Benchmarking Large Vision-Language Models via Directed Scene Graph for Comprehensive Image Captioning

Supplementary Material

A. More Experimental Analysis

A.1. Analysis on the effect of components of LVLMs

We mainly analyze the effect of image resolution and base Large Language Model (LLM) on the quality of generated detailed captions and the perception of tiny objects in Tab. S1 and Tab. S2, respectively. Compared with LLaVA-1.5-13B [31], LLava-Next-13B [30] increases the input image resolution. The performance comparison of generating detailed captions between these two LVLMs indicates that increasing the input image resolution can enable the LVLM to recognize more objects, and describe the objects along with the key relations between them in a more precise way. When upgrading the base LLM to the larger and more powerful Yi-34B [48], LLava-Next-34B [30] further widened its advantage over LLava-Next-13B [31] in generating comprehensive captions as shown in Tab. S1. Moreover, we present the performance of 10 LVLMs in CompreQA-P and CompreQA-Cap shown as Tab. S2. We compare LLaVA-1.5-13B [31], LLava-Next-13B [31] and LLava-Next-34B [30] and observe that gains brought by the base LLM have given LLaVA-Next-34B[30] a significant advantage on the two fine-grained objects VQA metrics. While increasing image resolution often improves fine-grained object understanding, we find that it is not the sole contributor to this task. For example, although the input resolution of InternVL-Chat-V1-5 [10] is merely 448, not even one-third that of miniGemini-34B-HD [25], the former surpasses the latter by 4.79% and 3.54% in two metrics, respectively. The comparison between LLava-Next-13B [30] and LLaVA-1.5-13B [31] shows that the former features a several-fold increase in input resolution compared to the latter, and outperforms the latter on the CompreQA-P. However, LLava-Next-13B [30] falls short on the CompreQA-Cap evaluation. This reaffirms that merely increasing input resolution is insufficient to enhance the model's comprehension of fine-grained objects.

A.2. Consistency with human evaluation scores

We present the complete results of human evaluation scores across 10 LVLMs and human performance in Tab. S3. All our metrics on *CompreCap* dataset achieve a strong consistency with human evaluation scores across all LVLMs and human performance, emphasizing the credibility of our *CompreCap* and its promising prospects in LVLMs evaluation.

A.3. Compare with directly scoring with Llama3

We first organize the annotated objects and the descriptions of attributes and relationships into coherent ground-truth (GT) captions. Then, we leverage Llama3 [14] to directly score the generated detailed captions, with the prompt 'Please quantify the quality of the given ⟨generated caption⟩ on a score scale from 0 to 5 for 'object coverage', 'object attributes', 'relationships between objects' and 'overall quality', without any other explanation, using the given ⟨GT caption⟩ as a standard. The higher the score, the more the given ⟨generated caption⟩ matches the content of the given ⟨GT caption⟩.'. Then we will obtain the object score, attribute score, relation score as well as an overall score. The results are shown in Tab. S4. When directly assessing with Llama3 [14], the high-quality captions generated by human lag behind most LVLMs on scores of all dimensions, which is unexpected and unreasonable as shown in Fig. S5. This indicates that with the annotated directed scene graph of *CompreCap*, we can decouple the generated captions into a hierarchical structure, allowing for a more precise match with annotations at the levels of objects, attributes, and relationships.

We select the human performance, two top-performing models in Tab. 2 of the manuscript paper, GPT-4o and LLaVA-Next-34B, as well as a medium-performing model, ShareGPT4V13B, and a low-performing model, InstructBLIP7B, for comparative analysis. With the human assessment scores of the image captions generated by both LVLMs and human, the metrics derived from our method faithfully align with human evaluation scores compared to those directly obtained from Llama3, as shown in Fig. S1. The high consistency with human judgment emphasizes the credibility and practicality of our CompreCap, proving its promising prospects in LVLMs evaluation.

Table S1. Analysis on the influence of image resolution and base LLM on the generated detailed captions. 'Max Res.' denotes the maximum resolution.

Model	LLM	Max Res.	Caption Length	$S_{\text{object}}(\%)\uparrow$	$S_{\text{attribute}}\uparrow$	$S_{\text{relation}}\uparrow$	S-Cov.(%) $\uparrow$
LLaVA-1.5-13B [31]	Vicuna-13B	336	86.97	59.93	2.02$\pm$*0.01*	2.60$\pm$*0.00*	44.11$\pm$*0.01*
LLava-Next-13B [30]	Vicuna-13B	1008	172.19	70.55	2.50$\pm$*0.01*	2.73$\pm$*0.01*	56.68$\pm$*0.28*
LLava-Next-34B [30]	Yi-34B	1008	179.24	72.86	2.59$\pm$*0.00*	2.79$\pm$*0.00*	58.49$\pm$*0.15*

Table S2. Accuracy of LVLMs on CompreQA-P and CompreQA-Cap. The **best** results are in bold. The second and third best results are in underline and double underline, respectively. 'Max Res.' denotes the maximum resolution.

Model	Visual Encoder	LLM	Max Res.	CompreQA-P ACC(%) $\uparrow$	CompreQA-Cap ACC(%) $\uparrow$
InstructBLIP-7B [12]	EVA-ViT-G	Vicuna-7B	224	35.28$\pm$*0.00*	36.52$\pm$*0.00*
MiniGPT4-v2 [55]	EVA-ViT-G	Llama2-Chat-7B	448	51.06$\pm$*0.14*	40.78$\pm$*1.16*
LLaVA-1.5-13B [31]	CLIP-ViT-L/14	Vicuna-13B	336	82.45$\pm$*0.38*	84.87$\pm$*0.17*
LLava-Next-13B [30]	CLIP-ViT-L/14	Vicuna-13B	1008	84.81$\pm$*0.59*	81.68$\pm$*0.44*
ShareGPT4V-13B [9]	CLIP-ViT-L/14	Vicuna-13B	336	82.03$\pm$*0.44*	85.34$\pm$*0.33*
miniGemini-34B-HD [25]	CLIP-ViT-L/14 & CLIP-ConvNext-L	Yi-34B	1536	86.88$\pm$*0.90*	89.01$\pm$*0.29*
LLaVA-Next-llama3-8B [30]	CLIP-ViT-L/14	Llama3-8B-Instruct	1008	88.48$\pm$*0.52*	90.90$\pm$*0.17*
InternVL-Chat-V1-5 [10]	InternViT-6B-V1-5	InternLM2-Chat-20B	448	**91.67**$\pm$*0.00*	**94.33**$\pm$*0.00*
LLaVA-Next-34B [30]	CLIP-ViT-L/14	Yi-34B	1008	91.43$\pm$*0.22*	92.55$\pm$*0.00*
GPT-4o [35]	-	-	-	90.96$\pm$*0.38*	91.37$\pm$*0.33*

A.4. Ability of Llama3 to Distinguish Caption Quality

With the human annotation in the format of directed scene graph, we can apply Llama3 [14] to assess the attribute descriptions bound to objects and directional relation descriptions between objects, instead of individual matching the words of objects, attributes and relations. To verify the ability of Llama3 to distinguish caption quality, we construct 50 pairs of good and bad captions by randomly shuffling object attributes and swapping the subject and object in relation descriptions, as illustrated in Fig. 2 of the manuscript paper. We show some examples in Fig. S2. Then, we require Llama3 [14] to score these captions on a scale from 0 to 5 using the prompt similar to that in Fig. 3 of the manuscript paper. We compare the scores with those obtained from the evaluation method of DetailCaps [13], which isolately assess the words of objects, attributes and relations. The comparison in Tab. S5 reveals that Llama3 [14] effectively discerns factual errors, such as those illustrated by the bad captions when the correct structure of scene graph is disrupted.

