A AUTOMATIC FINE-GRAINED HALLUCINATION ANNOTATION PIPLINE (AFHA)

The existing multimodal hallucination research suffers from a lack of large-scale datasets with fine-grained annotations specific to hallucinations. To address this issue, we design an automatic fine-grained hallucination annotation pipline featuring annotations for four hallucination types and specific hallucination content. **Data Annotation.** We annotated image-text paired data using the GPT-4 prompt method. We initially established a rigorous definition for various types of hallucinations. Building upon this groundwork, we engaged GPT-4 to rephrase the collated image-text pairs in line with the diverse classifications of hallucinations. This step involved inject distinctive hallucinatory elements into the original captions. An example of a prompt designed to generate 'object' annotations is illustrated in Figure 8(b). The outcome of this procedure was a collection of image descriptions enriched with specified hallucination categories. Moreover, we delegated to GPT-4 the responsibility of concocting specific hallucinatory content. Consequently, this strategy yielded an extensively annotated dataset facilitated by GPT-4, with samples of annotations across different types of hallucinations displayed in Figure 8(c). For more details, please refer to the Appendix C.2. **Data Filtering.** Following the initial annotation phase, we identified that the quality of the labeled data remained unsatisfactory. Random sampling revealed that approximately 30% of the annotated dataset still harbored noise that failed to meet our stringent labeling criteria. Hence, we proceeded to craft a tailored prompt to commission GPT-4 for the task of purging and refining the noisy annotations, a process thoroughly outlined in the Appendix C.3. Subsequent to GPT-4's meticulous cleanup operation, a manual verification process ascertained that over 97% of the data accorded with the stipulated annotation standards.

B DATASET ANALYSIS FOR HAL-DATA

Name	COCO	LCS	ShareGPT4V
Hal-Data 130K	40%	40%	20%
Hal-Data 2M	0%	100 %	0%

Table 9: The proportion of data sources in Hal-Data 130K and Hal-Data 2M.

B.1 Data source

For the compilation of the Hal-Data 130K dataset, data were sourced from several established datasets, including: (1) COCO, (2) Conceptual Captions (CC), (3) SBU, (4) LAION, and (5) Share-GPT4. Table 9 and Table 10 provides the detailed distribution ratio of these sources. In the scale-up process of Hal-Data to 2 million entries, priority was given to incorporating 558,000 data points from the LLaVA dataset, which constitutes a curated collection derived from LAION, CC, and SBU. The remaining 1400K entries were likewise obtained from this trio of datasets, ensuring a consistent and diverse array of data for robust model training. As shown in Table 10, we also compared with other hallucinatory datasets like M-HalDetect[15] and HaELM [48].

C EXPERIMENTS SETTINGS

C.1 Training Setting for Hal-Evaluator and Hal-VL

We followed the original approach of LLaVA 1.5 [29], we used the complete pre-training dataset of LLaVA 1.5 during the first stage of pre-training. We also keeping the same hyperparameter settings with LLaVA 1.5 . Our experiments were conducted using 16 NVIDIA A100 GPUs with 80G of memory. We used Deepspeed [39] for Hal-Evaluator and Hal-VL, with a batch size of 64 on a single GPU.

C.2 Setting for GPT-4 Annotation of AFHA

As shown in the Figure 10, we present the prompts used during the annotation phase for the Hal-Data-130K dataset. We employed a unified prompt template for different types of hallucinations, only modifying the definition of hallucination within it. This standardized approach ensures consistency across the dataset annotations. Through this methodology, each instance of potential hallucination—be it an object, event, or relationship that does not exist within the image—is flagged with a corresponding definition tailored to the type of discrepancy encountered. This provides a structured framework for subsequently evaluating the frequency and nature of hallucinations produced by LVLMs when generating image descriptions.

