Why LVLMs Are More Prone to Hallucinations in Longer Responses: The Role of Context

Supplementary Material

This supplementary material provides further details on our findings, the specific prompts and configurations used in our experiments, additional quantitative and qualitative results, and a discussion of limitations.

Specifically, we first provide supplementary experimental settings used in our analysis experiments. (Sec. A). Next, we present complementary results to support our analysis (Sec. B). We then describe further implementation details and experimental setups for the experiments in the main paper (Sec. C). Additionally, we conduct ablation studies and evaluate HalTrapper on additional benchmarks to further validate its effectiveness (Sec. D). We also include visualizations to aid comparison and provide a clearer understanding of HalTrapper (Sec. E). Finally, we provide a discussion of the limitations of our work (Sec. F).

A. Supplementary Details on Exploratory Experiments and Analyses

A.1. Settings for Hallucinations Beyond Length

For the experiment of modifying image and text context (Sec. 3.2), since the image cropping experiment requires manual re-annotation of cropped part, we randomly sample 50 images from COCO dataset for this experiment.

A.2. Prompt Design for Completeness

In Fig. 4(a) of the paper, we demonstrate that the model is more prone to hallucinations when its content is incomplete by adjusting the amount of textual context inserted into the model. To eliminate the influence of length, we designed prompts of different lengths for different groups, ensuring that the total number of sentences in each prompt remains consistent (4 here). Although the prompt lengths varied in our design, we endeavored to maintain consistency in the information contained within them as much as possible. Below are the specific prompts we used, where {} are placeholders for sentences to be inserted:

- Group **w/o sentence**: *Please help me describe this image in detail. I'd like to hear more about it, even if it's just small things. Anything you can say about it would be useful in some way. It doesn't have to be important, just whatever comes to mind.*
- Group **+1 sentence**: *I already know that* {} *Could you describe any other details of the image for me? It doesn't have to be anything specific, just whatever else you can say about it. Even if it seems unimportant, it might still be worth mentioning.*
- Group **+2 sentences**: *I already know that* {} *Could you describe any other details of the image for me? Maybe there's something that hasn't been mentioned yet, or just anything that comes to mind.*

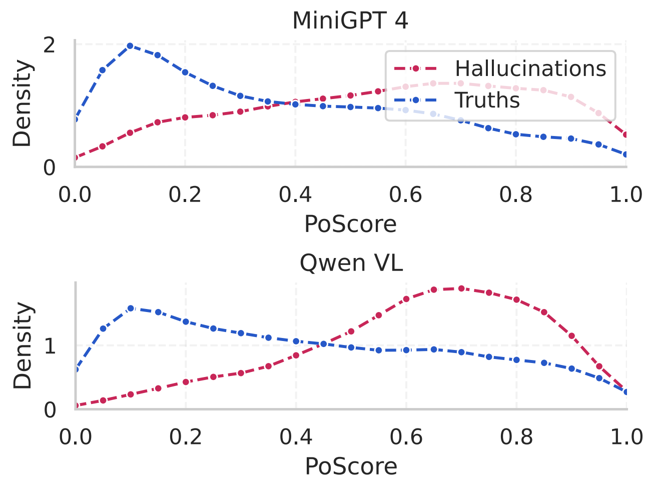

Figure 7. Distribution of hallucinated and non-hallucinated object positions in captions generated by different models.

B. Additional Results for Exploratory Experiments and Analysis

B.1. Additional Baseline Results for Hallucinations Linked to Length

We conduct the same experiments as Sec. 3.1 on Qwen VL Chat and MiniGPT-4. Results are shown in Fig. 7. The results demonstrate that Qwen VL Chat and MiniGPT-4 also show a pronounced tendency for increased hallucinations with longer input contexts.

B.2. Qualitative Support for Statistical Analysis

In the main paper, we conduct a series of statistical experiments to demonstrate that hallucinations in LVLMs are not solely related to input length, but also influenced by coherence and completeness. To facilitate understanding, we provide qualitative examples of the experiments here.

Illustrations for Hallucination Beyond to Length. Fig. 8 presents an example from the experiment described in Fig. 2 of the main text. It can be observed that both cropping the image and enriching the prompt lead to earlier occurrences of hallucinations.