A.5. Analysis on the influence of LLM evaluator

GPT4 [1] is known as the best close-source LLM. In this part, we employ GPT4 as the evaluator to assess the quality of detailed captions generated by 10 LVLMs and human. The prompt used for evaluation is the same as that in Fig. 3 of the manuscript paper. The evaluation results presented in Fig. S3 show that our metrics obtained by using GPT-4 also demonstrate a strong alignment with human evaluation scores. We choose LLama3 as the evaluator considering the advantages of non-API evaluation in terms of speed and stability.

It is important to note that, to fully leverage the discriminative power of Llama3 [14], it is necessary to decouple the detailed captions and match them with annotations at the levels of object, attribute, and relation. Otherwise, as highlighted in App. A.3, directly scoring the detailed captions with dense text using Llama3 [14] fails to yield solid evaluation results on *CompreCap*.

A.6. Analysis on the effect of parsers

Our evaluation method applies the spaCy [18] parser to parse nouns and decouple the generated captions into a hierarchical structure. We present the consistency of results using another well-known parser, *i.e.*, NLTK [33], with human evaluation scores in Fig. S6. The high consistency shows that our metrics are not sensitive to parsers.

A.7. Results of traditional caption metrics on CompreCap

We first evaluate 10 LVLMs and human performance with traditional caption metrics, including BLEU-4 [36], METEOR [3], ROUGE-L [27], CIDER [43], and CLIPScore [17], on our *CompreCap* dataset, then show the consistencies of these traditional metrics with human evaluation scores in Fig. S4. The scores of

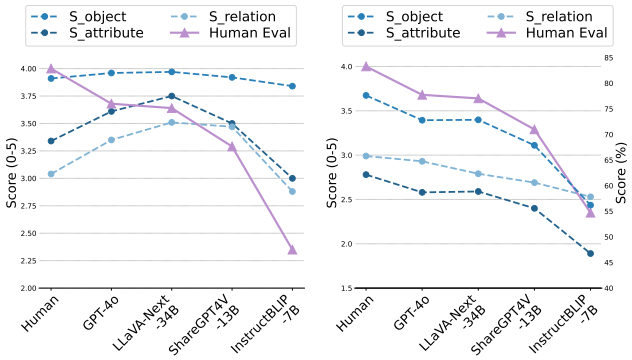
Table S3. Evaluation of the detailed captions generated by the 10 LVLMs on *CompreCap* benchmark, including the human evaluation score reported. The **best** results are highlighted in bold. The second and third best results are highlighted in underline and double underline, respectively.

Model	Caption Length	$S_{\text{object}}(\%)\uparrow$	$S_{\text{attribute}}\uparrow$	$S_{\text{relation}}\uparrow$	S-Cov.$(\%)\uparrow$	$S_{\text{unified}}\uparrow$	Human Evaluation
InstructBLIP-7B [12]	69.93	56.20	1.89±*0.00*	2.53±*0.00*	42.03±*0.14*	48.16	2.36
MiniGPT4-v2 [55]	350.42	56.74	1.86±*0.00*	1.88±*0.01*	43.03±*0.19*	42.28	1.83
LLaVA-1.5-13B [31]	86.97	59.86	2.01±*0.01*	2.59±*0.00*	43.81±*0.25*	50.32	2.39
ShareGPT4V-13B [9]	155.91	67.88	2.40±*0.01*	2.69±*0.00*	55.86±*0.17*	55.56	3.28
LLaVA-Next-llama3-8B [30]	168.99	70.22	2.48±*0.00*	2.72±*0.01*	56.95±*0.08*	56.91	3.34
miniGemini-HD-34B [25]	173.71	70.70	2.48±*0.00*	2.70±*0.00*	57.20±*0.11*	56.88	3.37
InternVL-Chat-V1-5 [10]	115.22	70.56	2.50±*0.00*	2.87±*0.00*	57.58±*0.09*	58.48	3.42
LLaVA-Next-34B [30]	179.24	**72.86**	**2.59**±*0.00*	2.79±*0.00*	**58.49**±*0.15*	58.85	3.64
GPT-4V [47]	202.06	72.31	2.52±*0.00*	2.73±*0.00*	57.27±*0.14*	57.74	3.63
GPT-4o [35]	108.20	72.78	2.58±*0.00*	**2.93**±*0.00*	57.54±*0.23*	**60.05**	3.68
Human	133.61	**77.62**	**2.78**±*0.00*	**2.99**±*0.00*	**59.58**±*0.16*	**62.99**	4.0

Table S4. Directly scoring the detailed captions generated by the 10 LVLMs with Llama3 [14] on *CompreCap* benchmark. The **best** results are highlighted in bold. The second and third best results are highlighted in underline and double underline, respectively. The scores directly output by Llama3 [14] across various dimensions show a weak consistency with human evaluation scores.

Model	Caption Length	Object Score	Attribute Score	Relation Score	Overall Score	Human Evaluation
InstructBLIP-7B [12]	69.93	3.84±*0.01*	3.00±*0.02*	2.88±*0.02*	3.18±*0.02*	2.36
MiniGPT4-v2 [55]	350.42	2.26±*0.01*	1.67±*0.02*	1.32±*0.02*	1.66±*0.02*	1.83
LLaVA-1.5-13B [31]	86.97	3.79±*0.00*	2.89±*0.01*	2.73±*0.01*	3.07±*0.01*	2.39
ShareGPT4V-13B [9]	155.91	3.92±*0.01*	3.50±*0.01*	3.47±*0.02*	3.68±*0.01*	3.28
LLava-Next-llama3-8B [30]	168.99	3.96±*0.00*	3.62±*0.00*	3.50±*0.02*	3.76±*0.01*	3.34
miniGemini-HD-34B [25]	173.71	3.97±*0.00*	**3.76**±*0.01*	3.50±*0.02*	**3.81**±*0.01*	3.37
InternVL-Chat-V1-5 [10]	115.22	**3.98**±*0.01*	3.62±*0.01*	3.32±*0.02*	3.69±*0.01*	3.42
LLava-Next-34B [30]	179.24	3.97±*0.01*	3.75±*0.01*	**3.51**±*0.00*	3.79±*0.01*	3.64
GPT-4V [47]	202.06	3.96±*0.00*	3.74±*0.01*	3.34±*0.01*	3.74±*0.01*	3.63
GPT-4o [35]	108.20	3.96±*0.00*	3.61±*0.01*	3.35±*0.01*	3.67±*0.00*	3.68
Human	133.61	3.91±*0.00*	3.34±*0.01*	3.04±*0.01*	3.40±*0.01*	4.0

captions generated by human do not outperform all LVLMs on all traditional metrics. And all traditional caption metrics fail to align with human judgment. The results show that existing traditional metrics cannot reasonably evaluate comprehensive image captions consisting of dense text.