C.3 Setting for GPT-4 Filtering of AFHA

Figure 11 provides a clear overview of the designed prompts that are instrumental in discerning and eliminating aberrant data points classified as noise within the hallucination annotations. These prompts are meticulously crafted to navigate through the complex nuances of linguistic and contextual interpretations that the model may generate, systematically identifying and separating those annotations that do not adhere to the objective standards of our dataset. This filtration process is crucial for maintaining the integrity and reliability of our annotations, thereby ensuring that subsequent analyses are based on high-quality, noise-free data.

C.4 Experiment Setting of Chain-of-Thought for Discriminative Evaluation.

In order to test the impact of Chain-of-Thought on discriminative evaluation, we have made a simple modification to the discriminative prompt template. The new prompt template is as follows:

<Image> I

Caption: $C \in \{C^T, C^O, C^R, C^E, C^A\}$.

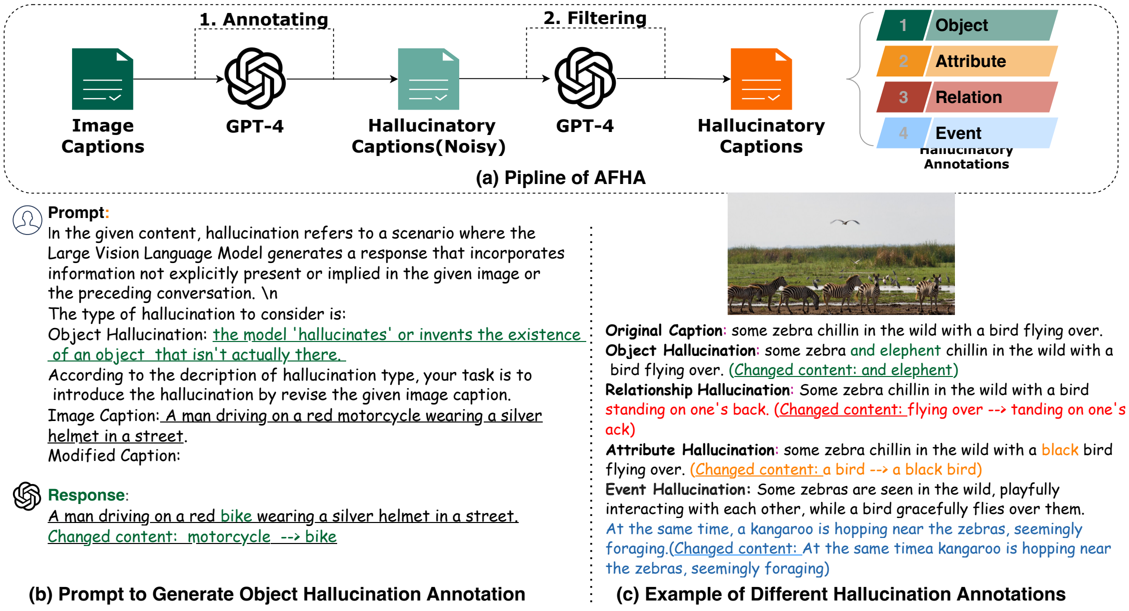
Question: Does the description in the caption accurately reflect the content of the image?

Please conduct a step-by-step analysis of the image and its associated caption, thereafter providing an answer to the query.

D MORE EXPERIMENTS

D.1 More result of Hal-VL.

In this section, we provide detailed evaluations of Hal-VL on more general benchmarks and Hal-Eval (Table **??** and Table 14). As



(a) Pipline of AFHA

Prompt:
In the given content, hallucination refers to a scenario where the Large Vision Language Model generates a response that incorporates information not explicitly present or implied in the given image or the preceding conversation. \n
The type of hallucination to consider is:
Object Hallucination: the model 'hallucinates' or invents the existence of an object that isn't actually there.
According to the decription of hallucination type, your task is to introduce the hallucination by revise the given image caption.
Image Caption: A man driving on a red motorcycle wearing a silver helmet in a street.
Modified Caption:

Response:
A man driving on a red bike wearing a silver helmet in a street.
Changed content: motorcycle --> bike