Illustrations for Coherence Analysis. Fig. 9 supplements the visualization on the right side of Fig. 3 with a complete

Model	θ_{IG}	θ_{EE}	N	T_{sep}
LLaVA v1.5 7B	0.75	1	10	‘ ’
MiniGPT 4	0.75	0	10	‘, ’
Qwen VL Chat	0.85	0	5	‘ ’
Qwen2 VL 7B	0.75	1	5	‘ ’
Janus Pro 7B	0.75	1	5	‘ ’

Table 5. Parameters used for hallucination suppression.

example, illustrating that hallucinated pairs exhibit significantly higher attention similarity scores.

Illustrations for Completeness Analyses. Fig. 10 and Fig. 11 visualize specific examples from the two experiments shown in Fig. 4(a) and (b) of the main text, respectively. Fig. 10 further demonstrates that hallucinations tend to appear earlier when more visual context is included, while Fig. 11 shows that similar hallucinations consistently emerge despite variations in prompts.

C. Detailed Implementation and Experimental Setup

C.1. Details of Datasets and Benchmarks.

COCO [44], the Common Objects in Context dataset is widely used in computer vision, providing detailed annotations for 80 object categories and serving as a valuable resource for evaluating hallucination detection and suppression.

AMBER [73], an LLM-free multi-dimensional benchmark, is also specifically designed to assess hallucinations in LVLMs. With 1004 images and more comprehensive annotations than COCO, AMBER enables the detection of hallucinations beyond the 80 COCO categories, offering a broader evaluation scope.

C.2. Prompt Design for EEScore

For hallucination detection, we employ a “reason-then-imagine” prompt to derive both the imagination and reasoning sets used in the computation of EEScore (Sec. 5.1.2). The specific prompt are presented as follows:

> Based on this image, please imagine what object might be in the {direction} outside the frame, and explain why. Specifically, your response should follow the following format:
>
> Imagination: <one imaginary object here>
> Reason: The image features <briefly describe this image, be careful to mention all objects related to your imagination>, which suggests that <your imagination here>.

C.3. Construction and Insertion of Contrastive Contextual Tokens (CCT)

After identifying the potential hallucinated objects $S_{induction}$ as described in the paper, we construct CCT by first truncating or padding the elements in this set to a fixed length N, yielding a new set S', and then encoding them using a text encoder.

Specifically, when $|S_{induction}| > N$, *i.e.* the number of elements in the potential hallucinated objects set exceeds N, the set is truncated based on the priority of each element, with the lowest-priority elements being removed. The priority assignment is determined as follows:

- If both elements are sourced from IG, the element exhibiting the higher similarity in attention score is assigned higher priority.
- If one element originates from IG and the other from EE, the element from IG is given precedence.
- If both elements are sourced from EE, they are deemed to have equal priority, and removal is determined by a random selection process.

On the other hand, when $|S_{induction}| < N$, we randomly select additional *unrelated* objects from a predefined object list to include in the set. Objects that have never appeared in our pipeline before, including the caption and EE responses, are considered unrelated.

To derive the CCT from S', we first concatenate all elements of S' into a single string using a predefined separator T_{sep}. This ensures a structured and well-defined representation for encoding:

$$T = s_1\, T_{sep}\, s_2\, T_{sep}\, \ldots\, T_{sep}\, s_N, \quad \text{where } s_i \in S'.$$

Finally, we apply the text encoder ϕ to generate the corresponding text embedding for the modified set S', which can be formally expressed as:

$$x_{cct} = \phi(T).$$

For the insertion of the CCT, we place it in the contrastive decoding branch immediately after image tokens.

C.4. Hyperparameters for Induction and Suppression

Hyperparameters for Induction. We consistently use greedy decoding when generating hallucination candidates. For the EE metric, we employed $|\mathcal{D}| = 8$. The directions are: “*top*”, “*bottom*”, “*left side*”, “*right side*”, “*top left corner*”, “*top right corner*”, “*bottom left corner*”, and “*bottom right corner*”.

Hyperparameters for Suppression. Across all experiments, the model is prompted with the instruction: “*Please help me describe the image in detail.*” to generate captions. For nucleus sampling, we set the temperature to 1.0 and top_p to 1.0. In beam search, we used a beam size of 5. We employed nucleus sampling when evaluating AMBER. For all suppression experiments, we adapt different hyperparameters for different models (See Table 5.)

