

# *HalLoc*: Token-level Localization of Hallucinations for Vision Language Models

## Supplementary Material

### A. Motivation for a Probabilistic Hallucination Detection Module

By developing a model that outputs calibrated token-level probabilities of uncertainty, we gain a fine-grained understanding of the model’s confidence in each word it produces. This calibration ensures that the probability estimates accurately reflect the true likelihood of correctness. When the model predicts a token with high uncertainty, it may indicate a higher risk of hallucination at that point in the text.

Even though we have token-level log probabilities, they often correlate poorly with actual error rates due to miscalibration. A log probability might suggest high confidence numerically, but without calibration, it doesn’t guarantee this confidence is justified. Calibrated uncertainties adjust these probabilities to align better with real-world correctness.

*HalLoc* offers a valuable opportunity to develop an external token-level hallucination detection model that produces well-calibrated uncertainty probabilities. Calibrating token probabilities across a large vocabulary is inherently challenging due to the sheer number of possible tokens and the complexity of accurately estimating their individual probabilities. In contrast, a token-level hallucination detection model trained on *HalLoc* only needs to be calibrated at a binary level for each token—simply determining whether a token is hallucinated. This reduction to a binary classification task makes the calibration process more accessible and more intuitive. Moreover, since this model operates externally, it does not interfere with the language model’s generation process, allowing us to enhance uncertainty estimation and hallucination detection without impacting the quality of text generation.

### B. Calibration Results and Analysis

To illustrate that an external hallucination detection model trained on *HalLoc* can enhance the calibration quality of token-level probabilities, we conduct a comprehensive calibration analysis using the Expected Calibration Error (ECE) [4]<sup>1</sup> and Adaptive Calibration Error (ACE) [5]<sup>2</sup> met-

<sup>1</sup>ECE measures the difference between predicted confidence and accuracy over a set of equally-spaced probability bins  $B_m$ , providing a metric for model calibration. Formally:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|$$

<sup>2</sup>ACE is similar to ECE but uses adaptive binning to ensure each bin has an equal number of samples.

rics. Calibration measures the agreement between predicted probabilities and actual outcomes, with lower ECE and ACE values indicating better-calibrated models. Tables 3, 4, and 2 present the calibration results across three datasets: **HalLoc-Instruct**, **HalLoc-Caption**, and **HalLoc-VQA**.

The models evaluated include:

- **InternVL [1]**: A strong Vision Language Model providing log probabilities.
- **HalLocalizer (InternVL + VisualBERT)**: Our model trained on *HalLoc* that combines InternVL embeddings with VisualBERT.
- **HalLocalizer (VisualBERT)**: Our model utilizing only VisualBERT.

We report the ECE and ACE values as percentages (%). Avg represents the macro-average of calibration errors for positive and negative labels<sup>3</sup>.

#### B.1. Improvement in Calibration with *HalLocalizer*

Across all datasets, both versions of *HalLocalizer* exhibit substantially lower ECE and ACE values than the baseline InternVL model, indicating superior calibration and more reliable probability estimates.

#### B.2. Performance Across Different Hallucination Types

*HalLocalizer* versions maintain consistent and robust performance across various hallucination categories, including object, attribute, relationship, and scene. Notably, low ECE and ACE values in *attribute* and *relationship* categories suggest that *HalLocalizer* effectively identifies complex hallucinations related to attributes and relationships. This consistent performance indicates that *HalLocalizer* can effectively identify subtle and complex hallucinations, enhancing its applicability in diverse scenarios.

#### B.3. Impact of Temperature Scaling

Applying temperature scaling further reduces the ECE and ACE values for both InternVL and *HalLocalizer*, enhancing their calibration. The reduction is more pronounced for InternVL, indicating that it benefits more from calibration techniques but still doesn’t match the baseline calibration of *HalLocalizer*. This suggests that while temperature scaling is beneficial, *HalLocalizer* models inherently possess better calibration than the baseline.

<sup>3</sup>Due to the natural imbalance in positive labels (hallucinated tokens), it is helpful to analyze each label separately.

Table 1. Avg Prob (Original) refers to average hallucination probability before adding Gaussian Noise. Avg Prob ( $\sigma$ ) refers to average hallucination probability after applying Gaussian Noise with different blur intensities ( $\sigma=5$ ,  $\sigma=20$ ).

	Avg Prob (Original)	Avg Prob ( $\sigma=5$ )	Avg Prob ( $\sigma=20$ )
Overall	19.70	19.99	20.81

#### B.4. Challenge of Detecting Positive Instances

Our analysis reveals that calibration errors are consistently higher for positive instances—specifically, hallucinated tokens—across all models and datasets. This significant gap between the calibration errors of positive and negative instances highlights a critical obstacle in accurately detecting hallucinated tokens. The challenge is exacerbated by the varying distribution of hallucinated tokens across different tasks (with almost 100% in *HalLoc-VQA*, 25.37% in *HalLoc-Instruct*, and 5.35% in *HalLoc-Caption* for hallucinated samples), making it difficult to train a granular hallucination detection model effectively. Addressing this issue is imperative for advancing the field, and future work must focus on improving the sparse positive labels to enhance detection accuracy and model reliability.

### C. Analyzing the Grey Area

In real-world applications, visual data is rarely precise—it often suffers from noise, distortions, or ambiguities caused by factors like poor lighting, motion blur, or environmental interference. We introduce Gaussian noise into images during our experiments to simulate these imperfections, aiming to evaluate how *HalLocalizer* handles such visual noise. This approach is critical as it probes the grey area where the model’s interpretations oscillate.

In our framework, object hallucination includes cases where descriptions—such as attributes and relationships—pertain to nonexistent objects. To explore this, we selectively apply Gaussian noise: to the object bounding boxes pertaining to truthful objects (including those describing attributes and relationships). For scene tokens, Gaussian noise is added to the entire image.

The results, summarized in Table 1, reveal a gradual increase in the likelihood of hallucinated tokens as image noise intensifies, rather than an abrupt shift from non-hallucination to hallucination. This nuanced progression underscores the limitations of binary classification in capturing such subtle transitions. Consequently, traditional binary metrics may fail to reflect the model’s performance under varying noise conditions adequately. To address this, incorporating probabilistic or spectrum-based evaluation methods could provide a more detailed understanding

of the model’s behavior in the face of visual uncertainty.

Probability Model	Calibration Techniques	Label	total		object		attribute		relationship		scene	
			ECE	ACE	ECE	ACE	ECE	ACE	ECE	ACE	ECE	ACE
InternVL	original	pos	73.72	67.21	-	-	-	-	-	-	-	-
		neg	91.60	78.17	-	-	-	-	-	-	-	-
		avg	82.66	72.69	-	-	-	-	-	-	-	-
	+TS	pos	39.25	38.67	-	-	-	-	-	-	-	-
		neg	53.31	44.47	-	-	-	-	-	-	-	-
		avg	46.28	41.57	-	-	-	-	-	-	-	-
HalLocalizer (InternVL + VisualBERT)	original	pos	-	-	20.92	20.80	0.86	1.37	28.71	28.48	3.23	3.02
		neg	-	-	21.98	22.03	12.13	12.19	11.50	11.53	1.67	1.75
		avg	-	-	21.45	21.41	6.50	6.78	20.11	20.00	2.45	2.38
	+TS	pos	-	-	18.80	18.36	6.80	7.00	24.83	24.66	2.23	2.20
		neg	-	-	20.64	20.87	8.86	11.25	9.43	9.82	1.22	1.69
		avg	-	-	19.72	19.62	7.83	9.12	17.13	17.24	1.73	1.95
HalLocalizer (VisualBERT)	original	pos	-	-	30.97	30.68	17.96	18.17	27.51	27.64	20.42	19.20
		neg	-	-	5.36	5.41	6.12	6.13	7.10	7.16	0.96	0.97
		avg	-	-	18.16	18.05	12.04	12.15	17.30	17.40	10.69	10.08
	+TS	pos	-	-	27.34	27.29	13.15	13.89	24.03	23.99	17.72	16.11
		neg	-	-	4.28	4.66	3.45	5.30	5.77	6.15	0.68	1.02
		avg	-	-	15.81	15.97	8.30	9.60	14.90	15.07	9.20	8.56

Table 2. Probability Calibration of HalLocalizer on **HalLoc-VQA**. TS stands for Temperature Scaling.