(b) Prompt to Generate Object Hallucination Annotation

Original Caption: some zebra chillin in the wild with a bird flying over.
Object Hallucination: some zebra and elephant chillin in the wild with a bird flying over. (Changed content: and elephent)
Relationship Hallucination: Some zebra chillin in the wild with a bird standing on one's back. (Changed content: flying over --> tanding on one's back)
Attribute Hallucination: some zebra chillin in the wild with a black bird flying over. (Changed content: a bird --> a black bird)
Event Hallucination: Some zebras are seen in the wild, playfully interacting with each other, while a bird gracefully flies over them. At the same time, a kangaroo is hopping near the zebras, seemingly foraging.(Changed content: At the same timea kangaroo is hopping near the zebras, seemingly foraging)

(c) Example of Different Hallucination Annotations

Figure 8: The sub-figure (a) illustrates the pipeline of AFHA. The lower left sub-figure (b) visualizes the prompt used for generating Object Hallucination Annotations (Please refer to Appendix C.2 for the prompt of other hallucination annotations.), while the lower right sub-figure visualizes examples of annotations for four types of hallucinations.

Name	Data Source	Visible	Annotated by	Samples
M-HalDetect	COCO	✓	Human	4K
HaELM	COCO	✗	Human,GPT3.5	15K
Hal-Data 130K	LCS, COCO, ShareGPT4V	✓	GPT4	130K
Hal-Data 2M	LCS	✓	Hal-Annotator	**1958K**

Table 10: Comparison of hallucinatory datasets and Hal-Data. 'LCS' abbreviates the LAION [41], CC [6], and SBU [37] datasets. The 'Visible' column denotes the image visibility during captioning, and the last column shows the average character number of the caption.

Method	**Overall Score ↑**	**Hallucination Rate ↓**	**Score in Each Question Type ↑**							
			Attribute	Adversarial	Comparison	Counting	Relation	Environment	Holistic	Other
LLaVA-RLHF$_{7B}$ [44]	2.05	0.68	2.92	1.83	**2.42**	1.92	2.25	2.25	1.75	1.08
LLaVA$_{7B}$ [31]	1.55	0.76	1.33	0.00	1.83	1.17	2.00	2.58	1.67	1.83
LLaVA$_{7B}$-Hal-Evaluator	2.12 (↑ 0.56)	0.60 (↓ 0.16)	2.78	2.14	2.12	1.79	1.89	2.32	1.63	1.61
miniGPT-4$_{7B}$ [31]	1.39	0.71	0.75	1.83	2.16	0.91	1.25	1.33	0.91	1.91
miniGPT-4$_{7B}$-Hal-Evaluator	1.89 (↑ 0.40)	0.56 (↓ 0.15)	1.23	2.05	2.43	1.84	2.21	2.38	1.13	1.88

Table 11: Evaluation results for different MLLMs on MMHal-Bench.

Model	**In-domain**					**Out-of-domain**					**Length**
	Object Ratio	Relation Ratio	Attribute Ratio	Event Ratio	**Acc**	Object Ratio	Relation Ratio	Attribute Ratio	Event Ratio	**Acc**	
LLaVA1.5	23.7	58.8	10.6	7.0	55.7	30.0	48.4	11.6	10.2	49.5	10.3
LLaVA1.5	42.2	13.0	3.6	41.4	44.6	34.6	8.8	2.7	54.3	46.4	84.5
Hal-VL	35.0	37.5	21.5	6.2	70.9	29.4	30.4	20.4	10.1	60.4	10.5
Hal-VL	31.3	22.9	23.8	22.2	64.1	27.8	17.7	22.2	32.5	56.9	85.5

Table 12: Generative Hallucination Evaluation for Hal-VL and LLaVA 1.5.

