CAPability: A Comprehensive Visual Caption Benchmark for Evaluating Both Correctness and Thoroughness



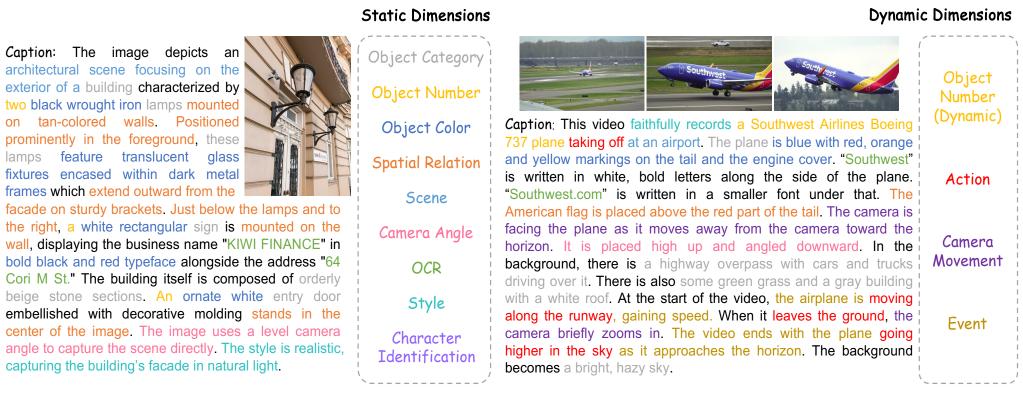






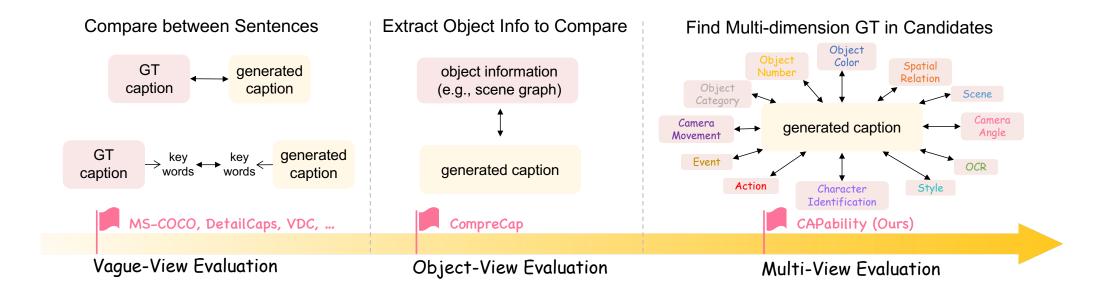
Zhihang Liu¹, Chen-Wei Xie², Bin Wen², Feiwu Yu², Jixuan Chen², Pandeng Li^{1,2}, Boqiang Zhang¹, Nianzu Yang³, Yinglu Li¹, Zuan Gao¹, Yun Zheng², Hongtao Xie¹ ¹University of Science and Technology of China, ²Tongyi Lab, Alibaba Group, ³Shanghai Jiao Tong University

Introduction



An example of image caption (left) and video caption (right) task. By analyzing the components of captions, we conclude 12 dimensions (9 static dimensions and 4 dynamic dimensions with object number shares on both static and dynamic), which all contribute to a detailed and comprehensive caption. The static dimensions are shared in both images and videos. Video data has additional dynamic dimensions that need to be judged with temporal relations.

- > We introduce CAPability, a comprehensive visual caption benchmark featuring 6 views and 12 dimensions, based on a new human-annotated dataset of nearly 11K images and videos.
- > We propose a novel evaluation framework that assesses both correctness and thoroughness of captions by using precision and hit metrics.
- \triangleright We assess an additional capability via the $K\overline{T}$ metric, which indicates the performance gap between QA and the captioning task.



Existing visual captioning benchmarks are inadequate for modern MLLMs, as their reliance on brief ground truths, unreliable metrics, and incomplete visual coverage fails to assess caption correctness and thoroughness.