(a) Directly scoring with Llama3 (b) **Our evaluation method**

Figure S1. Our method demonstrates a high consistency with human evaluation scores across all models, whereas the results directly produced by Llama3 exhibit conflicts. This indicates that the annotations of the directed scene graph facilitate more accurate matching at the levels of objects, attributes and relations during comprehensive caption assessment.

Figure S2. We construct the bad captions through shuffling the attributes and swapping the subject and object in relation descriptions. The correct and incorrect components compared with the reference caption are marked in green and red, respectively.

Reference: A blue refrigerator is next to a white cabinet.
Good Caption: There are a blue refrigerator and a white cabinet.
Bad Caption: There are a white refrigerator and a blue cabinet.

Reference: A woman is in front of the panda.
Good Caption: There is a woman in front of the panda.
Bad Caption: There is a panda in front of the woman.

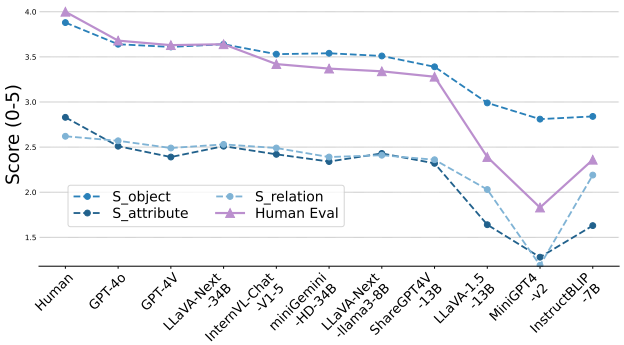
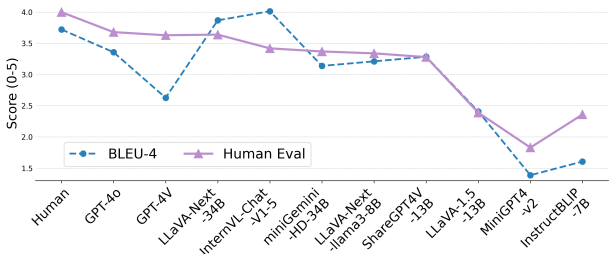
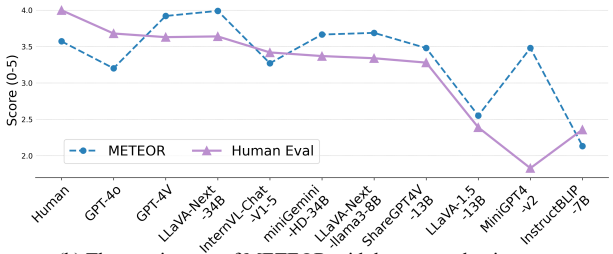


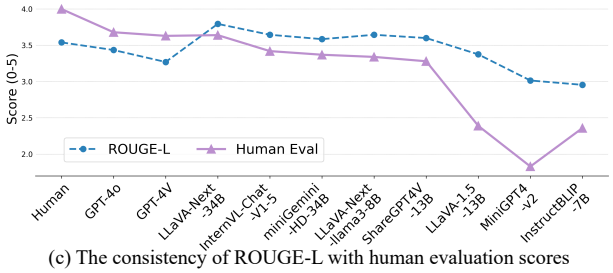
Figure S3. The consistency of results using GPT4 [1] with human evaluation scores across 10 LVLMs and human performance. All our metrics obtained using GPT4 [1] align with human judgment.



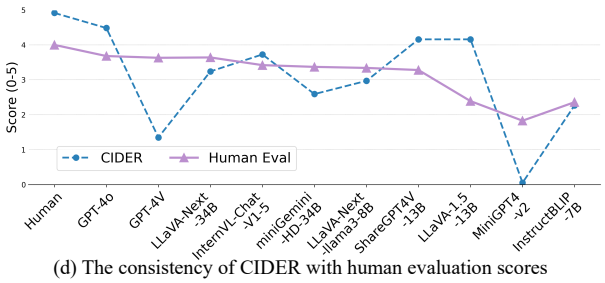
(a) The consistency of BLEU-4 with human evaluation scores



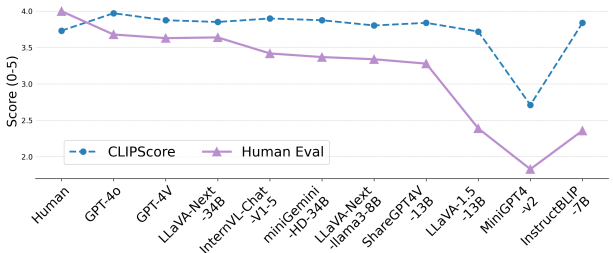
(b) The consistency of METEOR with human evaluation scores



(c) The consistency of ROUGE-L with human evaluation scores

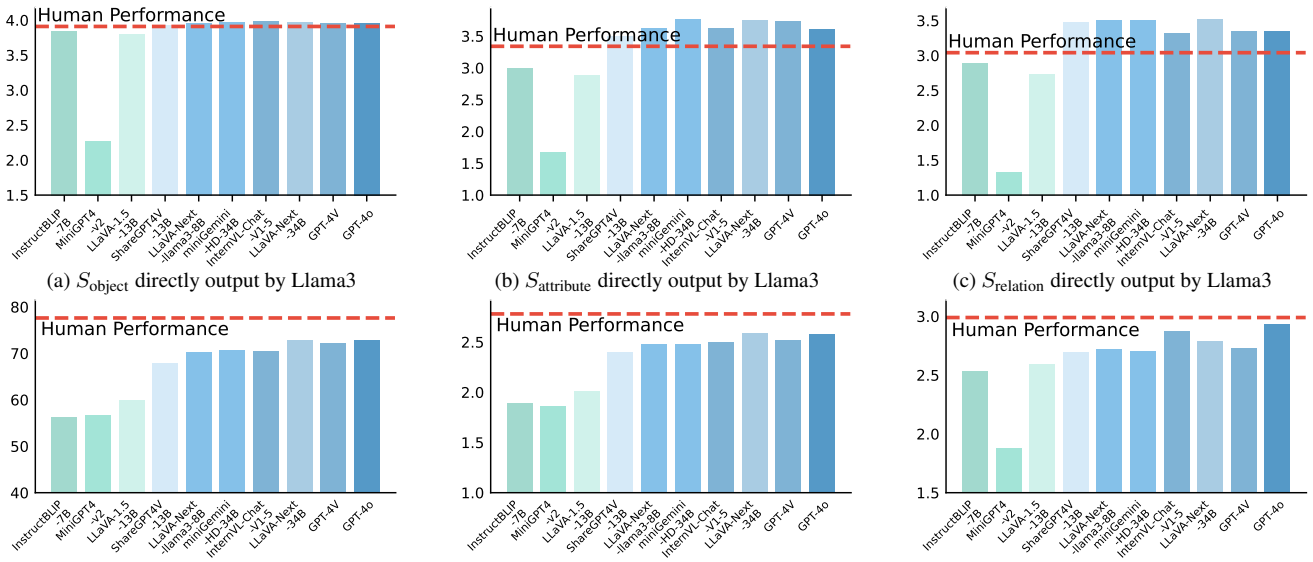


(d) The consistency of CIDER with human evaluation scores



(e) The consistency of CLIPScore with human evaluation scores

Figure S4. The consistency of traditional caption metrics with human evaluation scores across 10 LVLMs and human performance. All traditional caption metrics fail to align with human judgment. The scores of all traditional metrics are linearly scaled to 0-5.