Probability Model	Calibration Techniques	Label	total		object		attribute		relationship		scene	
			ECE	ACE	ECE	ACE	ECE	ACE	ECE	ACE	ECE	ACE
InternVL	original	pos	64.44	64.44	-	-	-	-	-	-	-	-
		neg	78.90	72.71	-	-	-	-	-	-	-	-
		avg	71.67	68.57	-	-	-	-	-	-	-	-
	+TS	pos	31.79	32.01	-	-	-	-	-	-	-	-
		neg	43.08	37.37	-	-	-	-	-	-	-	-
		avg	37.44	34.66	-	-	-	-	-	-	-	-
HalLocalizer (InternVL + VisualBERT)	original	pos	-	-	22.93	22.95	1.34	1.19	8.83	8.61	1.65	1.78
		neg	-	-	0.11	0.13	1.95	2.00	5.83	5.89	0.29	0.36
		avg	-	-	11.52	11.54	1.65	1.59	7.33	7.25	0.97	1.07
	+TS	pos	-	-	21.75	21.70	2.56	2.41	7.55	7.34	1.11	2.00
		neg	-	-	0.33	0.34	0.97	2.39	5.51	5.73	0.26	0.48
		avg	-	-	11.04	11.02	1.77	2.40	6.53	6.54	0.69	1.24
HalLocalizer (VisualBERT)	original	pos	-	-	20.42	20.40	0.78	0.44	8.94	8.95	2.43	2.09
		neg	-	-	0.19	0.16	1.73	1.76	6.32	6.33	0.23	0.28
		avg	-	-	10.31	10.28	1.25	1.10	7.63	7.64	1.33	1.19
	+TS	pos	-	-	19.69	19.63	1.07	0.85	2.31	2.16	3.87	3.77
		neg	-	-	0.14	0.18	1.48	1.79	4.16	5.83	0.24	0.36
		avg	-	-	9.92	9.90	1.27	1.32	3.24	4.00	2.06	2.06

Table 3. Probability Calibration of HalLocalizer on **HalLoc-Instruct**. TS stands for Temperature Scaling.

Probability Model	Calibration Techniques	Label	total		object		attribute		relationship		scene	
			ECE	ACE	ECE	ACE	ECE	ACE	ECE	ACE	ECE	ACE
InternVL	original	pos	46.47	46.47	-	-	-	-	-	-	-	-
		neg	64.86	46.61	-	-	-	-	-	-	-	-
		avg	55.66	46.54	-	-	-	-	-	-	-	-
	+TS	pos	16.51	15.57	-	-	-	-	-	-	-	-
		neg	31.63	14.16	-	-	-	-	-	-	-	-
		avg	24.07	14.87	-	-	-	-	-	-	-	-
HalLocalizer (InternVL + VisualBERT)	original	pos	-	-	37.29	37.13	51.34	50.80	44.07	49.00	81.98	81.88
		neg	-	-	0.04	0.08	0.57	0.54	0.17	0.13	0.10	0.10
		avg	-	-	18.66	18.61	25.96	25.67	22.12	24.57	41.04	40.99
	+TS	pos	-	-	34.34	34.32	51.22	50.68	40.68	45.81	79.86	79.76
		neg	-	-	0.29	0.27	0.59	0.56	0.63	0.60	0.39	0.39
		avg	-	-	17.32	17.30	25.91	25.62	20.66	23.21	40.12	40.08
HalLocalizer (VisualBERT)	original	pos	-	-	21.61	21.41	31.30	31.30	21.31	19.46	22.02	24.05
		neg	-	-	0.14	0.18	0.11	0.16	0.20	0.24	0.02	0.02
		avg	-	-	10.88	10.79	15.71	15.73	10.75	9.85	11.02	12.04
	+TS	pos	-	-	17.03	16.95	28.98	28.99	18.77	17.22	20.45	21.46
		neg	-	-	0.45	0.40	0.24	0.21	0.30	0.26	0.17	0.16
		avg	-	-	8.74	8.67	14.61	14.60	9.54	8.74	10.31	10.81

Table 4. Probability Calibration of HalLocalizer on **HalLoc-Caption**. TS stands for Temperature Scaling.

## D. Illustrative Examples of *HalLoc*

We provide examples of data points in *HalLoc* for each of *HalLoc-VQA*, *HalLoc-Instruct*, and *HalLoc-Caption* in Figures 1, 2, and 3. The general annotation format of *HalLoc* is illustrated in Figure 4.


	<div>source textchurch</div> <div>source metadata{id: '\_03634994', source: 'GQA'}</div> <div>qa metadata{ }</div> <div>qa ids[\_03634994]</div> <div>prompt&lt;image&gt; Question: Which place is it?</div> <div>image_id2367212</div> <div>hallucinated_textcastle</div> <div>           annotations           {             object : [ ],             attribute : [ ],             relationship : [ ],             scene : [               {                 scene : {                   name : castle,                   word_index : '0',                   char_index : '0:6',                 }               }             ]           }         </div> <div>splittrain</div> <div>idvqa_18875</div>
------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 1. Example of an entry in *HalLoc-VQA* of hallucination type **scene**.


	<div>source textThe man is to the right of the refrigerator.</div> <div>source metadata{id: '00142745', source: 'GQA'}</div> <div>qa metadata{ }</div> <div>qa ids[00142745]</div> <div>prompt&lt;image&gt; What is the position of the man relative to the refrigerator?</div> <div>image_id2336351</div> <div>hallucinated_textThe man is to the <b>left</b> of the <b>refrigerator</b>.</div> <div>           annotations           {             object : [ ],             attribute : [ ],             relationship : [               {                 subject : {                   id : 960534,                   name : man,                   word_index : '1',                   char_index : '4:7',                 },                 predicate : {                   name : left,                   word_index : '1',                   char_index : '4:7',                   category : preposition,                 },                 object : {                   id : 960545,                   name : refrigerator,                   word_index : '8',                   char_index : '30:42',                 }               }             ],             scene : [ ]           }         </div> <div>splittrain</div> <div>idinstruct_53</div>
-------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 2. Example of an entry in *HalLoc-Instruct* of hallucination type **relationship**.