> MS-COCO[5] MSRVTT [7]

Dream-1K [21]

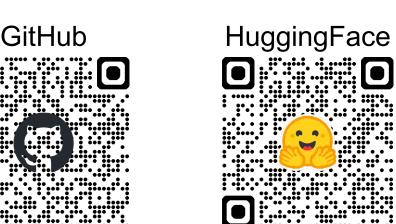
DetailCaps [22]

CompreCap [20]

Views	Dimensions				
Object Deleted	Object Category, Object Color,				
Object-Related	(Dynamic) Object Number, Spatial Relation				
Global-Related	Scene, Style				
Text-Related	OCR				
Camera-Related	Camera Angle, Camera Movement				
Temporal-Related	Action, Event				
Knowledge-Related	Character Identification				

Our designed views and more detailed dimensions. We can treat a caption from the listed six views, and then split them

oject Page	ArXiv	GitHub	HuggingFace						
caption from the listed everal dimensions.	six views, and then split them	benchmarks in different aspects. We are the most comprehensive with image and video data, multi-view annotations, and new thoroughness evaluation methods proposed.							



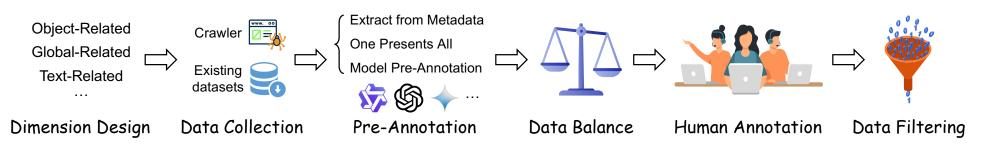
Sentences

Object Info

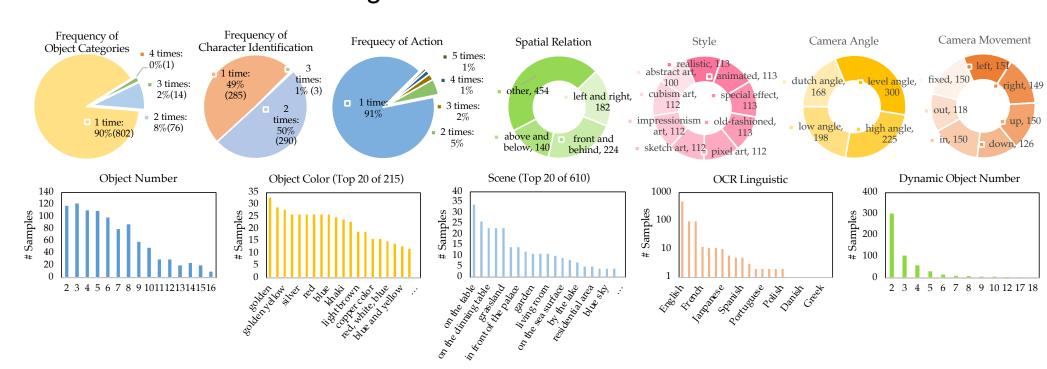
Comparison of our CAPability and other visual caption

