

CapArena: Benchmarking and Analyzing Detailed Image Captioning in the LLM Era

Kanzhi Cheng^{◇ †} Wenpo Song^{◇ †} Jiaxin Fan^{◇ †} Zheng Ma[◇] Qiushi Sun[♣]
Fangzhi Xu[♡] Chenyang Yan[◇] Nuo Chen[◇] Jianbing Zhang^{◇*} Jiajun Chen[◇]

[◇]National Key Laboratory for Novel Software Technology, Nanjing University

[♣]The University of Hong Kong [♡]Shanghai Artificial Intelligence Laboratory

chengkz@smail.nju.edu.cn zjb@nju.edu.cn

Abstract

Image captioning has been a longstanding challenge in vision-language research. With the rise of LLMs, modern Vision-Language Models (VLMs) generate detailed and comprehensive image descriptions. However, benchmarking the quality of such captions remains unresolved. This paper addresses two key questions: (1) How well do current VLMs actually perform on image captioning, particularly compared to humans? We built *CapArena*, a platform with over 6000 pairwise caption battles and high-quality human preference votes. Our arena-style evaluation marks a milestone, showing that leading models like GPT-4o achieve or even surpass human performance, while most open-source models lag behind. (2) Can automated metrics reliably assess detailed caption quality? Using human annotations from *CapArena*, we evaluate traditional and recent captioning metrics, as well as VLM-as-a-Judge. Our analysis reveals that while some metrics (e.g., METEOR) show decent caption-level agreement with humans, their systematic biases lead to inconsistencies in model ranking. In contrast, VLM-as-a-Judge demonstrates robust discernment at both the caption and model levels. Building on these insights, we release *CapArena-Auto*, an accurate and efficient automated benchmark for detailed captioning, achieving 94.3% correlation with human rankings at just \$4 per test. Data and resources are open-sourced at [CapArena](#).

1 Introduction

Image captioning, the task of generating textual descriptions for images, has long been a fundamental challenge in both the computer vision and natural language processing communities (Vinyals et al., 2015; Anderson et al., 2018; Cornia et al., 2020). It has broad and valuable applications, such as assisting visually impaired individuals and supporting

*Correspondence to: Jianbing Zhang

[†]Equal contribution.

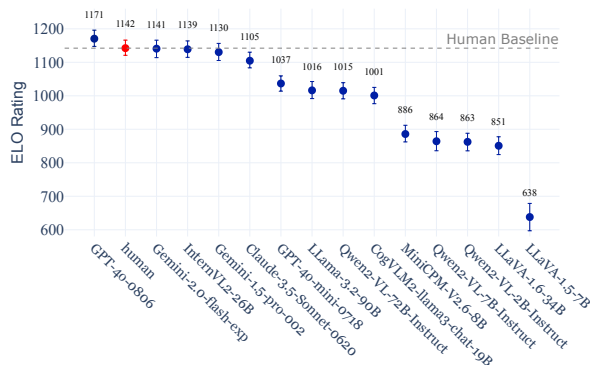


Figure 1: Model rankings from *CapArena* in detailed captioning. Top models are comparable to humans, while most open-source models lag behind.

multimedia retrieval. Driven by progress in LLMs, modern Vision-Language Models (VLMs) are capable of generating long, detailed descriptions of image content (OpenAI, 2023; Chen et al., 2024b), moving beyond the short captions of traditional methods. The proliferation of VLMs presents new opportunities for the image captioning field.

However, image captioning has not advanced as expected. Current VLMs focus on tasks like Visual Question Answering (Yue et al., 2024) and multimodal reasoning (Cheng et al., 2024a), bypassing the essential task of image captioning. A few works still rely on MSCOCO (Lin et al., 2014)—a dataset with an average caption length of 10 words, which is clearly outdated for evaluating advanced VLMs. This obstacle stems from the inherent difficulty in evaluating detailed captions. Unlike multiple-choice questions or mathematical reasoning, captioning lacks explicit answers, resulting in the absence of reliable evaluation benchmarks. Researchers are unable to assess the captioning capabilities of existing VLMs, nor effectively evaluate and improve their own models.

This paper addresses two key questions to drive the evolution of image captioning in the LLM era: (1) *How do existing VLMs perform in detailed cap-*

tioning? Do top models achieve human-level performance, and how do they compare against each other? We conduct the first large-scale human evaluation to benchmark current VLMs. Further, using resulting annotations from (1) as ground truth, we analyze existing captioning metrics and ask: (2) *How can we develop automated evaluation methods that reliably measure detailed caption quality and align with human preferences?*

To explore the first question, we developed *CapArena*, which includes over 6000 high-quality human annotations to evaluate the detailed captioning capabilities of 14 advanced VLMs and humans. In our preliminary study, we found that traditional scoring methods are unsuitable for annotating detailed captions’ quality due to their fine-grained intricacy and diversity. Inspired by LLM evaluation (Chiang et al., 2024), we adopted a pairwise caption battles paradigm. The resulting model rankings are shown in Figure 1. For the first time, we observe that state-of-the-art models, such as GPT-4o, are comparable to or even surpass human-level performance, marking a pivotal milestone in image captioning. While open-source VLMs achieve competitive results on general benchmarks, *CapArena* reveals a persistent performance gap between them and commercial models. An exception is InternVL2-26B (Chen et al., 2024b), a smaller open-source model that stands out with its strong performance, underscoring the potential of compact and efficient VLMs for detailed captioning.