(a) S_{object} directly output by Llama3 (b) $S_{\text{attribute}}$ directly output by Llama3 (c) S_{relation} directly output by Llama3

(d) S_{object} from **our method** (e) $S_{\text{attribute}}$ from **our method** (f) S_{relation} from **our method**

Figure S5. The human performance exceed all LVLMs on the scores of objects, attributes and relations with our evaluation method. Compared to directly scoring detailed captions with Llama3, the annotations of the directed scene graph provide more precise references across various dimensions.

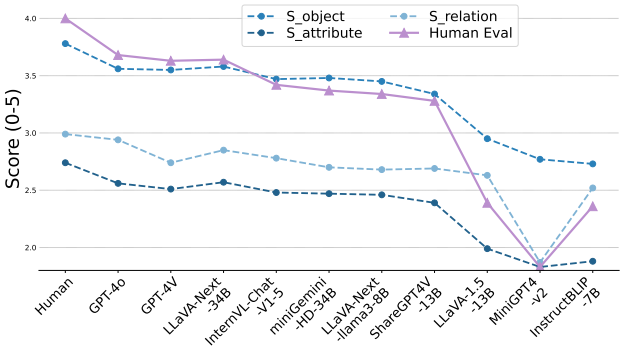


Figure S6. The consistency of results based on NLTK [18] with human evaluation scores across 10 LVLMs and human performance. All our metrics obtained based on NLTK [18] align with human judgment.

Table S5. The score gaps between good and bad captions. The comparison reveals that Llama3 [14] effectively distinguishes the quality difference between good and bad captions. 'Good' and 'Bad' denote 'good captions' and 'bad captions', respectively. We conducted three tests repeatedly, and the mean along standard deviation are reported.

DetailCaps (F1-score)		Llama3 (0∼5)	
Good	Bad	Good	Bad
$83.15_{\pm 0.00}$	$82.83_{\pm 0.00}$	$4.75_{\pm 0.00}$	$1.98_{\pm 0.02}$

A.8. Analysis on the stability of generated detailed caption

In Tab. 2 of the manuscript, we include error bars in the evaluation results to certify the reliability and consistency of our evaluation methodology. However, LVLMs can generate different detailed captions even we use the same prompt for guidance. To investigate the fluctuations in the quality of detailed captions, we repeatedly utilize GPT-4o [35] and LLava-Next-34B [30] to generate detailed captions three times and evaluate their quality. The average performances along the error bars are reported in Tab. S6. Although the averaged length of the generated detailed captions differs each time, the evaluation metric scores remain relatively consistent.

A.9. Analysis of prompts for caption generation

We analyze the effects of prompts used for caption generation with three different prompts in Tab. S7. The compared results between prompts #1 and #2 indicate that emphasizing the descriptions of objects can enhance the quality of generated captions across all metrics. And prompt #3 which highlights both objects and relations further improves the comprehensiveness of generated detailed captions. Overall, more explicit and detailed prompts contribute to the generation of higher-quality captions by LVLMs.

A.10. Analysis on the failure cases in the fine-grained object VQA

We analyze the segmentation map coverage distributions of 9 LVLMs' error objects in CompreQA-for-Caption and show the distributions in Fig. S7. All 9 LVLMs tend to choose inaccurate captions for tiny objects whose segmentation map coverage $<2\%$, indicating that the smaller the object, the greater the challenge for LVLMs to accurately describe it.

We present comparison examples mong LLava-Next-34B [30], InternVL-Chat-V1-5 [10] and GPT-4o [35] on CompreQA-for-Presence and CompreQA-for-Caption in Fig. S8 and show cases where GPT-4o [35] fails in. We observe GPT-4o [35] fails to accurately understand fine-grained objects in the background (*e.g.*,

Table S6. Investigate the fluctuations in the detailed captions generated by GPT-4o and LLava-Next-34B.

Model	Visual Encoder	LLM	Caption Length	$S_{\text{object}}(\%)\uparrow$	$S_{\text{attribute}}\uparrow$	$S_{\text{relation}}\uparrow$	S-Cov.$(\%)\uparrow$
LLava-Next-34B [30]	CLIP-ViT-L/14	Yi-34B	179.87$\pm$*0.67*	72.46$\pm$*0.30*	2.57$\pm$*0.01*	2.79$\pm$*0.01*	58.47$\pm$*0.41*
GPT4o [47]	-	-	108.06$\pm$*0.21*	72.79$\pm$*0.12*	2.58$\pm$*0.00*	2.92$\pm$*0.02*	57.62$\pm$*0.18*

Table S7. Analysis on the effect of prompts used for detailed caption generation. Prompt #1: 'Please describe the image in detail.'. Prompt #2: 'Please describe the image in detail, focusing on the visible objects.' Prompt #3: 'Please describe the image in detail, focusing on the visible objects and the relationships among these objects.'

Model	Prompt	Caption Length	$S_{\text{object}}(\%)\uparrow$	$S_{\text{attribute}}\uparrow$	$S_{\text{relation}}\uparrow$	S-Cov$(\%)\uparrow$
GPT4o [20]	#1	90.46	70.22	2.48$\pm$*0.00*	2.89$\pm$*0.00*	56.74$\pm$*0.18*
	#2	91.05	71.38	2.54$\pm$*0.00*	2.92$\pm$*0.01*	57.20$\pm$*0.07*
	#3	108.20	72.78	2.58$\pm$*0.00*	2.93$\pm$*0.00*	57.54$\pm$*0.23*
LLava-Next-34B [30]	#1	153.38	69.38	2.45$\pm$*0.00*	2.71$\pm$*0.00*	56.96$\pm$*0.05*
	#2	154.22	70.90	2.53$\pm$*0.01*	2.78$\pm$*0.00*	57.62$\pm$*0.11*
	#3	179.24	72.86	2.59$\pm$*0.00*	2.79$\pm$*0.00*	58.49$\pm$*0.15*

the 'bicycle'). Then, we directly ask GPT-4o [35] to describe the tiny objects which are misinterpreted in CompreQA. The illustration in Fig. S8 shows that the descriptions generated by GPT-4o [35] conflict the visual content, as it overlooks the 'tv' within the carriage and mischaracterizes the attributes of the 'telephone' and 'bicycle'.

Table S8. Comparison with similar benchmarks. Visual-Genome includes VG-Attributes and VG-Relations.