	<div>source textThere are two people sitting at the table. Both of them are studying. There are two cups on the table. On the table is a black open laptop. The cups are white. There is a binder on the table. The woman is holding a highlighter. The table behind them is empty. It has blue seats. The umbrella is attached to a pole.</div> <div>source metadata{id: 'sp_35891', source: 'stanford'}</div> <div>qa metadata{ }</div> <div>qa ids[08452448, 08452056, 08451948, 08451888, 08451933]</div> <div>prompt&lt;image&gt; Can you describe the main features of this image for me?</div> <div>image_id2381770</div> <div>hallucinated_textThere are two people sitting at the table. Both of them are studying. There are two cups on the table. On the table is a black open laptop. The <b>laptop</b> is on the <b>chair</b>. The cups are white. There is a binder on the table. The woman is holding a highlighter. The <b>guy</b> is to the <b>left</b> of the <b>book</b>. The table behind them is empty, except for a <b>bottle</b>. It has blue seats. The <b>umbrella</b> is attached to a pole and is <b>yellow</b>. To the <b>left</b> of the image, there is a <b>car</b> positioned under a tree.</div> <div>           annotations           {             object : [               {                 obj : {                   name : bottle,                   char_index : '337:343',                   question_id : 08451933,                 },                 attribute : [                   {                     attribute : {                       name : left,                       char_index : '421:425',                       question_id : 08451948,                     },                     obj : {                       name : car,                       char_index : '451:454',                       question_id : 08451948,                     },                   },                   {                     attribute : {                       name : yellow,                       char_index : '412',                       question_id : 08451888,                     },                     obj : {                       name : umbrella,                       char_index : '368:376',                       question_id : 08451948,                     },                   },                 ],                 relationship : [                   {                     subject : {                       name : guy,                       char_index : '260:263',                       question_id : 08752448,                     },                     object : {                       name : book,                       char_index : '286:290',                       question_id : 08752448,                     },                     predicate : {                       name : left,                       char_index : '274:278',                       question_id : 08752448,                     },                   },                   {                     subject : {                       name : laptop,                       char_index : '144:150',                       question_id : 08452056,                     },                     object : {                       name : chair,                       char_index : '161:166',                       question_id : 08452056,                     },                     predicate : {                       name : on,                       char_index : '274:278',                       question_id : 08452056,                     },                   },                 ],                 scene : [ ]               }             ]           }         </div> <div>splittrain</div> <div>idcaption_18232</div>
------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 3. Example of an entry in *HalLoc-Caption* comprising of multiple types of hallucination; **object**, **attribute**, **relationship**.

source text	"Ground truth text of the hallucinated text (VQA/Instruct/Caption)"
source metadata	{           source           {             id             id of the source_text in the source dataset           }         }
qa metadata	{ }
qa ids	Question ids from the HQA Database used to create the hallucinated text
prompt	Instruction
image_id	hallucinated_text
annotations	{           object : [             {               obj : {                 name                 token                 word_index                 word index in the hallucinated text                 char_index                 character index in the hallucinated text               },               attribute : [                 {                   attribute : {                     name : "",                     word_index : "",                     char_index : "",                   },                   obj : {                     name : "",                     word_index : "",                     char_index : "",                   },                 },               ],               relationship : [                 {                   subject : {                     name : "",                     word_index : "",                     char_index : "",                   },                   object : {                     name : "",                     word_index : "",                     char_index : "",                   },                   predicate : {                     name : "",                     word_index : "",                     char_index : "",                   },                 },               ],               scene : [                 {                   scene : {                     name : "",                     word_index : "",                     char_index : "",                   },                 },               ],             }           ],         }
split	benchmark_type (VQA, Instruct, Caption)_id

Figure 4. Annotation format of *HalLoc*

## E. HalLoc Instruction Details

Each entry of *HalLoc* consists of an {**Instruction**}-{**Response**} pair. We show the details of the templates we used for the instructions in *HalLoc-VQA* in Figure 5, *HalLoc-Instruct* in Figure 6 and *HalLoc-Caption* in Figure 7.

<image>	{Q}
<image>	Question: {Q}
<image>	{Q} A short answer to the question is
<image>	Q: {Q}, A:
<image>	Given the image, answer the following question with no more than three words {Q}
<image>	Based on the image, respond to this question with a short answer: {Q}, Answer:
<image>	Use the provided image to answer the question: {Q}. Provide your answer as short as possible.
<image>	What is the answer to the following question? : {Q}
<image>	The question {Q} can be answered using the image. A short answer is:
<image>	Question: {Q}, A:

Figure 5. Examples of instructions for model responses in *HalLoc-VQA*. Q is a placeholder for the visual question.

<image>	Explain {Q}
<image>	Describe {Q}
<image>	Discuss {Q}
<image>	Question: {Q}, Long answer:
<image>	Based on the image, respond to this question with a full sentence {Q}. Answer:
<image>	{Q}, A long answer to the question is:
<image>	What is the answer to the following question? : {Q}
<image>	Q: {Q}, A:

Figure 6. Examples of instructions for model responses in *HalLoc-Instruct*. Q is a placeholder for the visual question.

Write a detailed description of the given image	<image>
Can you describe the main features of this image for me?	<image>
<image>	Describe the following image
<image>	What are the key elements in this picture?
<image>	What's happening in this scene?
Can you elaborate on the elements of the picture provided?	<image>
Describe the following image	<image>
<image>	What do you see happening in this image?
<image>	Can you elaborate on the elements of the picture provided?
<image>	Can you describe the main features of this image for me?
<image>	What is this photo about?
Explain the visual content of the image in great detail.	<image>
Analyze the image in a comprehensive and detailed manner.	<image>
What do you see happening in this image?	<image>
What are the key elements in this picture?	<image>
<image>	Explain the visual content of the image in great detail
<image>	What do you think is going on this snapshot?
<image>	Analyze the image in a comprehensive manner.
What is happening in the scene?	<image>

Figure 7. Examples of instructions for model responses in *HalLoc-Caption*.

## F. HQA Injection Pipeline Details

### F.1. Annotating QA pairs

*HalLoc* utilizes the GQA dataset [2] as the foundation for its questions and answers. In particular, we take advantage of the questions and scene graphs provided by GQA to achieve detailed annotations of the components within each question and answer. Figures 8 and 9 show examples of how we annotate a GQA question and answer to save them in the HQA Database. The hallucinated answer in Figure 8 is directly derived from the GQA question (the other choice). The hallucinated answer in Figure 9 comes from crafted hallucinated candidates.


image_id	2374937	
question_id	1485806	
semantic_type	relationship	
detailed_type	predicate	
hallucination_type	other	
question	Is the open drawer to the right or to the left of the freezer where the magnet is on?	
original_answer	left	
hallucinated_answer	right	
hallucinated_trait	<pre>{ "subject": { "id": 2999233, "name": "drawer", "detailed_description": { "name": "open drawer", "word_index": "2:4", "detail_type": "attribute", "other_id": "None", "annotations": { "name": "drawer", "word_index": "1" } } }, "object": { "id": 2216487, "name": "fridge", "detailed_description": { "name": "freezer the magnet is on", "word_index": "13:18", "detail_type": "relation", "other_id": "3696339", "annotations": { "name": "freezer", "word_index": "0" } } }, "predicate": { "name": "right", "category": "preposition" } }</pre>	
metadata	<pre>{ "source": "GQA", "hallucinated_answer_source": "GQA", "gqa_types": { "detailed": "relChooser", "semantic": "rel", "structure": "choose" } }</pre>	

Figure 8. Example of how each component in a single visual QA pair in the GQA dataset is annotated with hallucinating question and answer in *HalLoc*’s HQA Database.

