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

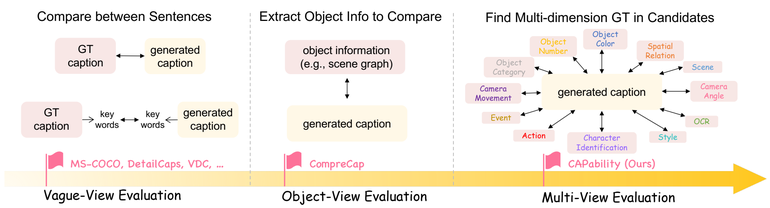
Zhihang Liu[1], Chen-Wei Xie[2], Bin Wen[2], Feiwu Yu[2], Jixuan Chen[2], Pandeng Li[1,2], Boqiang Zhang[1], Nianzu Yang[3], Yinglu Li[1], Zuan Gao[1], Yun Zheng[2], Hongtao Xie[1]

[1]University of Science and Technology of China, [2]Tongyi Lab, Alibaba Group, [3]Shanghai 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.
- We assess an additional capability via the $K\bar{T}$ 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.

Views	Dimensions
Object-Related	Object Category, Object Color, (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 into several dimensions.

Benchmark	Data Type		Anno-tations	Evaluation	
	Image	Video		Thoroughness	$K\bar{T}$
MS-COCO[5]	✓	-	Sentences	-	-
MSRVTT [7]	-	✓	Sentences	-	-
Dream-1K [21]	-	✓	Sentences	single dim	-
VDC [23]	-	✓	Sentences	-	-
DetailCaps [22]	✓	-	Sentences	-	-
CompreCap [20]	✓	-	Object Info	single dim	-
CAPability (Ours)	✓	✓	Multi-view Elements	✓	✓

Comparison of our CAPability and other visual caption benchmarks in different aspects. We are the most comprehensive with image and video data, multi-view annotations, and new thoroughness evaluation methods proposed.

Project Page

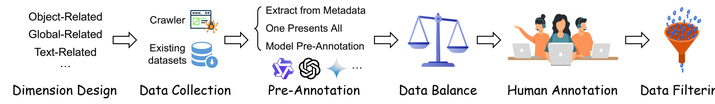
ArXiv

GitHub

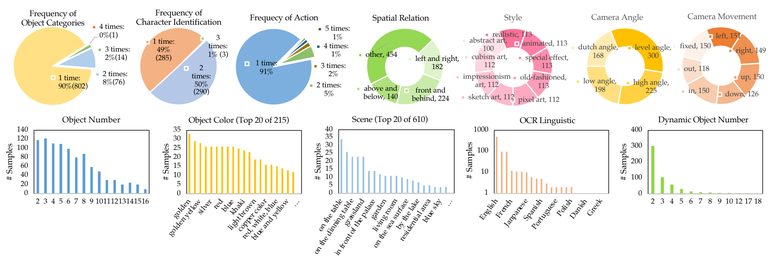
HuggingFace

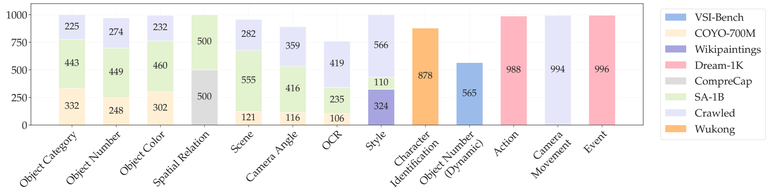
CAPability

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.

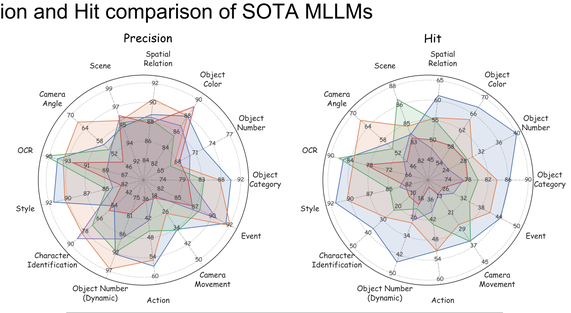
$$Precision = \frac{|S(COR)|}{|S(COR) \cup S(INC)|}$$

$$Hit = \frac{|S(COR)|}{|S(ALL)|}$$

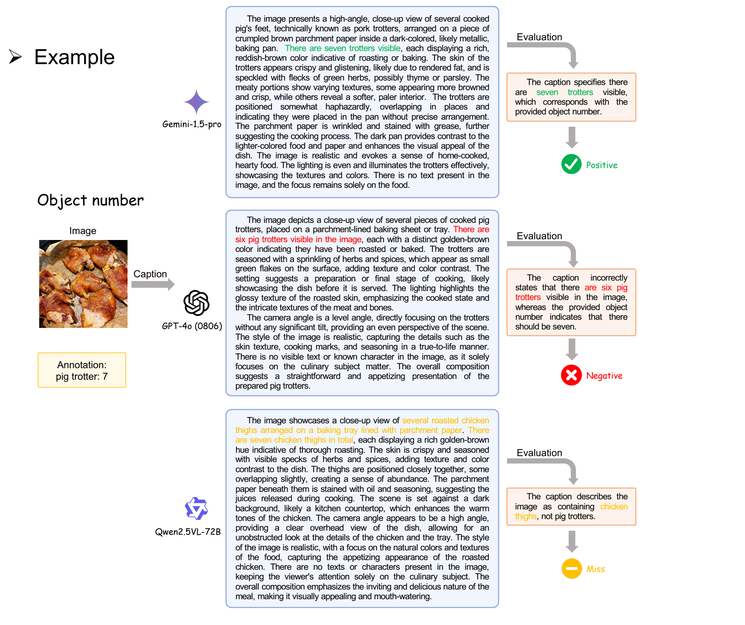
$$K\bar{T} = \frac{|S_{qa}(COR) \cap [S(INC) \cup S(MIS)]|}{|S_{qa}(COR)|}$$

Experiments

Precision and Hit comparison of SOTA MLLMs



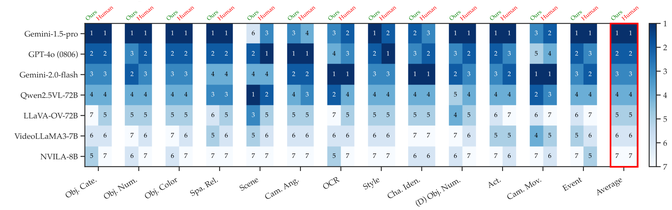
Example



QA-based evaluation and the $K\bar{T}$ 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.
QA Acc	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
	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
	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
	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
$K\bar{T}$	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
	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
	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