	Length	Mask	Object	Attribute	Relation
COCO-Stuff	10	✓	✓	-	-
COCO-OOD	10	✓	✓	-	-
SugarCrepe	11	-	✓	-	-
Visual-Genome	92	-	✓	✓	✓
CompreCap	**172**	✓	✓	✓	✓

B. Comparison with Similar Benchmarks

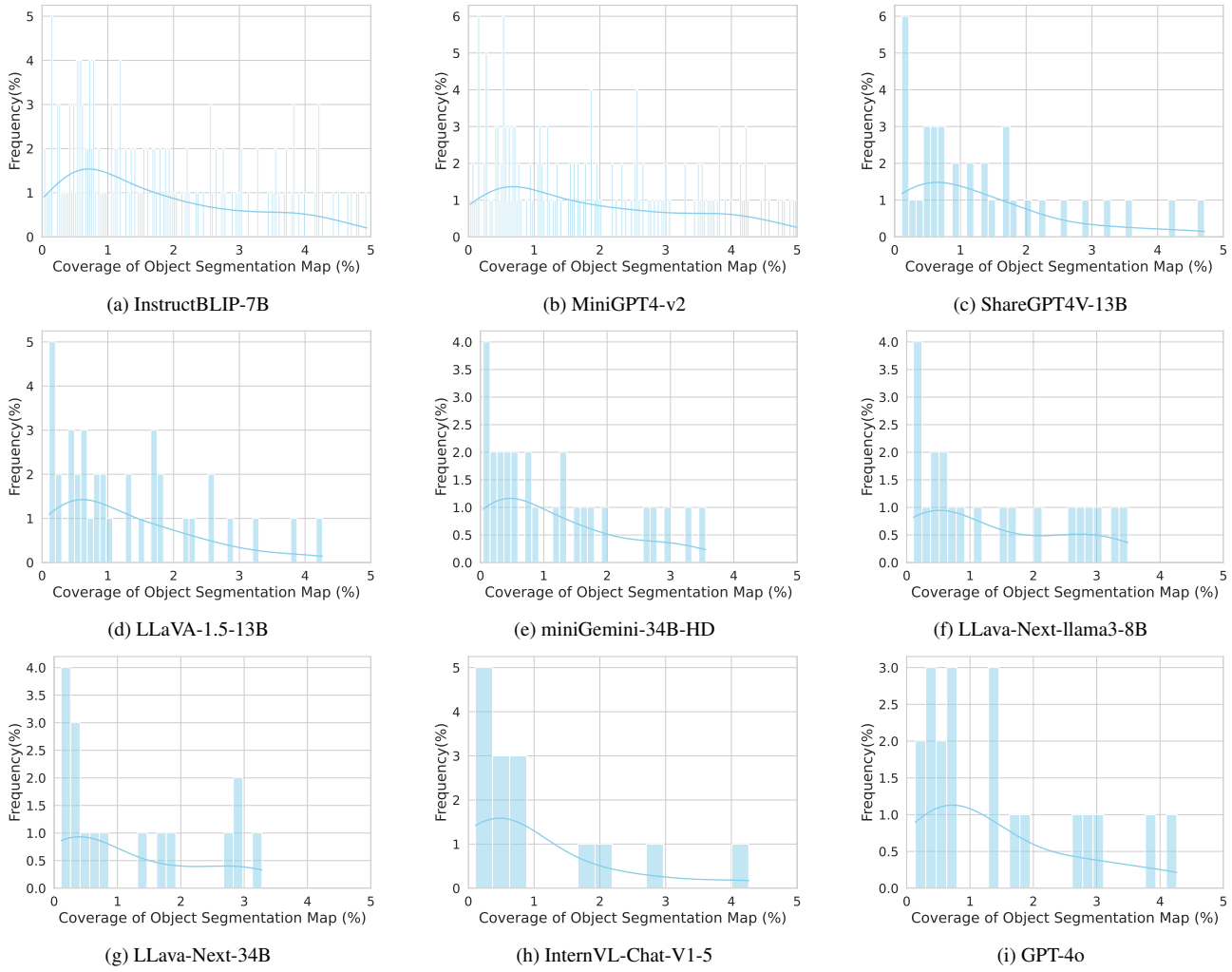
Here we compare several related benchmarks in Tab. S8, especially Visual-Genome (VG) [21] and other COCO-based benchmarks. COCO-Stuff/OOD [7, 19] have object and mask annotations, while SugarCrepe generates various forms of hard negatives based on COCO's short captions (averaged length is about 10 words). VG has object, attribute, and relation annotation with a bounding box, which has about 92 lengthy captions per image. Our CompreCap has a more detailed caption (about 172) and more dense mask annotation than VG, where the bounding box coverage has about 99.23% *vs.* 85.26% of VG. For comprehensive image captioning, a dataset with wider coverage and a more detailed description is important. Moreover, based on our dataset, we also propose an evaluation metric for comprehensive image captioning.

They all lack annotations of object-bound attributes and directional relations. VG is the closest to our dataset, but its relatively brief annotations limit the diversity of attribute and relation descriptions, and its lower coverage cannot ensure the inclusion of all main objects in images, hindering the comprehensive evaluation of detailed captions. Moreover, VG does not propose a method to assess coherent long image captions. Instead, it makes the model output captions for different image regions sequentially and still uses traditional n-gram evaluation metrics.

The characteristics of *CompreCap* are highlighted as follows: **1. Comprehensive annotation:** *CompreCap* has object, attribute, and relation annotations with pixel-level masks; **2. More detailed caption:** *CompreCap* has more detailed human-annotated captions, with an average of 172 words *vs.* VG's 92; **3. Wider coverage:** The pixel coverage of *CompreCap* is 95.83%. Its bounding box coverage is 99.23% *vs.* 85.26% of VG; **4. Evaluation method for long caption of image:** Compared to other benchmarks, *CompreCap* proposes a method to comprehensively evaluate the visual context in detailed captions.

C. Data Comparison with MSCOCO

In Sec. 3.1 of the manuscript paper, we describe the process of sampling 560 images from the MSCOCO [28] Panoptic Segmentation dataset. MSCOCO Panoptic Segmentation dataset includes instance segmentation annotations for objects within images. Each segmentation map corresponds to an object in the image. However, we observed that segmentation maps for certain visible objects are missing. These missing objects are beyond the common objects listed in MSCOCO Panoptic Segmentation dataset. Additionally, we noticed that some of the segmentation maps for common objects overlapped with other visible objects in the image. Thus, we re-annotate the 560 images. Concretely, we first collect a common categories vocabulary from several well-known datasets including RAM [53], COCO [28], OpenImagesV4 [22], and Object365 [39]. Then, we re-annotate category labels and more precise segmentation maps for common objects in images within the vocabulary. The examples of re-annotated segmentation maps are demonstrated in Fig. S9. Moreover, we present a comparison on the data statistics of the *CompreCap*



(a) InstructBLIP-7B (b) MiniGPT4-v2 (c) ShareGPT4V-13B

(d) LLaVA-1.5-13B (e) miniGemini-34B-HD (f) LLava-Next-llama3-8B

(g) LLava-Next-34B (h) InternVL-Chat-V1-5 (i) GPT-4o

Figure S7. We present the error objects' segmentation map coverage in CompreQA-for-Caption of 9 LVLMs. The distribution shows that LVLMs tend to choose inaccurate captions for tiny objects whose segmentation map coverage $<2\%$.

Table S9. The data statistics of *CompreCap* and a subset of MSCOCO panoptic segmentation validation dataset ($\text{MSCOCO}_{\text{sub}}$). $\text{MSCOCO}_{\text{sub}}$ contains the same images as *CompreCap*.

Dataset	# Images	# Objects Categories	# Objects per Image	# Relations per Image	Averaged Text Length per Image	# Q/As
$\text{MSCOCO}_{\text{sub}}$ [28]	560	131	5.46	-	10	-
CompreCap	560	412	6.26	8.84	172	846

dataset with a subset of the MSCOCO Panoptic Segmentation dataset ($\text{MSCOCO}_{\text{sub}}$) in Tab. S9. Both datasets contain the same set of images, but *CompreCap* dataset includes additional annotations such as more types of objects, longer texts, and relations between objects.

