model’s predictions, and how does it compare to incorrect visual features? (2) To what extent does the model rely on the temporal order of the images in the sequential dataset for predictions? (3) How important is the presence of rich input for the model, and how does simplifying it by removing

some information impacts the model’s comprehension? (4) How effective are training-free mitigation techniques like MemVR and HALC in preventing hallucinations using image sequences as input, and how do noisy text inputs interfere these techniques? By addressing these aspects, we aim to understand the model’s sensitivity to temporal information and noise in both text and images.

## 3 Related Work

Mementos is a novel benchmark with a diverse and unique structure. It relies on image sequences, reinforces the need for reasoning skills, and better represents our ever-changing world (Wang et al., 2024). Its design makes it easy to manipulate inputs and compare the response of the model to a baseline, helping us identify key factors that influence hallucinations.

In addition, one of our experiments was inspired by the Hallucination Snowballing paper (Zhong et al., 2024). It indicated that hallucinations from the previous LVLMs interactions can affect LVLMs results, leading to snowballed hallucinations.

To mitigate the hallucinations, various techniques have been introduced, including MemVR (Zou et al., 2024) and HALC (Chen et al., 2024), which are plug-and-play modules that integrate into the LVLM without any additional training. These two approaches were chosen for some of the experiments, as they enhance the reliance of the model on visual features. MemVR achieves this by reinjecting visual information when hallucination occurs ruled by an entropy-based metric, while HALC leverages distinct fine-grained visual information using object detection to correct hallucinated visual tokens.

## 4 Experiment Design and Evaluation Metrics

In this experiment, we will use the *daily-life* part of the Mementos benchmark, which consists of

selected images derived from video clips in the NextQA dataset, representing a range of everyday life scenarios (Wang et al., 2024). In addition, the study will be divided into two main categories, sequential and combined input. Despite the division, the dataset for both remains the same, the only difference relies on the input format. While in the sequential the examples are folders containing images representing daily life scenes, in which each one of them is individually given to the model, the combined is an adapted input that combines all frames from an image sequence into one composite image. With respect to the models, we conducted the experiments using the best open source models, respectively, Chat-UniVi for sequential input and LLaVA-1.5 for combined input as shown in (Wang et al., 2024). These models will serve as a baseline and for paper recreation.

We organized the sequential and combined input with different approaches as explained in Section 4.1 and 4.2. The information about image sequence length distribution is shown in Table 3. Once the dataset is properly organized, the mementos benchmark will take place. The models will be instructed to generate a description for the given images, afterwards, the output will be compared to the ground truth, which was manually built as provided in Mementos repository<sup>1</sup>. This comparison will be with respect to the object and action keywords from the images episode. The extraction will be assisted by the GPT-4 (OpenAI, 2023), then they will be matched with a synonym graph, replacing the matched keywords with the root word from the synonym graph. Finally, with the keywords lists ready, the Recall, Precision and F1 scoring methods will be calculated.

## 4.1 Sequential

In the sequential dataset, the chosen tests relies on shuffling the order of the images, deleting some images, and introducing noise by adding an incorrect description of the episode by text or random images. For the incorrect images and the reduced example, the number of changes, either by adding or reducing the number of images, will be proportional to the size of the example. We gradually increased the rate of these changes from 20%, 40%, 60% and 80%.

---

<sup>1</sup><https://github.com/umd-huang-lab/Mementos>

### 4.1.1 Shuffled

The first thing we want to analyze is the temporal information. For each example, the images are provided to the model sequentially, one by one following a specific chronological order. Therefore, by shuffling the images of each example, we can see how much the model indeed needs the chronological order of the event in order to comprehend the episode and provide the correct objects and actions happening.

### 4.1.2 Reduced example

The idea behind this test is to gradually reduce the number of images given to the model for each example. This will be a good way to analyze the necessity of overly descriptive input given by multiple images describing the action, in order to comprehend the idea behind the episode, while, on the other side, excess information might be confusing for the model.

### 4.1.3 Incorrect images

This time the aim was to assess how much the model can be affected by the existence of visual noise. Some images not related to the episode will be added to the example sequence. This can indicate about the model's capability of focusing on the crucial and important information about what is happening ignoring possible distractions about unimportant factors. The incorrect images are the first images from the last 30 examples of the sorted dataset, that were randomly chosen and placed inside each sequence of images.