### F.2. Crafting Hallucinated Answers

After we generate a set of hallucinated answer candidates that reflect common causes and patterns of hallucinations in Large Vision-Language Models (LVLMS), we use GPT-4 [6] to choose a hallucinated answer from these candidates. Specifically, GPT-4 is employed to assess hallucinated answer candidates for attribute and relationship questions, which demand advanced reasoning to avoid generating nonsensical or overly similar responses. Figures 10 and 11 show the prompts used to evaluate the hallucinated answers. In practice, each prompt is associated with in-context examples to demonstrate how the process works.


image_id	2374937	
question_id	1485806	
semantic_type	attribute	
detailed_type	general	
hallucination_type	Concept association bias	
question	What color are the mirrors?	
original_answer	black	
hallucinated_answer	blue	
hallucinated_trait	<pre>{ "obj": { "id": 1130848, "name": "mirrors", "attribute": { "name": "blue", "category": "color" } } }</pre>	
metadata	<pre>{ "source": "GQA", "hallucinated_answer_source": "GPT-4", "gqa_types": { "detailed": "directWhich", "semantic": "attr", "structure": "query" } }</pre>	

Figure 9. Example of how each component in a single visual QA pair in the GQA dataset is annotated with hallucinating question and answer in *HalLoc*’s HQA Database.

### F.3. Determining Injection Points

We pose the question to the paragraph to identify where to inject the hallucinated answers. Specifically, subsections of the paragraph that share similarities with the hallucinated elements the question seeks become ideal candidates for injection. A brief reminder that the hallucinated elements for each question type are:

Object: <obj>

Attribute: <attr><obj>

Relationship: <obj1><rel><obj2>

Scene: <sce>

Figures 12, 13, and 14 illustrate the prompts used to guide GPT-4 in determining injection points for attribute, relationship, and scene type questions. Note that identifying injection points for object type questions is unnecessary, as it is evident that the paragraphs do not include the hallucinated object.

### F.4. Injecting Hallucinated Answers

After identifying possible injection points, we prompt GPT-4 to inject the hallucinated answer to the paragraph.

Figures 15, 16, 17, and 18 illustrate the prompts used to inject hallucinated answers for object, attribute, relationship, and scene type questions, respectively.

### F.5. Verifying Injection

After each injection, we verify whether the injection pipeline properly inserted the hallucinated answer. This step is necessary because, despite specific instructions and rule-based filtering, we have observed several instances where GPT-4 fails to inject hallucinated answers correctly. In particular, while creating *HalLoc-Caption*, which requires multiple rounds of HQA injection, we observed success rates of 57%, 33%, 30%, and 61% for object, attribute, relationship, and scene questions, respectively.

Figures 19, 20, 21, and 22 show the prompts used to verify each injection step for object, attribute, relationship, and scene questions. Refer to Algorithm 1 for a complete algorithmic overview of the HQA injection.

---

**Algorithm 1** HQA Injection

---

**Require:** Paragraph  $P$ , Number of HQA pairs  $n$ , QA database  $Q$

```

1: Select  $n$  HQA pairs from  $Q$ 
2: Initialize  $curr\_paragraph \leftarrow P$ 
3: Initialize  $coarse\_annotations \leftarrow \emptyset$ 
4: for each question-hallucinated answer pair  $i \in [1, n]$ 
   do
5:   Extract components  $a, b, c$  from  $i$ 
6:   if answer  $c$  in  $curr\_paragraph$  then
7:      $injection\_point \leftarrow c$ 
8:   else if component  $a$  or  $b$  in  $curr\_paragraph$  then
9:      $injection\_point \leftarrow$  phrase surrounding  $a$  or  $b$ 
10:  else
11:     $injection\_point \leftarrow$  selected by GPT-4
12:  end if
13:  Use GPT-4 to inject the hallucinated answer at the
     $injection\_point$ 
14:  Obtain  $modified\_paragraph$ ,  $phrase$ , and
     $components$ 
15:  if  $modified\_paragraph$  does not contain phrases in
     $coarse\_annotations$  then
16:    Continue to next question
17:  end if
18:  if  $phrase$  not unique in  $modified\_paragraph$  then
19:    Continue to next question
20:  end if
21:  if  $phrase$  does not contain components then
22:    Continue to next question
23:  end if
24:  if additional hallucinations introduced then
25:    Continue to next question
26:  end if
27:  if hallucinated answer differs from hallucination in
     $phrase$  then
28:    Continue to next question
29:  end if
30:  Update  $curr\_paragraph$  to  $modified\_paragraph$ 
31:  Add  $phrase, components$  to  $coarse\_annotations$ 
32: end for
33: for each  $ann$  in  $coarse\_annotations$  do
34:   Find index of  $phrase$  in  $modified\_paragraph$ 
35:   Find index of  $components$  in  $phrase$ 
36:   Annotate each component with index
37: end for

```

---



<p><b>Task:</b>          Modify an Answer by Replacing the Correct Attribute with a False One</p>
<p><b>Provided Information:</b></p> <ol style="list-style-type: none"> <li>1. Question: The question asks for the attribute of an object.</li> <li>2. Answer: The correct attribute of the object.</li> <li>3. Components: The &lt;attribute&gt;&lt;object&gt; relationship that the question asks for.</li> <li>4. Hallucination Candidates: A list of possible incorrect attributes.</li> <li>5. Hard Danger Zone: Other correct attributes that could replace the original correct attribute in the &lt;attribute&gt;&lt;object&gt; relationship.</li> <li>6. Image-Region Descriptions: Contextual descriptions of the image.</li> </ol>
<p><b>Goal:</b>          Replace the correct attribute in the answer with a false attribute from the list of Hallucination Candidates.</p>
<p><b>Process:</b></p> <ol style="list-style-type: none"> <li>1. Remove irrelevant attributes that do not answer the question.</li> <li>2. Remove attributes synonymous with the correct answer.</li> <li>3. Remove attributes synonymous with those in the Hard Danger Zone.</li> <li>4. Remove attributes that are contextually implausible              (e.g., avoid illogical pairings like "delicious car").</li> <li>5. If valid candidates remain, randomly select one as the false attribute.</li> <li>6. If no valid candidates remain, generate a plausible but incorrect attribute.</li> </ol>
<p><b>Output:</b></p> <ul style="list-style-type: none"> <li>- correct_attribute: The original correct attribute</li> <li>- false_attribute: The selected false attribute</li> <li>- eliminated_candidates: [name: candidate name,              reason: one of &lt;'irrelevant', 'synonym', 'hard dang', 'implausible'&gt;,              detailed_reason: a detailed reason for the elimination              ]</li> </ul>

Figure 10. Prompt used to craft hallucinated answers for hallucination type attribute.

