Panting With Words: Elevating Detailed Image Captioning with Benchmark and Alignment Learning

Qinghao Ye* , Xianhan Zeng* , Fu Li, Chunyuan Li, Haoqi Fan

Background

- Detailed image captioning has long been a pivotal task in visual understanding.
- However, the evaluation of detailed image captioning remains underexplored due to outdated evaluation metrics and coarse annotations.
- Our goal is to **evaluate** and **elevate** the captioning capability of modern VLMs accurately and comprehensively.

DCScore & DeCapBench

- DCScore: a fine-grained metric evaluating detailed captions by generating and assessing **primitive information units**.

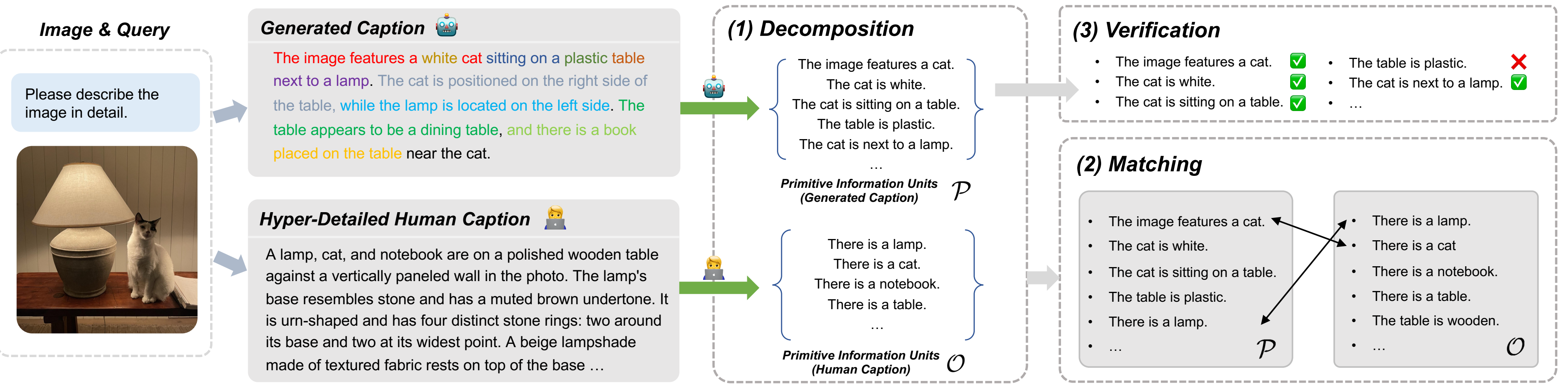


Image & Query

Please describe the image in detail.

Generated Caption

The image features a white cat sitting on a plastic table next to a lamp. The cat is positioned on the right side of the table, while the lamp is located on the left side. The table appears to be a dining table, and there is a book placed on the table near the cat.

Hyper-Detailed Human Caption

A lamp, cat, and notebook are on a polished wooden table against a vertically paneled wall in the photo. The lamp's base resembles stone and has a muted brown undertone. It is urn-shaped and has four distinct stone rings: two around its base and two at its widest point. A beige lampshade made of textured fabric rests on top of the base …

(1) Decomposition

The image features a cat.
The cat is white.
The cat is sitting on a table.
The table is plastic.
The cat is next to a lamp.
…