Method	MME	MMBench	MM-Vet	SEED-Bench
BLIP-2 [22]	1293.84	-	22.4	46.4
mPLUG-Owl [50]	967.34	46.6	-	34.0
InstructBLIP [10]	1212.82	36.0	26.2	53.4
Otter [21]	1292.26	48.3	24.6	32.9
Qwen-VL-Chat [4]	1487.58	60.6	-	58.2
LLaVA [30]	502.82	36.2	28.1	33.5
MiniGPT-4 [53]	581.67	23.0	22.1	42.8
LLaVA-1.5 [28]	1510.70	64.3	30.5	58.6
Hal-VL	**1620.10**	**66.3**	**31.3**	**59.4**

Table 13: Zero-shot multi-modal evaluation on multi-modal benchmarks including MME [12], MMBench [33], MM-Vet [51], SEED-Bench [20]. The overall scores are reported for evaluation. For MMBench, we report test results.

Model	**In-domain**					**Out-of-domain**					**Length**
	Object Ratio	Relation Ratio	Attribute Ratio	Event Ratio	**Acc**	Object Ratio	Relation Ratio	Attribute Ratio	Event Ratio	**Acc**	
LLaVA1.5	23.7	58.8	10.6	7.0	55.7	30.0	48.4	11.6	10.2	49.5	10.3
LLaVA1.5	42.2	13.0	3.6	41.4	44.6	34.6	8.8	2.7	54.3	46.4	84.5
Hal-VL	35.0	37.5	21.5	6.2	70.9	29.4	30.4	20.4	10.1	60.4	10.5
Hal-VL	31.3	22.9	23.8	22.2	64.1	27.8	17.7	22.2	32.5	56.9	85.5

Table 14: Generative Hallucination Evaluation for Hal-VL and LLaVA 1.5.

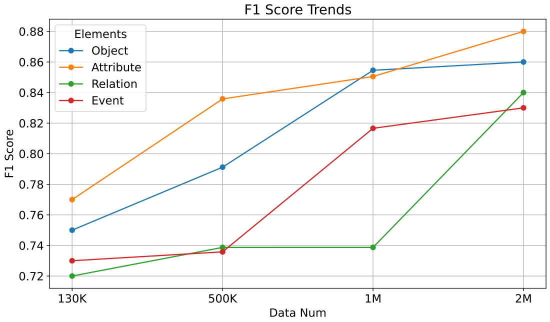


Figure 9: Evaluation results on Hal-Eval-Dis after fine-tuning Hal-Evaluator using SFT data of varying sizes. We have calculated the average F1 value across all types of hallucinations.

indicated in the Table 8, we found that Hal-VL achieved significant advantages on the vast majority of these benchmarks. This demonstrates its robustness and generalizability across different tasks and datasets, underlining its effectiveness in handling hallucination problems in LVLMs.

D.2 Analysis of Different Types of Hallucinations Based on GPT-4

We further investigated the proportion of different types of hallucinations present within the output of LVLMs. As illustrated in Figure 2, we collected 5,000 image-caption pairs from COCO [26]. These were described by mPLUG-owl [50] and LLaVA [30], respectively. Subsequently, we provided both the ground truth image descriptions and the model-generated descriptions to GPT-4 [35]. We prompted it to evaluate whether these descriptions included hallucinations and to categorize them based on Object, Attribute, Relationship, and Event hallucinations.

Using the definitions of different types of hallucinations given in Section 2, we asked GPT-4 to first determine whether the model's output contained hallucinations. If hallucinations were present, GPT-4 was instructed to classify them according to the definitions of different types of hallucinations. We tabulated the proportions of different types of hallucinations at varying description lengths. We found a significant increase in the proportion of event hallucinations as the length of the description extended.

D.3 Impact of SFT Data Scales for Hal-Evaluator

To investigate the effect of different data scales on the efficacy of Hal-Evaluator, we enlarged the scale of the SFT fine-tuning data from 130K to 2M and evaluated Hal-Evaluator on Hal-Eval-Dis. The experimental outcomes reveal that the application of 2M-scale data resulted in the highest accuracy in identifying hallucinations. This highlights the importance of data scale in enhancing the performance of such evaluators.