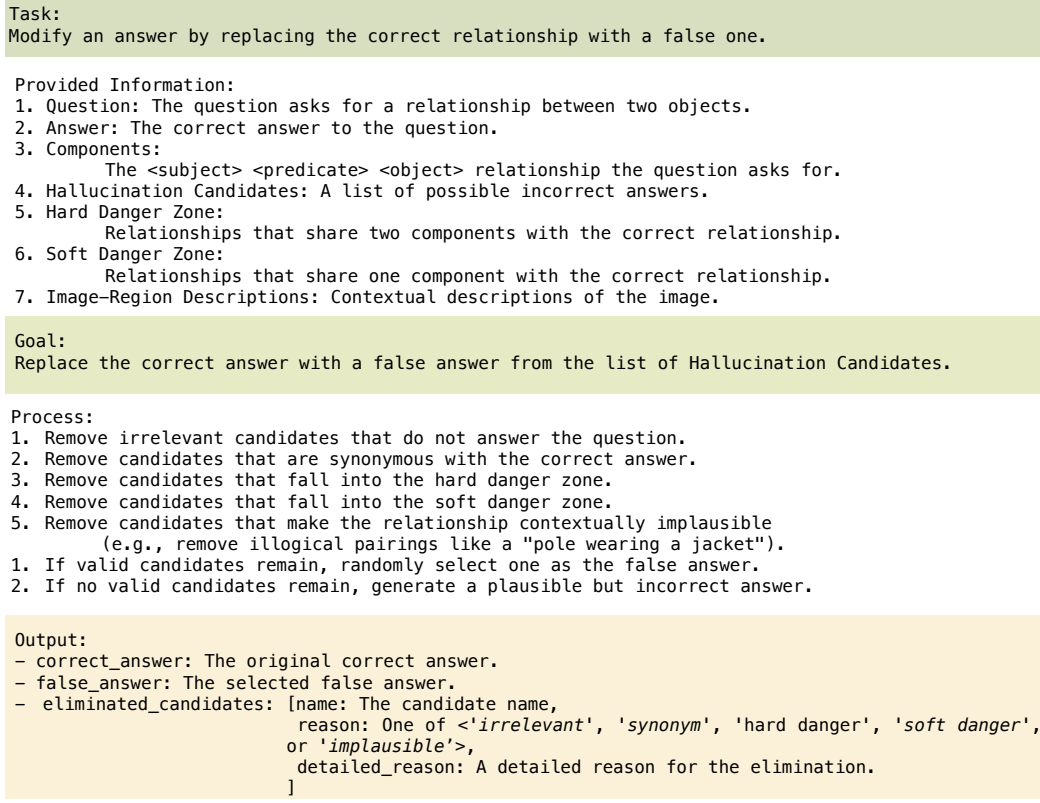


Figure 11. Prompt used to craft hallucinated answers for hallucination type relationship.

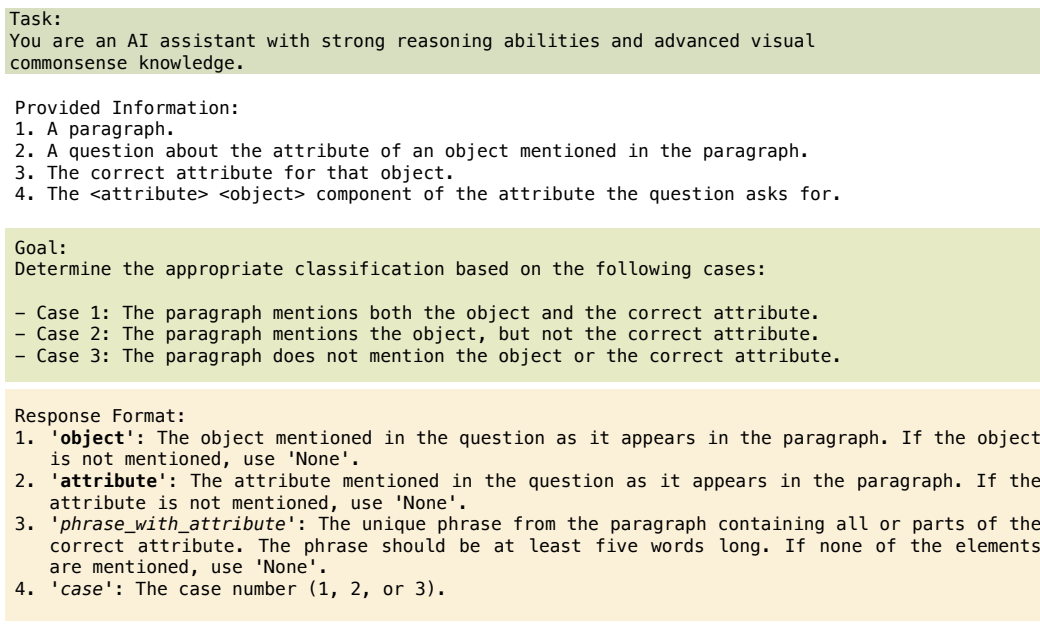


Figure 12. Prompt used to determine the injection points of the hallucinated answers to source texts for hallucination type attribute.

<p><b>Task:</b> You are an AI assistant with strong reasoning abilities and advanced visual commonsense knowledge.</p> <p><b>Provided Information:</b> 1. A paragraph. 2. A question about the relationship between two objects. 3. The correct answer to the question. 4. The &lt;subject&gt; &lt;predicate&gt; &lt;object&gt; component of the relationship the question asks for.</p> <p><b>Goal:</b> Determine the appropriate classification based on the following cases:</p> <ul style="list-style-type: none"> <li>- Case 1: All three elements (subject, predicate, object) of the correct relationship are present in the paragraph. The correct answer to the question can be found in the paragraph.</li> <li>- Case 2: One or two elements (subject, predicate, object) of the correct relationship are present in the paragraph.</li> <li>- Case 3: None of the elements (subject, predicate, object) of the correct relationship are present in the paragraph.</li> </ul> <p><b>Response Format:</b> 1. '<b>subject</b>': The subject of the correct relationship as it appears in the paragraph. If it isn't mentioned in the paragraph, use 'None'. 2. '<b>predicate</b>': The predicate of the correct relationship as it appears in the paragraph. If it isn't mentioned in the paragraph, use 'None'. 1. '<b>object</b>': The object of the correct relationship as it appears in the paragraph. If it isn't mentioned in the paragraph, use 'None'. 2. '<b>phrase_with_relationship</b>': The unique phrase from the paragraph containing all or parts of the correct relationship. The phrase should be at least five words long. If none of the elements are mentioned, use 'None'. 3. '<b>case</b>': The case number (1, 2, or 3).</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 13. Prompt used to determine the injection points of the hallucinated answers to source texts for hallucination type relationship.

<p><b>Task:</b> You are an AI assistant with strong reasoning abilities and advanced visual commonsense knowledge.</p> <p><b>Provided Information:</b> 1. A paragraph. 2. A question about the place, location, or weather described in the paragraph. 3. The correct answer to the question.</p> <p><b>Goal:</b> Determine the appropriate classification based on the following cases:</p> <ul style="list-style-type: none"> <li>- Case 1: The paragraph mentions the correct answer to the question.</li> <li>- Case 2: The paragraph does not mention the correct answer to the question.</li> </ul> <p><b>Response Format:</b> 1. '<b>scene</b>': The correct answer as described in the paragraph. If it is not mentioned, use 'None'. 2. '<b>phrase_with_scene</b>': The unique phrase from the paragraph that includes the correct answer. The phrase should be at least five words long. If it is not mentioned, use 'None'. 3. '<b>case</b>': The case number (1 or 2).</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 14. Prompt used to determine the injection points of the hallucinated answers to source texts for hallucination type scene.

<p><b>Task:</b> You are an AI assistant with strong reasoning abilities and advanced visual commonsense knowledge.</p> <p><b>Provided Information:</b></p> <ol style="list-style-type: none"> <li>1. A paragraph describing an image.</li> <li>2. A question about the paragraph.</li> <li>3. An incorrect answer to the question.</li> <li>4. The non-existent object mentioned in the incorrect answer.</li> </ol> <p><b>Goal:</b> The incorrect answer includes a non-existent object. You need to add a phrase or sentence to the original paragraph that introduces the non-existent object.</p> <p><b>Requirements:</b></p> <ul style="list-style-type: none"> <li>- Do not introduce any additional objects beyond the one mentioned in the incorrect answer.</li> <li>- Make no other changes to the paragraph except for the addition of the phrase or sentence that introduces the non-existent object.</li> </ul> <p><b>Response Format:</b></p> <ol style="list-style-type: none"> <li>1. 'updated_paragraph': The paragraph with the added phrase or sentence introducing the non-existent object.</li> <li>2. 'modification_details': A dictionary containing:</li> <li>3. 'added_phrase': The phrase or sentence added to the paragraph.</li> <li>4. 'context_with_incorrect_answer': The incorrect answer within the added phrase. Should be at least five words long.</li> <li>5. 'non_existent_object': A dictionary with:</li> <li>6. 'original_term': The original name of the non-existent object as it appeared in the incorrect answer.</li> <li>7. 'updated_term': The name of the non-existent object as it appears in the updated paragraph.</li> </ol>
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 15. Prompt to inject hallucinated answers pertaining to <obj> hallucination.