### 4.1.4 Incorrect prompt

Finally, the last experiment in the sequential part aims to analyze how much the model can be influenced by adding a misleading text prompt in the input. Inspired by the Hallucination Snowballing paper (Zhong et al., 2024), the idea is to add a fake conversation history to the prompt and intentionally introduce incorrect information in it. The goal is to determine whether the model will be influenced by a wrong list of objects or it will rely solely on visual input, as required by the original task. In the prompt, we faked a conversation history in which the user asked the model to retrieve a list of objects from the image sequences, and the model responded with a list supplied by the GPT-4 which acts as the hallucinated objects as shown in Appendix F. This approach was inspired by the Mementos paper when doing the keyword extraction

evaluation. The difference is in our paper, we are not only extracting the object keywords found in the sequence but based on that we fabricated an object list that is not found in the sequence. We manually verified the generated incorrect object list and placed it as if it were the model’s response. The prompt used to generate is shown in Table 4.

## 4.2 Combined

Using combined dataset adapted into composite images, we investigate the model’s behaviour by using the same incorrect prompt like in sequential input as explained in Section 4.1.4. Besides that, instead of performing the incorrect images to compare the degree of influence between visual and text noise, we will apply MemVR (Zou et al., 2024) and HALC (Chen et al., 2024) as hallucination mitigation techniques. The goal here is to determine whether the incorporation of these mitigations techniques can counteract potential errors introduced by misleading textual descriptions in the prompt, since both of them were designed with the intention of reinforcing the visual information.

## 5 Results and Analysis

We replicated the Mementos paper as in Table 2 and we only found 448 sequences of ground truth instead of 450 as mentioned in (Wang et al., 2024). Our scores compared to the Mementos paper are not exactly the same, but similar considering that the ground truth object and behaviour list are generated by GPT, so they might change on different requests. It suggests that we replicated the paper properly, which allowed us to proceed with our experiments. Due to limited resources, all experiments were made with the first 50 examples of the sorted dataset. The low number of images might cause differences in some of the results compared to if we were to utilize the fully sampled data. The results of our experiments are shown in Table 1 and in Plot 1. The code of the experiments can be accessed in the repository<sup>2</sup>.

### 5.1 Sequential

In relation to the first sub-research question outlined in Section 2, at first glance, we observe a drastic drop around 8-13% in the object scoring metric for the incorrect prompt, mostly noticed in the precision and F1. This shows us how much the

model prefers to rely on the additional text information over the visual features, considering that the images totally contradict the text information while the task was purely to analyze the visual episode. It also aligned with the findings in (Zhong et al., 2024) where hallucinations in previous interactions affect the LVLMs ability to generate text correctly.

Another sensitivity of the model is to the insertion of incorrect images. Overall, the model demonstrated a trend in reduced performance. This suggests that there is a possibility that the occurrence of incorrect images caused the models to be confused, leading to an increase in hallucinations. When we examined in detail, we can see stronger hallucinations in the visual noise compared to the text noise in most metrics. This suggests that the insertion of incorrect images is causing hallucinations more than injecting incorrect prompt.

Despite the expected drop in the performance of incorrect images, it is still interesting to investigate the increase in some performance metrics on the higher rate of insertion of incorrect images. For instance, increase of 5% in object recall metric as the rate moved from 40% to 60%. This might suggest that as the rate of the incorrect images increased, the model appeared to find a compromise by incorporating both correct and incorrect visual context and being able to generate the text less hallucinated than when the incorrect image rate was lower. However, this remains a tentative explanation due to the limitation of the evaluation method used. In addition, we can see an unexpected relatively high performance for the rate of 80%, making it even comparable to the incorrect prompt, considering that the examples have been changed almost completely. This can be due to the evaluation method that relies on keyword matching which allows the model to achieve higher scores, even when the descriptions don’t accurately match the correct scene, as clearly shown in Table 5. The Mementos evaluation method appeared to face a limitation since it only considers the keywords extracted. A more improved evaluation approach for future work such that considers the meaning of the sentence may provide deeper insights and help clarify these findings.