In response to the second question, we conducted a comprehensive analysis of traditional and recent captioning metrics, as well as the ability of VLM-as-a-Judge (Chen et al., 2024a) to assess caption quality. We compared these metrics against human preferences from *CapArena*. Our results reveal that most metrics designed for short captions, such as CLIPScore (Hessel et al., 2021), fail entirely in the detailed captioning task. Although some rule-based metrics, such as METEOR (Banerjee and Lavie, 2005), exhibit decent agreement with human judgments at the caption level, they suffer from systematic biases across different VLMs. This results in low agreement at the model level, where the rankings produced by these metrics deviate significantly from human rankings. In contrast, we introduce VLM-as-a-Judge with reference captions, which demonstrates robust discernment for detailed captions. It achieves the highest alignment with human judgments at both the caption and model levels.

In light of these findings, we release *CapArena-*

Auto, an automated benchmark for detailed captioning. It comprises 600 samples and innovatively adopts the pairwise battle paradigm to improve evaluation reliability. VLM-as-a-Judge is employed to estimate human preferences by comparing captions against three baseline models. With 94.3% correlation to human rankings and just \$4 per test, *CapArena-Auto* offers a fast and robust pipeline for evaluating detailed captioning.

2 Related Work

2.1 Vision-Language Models

In recent years, Vision-Language Models (VLMs) have experienced rapid advances, achieving state-of-the-art performance across various multimodal tasks. Most VLMs integrate visual encoders (Radford et al., 2021; Zhai et al., 2023) with large language models (Bai et al., 2023; Touvron et al., 2023), allowing the latter to possess visual perception capabilities (Liu et al., 2024; Ye et al., 2023; Bai et al., 2023; Chen et al., 2024b, *inter alia*). To achieve this, VLMs are typically trained in two stages: pre-training on large-scale caption data to align visual and textual information, followed by supervised fine-tuning (SFT) with instruction-following capacity. Captioning plays a crucial role in building VLMs, as accurately describing image content forms the foundation for complex recognition and reasoning tasks (Cheng et al., 2024b; Huang et al., 2025). Current VLM evaluations mainly focus on visual question answering related to knowledge (Yue et al., 2024) and reasoning (Wu et al., 2024; Sun et al., 2024), while captioning is often overlooked due to evaluation difficulties. Our work is dedicated to benchmarking and analyzing the captioning ability of VLMs.

2.2 Image Captioning and Evaluation Metrics

Image captioning significantly progressed over the past decade (Vinyals et al., 2015; Xu, 2015; Anderson et al., 2018; Cheng et al., 2022; Wang et al., 2022; Ma et al., 2023; Cheng et al., 2023; Pi et al., 2024). These methods use human-annotated datasets like MSCOCO (Lin et al., 2014) and No-caps (Agrawal et al., 2019), evaluating generated captions by comparing them to reference sentences using rule-based metrics such as BLEU (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015). Recent research shows that CLIP-based metrics (Hessel et al., 2021) exhibit higher human consistency for short captions. However, current VLMs gener-



Image	Pairwise Caption Battle	Preference
	<p>Caption1 (Qwen2-VL-72B): ... The dog is holding a stick in its mouth, and the kitten is standing on its hind legs, reaching up to grab the stick. The kitten’s front paws are extended towards the stick, and its body is slightly arched as it tries to take the stick from the dog. ...</p> <p>Caption2 (Human): ... A gray and black kitten is leaping into the air at the dog’s face. It is facing the back at a right angle. Its right leg is extended out with its paw in front of the dog’s face. Its tail is down to the right in the air. ...</p>	Caption2 is better (more accurate and vivid description of the cat’s posture).
	<p>Caption1 (GPT-4o): ... The mural features a figure that appears to be inspired by traditional Asian art, possibly a deity or spiritual figure. The figure is depicted with a serene expression ... Surrounding the figure is a halo-like glow in warm tones of yellow and orange, adding a sense of divinity or spirituality.</p> <p>Caption2 (LLama-3.2-90B): The image depicts a vibrant mural of a woman on the side of a building, surrounded by various objects and features. **Mural:** The mural is painted in bright colors, featuring a woman with long dark hair wearing a blue robe ...</p>	Caption1 is better (<i>a spiritual figure</i> is much more informative than <i>a woman</i>).

Table 1: Examples of pairwise battles in *CapArena*. Red and green indicate less accurate and more preferable expressions, respectively. The evaluation guidelines are detailed in Section 3.1, and more examples are in Table 6.

ate significantly longer, detailed captions, where traditional metrics are not effective, posing challenges for evaluation.

Several recent works have focused on detailed captioning. Dong et al. (2024) introduced a new metric CAPTURE and improved VLM captioning performance. Lu et al. (2024) proposed a sophisticated scene graph-based metric. Differently, we first conducted a large-scale human-centered empirical study to systematically evaluate advanced VLMs. We then concentrated on the consistency between metrics and human preferences. An in-depth analysis using human annotated data (50 times larger than previous works) reveals the robustness and systematic biases of different metrics.

3 CapArena: Benchmarking VLMs in Detailed Image Captioning

While current VLMs excel in tasks like visual perception, question answering, and reasoning, their ability to generate long, detailed image descriptions remains unclear. In this section, we address this gap by introducing the first large-scale human-centered empirical study to benchmark VLMs in the context of detailed image captioning.

Next, we first present our evaluation protocol tailored for detailed captioning task (Section 3.1). Then, we describe the implementation of our annotation platform *CapArena* (Section 3.2). Finally, we highlight our key findings on the performance of existing advanced VLMs (Section 3.3).

3.1 Evaluation Protocol

We originally tried a scoring system (Hodosh et al., 2013), where annotators were asked to assign a

score from 1 to 5 to a single description. However, the task proved to be inherently complex and subjective. Since most generated captions cover the main content of the image, annotators found it difficult to assign precise grades, leading to low inter-annotator consistency. Inspired by open-domain evaluations of LLMs, we shifted to a pairwise comparison methodology (Chiang et al., 2024) to assess detailed captions, which was further validated in our preliminary study.