Primitive Information Units (Generated Caption) $\mathcal{P}$

There is a lamp.
There is a cat.
There is a notebook.
There is a table.
…

Primitive Information Units (Human Caption) $\mathcal{O}$

(2) Matching

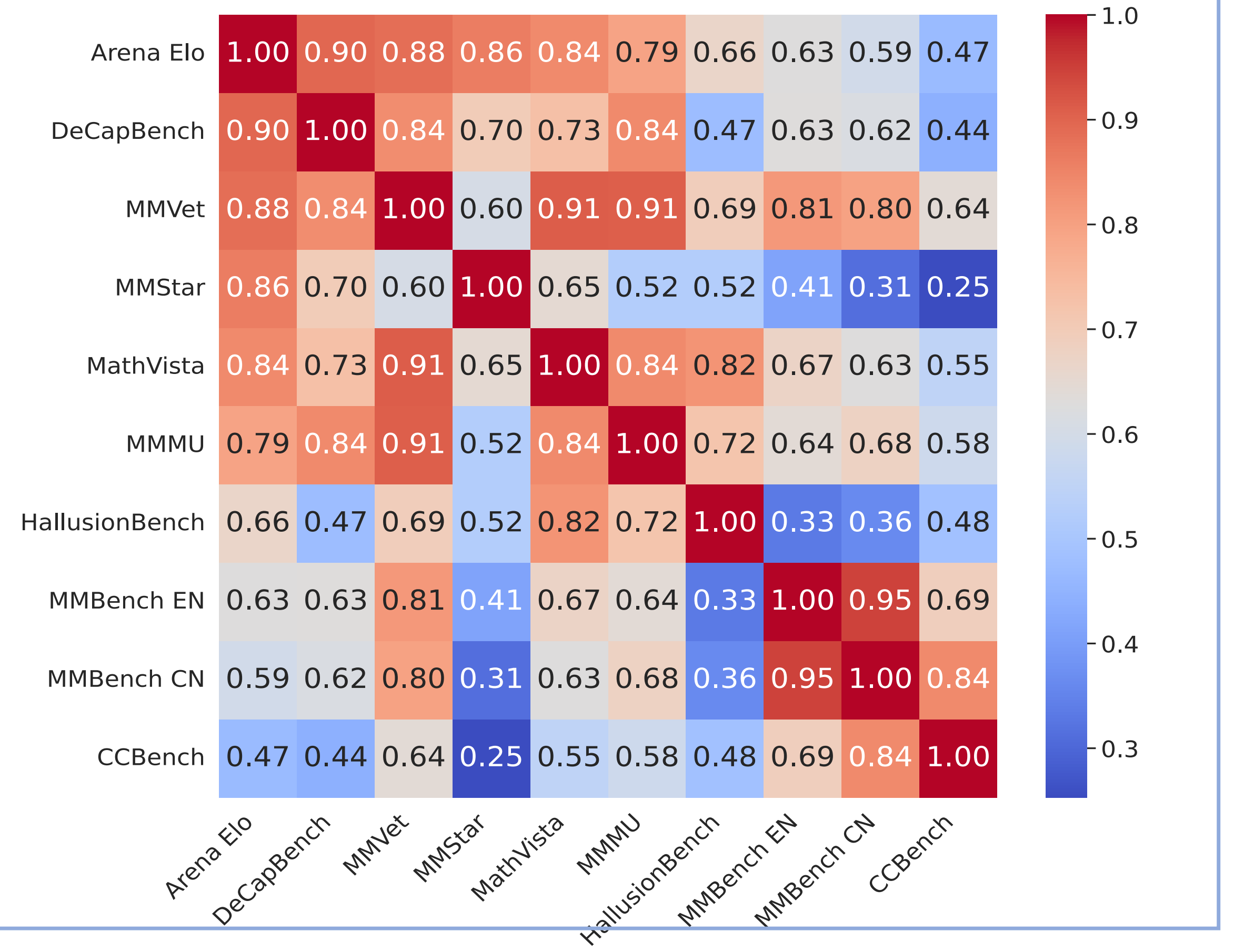
$\mathcal{P}$:
- The image features a cat.
- The cat is white.
- The cat is sitting on a table.
- The table is plastic.
- There is a lamp.
- …

$\mathcal{O}$:
- There is a lamp.
- There is a cat
- There is a notebook.
- There is a table.
- The table is wooden.
- …