Turning to the second sub-research question in Section 2, we can see no major differences when shuffling the order of the images in a sequence for this dataset which relies on short videos. For the Mementos method of evaluation, only feeding

<sup>2</sup><https://gitlab.com/nellazabrina/hallucinations-in-mementos-image-sequences>

Table 1: Performance Comparison of Sequential vs Combined Input Across Sequence Perturbations

| Input type | Variant                             | Object        |               |               | Behavior      |               |               |
|------------|-------------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|
|            |                                     | Recall        | Precision     | F1            | Recall        | Precision     | F1            |
| Sequential | Chat-UniVi (Paper replication)      | 43.21%        | <b>41.50%</b> | <b>40.57%</b> | 24.69%        | 22.50%        | 22.24%        |
|            | Incorrect Prompt                    | 39.03%        | 28.56%        | 31.98%        | 26.20%        | 26.22%        | 24.80%        |
|            | Incorrect Images rate=0.2           | 39.70%        | 33.57%        | 35.05%        | 24.39%        | 25.23%        | 23.64%        |
|            | Incorrect Images rate=0.4           | 30.40%        | 28.33%        | 28.14%        | 21.09%        | 21.62%        | 19.80%        |
|            | Incorrect Images rate=0.6           | 35.85%        | 30.62%        | 31.78%        | 18.65%        | 20.29%        | 18.40%        |
|            | Incorrect Images rate=0.8           | 32.53%        | 26.34%        | 28.18%        | 22.22%        | 24.02%        | 21.71%        |
|            | Reduced Example rate=0.2            | 40.91%        | 36.72%        | 37.54%        | 26.71%        | 25.77%        | 24.57%        |
|            | Reduced Example rate=0.4            | 41.13%        | 37.29%        | 38.22%        | <b>28.10%</b> | 26.60%        | 25.70%        |
|            | Reduced Example rate=0.6            | 41.35%        | 37.09%        | 37.90%        | 23.63%        | 23.66%        | 21.88%        |
|            | Reduced Example rate=0.8            | 41.57%        | 39.35%        | 38.97%        | 23.63%        | 26.37%        | 23.22%        |
| Combined   | Shuffled                            | <b>43.48%</b> | 39.53%        | 39.73%        | 26.67%        | 26.85%        | 25.29%        |
|            | LLaVA1.5 (Paper replication)        | 37.46%        | 37.58%        | 36.41%        | 24.18%        | 28.59%        | 24.52%        |
|            | LLaVA1.5 + Incorrect Prompt         | 37.83%        | 38.13%        | 36.80%        | 24.68%        | 29.65%        | 24.87%        |
|            | LLaVA1.5 + MemVR                    | 37.04%        | 37.41%        | 36.20%        | 24.89%        | 29.27%        | 25.05%        |
|            | LLaVA1.5 + MemVR + Incorrect Prompt | 37.41%        | 36.70%        | 35.42%        | 21.84%        | 26.73%        | 22.78%        |
|            | LLaVA1.5 + HALC                     | 40.41%        | 41.27%        | 39.84%        | 27.02%        | <b>33.59%</b> | <b>27.93%</b> |
|            | LLaVA1.5 + HALC + Incorrect Prompt  | 36.85%        | 36.78%        | 35.36%        | 25.19%        | 33.07%        | 26.81%        |

the model with the images is enough to make it comprehend what is happening in the scene, with no crucial temporal information.

Finally, in connection with the third sub-research question Section 2, the interesting and unexpected fact is that reducing the number of examples led to an increase in behaviour performance. This is likely because the dataset used consists of short videos where frames do not vary significantly. As a result, using multiple images as input may have introduced unnecessary complexity, potentially confusing the model. Less images have shown to be sufficient to describe the scene, suggesting a preference for simpler inputs even if a more complex input gives more details of the episode.

## 5.2 Combined

With respect to the first sub-research question in Section 2, the injection with the incorrect prompt did not appear to affect the performance compared to the combined input (paper replication) performance. Another point to consider is that, the performance is better than the sequential input injected with incorrect prompt, particularly in object and behaviour precision & F1 metrics. This suggested that the LLaVA1.5 model is able to ignore the incorrect information given from the incorrect prompt and focus on the visual information.