We believe a reliable evaluation protocol is crucial for accurate assessments. Inspired by Kasai et al. (2021), expert annotators drafted the initial guidelines, which were refined through in-house meetings with all annotators to ensure consistency. Finally, we established the following transparent evaluation protocol for detailed captioning.

Guidelines. Our guidelines primarily evaluate the quality of descriptions in terms of precision and informativeness. The full annotator guidelines can be found in Appendix H.

Precision: Precision measures how precise the content in the description is, i.e., whether the description aligns with the details in the image. For example, in the first case of Table 1, Qwen2-VL provides an inaccurate description of the cat’s posture, while the human description captures the crucial action of the cat pouncing toward the dog. Precision includes various aspects, such as objects, attributes, relationships, and positions.

Informativeness: Informativeness assesses how much of the key information in the image is comprehensively covered by the description, including the salient objects and important details. For example, in the second case of Table 1, Llama-3.2’s description of *a woman* is precise; however, it is

clearly less informative compared to GPT-4o’s description of *a spiritual figure*.

Hallucination: Hallucinations are considered an intractable flaw in VLMs (Li et al., 2023), where models generate objects that do not exist in the image. We instruct annotators to impose strict penalties for hallucinations, as they significantly harm the caption quality in real-world applications.

Additionally, previous studies show that human annotators are influenced by output length and response style (Chiang et al., 2024). To mitigate this, we asked annotators to focus solely on the quality of the descriptions, minimizing distractions from such aspects. Furthermore, for pairs of similar quality, we considered the longer description less favorable if it was noticeably too long.

3.2 CapArena: Pairwise Battle Platform

We developed *CapArena*, an annotation platform aimed at benchmarking VLMs’ performance in detailed captioning through anonymous pairwise battles. The platform covers a diverse range of image scenarios and evaluates a set of established VLMs through human annotator votes, providing reliable rankings between models.

Data Source. The test images are sourced from the recently proposed DOCCI dataset (Onoe et al., 2024), which includes high-resolution images of various real-life scenes. We provide detailed information about this dataset in Appendix A. Notably, each image in this dataset is paired by carefully crafted long, human-annotated descriptions, which are used as the human baseline on our platform.

For the tested models, we selected a diverse set of representative VLMs, including both commercial and open-source models, spanning a range of model sizes (the full list is provided in Appendix C). To minimize bias introduced by specific prompts, we crafted 10 prompts for detailed image captioning, such as *Describe this image in detail*. These prompts were manually reviewed to ensure they generated descriptions of similar quality and length (all prompts are provided in Appendix B). We show the caption length distribution of different VLMs in Figure 2, where GPT-4o is most similar to humans.

CapArena Infrastructure. We take N to denote the number of models. Let $n_1, n_2 \in [N]$ be the indices of the models. We define $A_t = (n_1, n_2)$ as the model pair compared at time t . The human response is indicated by $H_t \in \{0, 1\}$, where $H_t = 0$ indicates human preference for model n_1 , and $H_t = 1$ for model n_2 .

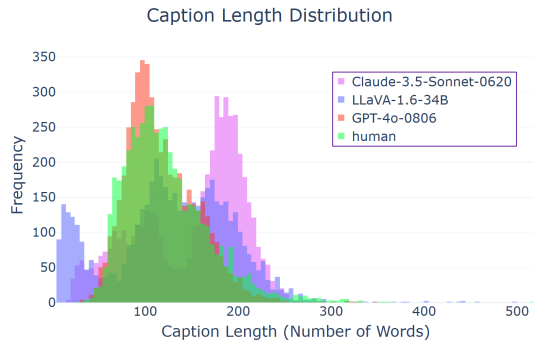


Figure 2: Caption length distribution of different VLMs.

To focus on comparing models with similar performance levels and thus accelerate the convergence of rankings, we adopt the probability update strategy from Chatbot Arena (Chiang et al., 2024). The sampling probability for each pair is proportional to the reduction in the size of the confidence interval:

$$P_t(a) \propto \sqrt{\frac{\hat{\Sigma}_{t,a,a}}{|\{t : A_t = a\}|}} - \sqrt{\frac{\hat{\Sigma}_{t,a,a}}{|\{t : A_t = a\}| + 1}},$$

where $P_t(a)$ is the probability of sampling pair a at time t , and $\hat{\Sigma}_t$ denotes the covariance matrix estimated from the t samples.

We then apply the Bradley-Terry (BT) model (Bradley and Terry, 1952) with the logistic form to calculate the scores of the models:

$$\hat{s} = \arg \min_{\xi} \sum_{t=1}^T \frac{1}{P(A_t)} \ell \left(H_t, \frac{1}{1 + e^{\xi_{A_t,2} - \xi_{A_t,1}}} \right),$$

where ξ is the vector of BT coefficients, which is an N -dimensional vector, and ℓ is the binary cross-entropy loss, defined as $\ell(h, p) = -(h \log(p) + (1 - h) \log(1 - p))$. The BT coefficients \hat{s} are used to create the ordered ranking of models. We bootstrap the BT rating estimate 1000 times to construct a confidence interval for each rating.

Annotator Training and Quality Control We made considerable efforts to enhance the reliability of the annotations. We conducted in-house annotations, with more than ten annotators being graduate students specializing in natural language processing, with no visual impairments, and familiar with the image captioning task.* We first conducted a preliminary annotation on 100 captions and developed the initial guidelines. In-house workshops followed to refine the process and ensure all annotators understood the task. Finally, we adopted the annotation methodology described in Section 3.1.

*The first four authors of participated in the annotation.