<p><b>Task:</b> You are an AI assistant with strong reasoning abilities and advanced visual commonsense knowledge.</p> <p><b>Provided Information:</b></p> <ol style="list-style-type: none"> <li>1. A paragraph describing an image.</li> <li>2. A question about the paragraph.</li> <li>3. An incorrect answer to the question.</li> <li>4. The hallucinated attribute in the incorrect answer.</li> <li>5. The image context.</li> <li>6. A possible insertion point: the phrase within the original paragraph where the hallucinated attribute could be inserted.</li> </ol> <p><b>Goal:</b> The incorrect answer contains a hallucinated attribute. Your job is to add a phrase or sentence to the original paragraph that introduces the hallucinated attribute.</p> <p><b>Requirements:</b></p> <ul style="list-style-type: none"> <li>- Do not introduce any additional objects or details except for the hallucinated attribute mentioned in the incorrect answer and the image context.</li> <li>- Make no other changes to the paragraph except for the addition of the phrase or sentence that introduces the hallucinated attribute.</li> <li>- Try to introduce the hallucinated attribute in the insertion point.</li> </ul> <p><b>Response Format:</b></p> <ol style="list-style-type: none"> <li>1. 'updated_paragraph': The paragraph with the added phrase or sentence introducing the hallucinated attribute.</li> <li>2. 'modification_details': A dictionary containing: <ul style="list-style-type: none"> <li>• 'added_phrase': The phrase or sentence added to the paragraph.</li> <li>• 'context_with_incorrect_answer': The phrase containing the full hallucinated attribute, with at least five words.</li> </ul> </li> <li>3. 'hallucinated_attribute': A dictionary with: <ul style="list-style-type: none"> <li>• 'hallucinated_trait': A dictionary with: <ul style="list-style-type: none"> <li>✓ 'original_term': The original term of the hallucinated trait from the incorrect answer.</li> <li>✓ 'updated_term': The term of the hallucinated trait as it appears in the updated paragraph.</li> </ul> </li> <li>• 'object': A dictionary with: <ul style="list-style-type: none"> <li>✓ 'original_term': The original name of the object associated with the hallucinated trait from the incorrect answer.</li> <li>✓ 'updated_term': The name of the object with the hallucinated trait as it appears in the updated paragraph.</li> </ul> </li> </ul> </li> </ol>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 16. Prompt to inject hallucinated answers pertaining to <attr> hallucination.

<p><b>Task:</b> You are an AI assistant with strong reasoning abilities and advanced visual commonsense knowledge.</p> <p><b>Provided Information:</b></p> <ol style="list-style-type: none"> <li>1. A paragraph describing an image.</li> <li>2. A question about the paragraph.</li> <li>3. An incorrect answer to the question.</li> <li>4. The hallucinated relationship in the incorrect answer.</li> <li>5. The image context.</li> <li>6. A possible insertion point: the phrase within the original paragraph where the hallucinated relationship could be inserted.</li> </ol> <p><b>Goal:</b> The incorrect answer contains a hallucinated relationship. Your job is to add a phrase or sentence to the original paragraph that introduces the hallucinated relationship.</p> <p><b>Requirements:</b></p> <ul style="list-style-type: none"> <li>- Do not introduce any additional objects or details except for the hallucinated relationship mentioned in the incorrect answer and the image context.</li> <li>- Make no other changes to the paragraph except for the addition of the phrase or sentence that introduces the hallucinated relationship.</li> <li>- Try to introduce the hallucinated relationship in the insertion point.</li> </ul> <p><b>Response Format:</b></p> <ol style="list-style-type: none"> <li>1. 'updated_paragraph': The paragraph with the added phrase or sentence introducing the hallucinated relationship.</li> <li>2. 'modification_details': A dictionary containing: <ul style="list-style-type: none"> <li>• 'added_phrase': The phrase or sentence added to the paragraph.</li> <li>• 'context_with_incorrect_answer': The phrase containing the full hallucinated relationship, with at least five words.</li> </ul> </li> <li>3. 'hallucinated_relationship': A dictionary with: <ul style="list-style-type: none"> <li>• 'subject': A dictionary with: <ul style="list-style-type: none"> <li>✓ 'original_term': The original term of the subject from the incorrect answer.</li> <li>✓ 'updated_term': The term of the subject as it appears in the updated paragraph.</li> </ul> </li> <li>• 'predicate': A dictionary with: <ul style="list-style-type: none"> <li>✓ 'original_term': The original name of the predicate from the incorrect answer.</li> <li>✓ 'updated_term': The term of the predicate as it appears in the updated paragraph.</li> </ul> </li> <li>• 'object': A dictionary with: <ul style="list-style-type: none"> <li>✓ 'original_term': The original name of the object from the incorrect answer.</li> <li>✓ 'updated_term': The term of the object as it appears in the updated paragraph.</li> </ul> </li> </ul> </li> </ol>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 17. Prompt to inject hallucinated answers pertaining to <rel> hallucination.

<p><b>Task:</b> You are an AI assistant with strong reasoning abilities and advanced visual commonsense knowledge.</p> <p><b>Provided Information:</b></p> <ol style="list-style-type: none"> <li>1. A paragraph describing an image.</li> <li>2. A question about the paragraph.</li> <li>3. An incorrect answer to the question.</li> <li>4. The incorrect place, location, or weather mentioned in the incorrect answer.</li> <li>5. A possible insertion point: the phrase within the original paragraph where the hallucinated place, location, or weather could be inserted.</li> </ol> <p><b>Goal:</b> The incorrect answer includes a incorrect place, location, or weather. You need to add a phrase or sentence to the original paragraph that introduces the incorrect place, location, or weather.</p> <p><b>Requirements:</b></p> <ul style="list-style-type: none"> <li>- Do not introduce any additional details beyond what is already in the paragraph.</li> <li>- Make no other changes to the paragraph except for the addition of the phrase or sentence that introduces the incorrect place, location, or weather.</li> <li>- Try to introduce the hallucinated place, location, or weather in the insertion point.</li> </ul> <p><b>Response Format:</b></p> <ol style="list-style-type: none"> <li>1. 'updated_paragraph': The paragraph with the added phrase or sentence introducing the non-existent object.</li> <li>2. 'modification_details': A dictionary containing: <ul style="list-style-type: none"> <li>• 'added_phrase': The phrase or sentence added to the paragraph.</li> <li>• 'context_with_incorrect_answer': The incorrect answer within the added phrase. Should be at least five words long.</li> </ul> </li> <li>3. 'incorrect_place_location_weather': A dictionary with: <ul style="list-style-type: none"> <li>• 'original_term': The original name of the incorrect place, location, or weather as it appeared in the incorrect answer.</li> <li>• 'updated_term': The name of the incorrect place, location, or weather as it appears in the updated paragraph.</li> </ul> </li> </ol>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 18. Prompt to inject hallucinated answers pertaining to <sce> hallucination.