EE	IG	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$	Precision	Recall	F1	Len
		58.6	18.8	68.1	76.4	72.0	105.2
✓		51.0	14.4	73.9	77.1	75.5	102.4
	✓	50.4	14.9	74.7	76.6	75.6	100.3
✓	✓	48.6	14.5	74.6	77.7	76.1	100.9

Table 6. Ablation study on CHAIR with LLaVA v1.5 7B

Dataset	Setting	+ours	Acc.↑	Prec.	Recall	F1↑
MSCOCO	*Random*	✗	85.0	97.5	71.8	82.7
		✓	**86.3**	**98.7**	**73.6**	**84.3**
	Popular	✗	81.7	89.5	71.9	79.7
		✓	**83.3**	**91.4**	**73.4**	**81.4**
	Adversarial	✗	80.5	86.8	72.1	78.7
		✓	**81.5**	**87.6**	**73.4**	**79.9**
A-OKVQA	*Random*	✗	78.8	96.3	59.9	73.9
		✓	**79.4**	**97.1**	**60.6**	**74.6**
	Popular	✗	76.1	88.5	60.0	71.5
		✓	**76.9**	**89.5**	**61.0**	**72.6**
	Adversarial	✗	72.5	80.2	59.9	68.5
		✓	**73.9**	**82.7**	**60.5**	**69.9**
GQA	*Random*	✗	75.5	94.1	54.4	58.9
		✓	**76.3**	**95.0**	**55.5**	**70.0**
	Popular	✗	71.2	82.0	54.3	65.3
		✓	**71.7**	**82.1**	**55.5**	**66.2**
	Adversarial	✗	69.6	78.1	54.5	64.2
		✓	**70.2**	**78.6**	**55.5**	**65.1**

Table 7. Results on POPE with LLaVA v1.5 7B. Acc. stands for accuracy, and prec. stands for precision. Higher scores indicate better performance and fewer hallucinations.

D. Supplementary Experiments for Suppression

Unless otherwise specified, all experimental results in this chapter are based on the LLaVA v1.5 7B model.

D.1. Ablation Study

In Table 6, we conduct an ablation study on the CHAIR benchmark to assess the contributions of different components in HalTrapper, namely External Expansion (EE) and Internal Grounding (IG). The baseline model without EE or IG achieves a CHAIR_S score of 58.6% and a CHAIR_I score of 18.8%. When adding EE alone, CHAIR_S reduces significantly to 51.0%, while CHAIR_I decreases to 14.4%. Precision improves to 73.9%, Recall to 77.1%, and F1 to 75.5%, indicating a clear enhancement in reducing hallucinations and improving response quality. Incorporating IG alongside EE further decreases CHAIR_S to 50.4% and slightly raises CHAIR_I to 14.9%, showing that IG helps maintain high response quality with moderate gains in hallucination reduction. Finally, using both EE and IG achieves the best results, with CHAIR_S and CHAIR_I reduced to 48.6% and 14.5%, respectively. These findings confirm that the combination of EE and IG maximizes performance by effectively balancing precision, recall, and hallucination reduction, achieving the highest overall reliability and accuracy in the responses.

MM-Vet gen. subset	Baseline	Ours
LLaVA v1.5 7B	23.2	25.5
Qwen VL Chat	30.7	31.1

Table 8. Results on MM-Vet [89] generation subset.

D.2. Additional Experiments on Adapted POPE

POPE [40], the Polling-based Object Probing Evaluation (POPE) is aimed at evaluating hallucinations in LVLMs. In a manner similar to the CHAIR benchmark, POPE addresses object hallucinations by querying the model with prompt "Is there a/an {object} in the image?" to determine whether the model can correctly identify specific objects within images. The full POPE evaluation consists of three distinct subsets: the "random" subset, which tests objects randomly chosen from the dataset; the "popular" subset, which focuses on commonly occurring objects; and the "adversarial" subset, which challenges the model's ability to identify objects that are closely related to those actually present in the image.