Focusing on the fourth sub-research question, the MemVR showed similar performance as the

combined input (paper replication) performance. However, HALC demonstrated the ability to mitigate the hallucination and even surpassed all other experiments in both sequential and combined input, particularly in behavior precision and F1 score. This suggests that even though the mitigation techniques commonly are effective for a single image, HALC can perform well in a composite image. Besides that, injecting incorrect prompt decreased the MemVR performance for less than 1% for the object metrics and around 2-3% for the behaviour metrics. On the other hand, the HALC performance dropped by 3.5-4.5% for the object metrics and 0.5-1.8% for the behavioural metrics. This suggests that even though both techniques enforce the visual information, they are still affected by the incorrect textual information given previously.

## 6 Limitations and further discussions

Our work was limited mainly by budget and resource constraints, as the evaluation method from the Mementos paper relies on ChatGPT, which comes with additional costs. Because of this, we had to use only 50 examples rather than the full dataset on our tests. Additionally, considering our budget, the paper was designed to focus solely on the *daily-life* dataset. It would be valuable to explore the other three domains from the Mementos paper to obtain more robust and reliable results. It would also be beneficial to try other types of data

manipulations, which could reveal more about the model’s behavior. The study can also be extended by refining the evaluation approach, not only limited to keyword-based metrics but also considering the meaning of the sentence.

## 7 Conclusion

The models’ reasoning comprehension are influenced by several key factors such as the total images in sequence, the presence of incorrect images as visual noise, and also the existence of the incorrect prompt as the textual noise. The incorrect text information interferes with the Chat-UniVi’s ability to generate text using sequential input, especially in object hallucinations. However, in the combined input using the LLaVA1.5, there is no significant difference when we inject an incorrect prompt. This implies that apart from the textual perturbation, the performance also depends on the model and input type used. A similar pattern was detected when we used incorrect images as noise, as it exhibited tendency to generate more hallucinations compared to the absence of noise data, particularly in object hallucinations. When comparing both text and visual perturbations specifically in inserting incorrect information, the incorrect images insertion generated more hallucinated text than the incorrect prompt injection.

The results highlighted that there is no major different when we shuffled the order of the image sequences. This means that temporal order did not give a significant impact for the text generation. It was also observed that the richer the inputs are, in terms of a longer sequence, the Chat-UniVi model faces increased difficulty in generating the text, leading to more hallucinations. This might be a result of the similarity between each image, as there are not a lot of changes in the *daily-life* dataset if we capture it in frames. Longer images introduced unnecessary complexity, hence a small number of images are sufficient for the model to generate the text.

Finally, even though hallucination mitigation techniques usually show effectiveness in a single image, only HALC performs reduced hallucinations in combined input. However, in the presence of the incorrect text prompts, both MemVR and HALC hallucinated more. This suggests that even though these techniques enforce the visual information to prevent hallucinations, the LLaVA1.5 model is still sensitive to textual perturbation.

## Acknowledgements

We would like to thank our professors, Raffaella Bernardi and Paolo Rota, for their support and guidance throughout this project. We also would like to thank the University of Trento for providing the resources to work on this research.

## References

- Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*.
- Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Xiayao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. 2024. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*.
- Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. *arXiv preprint arXiv:2407.00569*.
- Xin Zou, Yizhou Wang, Yibo Yan, Sirui Huang, Kenning Zheng, Junkai Chen, Chang Tang, and Xuming Hu. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.

## A Comparison Metrics between Mementos Paper and Our Paper Replication

Table 2: Performance Comparison of Mementos Paper and Our Replication Experiments Across Varying Total Images

| Input type              | Variant                  | Object |           |        | Behavior |           |        |
|-------------------------|--------------------------|--------|-----------|--------|----------|-----------|--------|
|                         |                          | Recall | Precision | F1     | Recall   | Precision | F1     |
| Sequential (Chat-UniVi) | 450 sequences (Mementos) | 40.74% | 40.78%    | 39.13% | 22.30%   | 31.10%    | 24.90% |
|                         | 448 sequences (ours)     | 42.04% | 42.35%    | 40.70% | 22.05%   | 30.62%    | 24.58% |
|                         | 100 sequences (ours)     | 43.95% | 42.83%    | 41.71% | 22.54%   | 24.80%    | 22.20% |
|                         | 50 sequences (ours)      | 43.21% | 41.50%    | 40.57% | 24.69%   | 22.50%    | 22.24% |
| Combined (LLaVA1.5)     | 450 sequences (Mementos) | 37.72% | 47.01%    | 40.18% | 22.17%   | 37.33%    | 26.65% |
|                         | 448 sequences (ours)     | 37.22% | 43.40%    | 38.53% | 22.72%   | 34.13%    | 26.09% |
|                         | 100 sequences (ours)     | 37.93% | 41.85%    | 38.33% | 23.85%   | 30.23%    | 25.20% |
|                         | 50 sequences (ours)      | 37.46% | 37.58%    | 36.41% | 24.18%   | 28.59%    | 24.52% |