(3) Verification

- The image features a cat. ✅
- The cat is white. ✅
- The cat is sitting on a table. ✅
- The table is plastic. ❌
- The cat is next to a lamp. ✅
- …

- Dataset: 400 images in ImageInWords with high-quality, human-curated detailed captions
- DeCapBench achieves the **highest** correlation with Arena Elo, with a Spearman's correlation of 0.90 among different VLM benchmarks



	Arena Elo	DeCapBench	MMVet	MMStar	MathVista	MMMU	HallusionBench	MMBench EN	MMBench CN	CCBench
Arena Elo	1.00	0.90	0.88	0.86	0.84	0.79	0.66	0.63	0.59	0.47
DeCapBench	0.90	1.00	0.84	0.70	0.73	0.84	0.47	0.63	0.62	0.44
MMVet	0.88	0.84	1.00	0.60	0.91	0.91	0.69	0.81	0.80	0.64
MMStar	0.86	0.70	0.60	1.00	0.65	0.52	0.52	0.41	0.31	0.25
MathVista	0.84	0.73	0.91	0.65	1.00	0.84	0.82	0.67	0.63	0.55
MMMU	0.79	0.84	0.91	0.52	0.84	1.00	0.72	0.64	0.68	0.58
HallusionBench	0.66	0.47	0.69	0.52	0.82	0.72	1.00	0.33	0.36	0.48
MMBench EN	0.63	0.63	0.81	0.41	0.67	0.64	0.33	1.00	0.95	0.69
MMBench CN	0.59	0.62	0.80	0.31	0.63	0.68	0.36	0.95	1.00	0.84
CCBench	0.47	0.44	0.64	0.25	0.55	0.58	0.48	0.69	0.84	1.00

Alignment Learning

- FeedQuill: a fine-grained caption preference collection pipeline
- RL: Simultaneously optimize signals from both two RMs
  - Precision RM: distill the precision score pipeline
  - Recall RM: distill the recall score pipeline

Model	AI2D	ChartQA	MMBench	SEEDBench	MME	MMMU	MMVet	MMStar	SciQA	LLaVA-W	WildVision	DECAPBENCH
Proprietary Model												
Claude-3.5-Sonnet	94.7	90.8	78.5	-	-/-	68.3	75.4	60.2	80.5	102.9	50.00	52.37
Gemini-1.5-Pro	94.4	87.2	73.9	-	-/-	62.2	64.0	58.7	-	-	35.45	46.34
GPT-4V	78.2	78.5*	79.8	49.9	1409/517	56.8	57.1	75.7	75.7	98.0	80.01	48.52
GPT-4o	94.2	85.7	80.5	76.2	-/-	69.1	76.2	59.8	83.5	106.1	89.41	53.44
Open-Source Model												
Cambrian-34B	79.7	73.8	81.4	-	-/-	49.7	53.2	85.6	67.8	-	-	35.12
VILA-40B	-	-	82.4	75.8	1762	51.9	51.2	54.2	-	-	-	38.02
XComposer-2.5-7B	81.5	82.2	82.2	75.4	2229	42.9	51.7	59.9	-	78.1	-	29.60
InternVL-2-8B	83.8	83.3	81.7	76.0	2210	49.3	60.0	59.4	97.0	84.5	-	45.55
InternVL-2-26B	84.5	84.9	83.4	76.8	2260	48.3	65.4	60.4	97.5	99.6	-	49.59
LLaVA-Onevision-7B	81.4	80.0	80.8	75.4	1580/418	48.8	57.5	61.7	96.0	90.7	54.50	43.49
FEEDQUILL-7B	81.3	80.3	80.5	75.8	1515/450	47.9	59.3	62.4	95.9	**100.5**	59.60	**55.65**

- Main Result

Our 7B model achieves state-of-the-art performance in detailed image captioning, surpassing GPT-4o

Explain the visual content of the image in great detail.