Different from the general POPE evaluation pipeline, since our method is specifically designed for hallucinations in the context of long text, we adapted it's pipeline by reframing it as an image captioning task. Specifically, we first prompt the model to generate a detailed caption for each image and subsequently use the GPT-4o-mini model to assess whether the specified queried object appears in the caption. We have retained POPE's original evaluation metrics, such as recall and F1 score.

Results. The results in Table 7 demonstrate that HalTrapper consistently enhances performance across all settings and datasets. For instance, on the MSCOCO [45] dataset, HalTrapper achieves up to a 1.7% improvement in F1 score in the "popular" setting, increasing from 79.7% to 81.4%. Similarly, on the A-OKVQA [60] dataset, the model shows a gain of 1.4% in the "adversarial" setting (from 68.5% to 69.9%). On the GQA [26] dataset, the method delivers substantial improvements, with the F1 score increasing by 1.3% in the "popular" setting (from 65.3% to 66.2%). These consistent gains highlight the effectiveness of HalTrapper in addressing hallucinations across various object recognition scenarios.

D.3. Additional Experiments on MM-Vet

MM-Vet [89] is a benchmark designed to evaluate the response quality of LVLMs on complex multi-modal tasks. Questions in MM-Vet requires models to integrate multiple core capabilities. Given that our HalTrapper is designed for long response scenarios, we evaluate only the subset of MM-Vet questions that are explicitly annotated as assessing

GPT-4o Prompt
You are required to score the performance of three AI assistants in describing a given image. You should pay extra attention to the hallucination, which refers to the part of descriptions that are inconsistent with the image content, such as claiming the existence of something not present in the image or describing incorrectly in terms of the counts, positions, or colors of objects in the image. Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria: 1: Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinations should be given higher scores. 2: Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not count as necessary details. 3: Fluency: whether the response sound natural and well-phrased. Responses that avoid excessive repetition and awkward phrasing should receive higher scores. Please output the scores for each criterion, containing only three values indicating the scores for Assistant 1, 2 and 3, respectively. The three scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment. [Assistant 1] {} [End of Assistant 1] [Assistant 2] {} [End of Assistant 2] [Assistant 3] {} [End of Assistant 3] Output format: Accuracy: <Scores of the three answers> Reason: Detailedness: <Scores of the three answers> Reason: Fluency: <Scores of the three answers> Reason:

Table 9. The prompt used for GPT-4o evaluation.

language generation and report the score. The evaluation is conducted using their official online evaluator.

Results. Table 8 presents the performance of our HalTrapper compared to the baseline on the MM-Vet [89] generation subset on both LLaVA v1.5 7B and Qwen VL. It can be observed that our HalTrapper achieves consistent improvements across two different models.

D.4. Additional Results of GPT-4o Assisted Evaluation

Since the CHAIR metric only evaluates object-level hallucinations while ignoring other types, such as colors and numbers, following prior work [25, 51], we adapt GPT-4o [1] for a more comprehensive assessment. GPT-4o's ability to perceive and interpret images allows it to evaluate hallucinations in longer responses, closely aligning with expert human judgment. Unlike previous studies that focused only on accuracy and detailedness, we expand the evaluation to include fluency, recognizing its importance in language generation. Specifically, we sample 50 images from COCO and prompt GPT-4o to score each generated text on a scale of 1-10. The exact prompt used is provided in Table 9.

GPT Eval	Baseline	PAI	Ours
Hal avg score	6.06	6.15	6.12
Det avg score	6.18	5.47	6.38
Flu avg score	7.56	7.38	7.59

Table 10. Comparison between PAI [51] and our HalTrapper on GPT-4o evaluation using the COCO [45] dataset with LLaVA v1.5 7B.

Results. Table 10 presents a comparison between our method and PAI [51] in three evaluation dimensions using GPT: hallucination (Hal), detail (Det), and fluency (Flu). Our findings indicate that PAI currently leads in terms of reducing hallucinations and providing detailed responses. However, we observed that PAI often attempts to repeat content in order to influence GPT's evaluation, leading to inflated Hal and Det scores that do not necessarily reflect genuine response quality. To address this, we introduced an additional Flu score to more comprehensively assess response quality and hallucination levels alongside Hal and Det scores. Our method achieves significantly more detailed and coherent text responses while maintaining a hallucination level comparable to that of PAI.