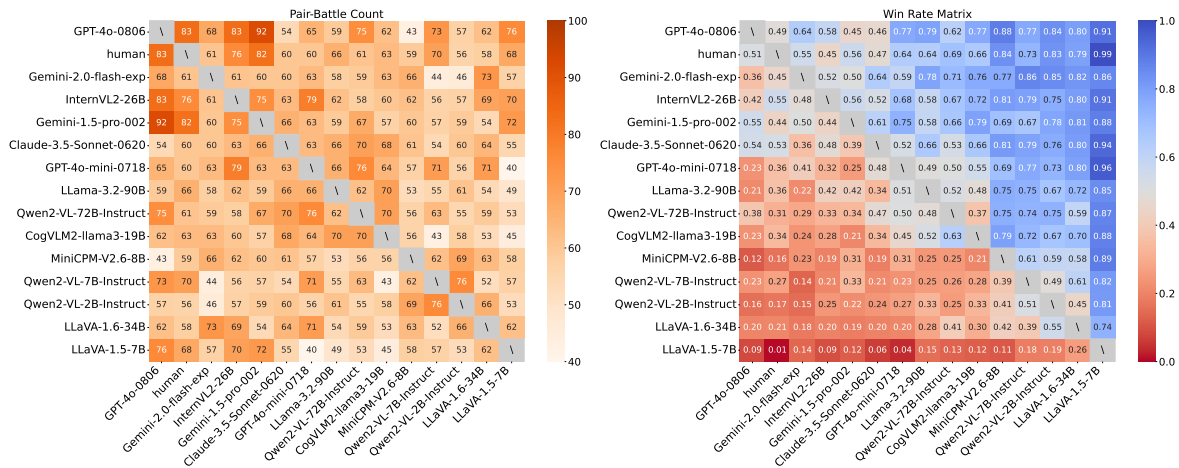


Figure 3: Battle counts and win rate matrix between models in *CapArena*.

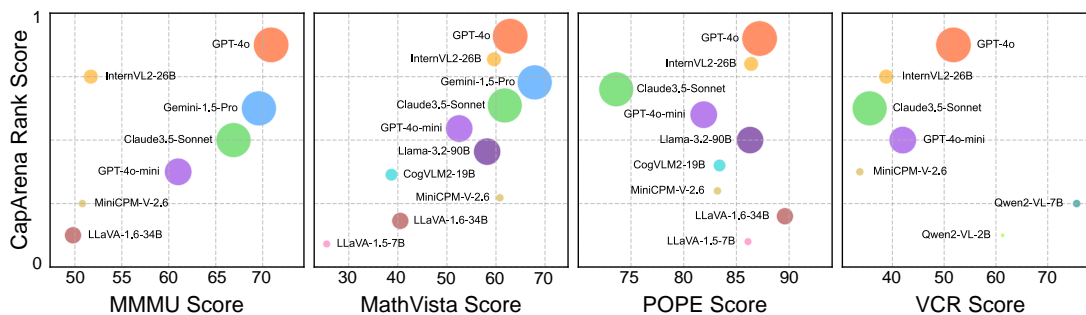


Figure 4: Correlation between vision-language benchmark scores and *CapArena* ranking. Models that perform well on general benchmarks do not necessarily excel in captioning. The size of each point represents the model size.

During the annotation process, we randomly selected 400 samples to be annotated by different annotators for inter-annotator validation. The results showed an agreement of 0.782, demonstrating the reliability of our manual labeling. Our empirical study found that the remaining 20% of disagreement stemmed not from annotator errors, but from the inherent subjectivity in interpreting which aspects of the image to emphasize. We also implemented measures to continuously monitor annotation quality during the labeling process (see Appendix E). Data collection began in October 2024, and by February 2025, a total of 6,522 annotation instances had been collected. The average time per annotation was 142 seconds.

3.3 VLMs Performance Analysis

The final ranking is shown in Figure 1. Figure 3 presents the battle count and win rate matrix.

Top models achieve human-level performance. Surprisingly, our evaluation reveals for the first time that leading models like GPT-4o have reached or even surpassed human-level performance in de-

tailed captioning tasks. This milestone demonstrates that machine-generated descriptions can now rival high-quality human descriptions. We conducted an empirical study in Appendix D, and the results show that even human expert annotators occasionally overlook image details, where GPT-4o’s descriptions are more comprehensive.

***CapArena* uncovers disparities in fine-grained visual perception across models.** As shown in Figure 1, most open-source models still lag significantly behind commercial models, even with large model sizes (e.g., Llama-3.2-90B). Despite competitive performance on general multimodal benchmarks, our results reveal the limitations of open-source alternatives in detailed captioning, where they lack stability in accurately perceiving image details. An exception is InternVL2-26B, a mid-sized model that demonstrates exceptional performance. We attribute this to its large-scale vision encoder, InternViT-6B, enabling strong visual understanding. These results highlight the growth potential of open-source models.

We present the correlation between model per-

formance on vision-language benchmarks and *CapArena* ranking in Figure 4. Strong performance on general benchmarks does not always translate to superior captioning ability. For example, despite its relatively low MMMU (Yue et al., 2024) score, InternVL2-26B excels in captioning. Moreover, the hallucination benchmark POPE (Li et al., 2023) lacks discriminative capacity for detailed captioning, underscoring the need for further exploration in assessing image comprehension.

Failure Cases. Besides weaker models missing image details or making errors, we identified several common mistakes: (1) failure to provide accurate descriptions for unusual scenes, (2) frequent neglect of subtle yet important details, (3) challenges in scenes requiring knowledge association, and (4) an inability of VLMs to accurately identify the time on clocks. Examples of these cases are provided in Table 7. These shortcomings highlight the need for further research in detailed captioning to advance real-world applications.

4 Analysis of Captioning Metrics

Thanks to its meticulously designed annotation process, *CapArena* can be regarded as a reliable evaluation system for detailed image captioning that reflects the performance of different models. However, it relies on costly human preference annotations, making annotating every model of interest impractical. Therefore, automatic metrics are crucial for evaluating and iterating on the development of detailed captioning capabilities in VLMs.