LLaVA-OneVision-7B

In the dimly lit room, a group of people are engrossed in their phones. The person **closest** to us, a man wearing a black shirt, is holding a phone with **a vibrant red screen**, which is prominently displayed in the foreground. The phone, a sleek device with a small screen, is held securely in their hand.

A little further away, another person is seen holding a phone with a **black** screen, which is positioned in the upper right corner of the image. This phone, slightly larger than the one in the foreground, is also held in a hand.

In the **background**, there's a person wearing a black and white **shirt**, their hand holding a phone with a white screen, which is positioned in the **upper left corner** of the image. The phone is slightly larger than the one in the foreground, and it's held in a hand.

The room they're in is dimly lit, creating an atmosphere of focus and concentration. The people are all dressed in dark clothing, adding to the overall mood of the scene. The image captures a moment of modern life, where technology and social interaction intersect.

FEEDQUILL-7B (This work)

The image is a photograph **taken from an overhead perspective**, focusing on the hands and upper bodies of two individuals. The main subjects are the hands of the individuals, which are holding **two different types of mobile devices**. The person **on the left is holding a flip phone, which is open, revealing its screen and keypad**. The person **on the right is holding a Blackberry device**, which is also open, displaying its screen and keyboard.

The person on the right is also holding a glass containing a red liquid, which appears to be wine. The glass is transparent and has a stem, suggesting it is a wine glass. The individual is wearing a black and white patterned **garment**, which could be a scarf or a piece of clothing with a distinctive design.

The background of the image is dark and out of focus, drawing attention to the hands and the objects they are holding. The lighting in the image is dim, with the main source of light coming from the screens of the mobile devices, which illuminate the hands and the glass of wine.

There are no visible texts or logos on the mobile devices, and the image does not provide any additional context about the location or the event taking place. The style of the photograph is candid and informal, capturing a moment in time without any staging or posing. The composition of the image emphasizes the interaction between the individuals and their devices, as well as the contrast between the modern technology and the traditional act of holding a glass of wine.

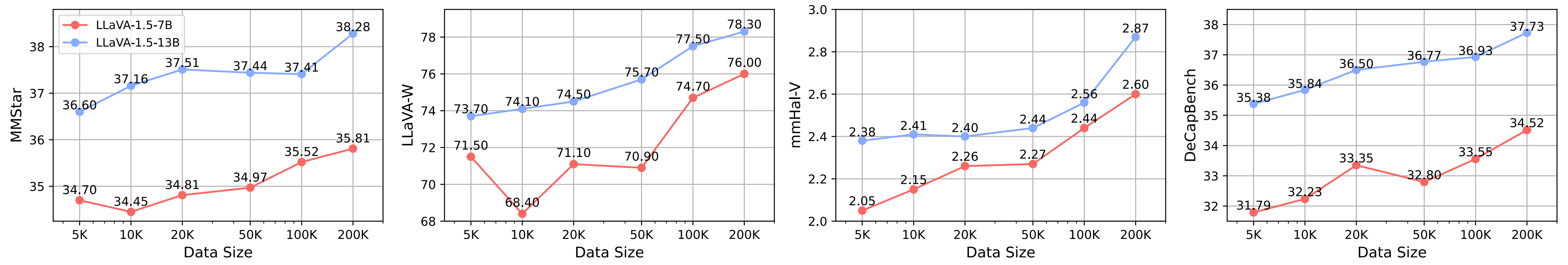
🟥 Wrong Elements (Hallucinations) 🟩 Correct Elements 🟦 Additional Correct Elements

Ablation Study

1. **Preference Data**: FeedQuill outperforms other held-out preference dataset after RL on the same SFT model