D. Data Samples from *CompreCap* dataset

We present more annotation examples of comprehensive caption evaluation and fine-grained objects VQA from *CompreCap* in Fig. S10, Fig. S11, and Fig. S12. The structure of directed scene graph illustrated in Fig. S10 and Fig. S11 is composed of the human annotation at the levels of object, attribute, and relation. As shown in Fig. S12, the CompreQA-for-Presence dataset includes an equal number of objects that are present and

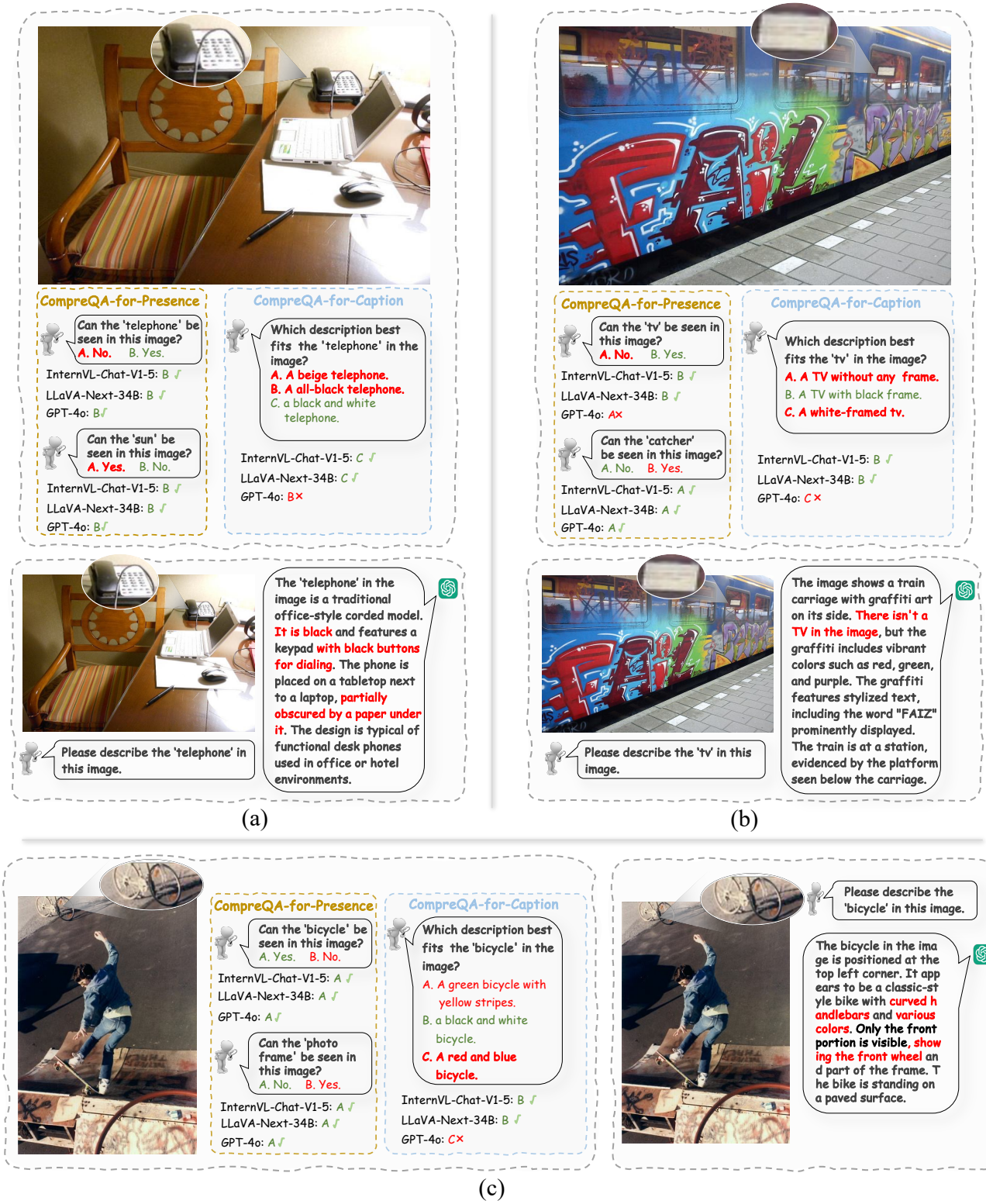


Figure S8. The answer is denoted in green, while error options are in red. The input images do not include the magnification effect, which is solely applied for clearer display purposes. We also ask GPT-4o [35] to describe the tiny objects and label incorrect descriptions with red. The examples show that GPT-4o [35] fails to comprehend the visual content of tiny objects.

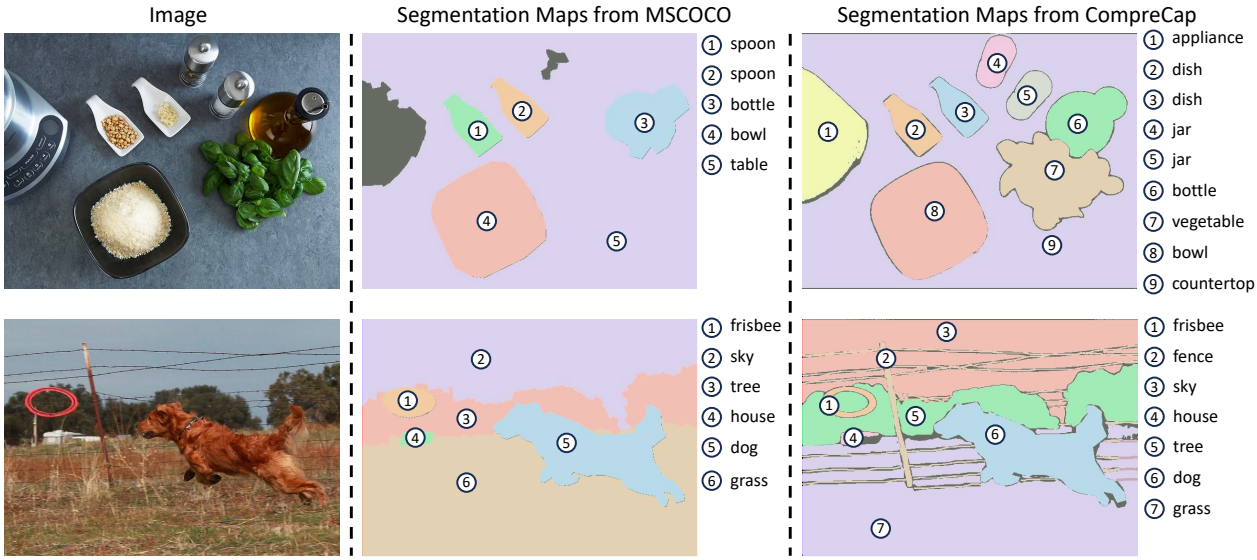


Figure S9. We refine the object and segmentation annotations based on MSCOCO panoptic segmentation dataset. Concretely, we add or re-annotate the class labels and more precise segmentation maps for the common objects within images, *e.g.*'jar' in the first raw, and 'fence' in the second raw. Besides, we improve the accuracy of pixel-level annotation, *e.g.*the segmentation map of 'frisbee' in the second raw.