This section evaluates a range of traditional and recently proposed captioning metrics, and the ability of VLM-as-a-Judge to assess the quality of detailed captions using the *CapArena* data. The over 6000 high-quality human annotations serve as the golden standard, providing the basis to analyze how these metrics correlate with human preferences and to identify their potential deficiencies.

4.1 Captioning Metrics

Traditional Metrics. We first consider rule-based metrics that are commonly used for MSCOCO evaluation: **BLEU** (Papineni et al., 2002) measures n-gram overlap between the generated and reference captions. **METEOR** (Banerjee and Lavie, 2005) incorporates synonym matching and stemming to improve recall. **SPICE** (Anderson et al., 2016) measures semantic consistency using scene

graph. **CIDEr** (Vedantam et al., 2015) evaluates consensus across multiple reference captions.

Next, we include a range of CLIP-based metrics, which claim higher alignment with human judgments (Hodosh et al., 2013). **CLIPScore** (Hessel et al., 2021) is a reference-free metric that measures the similarity using CLIP features, and **LongCLIPScore** (Zhang et al., 2024) is a variant adapted for longer text. **Polos** (Wada et al., 2024) enhances CLIPScore through supervised learning on human feedback. **FLEUR** (Lee et al., 2024) leverages a large multimodal model and achieves state-of-the-art human alignment on short captions.

Metrics Designed for Detailed Captions. To evaluate long, detailed descriptions, a few specialized metrics have been proposed. **CAPTURE** (Dong et al., 2024) extracts visual objects, attributes, and relationships, performing multi-stage matching between generated and reference captions. **VDCScore** (Chai et al., 2024) is a video captioning metric that decomposes the reference caption into question-answer pairs and utilizes an LLM to check their correspondence with the predicted caption.

VLM-as-a-Judge. Leveraging powerful LLMs to simulate human preferences has proven effective in open-ended scenarios (Zheng et al., 2023). While recent studies have explored VLM-as-a-Judge in general multimodal tasks (Chen et al., 2024a; Li et al., 2024b), we are the first to apply it to detailed caption comparison—a task that demands fine-grained discernment of vision-language semantic alignment. Similar to human judgment, given an image and two descriptions, VLMs determine which one is better. We employ several well-established VLMs (e.g., GPT-4o, Qwen2.5-VL (Bai et al., 2025), LLaVA-OneVision (Li et al., 2024a), LLaVA-Critic (Xiong et al., 2024)) as evaluator. We also introduce a reference-enhanced variant that incorporates human descriptions to assist the model’s judgment.

4.2 Experiment Settings

To evaluate these metrics’ ability to assess detailed caption quality, we compare metric-based judgments with human annotations on caption battle pairs in *CapArena* and analyze their correlation with human preferences. For scoring-based metrics like CIDEr and CLIPScore, scores are used to determine the winner in pairwise battles. For VLM-as-a-Judge, we use a prompt similar to human annotator guidelines (See Appendix I). Since some pairs are of similar quality, we allow ties to

Metrics	Need Ref?	Caption-level Agreement (Including Tie)					Model-level Agreement	
		Overall	Level1	Level2	Level3	Level4	Spearman	Kendall τ
Inter-Annotator	-	0.683	0.810	0.650	0.650	0.620	-	-
Output Length	No	0.585	0.672	0.593	0.552	0.521	0.710	0.582
Traditional Image Captioning Metrics								
BLEU-4	Yes	0.474	0.477	0.480	0.475	0.467	0.424	0.319
SPICE	Yes	0.417	0.441	0.422	0.415	0.387	0.275	0.231
CIDER	Yes	0.384	0.378	0.383	0.389	0.387	-0.279	-0.209
METEOR	Yes	0.576	0.657	0.582	0.536	0.530	0.785	0.582
Polos	Yes	0.479	0.526	0.467	0.462	0.462	0.420	0.363
CLIPScore	No	0.325	0.266	0.308	0.362	0.355	-0.574	-0.451
LongCLIPScore	No	0.400	0.422	0.404	0.384	0.395	-0.226	-0.121
FLEUR	No	0.458	0.513	0.462	0.444	0.414	0.393	0.297
Metrics Designed for Detailed Image Captioning								
CAPTURE	Yes	0.525	0.601	0.512	0.504	0.479	0.613	0.538
VDC-Score	Yes	0.557	0.687	0.579	0.496	0.460	0.890	0.736
VLM-as-a-Judge								
LLaVA-OneVision	No	0.606	0.709	0.642	0.541	0.537	0.921	0.780
LLaVA-Critic	No	0.609	0.735	0.631	0.544	0.530	0.903	0.736
Qwen2.5-VL	No	0.625	0.739	0.647	0.566	0.552	0.908	0.736
GPT-4o	No	0.628	0.740	0.647	0.572	0.557	0.930	0.802
GPT-4o (with ref)	Yes	0.627	0.733	0.663	0.559	0.560	0.943	0.846

Table 2: Evaluation of various metrics on detailed captioning tasks. *Need Ref* indicates whether human-written reference captions are required. Higher scores indicate better alignment with human preferences. The best results in each column are highlighted in **bold**. GPT-4o as the evaluator achieves the best performance.

align with human annotations. Scoring-based metrics use a threshold to determine draws, ensuring a similar occurrence rate as in human annotations.

We analyze the agreement between metrics and human preferences from two perspectives:

Caption-level Agreement. Caption-level agreement measures the consistency between metrics and human judgments for the same pairwise caption battle. It is calculated as the proportion of pairs in *CapArena* where the metric’s decision matches the human judgment (A wins / B wins / tie). To assess metric performance across caption pairs of varying distinction, we categorize all samples into four levels based on the ranking gap between models. Level 1 consists of the most easily distinguishable model pairs (e.g., GPT-4o vs. LLaVA-1.5), while Level 4 includes those with the most similar performance (e.g., GPT-4o vs. Gemini-1.5).