## B Distribution of Image Sequence Length in Our Experiment

Table 3: Data Composition for Sequential & Combined Input

|                                  | rate=0.2 | rate=0.4 | rate=0.6 | rate=0.8 |
|----------------------------------|----------|----------|----------|----------|
| Paper Replication (50 sequences) |          | 4-24     |          |          |
| Incorrect Prompt                 |          | 4-24     |          |          |
| Incorrect Images (s-input)       | 5-29     | 6-34     | 6-38     | 7-43     |
| Reduced Example (s-input)        | 3-19     | 2-14     | 2-10     | 1-5      |
| Shuffled (s-input)               | 4-24     |          |          |          |

## C Comparison of Metrics for Different Input Approaches

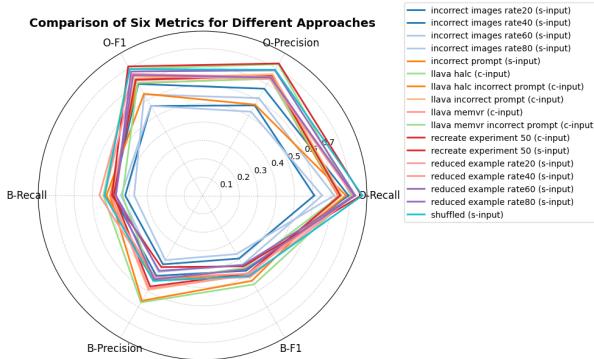


Figure 1: Chart Plot of 6 Metrics

## D All Prompts in Our Paper

Table 4: All Prompts Used in Our Paper

| Prompt   |
|--|
| <b>Tasks:</b> Generate description for the given image sequence & Extract object and behavior keywords<br>see Table 6 in (Wang et al., 2024) |
| <b>Task:</b> Extract object and behavior keywords from Ground Truth<br>inspired from Table 6 in (Wang et al., 2024)                          |

Text: <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. Please directly output the final object and action lists.

Here is an example:

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.

The lists are

Object list: [dog, toy, house]

Action list: [interaction, bark, express intimidation, move, lean backward, stop, look up]

Here is the paragraph:

**{Human-annotated description}.**

The lists are:>

#### **Task: Generate hallucinated object and behavior keywords**

Text: <The given lists are object list and action list consecutively. Please create a hallucinated list that contains object words that are not related with the given object list and a hallucinated list that contains action words that are not related with the given action list. However, I want the hallucinated object list still corresponds with the hallucinated action list, which means the objects in the hallucinated object list can do the actions in hallucinated action list. Also, you must convert the hallucinated actions to their complete root form.

The lists generated must be in list format [item1, item2, item3].

For example the lists are

Object list: [dog, toy, house]

Action list: [interaction, bark, express intimidation, move, lean backward, stop, look up]

The hallucinated lists are:

Object list: [cat, sofa, ball]

Action list: [play, nap, land, turn left]

Here are the given lists:

Object list:

**{object list}**

Action list:

**{action list}.**

The hallucinated lists are: >

## E Image Sample



Figure 2: (Sample) Data with Combined Representation

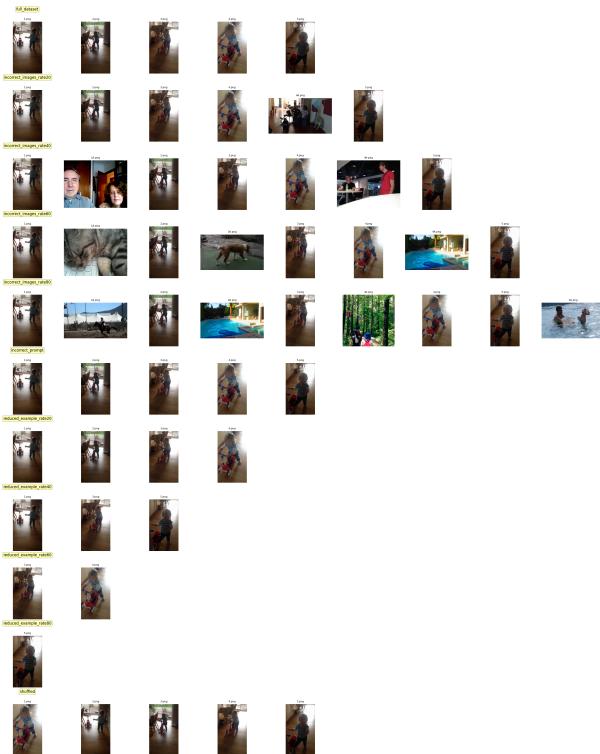


Figure 3: (Sample) Data with Sequential Representation

## F Incorrect Prompt Sample

User: Please create a list of the objects you see in these images.

LVLM: {**incorrect objects**} (eg. [boy, bicycle, gloves, shoes, hand, arm, chair, door])

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

LVLM:

## G Description Generation and Object & Behaviour Extraction Sample

Table 5: Generated Text and The Extracted Object & Action List from GPT

|  | <b>Generated Text (ours)</b>  | <b>Extracted List (GPT)</b>   |
|--|---|---|
| <b>Chat-UniVi (Paper Replication s-input)</b>    | In this image sequence, a young child is seen playing with a toy scooter on a hardwood floor. The child is wearing a striped shirt and riding the scooter, which has a blue design. The child is also holding onto a chair for support while playing. The scene takes place in a room with a dining table and multiple chairs. There are two bowls on the dining table, and a potted plant is placed nearby. The child appears to be having fun and enjoying their playtime with the scooter. | <b>object list (ground truth):</b> ['girl', 'scooter', 'kitchen', 'socks', 'leg', 'handle', 'wall']<br><b>object list (ours):</b> ['child', 'scooter', 'floor', 'shirt', 'chair', 'room', 'table', 'bowls', 'plant']<br><b>action list (ground truth):</b> ['interact', 'approach', 'ride', 'use', 'hold', 'concentrate', 'dismount', 'reach']<br><b>action list (ours):</b> ['play', 'wear', 'ride', 'hold', 'place', 'have fun', 'enjoy'] |
| Chat-UniVi + Incorrect Prompt (s-input)          | In the image sequence, a little girl is seen playing indoors with a toy bicycle. She is wearing shoes and gloves while she rides the bike, possibly enjoying herself in her imaginative play. The scene takes place in a room with a chair and a door visible in the background. The overall atmosphere appears to be one of fun and enjoyment for the little girl as she engages with her toy bicycle.   | <b>object list (ground truth):</b> ['girl', 'kitchen', 'scooter', 'leg', 'handle', 'wall']<br><b>object list (ours):</b> ['girl', 'bicycle', 'shoes', 'gloves', 'chair', 'door']<br><b>action list (ground truth):</b> ['interact', 'approach', 'ride', 'use', 'hold', 'concentrate', 'reach', 'dismount']<br><b>action list (ours):</b> ['play', 'ride', 'engage']   |
| Chat-UniVi + Incorrect Images rate=0.8 (s-input) | In this episode, a man is riding a small bike inside a swimming pool while a child, wearing a life jacket, is standing nearby. The child is also holding a water hose, possibly preparing to spray water at someone or something. The scene suggests a fun and playful atmosphere, with the man on the bike and the child with the water hose engaging in some interactive water play within the pool area.   | <b>object list (ground truth):</b> ['girl', 'kitchen', 'scooter', 'socks', 'leg', 'handle', 'wall']<br><b>object list (ours):</b> ['man', 'bike', 'swimming pool', 'child', 'life jacket', 'water hose']<br><b>action list (ground truth):</b> ['interact', 'approach', 'ride', 'wear', 'use', 'hold', 'direct', 'concentrate', 'reach', 'dismount']<br><b>action list (ours):</b> ['ride', 'stand', 'hold', 'spray', 'engage', 'play']     |
| Chat-UniVi + Reduced Example rate=0.4 (s-input)  | In this image sequence, a young child is playing with a toy scooter inside a kitchen. The child is standing on the blue scooter and pushing it along the hardwood floor. The kitchen features a dining table with chairs, a refrigerator, and a sink. The child appears to be having fun and exploring the environment on their scooter, creating a lively and playful atmosphere in the space.   | <b>object list (ground truth):</b> ['girl', 'scooter', 'kitchen', 'leg', 'handle', 'wall']<br><b>object list (ours):</b> ['child', 'scooter', 'kitchen', 'floor', 'table', 'chairs', 'refrigerator', 'sink']<br><b>action list (ground truth):</b> ['interact', 'approach', 'ride', 'wear', 'move', 'hold', 'concentrate', 'reach', 'dismount']<br><b>action list (ours):</b> ['play', 'stand', 'push', 'explore', 'create']                |