D.4 Effectiveness of Hall-evaluator for Mitigating Generative Hallucination.

The evaluation model Hal-Evaluator, utilized for generative hallucination assessment, can not only detect and evaluate model hallucinations but also modify the hallucinatory content within model outputs to aid in hallucination elimination. To validate the effectiveness of Hal-Evaluator in eliminating hallucinations, we conducted the following experiment: we used the output from MiniGPT-4 and LLaVA as the input for Hal-Evaluator, allowing Hal-Evaluator to detect and rectify any hallucinatory content. Subsequently, we re-evaluated the corrected output with MMHal-Bench [43] as shown in Table 7. Our findings indicate that Hal-Evaluator can effectively eliminate model hallucinations.

D.5 Compared with Other Hallucination Benchmarks

As illustrated in the Table 1, we have compared current mainstream LVLMs hallucination evaluation methods, highlighting that our approach ensures the most extensive coverage of hallucination.

D.6 Demo for Hal-Evaluator

As shown in Figure 13, through this visualization, we aim to provide a more intuitive understanding of the model's behavior and its ability to appropriately respond to different types of hallucinations.

E ANNOTATION PORTAL

The annotators are provided with an image, LM-generated detailed description of the image. For each description, the annotators mark parts of the sentence into appropriate cateogies: object, attribute, relation, event hallucination. We provide the explanation and examples for those annotators with Figure 13 and Figure 10

E.1 Annotation Agreement

The annotators involved in this research are internal members of our organization, with a good understanding of natural language processing and multimodal field studies. They have all consented to the use of their annotation information for this research and for it to be made public.

(1) Object Hallucination Annotatoin Prompt to GPT-4:

In the given content, hallucination refers to a scenario where the Large Vision Language Model generates a response that incorporates information not explicitly present or implied in the given image or the preceding conversation. \n
The type of hallucination to consider is:
Object Hallucination: the model 'hallucinates' or invents the existence of an object that isn't actually there.
According to the decription of hallucination type, your task is to introduce the hallucination by revise the given image caption.

(2) Attribute Hallucination Annotatoin Prompt to GPT-4:

In the given content, hallucination refers to a scenario where the Large Vision Language Model generates a response that incorporates information not explicitly present or implied in the given image or the preceding conversation. \n
The type of hallucination to consider is:
Attribute Hallucination: This happens when the object in the image is correctly identified, but its attributes, states, or behaviors are incorrectly described.
According to the decription of hallucination type, your task is to introduce the hallucination by revise the given image caption.

(3) Relation Hallucination Annotatoin Prompt to GPT-4:

In the given content, hallucination refers to a scenario where the Large Vision Language Model generates a response that incorporates information not explicitly present or implied in the given image or the preceding conversation. \n
The type of hallucination to consider is:
Relation Hallucination: This type of hallucination occurs when a LVLM incorrectly describes the relationship between objects. For example, if an image depicts a lamp on a desk and the system captions it as \'a lamp behind the desk,\' it has created a relationship hallucination. According to the decription of hallucination type, your task is to introduce the hallucination by revise the given image caption.

(4) Event Hallucination Annotatoin Prompt to GPT-4:

In the given content, hallucination refers to a scenario where the Large Vision Language Model generates a response that incorporates information not explicitly present or implied in the given image or the preceding conversation. \n
The type of hallucination to consider is:
Event Hallucination:This type of hallucination occurs when a LVLM not only describe a non-existent target but also construct complete events around the non-existent target, including its attributes, relations, and actions.
According to the decription of hallucination type, your task is to introduce the hallucination by revise the given image caption.

Figure 10: The image shows the prompt we used to elicit hallucination annotations from GPT-4.

I have a task that involves check the annotation of different types of hallucinations in image captions.