Method	MMBench ↑	VizWiz ↑	MMStar ↑	WildVision ↑	LLaVA-W ↑	DECAPBENCH ↑	mmHal-V ↑	CHAIR$_S$ ↓	CHAIR$_I$ ↓
LLaVA-1.5	64.8	50.0	33.1	14.48	65.3	24.50	1.85	47.8	25.3
w/ HA-DPO	64.3	54.1	33.5	15.17	65.1	22.45	2.12	49.3	25.5
w/ POVID	64.7	47.9	35.4	13.25	71.5	23.54	1.90	31.8	5.4
w/ CSR	64.2	52.8	33.8	13.85	70.3	23.70	2.12	15.7	7.9
w/ RLAIF-V	62.7	50.9	34.7	15.65	76.0	28.21	2.59	8.5	4.3
w/ FEEDQUILL	**66.3**	**55.2**	**35.8**	**19.68**	**76.0**	**34.52**	**2.60**	**5.1**	**2.6**

2. **Data Size**: As the size of preference data grows, the model's performance consistently improves



3. **Source of responses**

Source of Response Same Model	 Other Models	MMStar	LLaVA-W	mmHal-V	DECAPBENCH
		33.1	65.3	1.85	24.50
✓		37.6	75.1	2.74	26.32
	✓	38.0	71.5	2.53	34.84
✓	✓	**38.3**	**78.3**	**2.83**	**37.73**

Improvements arise from the model's ability across varying sources.

4. **Source of rewards**

Method	LLaVA-1.5-7B LLaVA-W	 DECAPBENCH	LLaVA-1.5-13B LLaVA-W	 DECAPBENCH
Base	65.3	24.50	72.8	25.55
Only c_p	67.3	25.21	74.3	26.23
Only c_r	46.2	10.03	56.9	15.11
$c_p + c_r$	**76.0**	**34.52**	**78.3**	**37.73**

Incorporating both precision reward and recall reward significantly improve model performance

5. **Compatibility Analysis**: FeedQuill is effective regardless of sft-model, consistently enhancing performance on downstream tasks

	Comprehensive Benchmark				Visual Hallucination	Visual Chat and Captioning		
Method	MMBench	MMStar	VizWiz	SciQA^I	mmHal-V	LLaVA-W	WildVision	DECAPBENCH
LLaVA-1.5-7B	64.8	33.1	50.0	66.8	1.85	65.3	14.48	24.50
+ FEEDQUILL	66.3 (+1.7)	35.8 (+2.7)	55.2 (+5.2)	68.9 (+2.1)	2.60 (+0.75)	76.0 (+10.7)	17.68 (+3.20)	34.52 (+10.02)
LLaVA-1.5-13B	68.7	34.3	53.6	71.6	2.33	72.8	16.17	25.55
+ FEEDQUILL	69.2 (+0.5)	38.3 (+4.0)	56.8 (+3.2)	73.4 (+1.8)	2.83 (+5.00)	78.3 (+5.5)	18.15 (+1.98)	37.73 (+12.18)
LLaVA-1.6-7B	67.1	37.6	57.6	70.2	2.58	79.8	26.15	35.74
+ FEEDQUILL	67.9 (+0.8)	38.6 (+1.0)	63.4 (+5.8)	70.3 (+0.1)	2.93 (+0.35)	82.4 (+2.6)	44.16 (+18.01)	52.69 (+16.95)
LLaVA-1.6-13B	69.3	40.4	60.5	73.6	2.95	85.2	33.69	36.28
+ FEEDQUILL	69.9 (+0.6)	41.1 (+0.7)	66.7 (+6.2)	73.5 (+0.1)	3.76 (+0.81)	87.1 (+1.9)	49.69 (+16.00)	53.26 (+16.98)
LLaVA-Onevision-7B	80.8	61.7	60.0	96.0	2.94	90.7	54.50	43.49
+ FEEDQUILL	80.5 (-0.3)	62.4 (+0.7)	60.4 (+0.4)	95.9(-0.1)	3.10 (+0.16)	100.5 (+9.8)	59.60 (+5.10)	55.65 (+12.16)