We also report human annotators’ internal consistency across the four levels for reference, manually verifying 100 samples per level.

Model-level Agreement. Caption-level agreement measures how often a metric matches human judgments on individual battles. However, a metric may

achieve high caption-level agreement while exhibiting bias toward certain models, leading to inaccurate performance estimation. To address this, we introduce model-level agreement, which measures the consistency between rankings derived from human judgments and those derived from metrics. Specifically, we replace human annotations with metrics for all pairwise battles in *CapArena* and compute rankings using the same ELO mechanism. We compute this agreement using Spearman’s rank coefficient and Kendall’s τ .

4.3 Results Analysis

Most traditional metrics fail in the detailed captioning task. Traditional metrics with high human agreement on short captions (Hodosh et al., 2013) deviate from human preferences when applied to detailed captioning task. As shown in Table 2, they exhibit low agreement both at the caption-level and model-level. Rule-based metrics, such as CIDEr, struggle due to the flexible nature of detailed captions, which complicates n-gram matching. For CLIP-based metrics like CLIPScore, our results reveal that current vision-language rep-

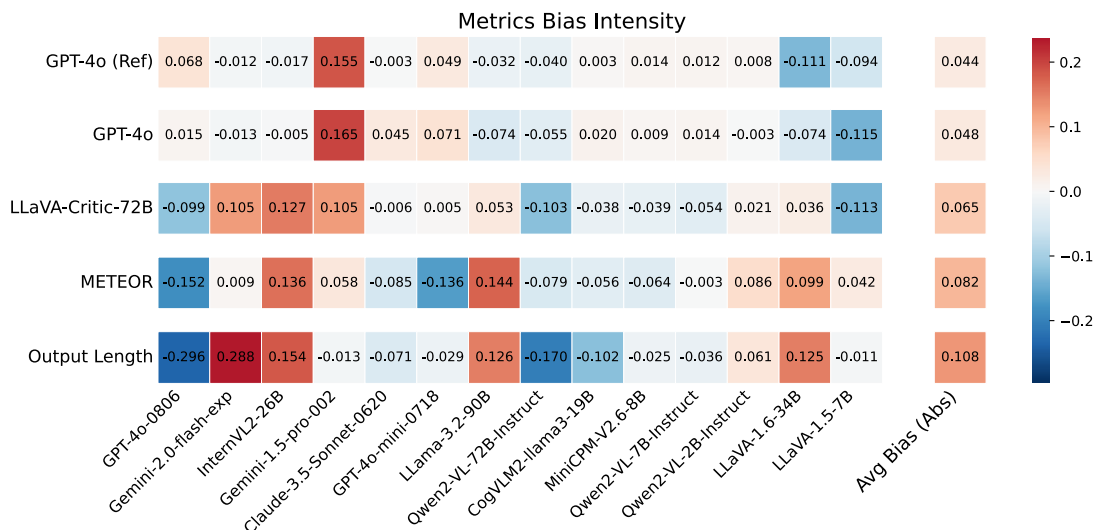


Figure 5: Metrics exhibit systematic biases—overestimating (positive values) or underestimating (negative values) certain models. Color saturation represents bias magnitude. Different metrics favor different models. GPT-4o exhibits lower biases (lighter overall colors), contributing to higher model-level agreement observed in Table 2.

resentation models fail to align fine-grained details in long descriptions to the image content.

VLM-as-a-Judge demonstrates stronger discernment for detailed captions. As shown in Table 1, accurate judgment for closely matched battles relies on a deep understanding of image content and a discriminative comparison between descriptions. Powerful VLMs with reasoning and fine-grained perception exhibit this capacity. No bells and whistles, GPT-4o as an evaluator, shows the highest consistency with human preferences, outperforming recently proposed metrics. Reference descriptions help the evaluator clarify uncertain image details, further improving model-level agreement.

Systematic biases—overestimating or underestimating certain models are a critical concern. As shown in Table 2, METEOR and Output Length (which simply favors longer captions) exhibit decent caption-level agreement, but their model-level agreement is notably lower. To investigate this issue, we assessed the metric’s intrinsic bias towards specific models, by calculating the model’s average win rate across all battles and comparing it with the golden win rate in Figure 3. A positive (negative) win rate difference indicates that the metric overestimates (underestimates) the model. As displayed in Figure 5, all metrics exhibit systematic biases. The degree of bias varies across metrics; Output Length shows a particularly strong bias, while GPT-4o-as-a-Judge has a lower bias than METEOR (average 4.4% vs. 8.2%). This suggests that the disagreement between GPT-4o-as-

a-Judge and humans is more likely due to random preferences per independent sample, rather than harmful bias towards specific models, leading to more accurate model ranking estimates.

Evaluating hard-to-distinguish caption pairs remains a challenge. Despite the reasonable performance of VLM-as-a-Judge, it still falls short of Inter-Annotator agreement, particularly on Level 3/4 samples which represent the most challenging model battles (e.g., 0.650/0.620 vs. 0.572/0.560). This suggests that the current VLM still has limitations in perceiving fine-grained image details and distinguishing subtle differences between captions, which in turn leaves room for further improvement in detailed caption evaluation.

5 CapArena-Auto: An Automated Benchmark for Detailed Captioning

CapArena relies on substantial human labor to estimate model performance, which is time-consuming and expensive. Therefore, developing an automated benchmark for detailed captioning is a desideratum to enable rapid evaluation and accelerate model development. In this section, based on the findings above, we introduce *CapArena-Auto*, a cheap and fast framework for detailed captioning evaluation. *CapArena-Auto* includes 600 evaluation images and assesses model performance through pairwise battles with baseline models.