Table S10. Comparison of *CompreCap* with long caption datasets.

Dataset	Averaged Text Length	Object	Segmentation Map	Attribute	Relation	Q/A	Answer Type
DOCCI [34]	136	-	-	-	-	-	-
IIW [16]	217	-	-	-	-	-	-
DCI [42]	148	✓	✓	✓	-	-	-
CompreCap	172	✓	✓	✓	✓	✓	A/B/C

absent, and CompreQA-for-Caption dataset contains one accurate description of an object and two inaccurate descriptions.

E. Discussion about More Long Caption Datasets

Considering there are several recent works on dense caption datasets such as DOCCI [34], ImageInWords(IIW) [16] and DCI [42], we show the comparison between these datasets and our *CompreCap* in Tab. S10. The IIW [16] and DOCCI [34] datasets do not provide special annotations for objects, attributes, relations, and segmentation maps. Each image within the two datasets is only annotated with one long text. Moreover, DCI [42] dataset provides semantic labels, segmentation maps and attribute descriptions for each object, but it doesn't consider the relationships between objects. Thus, the datasets in these works can not be utilized to comprehensively evaluate detailed captions on multiple levels as ours.

F. Information of Human Participants

In the construction and experimentation process of *CompreCap*, three stages required the involvement of human experts: data annotation, manual captioning of images for evaluating human performance, and manual scoring of captions generated by models and humans. We engaged a total of 50 participants from universities and a crowdsourcing platform, dividing them into groups of 20, 10, and 20 participants to respectively take part in these three stages, ensuring no overlap of personnel across the stages. These participants are aged 22 to 45, with backgrounds in linguistics, computer science, and mathematics.

During the manual captioning of images, the 10 human experts each described 56 different images as comprehensively as possible. They were not specifically instructed to focus on objects, attributes, or relations in their descriptions. During the manual evaluation scoring stage, the 20 experts independently scored all 6,160 captions generated by both models and humans according to their personal assessment criteria. The average score was used as the final score for both the models and human performance.

G. Limitations and Broader Impact

Broader Impact. We propose a human-annotated *CompreCap* benchmark, which is composed in the format of directed scene graph, for evaluating comprehensive captions generated by LVLMs. Using the *CompreCap* benchmark, we identify which LVLM is better at accurately describing text-rich visual content. Additionally, we design a vision question answering (VQA) task

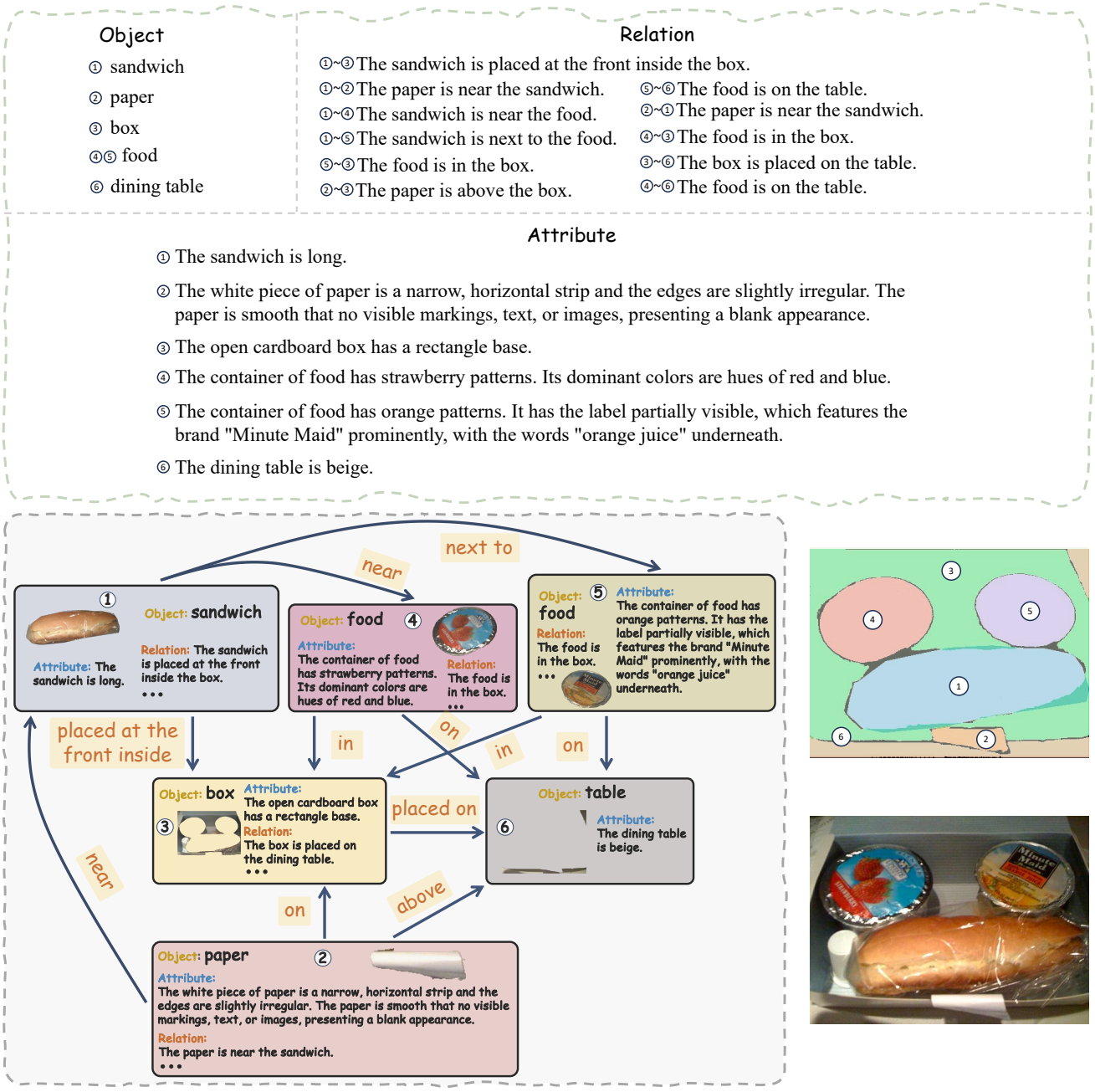


Figure S10. Data samples of *CompreCap* benchmark for evaluating comprehensive captions.

based on tiny objects to assess the fine-grained object perception ability of LVLMs.

Limitations. We have discussed that the qualities of detailed captions, including scores at the levels of object, attributes and relations, are not correlated to caption length generated by LVLMs. However, we do not quantitatively evaluate hallucinations (*i.e.*, incorrect descriptions and non-existent visual information) generated by LVLMs. We plan to assess hallucination components with error rates in future work.