E. Qualitative Results for Suppression

E.1. Comparison with PAI

Although PAI [51] demonstrates superior performance on hallucination benchmarks, its approach of directly enhancing attention scores adversely affects the model's language generation capabilities. Specifically, after applying the PAI method, LVLMs tend to produce redundant information. This issue is illustrated in Table 10, which presents evaluations using GPT-4o. We also present illustrative examples provided in Fig. 12.

We observe that PAI poses a risk of redundantly repeating image content when generating descriptions. For instance, details such as "boats docked at the harbor," "a red and white boat, a blue and white boat, and a blue and white ship," and "some boats are closer to the shore" are frequently reiterated across consecutive sentences. This redundancy compromises the coherence and logical structure of the generated output. In contrast, our model effectively mitigates such hallucinations, such as "a few people", while maintaining both the logical consistency and content integrity of the description.

E.2. Qualitative Results of Our HalTrapper

We provide additional visualizations to further demonstrate the effectiveness of our method, as shown in Fig. 13 and 14.

These results highlight the effectiveness of our proposed method. Specifically, the hallucinated objects generated by IG exhibit a notable overlap with the ground truth hallucinations in the caption, while our Contrastive Contextual Decoding (CCD) process effectively mitigates these hallucinations. In contrast, considering the issue of false positives, EE avoids the direct incorporation of hallucinated objects in captions. However, it still contributes to hallucination suppression. As demonstrated in the final example of Fig. 14, even though EE does not directly include the object "person," it extracts a latent, hallucinated object "cell phone," which is closely related to "person," thereby preventing the model from hallucinating "person."

F. Limitations

This work primarily addresses object-level hallucinations in long-form responses generated by large LVLMs. However, LVLMs are susceptible to a broader spectrum of hallucinations, including failures in instruction following and hallucinations at the attribute and relational levels. Moreover, our evaluations are mainly on image captioning benchmarks such as CHAIR and AMBER. While these benchmarks are widely used for evaluating hallucinations, they do not adequately cover more open-ended generative scenarios. Developing more comprehensive and standardized benchmarks for such settings represents a valuable direction for future research.

Figure 8. Illustrative example of hallucination positions under context modifications, corresponding to the mechanism shown in Fig. 2. Both cropping the image and enriching the prompt lead to earlier hallucination occurrences. Hallucinations are highlighted in **red**.

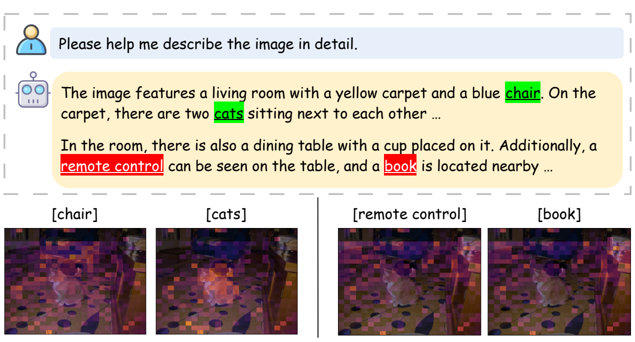

Figure 9. Illustrative example of attention similarity between hallucinated and non-hallucinated object pairs within the same caption, corresponding to the mechanism shown in Fig. 3. Hallucinated pairs exhibit significantly higher attention similarity scores. Hallucinations are highlighted in **red**.