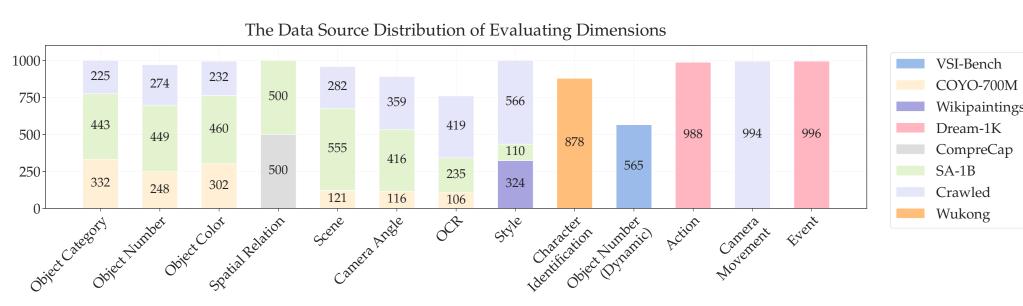
CAPability

Data Construction Pipeline



Distribution of Each Designed Dimension



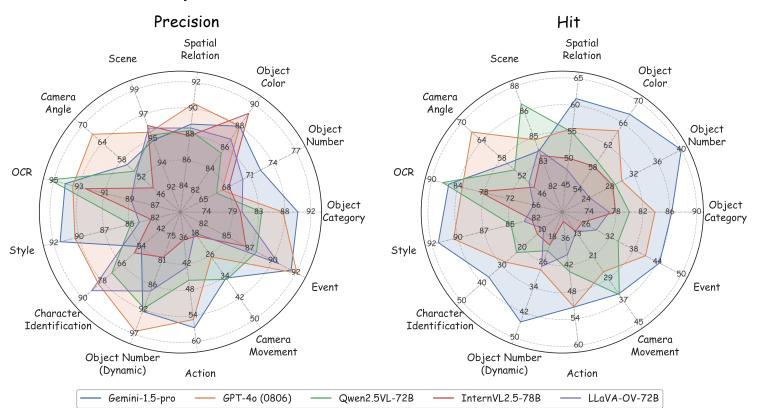


- Multiple Dimension Evaluation
 - > MIS: caption does not mention the corresponding content about the dimension.
 - > COR: caption mentions the corresponding content about the dimension, and describes it correctly compared with the annotations.
 - > INC: caption mentions the corresponding content about the dimension, but gives a wrong description compared with the annotations.

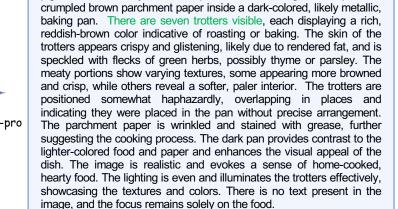


Experiments

Precision and Hit comparison of SOTA MLLMs



Example

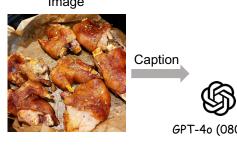


The image presents a high-angle, close-up view of several cooked pig's feet, technically known as pork trotters, arranged on a piece of

Object number

Annotation:

pig trotter: 7

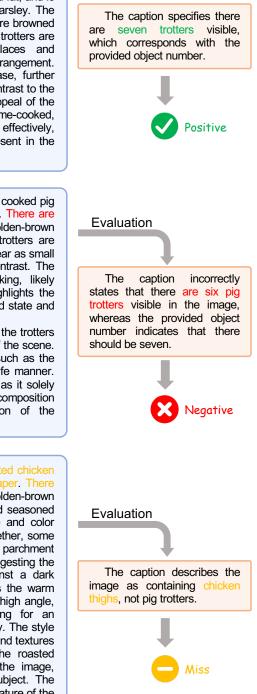


The image depicts a close-up view of several pieces of cooked pig trotters, placed on a parchment-lined baking sheet or tray. There are six pig trotters visible in the image, each with a distinct golden-brown color indicating they have been roasted or baked. The trotters are seasoned with a sprinkling of herbs and spices, which appear as small green flakes on the surface, adding texture and color contrast. The setting suggests a preparation or final stage of cooking, likely showcasing the dish before it is served. The lighting highlights the glossy texture of the roasted skin, emphasizing the cooked state and the intricate textures of the meat and bones.

The camera angle is a level angle, directly focusing on the trotters without any significant tilt, providing an even perspective of the scene. The style of the image is realistic, capturing the details such as the skin texture, cooking marks, and seasoning in a true-to-life manner. There is no visible text or known character in the image, as it solely focuses on the culinary subject matter. The overall composition suggests a straightforward and appetizing presentation of the