Object

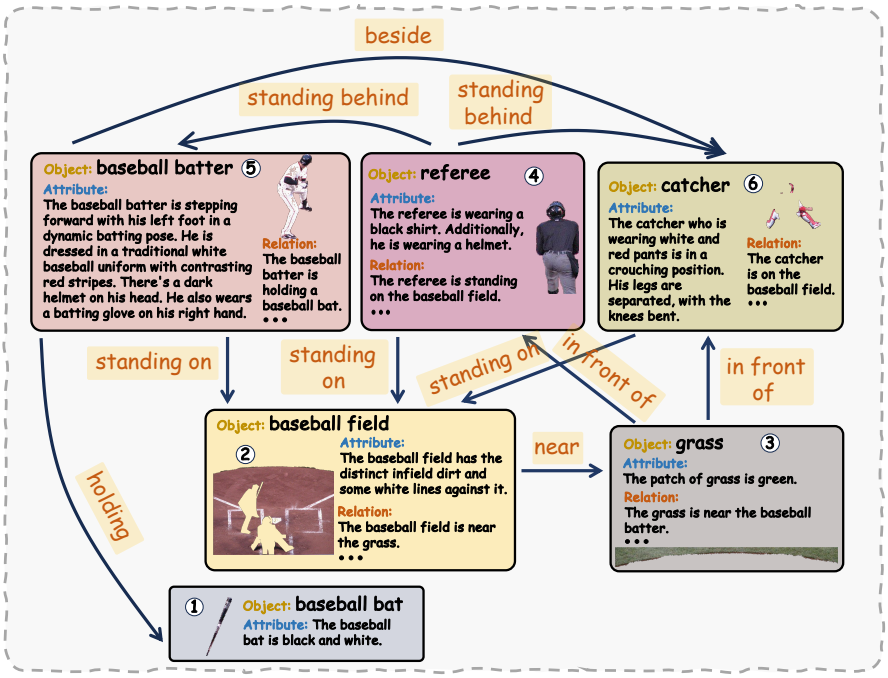
① baseball bat
② baseball field
③ grass
④ referee
⑤ baseball batter
⑥ catcher

Relation

②~③ The baseball field is near the grass.
③~⑥ The grass is in front of the catcher.
③~⑤ The grass is near the baseball batter.
④~⑤ The referee is standing behind a baseball batter.
②~④ The referee is standing on the baseball field.
④~⑥ The referee is standing behind the catcher.
①~⑤ The baseball batter is holding a baseball bat.
⑤~⑥ The baseball batter is beside the catcher.
②~⑤ The baseball batter is standing on the baseball field.
②~⑥ The catcher is on the baseball field.

Attribute

① The baseball bat is black and white.
② The baseball field has the distinct infield dirt and some white lines against it.
③ The patch of grass is green.
④ The referee is wearing a black shirt. Additionally, he is wearing a helmet.
⑤ The baseball batter is stepping forward with his left foot in a dynamic batting pose. He is dressed in a traditional white baseball uniform with contrasting red stripes. There's a dark helmet on his head. He also wears a batting glove on his right hand.
⑥ The catcher who is wearing white and red pants is in a crouching position. His legs are separated, with the knees bent.



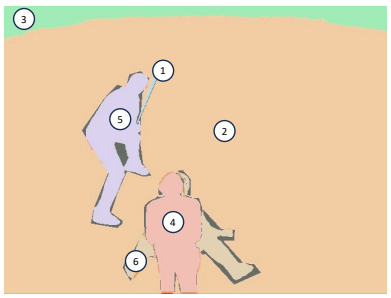


Figure S11. Data samples of *CompreCap* benchmark for evaluating comprehensive captions.

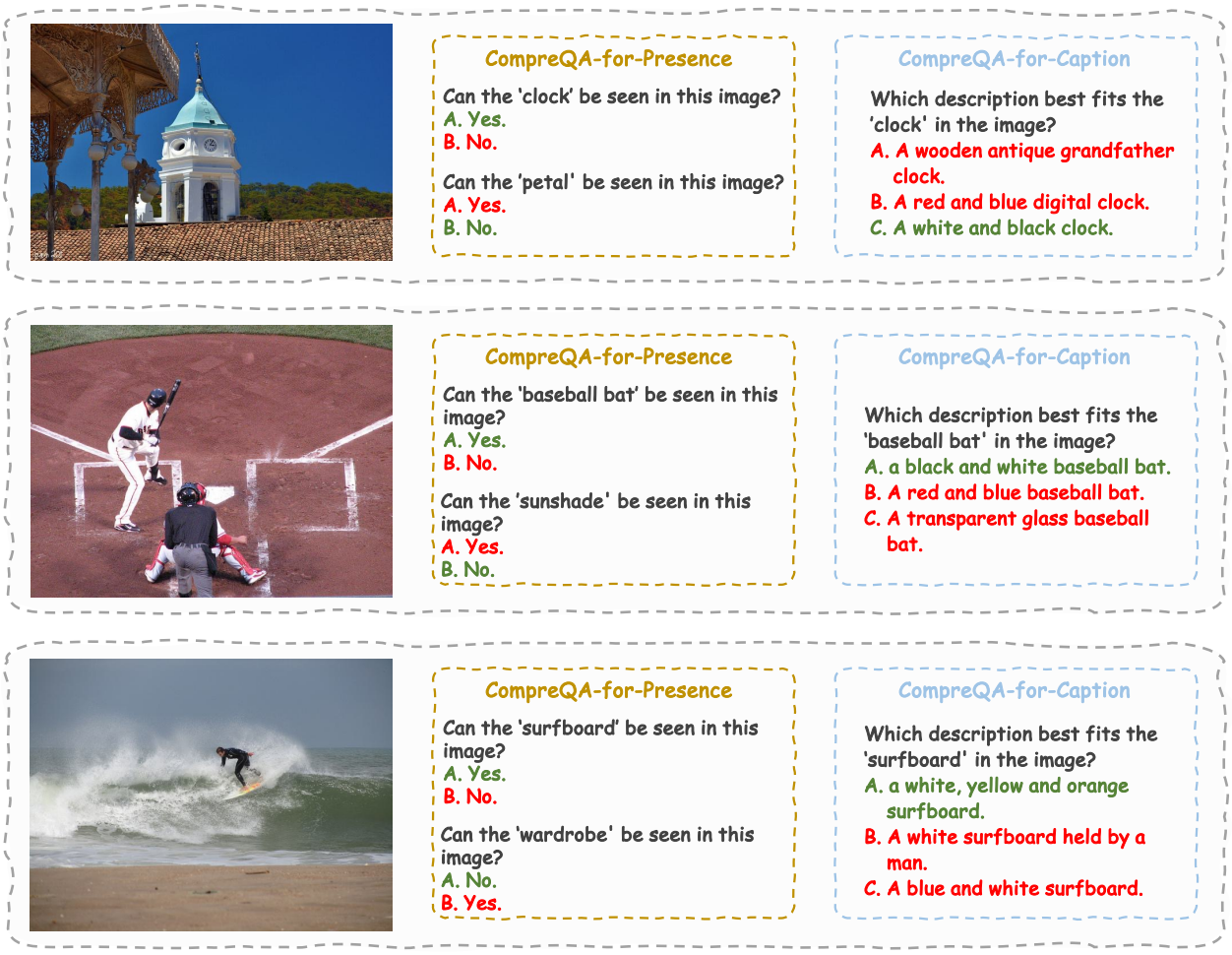


Figure S12. Data samples of *CompreCap* benchmark for fine-grained object VQA. The answer is denoted in green, while the error options are in red. In the CompreQA-for-Presence task, an equal number of objects that are present and absent are included.