|   |   | <b>Generated Text (ours)</b>  | <b>Extracted List (GPT)</b>   |
|---|---|---|---|
| Chat-UniVi<br>Shuffled (s-input)                  | + | In this image sequence, a young child is playing indoors with a toy scooter. The child is standing on the scooter and pushing it along the floor, enjoying their playtime. The room features a dining table and chairs, and there is a dog present in the scene. The child seems to be having fun and exploring their surroundings with the scooter.  | <p><b>object list (ground truth):</b> ['girl', 'kitchen', 'scooter', 'socks', 'leg', 'handle', 'wall']</p> <p><b>object list (ours):</b> ['child', 'scooter', 'floor', 'dining table', 'chairs', 'dog']</p> <p><b>action list (ground truth):</b> ['interact', 'approach', 'ride', 'wear', 'move forward', 'hold', 'concentrate', 'reach', 'dismount']</p> <p><b>action list (ours):</b> ['play', 'stand', 'push', 'enjoy', 'explore']</p>                    |
| <b>LLaVA1.5 (Paper replication c-input)</b>       |   | The image sequence shows a young girl playing with a toy car on a wooden floor. In each frame, the girl is seen in different positions, such as sitting, standing, and playing with the toy car. The sequence captures her various stages of play and interaction with the toy car, providing a glimpse into her playful activities.  | <p><b>object list (ground truth):</b> ['girl', 'scooter', 'kitchen', 'socks', 'feet', 'leg', 'handle', 'wall']</p> <p><b>object list (ours):</b> ['girl', 'toy car', 'floor', 'frame']</p> <p><b>action list (ground truth):</b> ['capture', 'interact', 'approach', 'wear', 'move forward', 'hold onto', 'direct', 'concentrate', 'ride', 'reach', 'dismount']</p> <p><b>action list (ours):</b> ['show', 'play', 'sit', 'stand', 'capture', 'interact']</p> |
| LLaVA1.5<br>MemVR + Incorrect Prompt<br>(c-input) | + | The image sequence shows a young child playing with a toy bicycle indoors. In the first picture, the child is holding the handlebars of the bicycle, while in the second picture, the child is riding the toy bicycle. The third picture shows the child standing next to the bicycle, and the fourth picture captures the child sitting on the floor with the bicycle. The child is also wearing a striped shirt in the last picture.              | <p><b>object list (ground truth):</b> ['girl', 'scooter', 'kitchen', 'socks', 'feet', 'wall']</p> <p><b>object list (ours):</b> ['child', 'bicycle', 'handlebars', 'floor', 'shirt']</p> <p><b>action list (ground truth):</b> ['interact', 'approach', 'ride', 'use', 'hold', 'concentrate', 'reach', 'dismount']</p> <p><b>action list (ours):</b> ['play', 'hold', 'ride', 'stand', 'sit', 'wear']</p>   |
| LLaVA1.5<br>HALC (c-input)                        | + | The image sequence shows a young child playing with a toy tricycle in a room. In the first picture, the child is standing next to the tricycle, while in the second picture, the child is riding the tricycle. The third picture captures the child falling off the tricycle, and the fourth picture shows the child sitting on the floor next to the tricycle. The sequence captures the child's playful and active moments with the toy tricycle. | <p><b>object list (ground truth):</b> ['girl', 'scooter', 'kitchen', 'socks', 'leg', 'handle', 'wall']</p> <p><b>object list (ours):</b> ['child', 'toy tricycle', 'room', 'floor']</p> <p><b>action list (ground truth):</b> ['interact', 'approach', 'ride', 'wear', 'move', 'hold', 'concentrate', 'reach', 'dismount']</p> <p><b>action list (ours):</b> ['play', 'stand', 'ride', 'fall', 'sit', 'capture']</p>  |