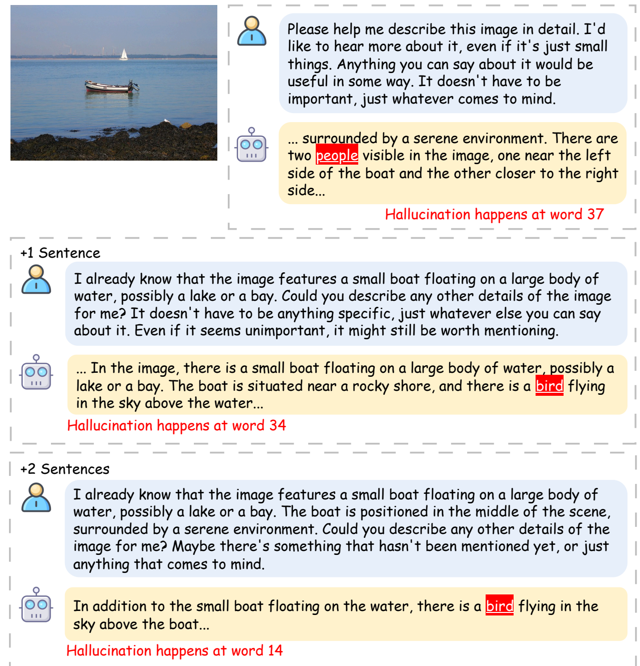

Figure 10. Illustrative example of the positions where hallucinations occur as the visual information in prompts gradually increases, while the prompt length remains similar, corresponding to the mechanism shown in Fig. 4(a). Hallucinations tend to appear earlier when more visual context is included. Hallucinations are highlighted in **red**.

Figure 11. Illustrative example of the repetition frequency distribution of hallucinated objects across different prompts for the same image, corresponding to the mechanism shown in Fig. 4(b). Similar hallucinations consistently appear despite changes in prompts. Hallucinations are highlighted in **red**.

LLaVA-v1.5-7b

The image features a large body of water with a marina filled with numerous boats of various sizes. The boats are docked in the harbor, creating a picturesque scene. The boats are arranged in a way that they are visible from different angles, showcasing the diversity of the vessels.

In addition to the boats, there are a few people scattered throughout the scene, likely enjoying the view or attending to their boats. The overall atmosphere of the image is serene and inviting, with the boats and people creating a sense of community and leisure.

Please help me describe the image in detail.

LLaVA-v1.5-7b + PAI

The image features a large body of water with numerous boats docked at the harbor. There are several boats of various sizes and colors, including a red and white boat, a blue and white boat, and a blue and white boat. The boats are docked in the water, with some of them being closer to the shore.

The scene is set in a harbor with numerous boats of different sizes and colors, including a red and white boat, a blue and white boat, and a blue and white boat. The boats are docked at the harbor, with some of them being closer to the shore. The scene is a beautiful representation of a marina filled with various boats.

LLaVA-v1.5-7b + Ours

The image features a large body of water with a marina filled with numerous sailboats. The boats are docked in a harbor, with some of them positioned closer to the shore and others further out in the water. The sailboats vary in size and are arranged in a way that fills the entire marina.

The scene is picturesque, with the boats floating on the water and the sky visible above. The marina is bustling with activity, showcasing the popularity of sailing in this area.

Figure 12. A qualitative comparison between baseline, PAI and our HalTrapper. Hallucinations are highlighted in **red** and the repetitions are highlighted in **blue**.

Please help me describe this image in detail.

LLaVA-v1.5-7b

The image features a picturesque scene of two horses grazing in a grassy field near the ocean. The horses are positioned close to each other, with one horse on the left side and the other on the right side of the field. The field is situated next to a lighthouse, which stands tall and majestic in the background.

In addition to the horses and the lighthouse, there are a few other elements in the scene. A car can be seen parked in the background, and a person is visible near the left side of the image. The combination of the horses, the lighthouse, and the ocean creates a serene and peaceful atmosphere.

LLaVA-v1.5-7b + Ours

The image features a picturesque scene of two horses grazing in a lush green field near the ocean. The horses are positioned close to each other, with one horse on the left side and the other on the right side of the field. The field is filled with grass, providing an ideal environment for the horses to graze.

In the background, there is a lighthouse, adding a sense of depth and beauty to the scene. The lighthouse stands tall and proud, overlooking the serene landscape where the horses are enjoying their time.

IG: car, person
EE: boat

Please help me describe this image in detail.

LLaVA-v1.5-7b

The image features a museum with a large clock mounted on the wall. The clock is positioned in the center of the scene, and it appears to be a prominent feature of the museum. The clock is surrounded by several signs, which are likely providing information about the exhibits or rules for visitors.

In addition to the clock and signs, there are a few other objects in the scene. A chair is located near the bottom left corner of the image, and a bottle can be seen on the right side of the scene. The overall atmosphere of the museum suggests that it is a place for visitors to learn and explore.