The types of hallucinations to consider are:

Relationship Hallucination: This type of hallucination occurs when an AI model incorrectly describes the spatial relationship between objects. For example, if an image depicts a lamp on a desk and the system captions it as "a lamp behind the desk," it has created a relationship hallucination because it inaccurately described the position of the lamp in relation to the desk.

Objective Hallucination: the LVLMs invents the existence of an object that isn't actually there.

Attributive Hallucination: This occurs when the system correctly identifies the objects in an image but inaccurately describes their attributes, states, or actions. The model might recognize an object (like a car or a cat) but then attribute to it colors, actions, or qualities that aren't actually present or true.

Event Hallucination: This type of hallucination occurs when a LVLM not only describe a non-existent target but also construct complete events around the non-existent target, including its attributes, relations, and actions.

Detailed Task Description: Below are two captions ("Original Caption" and "Hallucination Caption") that describe the same image, along with an annotation indicating a potential hallucination. You are an assistant, and your task is to compare the provided "Original Caption" with the "Hallucination Caption" and check for the following:

(1) **Determine if the "Hallucination Caption" truly contains one of the four types of hallucinations listed above.**

(2) Accuracy of Hallucination Type: Ensure that the type of hallucination annotated in the "Hallucination Annotation" is correct.

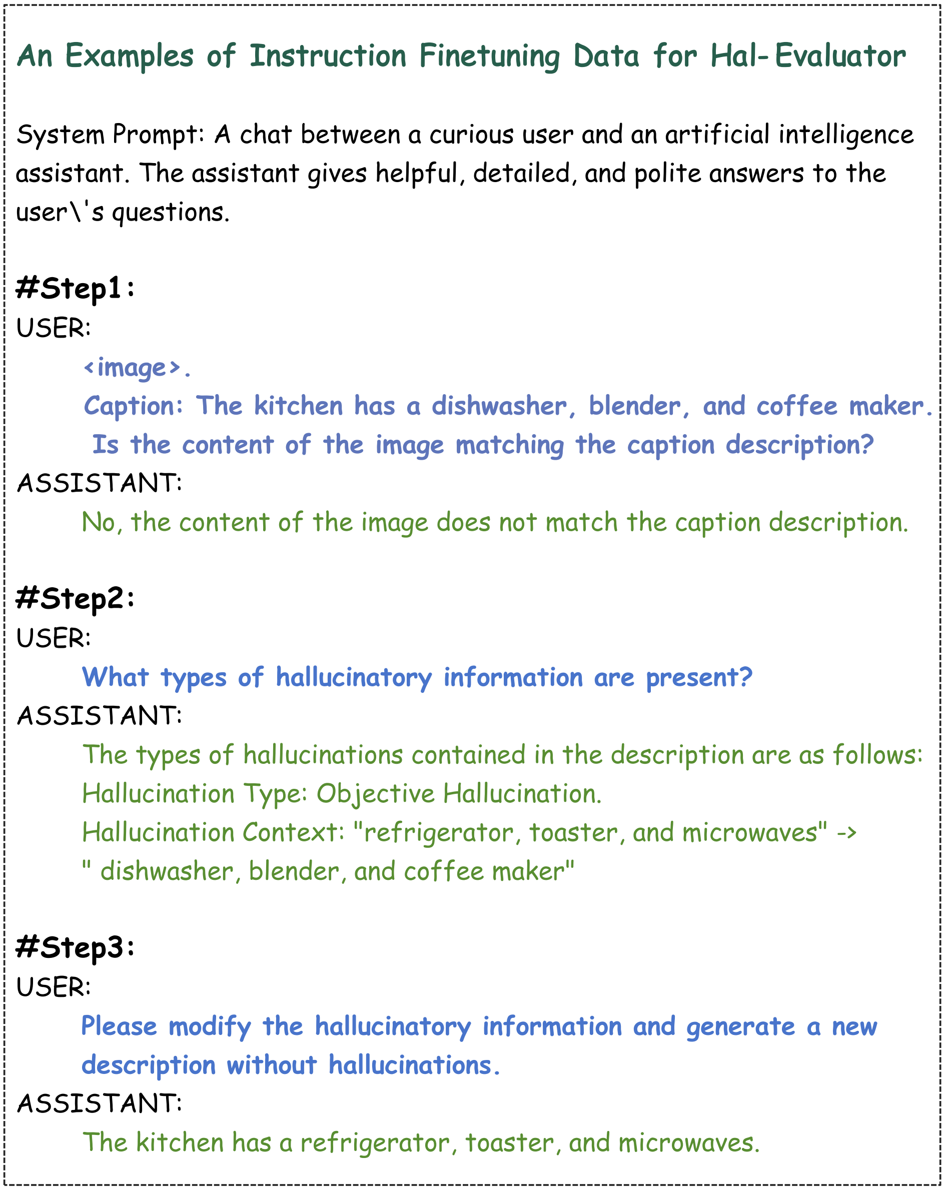
(3)Accuracy of Context in Hallucination: Confirm that the context mentioned in the "Hallucination Annotation" accurately reflects a discrepancy from the actual context of the original image or caption.