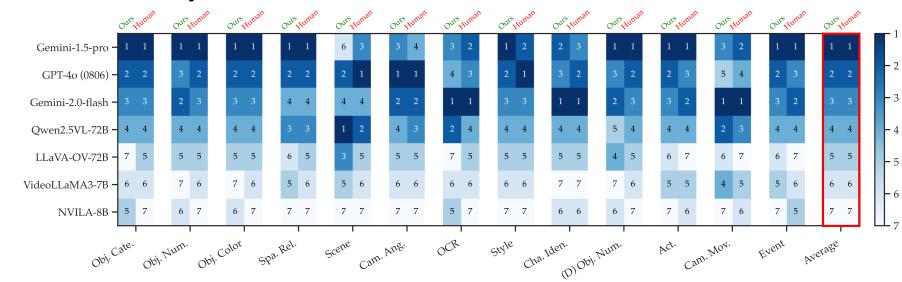
The image showcases a close-up view of se hue indicative of thorough roasting. The skin is crispy and seasoned with visible specks of herbs and spices, adding texture and color contrast to the dish. The thighs are positioned closely together, some overlapping slightly, creating a sense of abundance. The parchment paper beneath them is stained with oil and seasoning, suggesting the juices released during cooking. The scene is set against a dark background, likely a kitchen countertop, which enhances the warm tones of the chicken. The camera angle appears to be a high angle, providing a clear overhead view of the dish, allowing for an unobstructed look at the details of the chicken and the tray. The style of the image is realistic, with a focus on the natural colors and textures of the food, capturing the appetizing appearance of the roasted chicken. There are no texts or characters present in the image, keeping the viewer's attention solely on the culinary subject. The overall composition emphasizes the inviting and delicious nature of the meal, making it visually appealing and mouth-watering.



\triangleright QA-based evaluation and the $K\overline{T}$ metric

	Methods	Obj. Cate.	Obj. Num.	Obj. Color	Spa. Rel.	Scene	Cam. Ang.	OCR	Style	Cha. Iden.	(D) Obj. Num.	Act.	Cam. Mov.	Event	Avg.
	LLaVA-OV-72B	95.0	54.6	63.8	94.0	96.2	60.6	66.3	82.0	32.1	52.2	75.5	15.7	85.3	67.2
	Qwen2VL-72B	94.7	56.1	68.6	90.7	94.0	65.0	82.4	86.6	31.3	48.9	73.0	34.1	72.7	69.1
cc	InternVL2.5-78B	95.5	56.9	67.1	90.0	91.2	54.1	79.5	82.5	19.1	49.7	79.1	23.3	81.7	66.9
¥ A	Qwen2.5VL-72B	92.7	58.2	67.4	84.4	88.7	63.9	87.4	87.3	33.4	41.4	75.8	39.5	85.8	69.7
Q	GPT-4o (0806)	94.5	47.2	72.5	79.5	84.5	71.6	80.5	79.3	37.2	46.2	81.1	20.5	78.6	67.2
	Gemini-1.5-pro	97.3	51.6	78.8	94.4	87.1	56.8	84.8	84.2	41.2	51.2	74.4	32.2	82.8	70.5
	Gemini-2.0-flash	98.3	46.8	73.3	93.4	95.2	57.6	84.8	74.5	49.1	44.2	81.6	24.8	86.6	70.0
	LLaVA-OV-72B	20.3	64.3	39.6	49.6	13.9	33.0	13.7	9.4	74.8	66.1	53.2	31.4	67.1	41.3
	Qwen2VL-72B	16.7	62.1	37.0	46.0	12.6	35.4	10.2	10.4	84.7	78.3	47.6	53.1	60.4	42.6
ıt .	InternVL2.5-78B	19.1	57.8	35.4	45.2	11.4	21.4	11.0	47.0	8.2	73.0	62.4	68.1	69.9	40.8
$Kar{T}$	Qwen2.5VL-72B	15.3	60.7	33.5	37.2	9.1	24.3	5.9	8.5	47.4	66.2	49.3	38.2	61.1	35.1
	GPT-4o (0806)	13.1	55.2	26.6	34.7	7.8	16.1	6.7	3.5	30.9	64.8	41.7	53.9	51.7	31.3
	Gemini-1.5-pro	11.9	41.0	24.1	36.4	9.6	19.2	5.5	3.1	18.0	51.6	36.1	23.4	47.9	25.2
	Gemini-2.0-flash	16.3	52.6	32.8	45.1	12.6	9.0	4.5	3.2	25.7	58.4	45.9	23.1	54.5	29.5

Consistency with human evaluation



Comparison with other metrics