<p><b>Task:</b> You are an AI assistant with strong reasoning abilities and advanced visual commonsense knowledge.</p> <p><b>Provided Information:</b></p> <ol style="list-style-type: none"> <li>1. The original paragraph describing an image.</li> <li>2. The hallucinated object that was added to the paragraph.</li> <li>3. The modified paragraph after the addition.</li> <li>4. The specific phrase that was added.</li> <li>5. The object mentioned in the added phrase.</li> <li>6. A list of objects actually present in the image.</li> </ol> <p><b>Task:</b> Decide whether the modified paragraph should be kept or discarded. Discard the modified paragraph if:</p> <ul style="list-style-type: none"> <li>- The hallucinated object is a synonym of an existing object in the image.</li> <li>- The added phrase introduces additional hallucinated objects not present in the image.</li> <li>- The hallucinated object and the object mentioned in the added phrase refer to completely different things.</li> </ul> <p><b>Response Format:</b></p> <ol style="list-style-type: none"> <li>1. 'decision': 'keep' or 'discard'</li> <li>2. 'reason': 'is_a_synonym', 'introduced_additional_hallucinated_objects', 'different_terms', or 'keep'</li> <li>3. 'detailed_reason': 'A detailed explanation of the decision.'</li> </ol>
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 19. Prompt to verify each injection step of injecting hallucinated answers for <obj> hallucination.

<p><b>Task:</b> You are an AI assistant with strong reasoning abilities and advanced visual commonsense knowledge.</p> <p><b>Provided Information:</b></p> <ol style="list-style-type: none"> <li>1. The original paragraph describing an image.</li> <li>2. The hallucinated attribute and corresponding object that was added to the paragraph.</li> <li>3. The modified paragraph after the addition.</li> <li>4. The specific phrase that was added.</li> <li>5. The mentioned hallucinated attribute and object in the modified paragraph.</li> <li>6. A list of actual attributes of the object.</li> <li>7. A list of additional details of the image.</li> </ol> <p><b>Task:</b> Decide whether the modified paragraph should be kept or discarded. Discard the modified paragraph if:</p> <ul style="list-style-type: none"> <li>- The hallucinated attribute is a synonym of an actual attribute of the object.</li> <li>- The added phrase introduces additional hallucination apart from the intended hallucinated attribute.</li> <li>- The intended hallucinated attribute and object differ from the mentioned hallucinated attribute and object in the modified paragraph.</li> </ul> <p><b>Response Format:</b></p> <ol style="list-style-type: none"> <li>1. 'decision': 'keep' or 'discard'</li> <li>2. 'reason': 'is_a_synonym', 'introduced_additional_hallucinated_objects', 'different_terms', or 'keep'</li> <li>3. 'detailed_reason': 'A detailed explanation of the decision.'</li> </ol>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 20. Prompt to verify each injection step of injecting hallucinated answers for <attr> hallucination.

<p><b>Task:</b> You are an AI assistant with strong reasoning abilities and advanced visual commonsense knowledge.</p> <p><b>Provided Information:</b></p> <ol style="list-style-type: none"> <li>1. The original paragraph describing an image.</li> <li>2. The hallucinated relationship that was added to the paragraph.</li> <li>3. The modified paragraph after the addition.</li> <li>4. The specific phrase that was added.</li> <li>5. The mentioned hallucinated relationship in the modified paragraph.</li> <li>6. A list of additional details about the image.</li> </ol>
<p><b>Task:</b> Decide whether the modified paragraph should be kept or discarded based on the following criteria:</p> <ul style="list-style-type: none"> <li>- Discard if the added phrase introduces other hallucinations beyond the intended hallucinated relationship.</li> <li>- Discard if the intended hallucinated relationship differs from the mentioned hallucinated relationship in the modified paragraph.</li> </ul>
<p><b>Response Format:</b></p> <ol style="list-style-type: none"> <li>1. 'decision': 'keep' or 'discard'</li> <li>2. 'reason': 'is_a_synonym', 'introduced_additional_hallucinated_objects', 'different_terms', or 'keep'</li> <li>3. 'detailed_reason': 'A detailed explanation of the decision.'</li> </ol>

Figure 21. Prompt to verify each injection step of injecting hallucinated answers for <rel> hallucination.

<p><b>Task:</b> You are an AI assistant with strong reasoning abilities and advanced visual commonsense knowledge.</p> <p><b>Provided Information:</b></p> <ol style="list-style-type: none"> <li>1. The original paragraph describing an image.</li> <li>2. The hallucinated place, location, or weather that was added to the paragraph.</li> <li>3. The modified paragraph after the addition.</li> <li>4. The specific phrase that was added.</li> <li>5. The place, location, or weather mentioned in the added phrase.</li> </ol>
<p><b>Task:</b> Decide whether the modified paragraph should be kept or discarded.</p> <p>Discard the modified paragraph if:</p> <ul style="list-style-type: none"> <li>- The added phrase introduces additional hallucination apart from the intended hallucinated place, location, or weather.</li> <li>- The hallucinated place, location, or weather and the place, location, or weather mentioned in the added phrase refer to completely different things.</li> </ul>
<p><b>Response Format:</b></p> <ol style="list-style-type: none"> <li>1. 'decision': 'keep' or 'discard'</li> <li>2. 'reason': 'is_a_synonym', 'introduced_additional_hallucinated_objects', 'different_terms', or 'keep'</li> <li>3. 'detailed_reason': 'A detailed explanation of the decision.'</li> </ol>

Figure 22. Prompt to verify each injection step of injecting hallucinated answers for <sce> hallucination.

## G. Qualitative Analysis of *HalLocalizer*

We provide a qualitative analysis of the performance of *HalLocalizer*. We use the *HalLocalizer* trained with LLaVA [3] embeddings to conduct the analysis.

### G.1. Success Scenarios

Figure 23 presents examples where *HalLocalizer* successfully localized hallucinations and correctly identified their types. Figure 23 (a) and (b) show successful cases from *HalLoc-VQA*, while (c) and (d) depict successful examples from *HalLoc-Instruct*. Figure 23 (e) illustrates a successful case from *HalLoc-Caption*.

### G.2. Failure Scenarios

Figure 24 illustrates several instances where *HalLocalizer* did not perform as expected on the test cases from *HalLoc*.

Figure 24 (a) depicts a frequent failure pattern where *HalLocalizer* incorrectly identifies a non-existent object in the image as a relationship type. This error occurs because the question prompts for a relationship, but *HalLoc* classifies any non-existent object as an object type rather than a relationship.

Figure 24 (b) presents another common failure, where *HalLocalizer* only partially identifies components of hallucination. Such failure cases show that *HalLocalizer* is imperfect in identifying word boundaries, which should be explored further in future research.

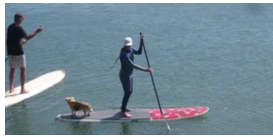
Figure 24 (c) highlights a recurring issue, particularly with *HalLoc-Caption*, where *HalLocalizer* fails to detect any hallucinated tokens. This problem likely stems from the sparse presence of hallucinated tokens in lengthy responses. Future research should explore innovative approaches to address this challenge.

### G.3. *HalLocalizer* in Real-World Scenarios

We evaluate *HalLocalizer* on free-form responses generated by real-world VLMs to assess its performance in practical settings beyond *HalLoc*. Figure 25 shows case studies where *HalLocalizer* is applied to detailed descriptions generated by LLaVA [3]. These examples highlight the robustness and adaptability of *HalLocalizer* when dealing with the complexity and variability of real-world data.



- (a) **Prompt** Given the image, answer the following question with no more than three words. What is the dog standing on?



Prediction:

**Relation** paddle ✓

Correct Hallucination Label:

**Relation** paddle

- (c) **Prompt** Explain: Is the girl wearing boots?



Prediction:

**Relation** No, the girl is carrying the boots ✓

Correct Hallucination Label:

**Relation** No, the girl is carrying the boots

- (b) **Prompt** Question: Which color is the skirt?