IG: chair, bottle
EE: horse

LLaVA-v1.5-7b + Ours

The image features a large wooden clock mounted on a wall, surrounded by several signs. The clock is positioned in the center of the scene, with the signs placed around it. The clock is quite prominent, taking up a significant portion of the wall space.

The signs surrounding the clock are likely related to the museum or the exhibit, providing information or instructions to visitors. The combination of the clock and the signs creates an informative and engaging atmosphere for those visiting the museum.

Please help me describe this image in detail.

LLaVA-v1.5-7b

The image features a cozy bedroom with a bed situated under a window. The window is adorned with wooden shutters, and it is open, allowing natural light to enter the room. The bed is covered with a plaid blanket, adding a touch of warmth and comfort to the space.

In the room, there is also a chair positioned near the bed, providing additional seating or a place to rest. The overall atmosphere of the room is inviting and relaxing, perfect for a peaceful night's sleep.

LLaVA-v1.5-7b + Ours

The image features a cozy bedroom with a bed situated under a window. The window is adorned with white wooden shutters, allowing natural light to enter the room. The bed is covered with a plaid quilt, adding a touch of warmth and comfort to the space. The room appears to be in a cabin, giving it a rustic and inviting atmosphere.

IG: chair
EE: bird

Figure 13. Examples of hallucination suppression using our HalTrapper. Hallucinations are highlighted in **red**.

Please help me describe this image in detail.

LLaVA-v1.5-7b

The image features a group of people playing frisbee on a grassy field. A man in a blue shirt is standing over a woman who is kneeling on the ground, possibly teaching her how to play. The man is holding a white frisbee, and the woman is looking up at him, attentively listening to his instructions.

There are several other people in the scene, some of them standing and others sitting on the grass. A backpack can be seen placed on the ground, possibly belonging to one of the players. The overall atmosphere appears to be friendly and fun, with everyone enjoying the outdoor activity.

LLaVA-v1.5-7b + Ours

The image features a group of people playing frisbee on a grassy field. A man in a blue shirt is standing over a woman who is sitting on the ground, possibly taking a break or waiting for her turn to play. The man appears to be teaching her how to play frisbee, as he is holding a frisbee in his hand.

There are several other people in the scene, some of them standing and others sitting on the ground, all engaged in the game or observing the action. The atmosphere seems to be lively and fun, with everyone enjoying the outdoor activity.

IG: backpack
EE: (empty)

IG: car
EE: (empty)

Please help me describe this image in detail.

LLaVA-v1.5-7b

The image features a young girl standing in a grassy field, holding a colorful kite with a striped pattern. She is smiling and appears to be enjoying her time flying the kite. There are several other people in the background, some of them standing closer to the girl and others further away.

In the scene, there is also a car visible in the background, likely parked near the field where the girl is flying her kite.

LLaVA-v1.5-7b + Ours

The image features a young girl standing in a grassy field, holding a colorful kite with a striped pattern. She is smiling and appears to be enjoying her time flying the kite. There are several other people in the background, some of them standing closer to the girl and others further away.

In the scene, there is also a carousel visible in the background, adding to the lively atmosphere of the park. The girl's kite is the main focus of the image, capturing the essence of a fun and carefree day spent outdoors.

Please help me describe this image in detail.

LLaVA-v1.5-7b

The image features a white teddy bear sitting on top of a camera, which is placed on a bed. The teddy bear appears to be posing for a picture, as if it's holding the camera. The camera is positioned in the center of the scene, with the teddy bear sitting on top of it.

In addition to the teddy bear and camera, there is a person partially visible in the background, likely taking the picture or observing the scene. The overall atmosphere of the image is playful and lighthearted, as the teddy bear and camera create a fun and whimsical composition.

LLaVA-v1.5-7b + Ours

The image features a white teddy bear sitting on top of a camera, which is placed on a bed. The teddy bear is positioned in the center of the camera, creating a playful and cute scene. The camera appears to be a Nikon model, and it is placed on a blue surface, possibly a blanket or a sheet. The overall atmosphere of the image is warm and inviting, with the teddy bear and camera creating a cozy and nostalgic scene.

IG: (empty)
EE: cell phone

Figure 14. Examples of hallucination suppression using our HalTrapper. Hallucinations are highlighted in **red**.