If the Hallucination Caption is same with Original Caption or does not contain hallucinations, please respond with "Bad Caption."

If the Hallucination Type or Context in Hallucination is wrong and need to revise, please answer: Wrong and give the revised annotation.

Otherwise, please just answer: Good annotation.

Figure 11: The figure illustrates the tailored prompt designed to guide GPT-4 in distinguishing and discarding labels that are inconsistent with the predefined criteria of hallucination types or are marked inaccurately.



An Examples of Instruction Finetuning Data for Hal-Evaluator

System Prompt: A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user\'s questions.

#Step1:

USER:
<image>.
Caption: The kitchen has a dishwasher, blender, and coffee maker.
Is the content of the image matching the caption description?
ASSISTANT:
No, the content of the image does not match the caption description.

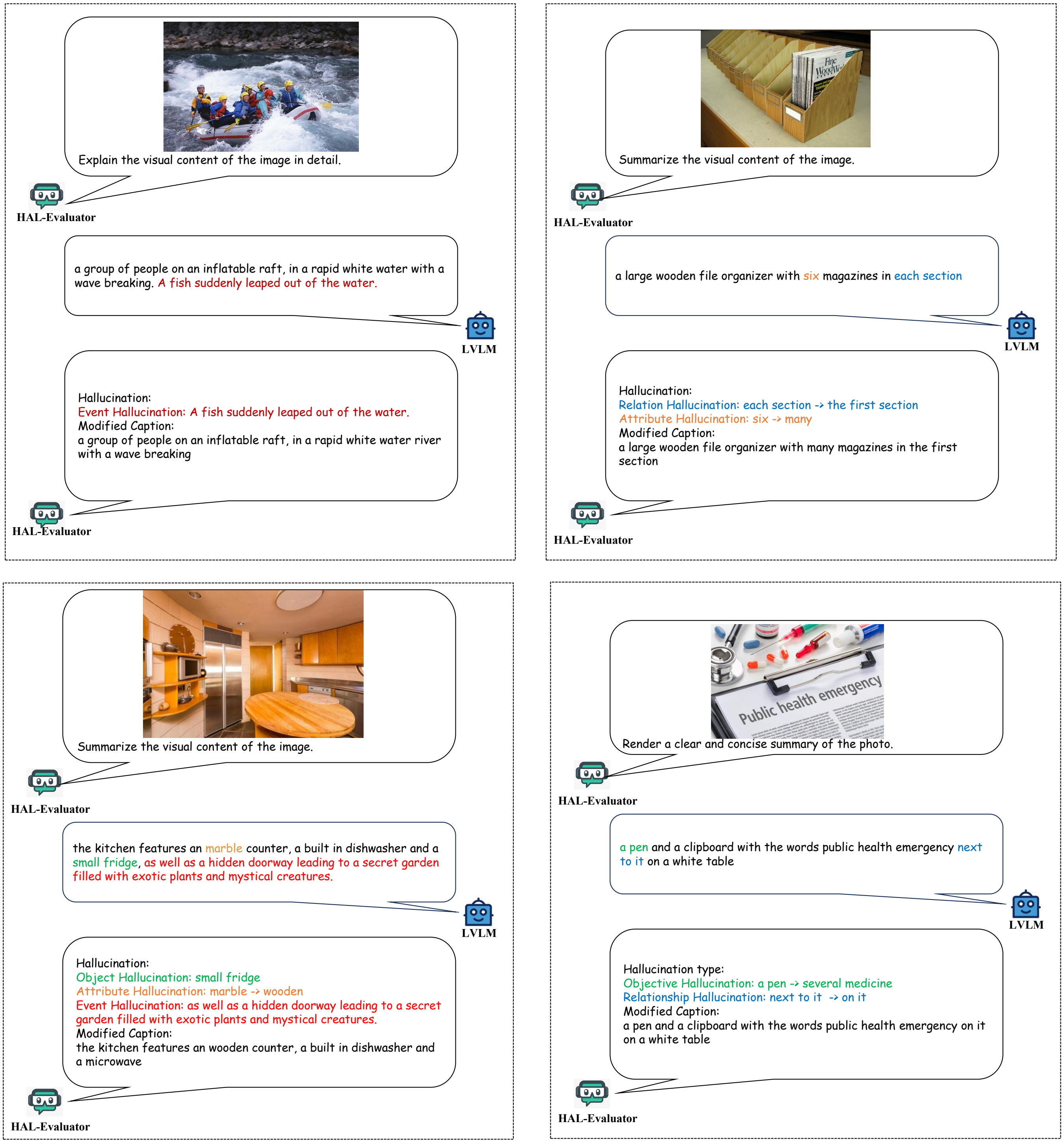
#Step2:

USER:
What types of hallucinatory information are present?
ASSISTANT:
The types of hallucinations contained in the description are as follows:
Hallucination Type: Objective Hallucination.
Hallucination Context: "refrigerator, toaster, and microwaves" -> " dishwasher, blender, and coffee maker"

#Step3:

USER:
Please modify the hallucinatory information and generate a new description without hallucinations.
ASSISTANT:
The kitchen has a refrigerator, toaster, and microwaves.

Figure 12: The provided image displays an example utilized in the construction of the instruction tuning dataset, which we refer to as Hal-Data-SFT. The highlighted portions in green represent the elements that the model needs to predict and compute loss during training.



Explain the visual content of the image in detail.

a group of people on an inflatable raft, in a rapid white water with a wave breaking. A fish suddenly leaped out of the water.

Hallucination:
Event Hallucination: A fish suddenly leaped out of the water.
Modified Caption:
a group of people on an inflatable raft, in a rapid white water river with a wave breaking

Summarize the visual content of the image.

a large wooden file organizer with six magazines in each section

Hallucination:
Relation Hallucination: each section -> the first section
Attribute Hallucination: six -> many
Modified Caption:
a large wooden file organizer with many magazines in the first section

Summarize the visual content of the image.

the kitchen features an marble counter, a built in dishwasher and a small fridge, as well as a hidden doorway leading to a secret garden filled with exotic plants and mystical creatures.

Hallucination:
Object Hallucination: small fridge
Attribute Hallucination: marble -> wooden
Event Hallucination: as well as a hidden doorway leading to a secret garden filled with exotic plants and mystical creatures.
Modified Caption:
the kitchen features an wooden counter, a built in dishwasher and a microwave

Render a clear and concise summary of the photo.

a pen and a clipboard with the words public health emergency next to it on a white table

Hallucination type:
Objective Hallucination: a pen -> several medicine
Relationship Hallucination: next to it -> on it
Modified Caption:
a pen and a clipboard with the words public health emergency on it on a white table

Figure 13: This figure presents the responses of the Hal-Evaluator when confronted with various descriptions of hallucinations.