Prediction:

**Attribute** black ✓

Correct Hallucination Label:

**Attribute** black

- (d) **Prompt** What is the answer to the following question: "What color is the blanket at the bottom?"



Prediction:

**Attribute** The blanket is brown. ✓

Correct Hallucination Label:

**Attribute** The blanket is brown.

- (e) **Prompt** Write a detailed description of the given image.



Prediction: **Object**

This is an image taken in a restaurant. In the foreground of the picture, there is a table; on the table, there is bowl, spoon, and dish, and in front of the table, there is a person sitting in a chair. Behind him, there are many people sitting in chairs, and there are many tables. On the tables, there are candles, flower vase, and other objects. There is a **racket** placed between the tables. On the left there is a light. Background is little dark.

Correct Hallucination Label: **Object**

This is an image taken in a restaurant. In the foreground of the picture, there is a table; on the table, there is bowl, spoon, and dish, and in front of the table, there is a person sitting in a chair. Behind him, there are many people sitting in chairs, and there are many tables. On the tables, there are candles, flower vase, and other objects. There is a **racket** placed between the tables. On the left there is a light. Background is little dark.

Figure 23. Success cases of HalLoc: Correctly localizes and identifies hallucination type (a) relationship and (b) attribute in *HalLoc-VQA*. Correctly localizes and identifies hallucination type (c) relationship and (d) attribute in *HalLoc-Instruct*. Correctly localizes and identifies hallucination type (e) object in *HalLoc-Caption*.

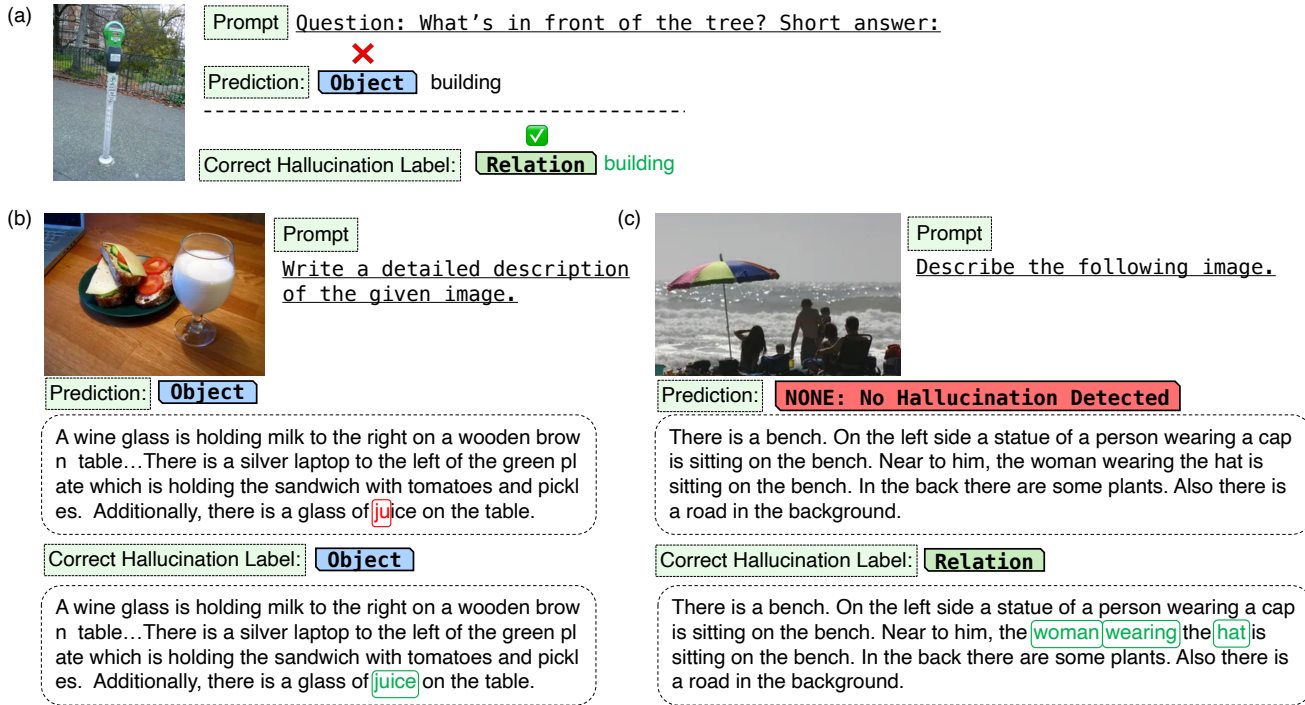


Figure 24. Failure cases: (a) Correctly detects hallucination but assigns the wrong hallucination type. (b) Correctly assigns the hallucination type but incorrect token-level detection. (c) Case where no hallucination was detected when there is a relationship hallucination present in the hallucinated text.

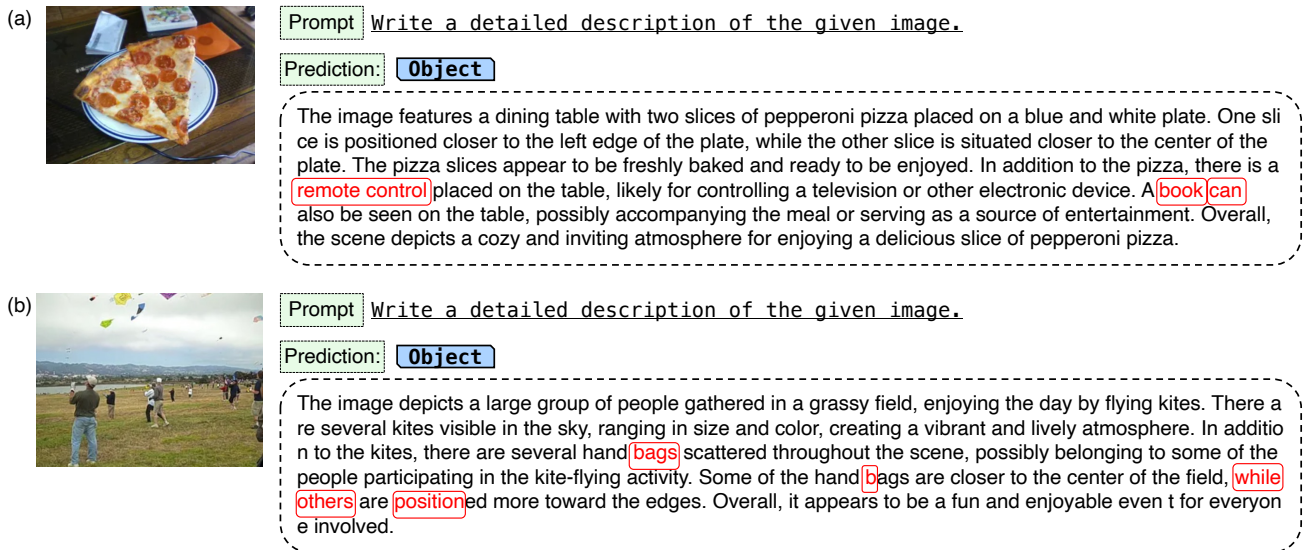


Figure 25. Examples where *HalLocalizer* detects and identifies hallucinations in image captions generated by a vision language model. We used LLaVA [3] for our example.

## References

- [1] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. 1
- [2] Drew A Hudson and Christopher D Manning. 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, 2019. 7
- [3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 16, 18
- [4] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015:2901–2907, 2015. 1
- [5] Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning, 2020. 1
- [6] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. 7