Curation of Test Samples. We selected images from the DOCCI test split (Onoe et al., 2024) as candidates. The images, photographed by diverse

annotators in everyday contexts, encompass a wide range of scenes and are of high resolution. Given that these images were newly captured and released after the timestamp of most existing VLMs’ training, we believe there is minimal risk of data contamination. Next, to sample a diverse test set, we adopted the image feature clustering from DOCCI, uniformly sampling 600 samples from 149 clusters. Additionally, we applied CLIP feature-based filtering to remove overly similar samples, ensuring the quality of the final selection.

Evaluation Protocol. We employ a pairwise battle paradigm and use VLM-as-a-Judge for comparison (Li et al., 2024c; Chou et al., 2024). For each of the 600 test samples, we compare captions from the test model and a baseline model to determine the better one. To reduce potential noise and bias from a single baseline (Lin et al., 2024), we use three baseline models with different performance levels: GPT-4o, CogVLM-19B, and MiniCPM-8B. We use GPT-4o as the judge due to its high agreement with human preferences and provide human reference captions as additional support.

To compute the final score for the test model, we assign +1 for a win, -1 for a loss, and 0 for a draw in each pairwise comparison. The model’s score in *CapArena-Auto* is the total sum of its scores across the 600 test samples. We provide the current leaderboard of *CapArena-Auto* in Table 8.

	Spearman	Kendall τ
DOCCI (with BLEU-4)	0.341	0.275
DOCCI (with METEOR)	0.859	0.648
CAPTURE	0.763	0.604
<i>CapArena-Auto</i>	0.943	0.824

Table 3: Correlation between the automated benchmark and *CapArena*’s golden ranking. *CapArena-Auto* exhibits the highest alignment with human preferences.

To validate the effectiveness of *CapArena-Auto*, we compare it with several recent detailed captioning benchmarks. Table 3 presents the correlation coefficients between the model rankings from these benchmarks and the golden ranking provided by *CapArena*. *CapArena-Auto* outperforms existing benchmarks by a large margin, better aligning with human preferences. Additionally, its streamlined design allows each evaluation to cost only \$4, making it an effective and affordable evaluation benchmark for detailed captioning.

6 Conclusion

In this paper, we explore the task of detailed image captioning in the LLM era. We conducted the first large-scale human-centered empirical study to benchmark advanced VLMs performance. The results demonstrate that, for the first time, leading models (e.g., GPT-4o) achieve or surpass human-level performance, while open-source models lag behind. Next, we provided an in-depth analysis of existing captioning metrics. Our experiments reveal that VLM-as-a-Judge sets a new standard, highlighting that systematic biases in existing metrics are the key factor affecting alignment with human preferences. Finally, we release *CapArena-Auto*, an automated benchmark that closely aligns with human preferences, offering a cost-effective and efficient tool for detailed captioning evaluation. Our findings lay the foundation for the future development of image captioning.

Limitations

One limitation of *CapArena* is the scope of models and the domains covered. Current pairwise battle evaluation includes 14 representative Vision-Language Models (VLMs), but this selection is constrained by the available annotation resources. As a result, several recent models released after our dataset was compiled have not been considered in the evaluation. Additionally, the images used mostly focus on everyday life scenarios, which means that other domains, such as artwork or medical images, are not yet represented. Extending the evaluation to these domains could provide a more complete picture of VLM performance.

Acknowledgments

This work is supported by the Natural Science Foundation of China (NSFC, Grant No. 62176115).

References

- Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. **No-caps: Novel object captioning at scale**. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.
- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. **Spice: Semantic propositional image caption evaluation**. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. **Bottom-up and top-down attention for image captioning and visual question answering**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. **Qwen-vl: A frontier large vision-language model with versatile abilities**. *arXiv preprint arXiv:2308.12966*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. **Qwen2.5-vl technical report**. *arXiv preprint arXiv:2502.13923*.
- Satanjeev Banerjee and Alon Lavie. 2005. **Meteor: An automatic metric for mt evaluation with improved correlation with human judgments**. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Ralph Allan Bradley and Milton E Terry. 1952. **Rank analysis of incomplete block designs: I. the method of paired comparisons**. *Biometrika*, 39(3/4):324–345.
- Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jeng-Neng Hwang, Saining Xie, and Christopher D Manning. 2024. **Auroracap: Efficient, performant video detailed captioning and a new benchmark**. *arXiv preprint arXiv:2410.03051*.
- Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinyu Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. **Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark**. *arXiv preprint arXiv:2402.04788*.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. **How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites**. *Science China Information Sciences*, 67(12):220101.
- Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. 2024a. **Vision-language models can self-improve reasoning via reflection**. *arXiv preprint arXiv:2411.00855*.
- Kanzhi Cheng, Zheng Ma, Shi Zong, Jianbing Zhang, Xinyu Dai, and Jiajun Chen. 2022. **Ads-cap: A framework for accurate and diverse stylized captioning with unpaired stylistic corpora**. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 736–748. Springer.
- Kanzhi Cheng, Wenpo Song, Zheng Ma, Wenhao Zhu, Zixuan Zhu, and Jianbing Zhang. 2023. **Beyond generic: Enhancing image captioning with real-world knowledge using vision-language pre-training model**. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5038–5047.
- Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024b. **Seeclick: Harnessing gui grounding for advanced visual gui agents**. *arXiv preprint arXiv:2401.10935*.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. **Chatbot arena: An open platform for evaluating llms by human preference**. *arXiv preprint arXiv:2403.04132*.
- Christopher Chou, Lisa Dunlap, Koki Mashita, Krishna Mandal, Trevor Darrell, Ion Stoica, Joseph E Gonzalez, and Wei-Lin Chiang. 2024. **Visionarena: 230k real world user-vlm conversations with preference labels**. *arXiv preprint arXiv:2412.08687*.
- Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. **Meshed-memory transformer for image captioning**. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587.
- Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. 2024. **Benchmarking and improving detail image caption**. *arXiv preprint arXiv:2405.19092*.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. **Clipscore: A reference-free evaluation metric for image captioning**. *arXiv preprint arXiv:2104.08718*.
- Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. **Framing image description as a ranking task: Data, models and evaluation metrics**. *Journal of Artificial Intelligence Research*, 47:853–899.
- Muye Huang, Lingling Zhang, Jie Ma, Han Lai, Fangzhi Xu, Yifei Li, Wenjun Wu, Yaqiang Wu, and Jun Liu. 2025. **Chartsketcher: Reasoning with multimodal feedback and reflection for chart understanding**. *arXiv preprint arXiv:2505.19076*.

- Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A Smith. 2021. [Transparent human evaluation for image captioning](#). *arXiv preprint arXiv:2111.08940*.
- Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. [Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model](#). *arXiv preprint arXiv:2406.06004*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024a. [Llava-onevision: Easy visual task transfer](#). *arXiv preprint arXiv:2408.03326*.
- Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. 2024b. [Vlrewardbench: A challenging benchmark for vision-language generative reward models](#). *arXiv preprint arXiv:2411.17451*.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024c. [From crowdsourced data to high-quality benchmarks: Arena-hard and benchmark pipeline](#). *arXiv preprint arXiv:2406.11939*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. [Evaluating object hallucination in large vision-language models](#). *arXiv preprint arXiv:2305.10355*.
- Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. [Wildbench: Benchmarking llms with challenging tasks from real users in the wild](#). *arXiv preprint arXiv:2406.04770*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. [Microsoft coco: Common objects in context](#). In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. [Visual instruction tuning](#). *Advances in neural information processing systems*, 36.
- Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. 2024. [Benchmarking large vision-language models via directed scene graph for comprehensive image captioning](#). *arXiv preprint arXiv:2412.08614*.
- Zheng Ma, Mianzhi Pan, Wenhan Wu, Kanzhi Cheng, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2023. [Food-500 cap: A fine-grained food caption benchmark for evaluating vision-language models](#). In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5674–5685.
- Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. 2024. [Docci: Descriptions of connected and contrasting images](#). In *European Conference on Computer Vision*, pages 291–309. Springer.
- OpenAI. 2023. [GPT-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. 2024. [Image textualization: An automatic framework for creating accurate and detailed image descriptions](#). *arXiv preprint arXiv:2406.07502*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. [Learning transferable visual models from natural language supervision](#). In *International conference on machine learning*, pages 8748–8763. PMLR.
- Qiushi Sun, Kanzhi Cheng, Zichen Ding, Chuanyang Jin, Yian Wang, Fangzhi Xu, Zhenyu Wu, Chengyou Jia, Liheng Chen, Zhoumianze Liu, et al. 2024. [Os-generation: Automating gui agent trajectory construction via reverse task synthesis](#). *arXiv preprint arXiv:2412.19723*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. [Llama: Open and efficient foundation language models](#). *arXiv preprint arXiv:2302.13971*.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. [Cider: Consensus-based image description evaluation](#). In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. [Show and tell: A neural image caption generator](#). In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.
- Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. [Polos: Multimodal metric learning from human feedback for image captioning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13559–13568.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. [Opa: Unifying architectures, tasks, and modalities through a simple](#)

sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. 2024. *Os-atlas: A foundation action model for generalist gui agents*. *arXiv preprint arXiv:2410.23218*.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. *Llava-critic: Learning to evaluate multimodal models*. *arXiv preprint arXiv:2410.02712*.

Kelvin Xu. 2015. *Show, attend and tell: Neural image caption generation with visual attention*. *arXiv preprint arXiv:1502.03044*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. *mplug-owl: Modularization empowers large language models with multimodality*. *arXiv preprint arXiv:2304.14178*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. *Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. *Sigmoid loss for language image pre-training*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. *Long-clip: Unlocking the long-text capability of clip*. In *European Conference on Computer Vision*, pages 310–325. Springer.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. *Judging llm-as-a-judge with mt-bench and chatbot arena*. *Preprint*, arXiv:2306.05685.

A DOCCI Dataset

The DOCCI dataset consists of long, human-annotated English descriptions for images. These images were collected, curated, and donated for research purposes. Most images in DOCCI are natural scenes captured in both indoor and outdoor settings. The researcher aimed to capture key challenges, such as spatial relations, counting, text rendering, and world knowledge, among others. DOCCI captions require detailed, accurate

descriptions of objects, attributes, and actions, with clear and grammatically correct language. The annotations ensure high quality and consistency for evaluating image captioning models. The dataset’s diverse scenes and high-quality human-annotated captions make it ideal for our benchmark.

B Prompts for generating detailed caption

Table 4 provides the list of prompts used to generate detailed captions for images. Figure 6 shows captions generated with different prompts by the same VLM.

Prompts for Caption Generating

Describe this image in detail.
Generate a detailed caption for this image.
Explain the contents of this photo with specifics.
Create a rich and descriptive caption for this visual content.
Give this image a detailed image caption.
Write a detailed explanation of the scene depicted here.
Provide a detailed description of this photo.
Provide a thorough description of this image.
Describe this scene with specifics.
Provide a descriptive caption for this image.

Table 4: Prompts for detailed captioning.

C Model List used in annotation

The models selected for our experiments include both commercial and open-source models. Among the commercial models, we utilize GPT-4o-0806, GPT-4o-mini-0718, Gemini-2.0-flash-exp, Gemini-1.5-pro-002, and Claude-3.5-Sonnet-0620. On the other hand, we also incorporate several open-source models, including LLaMA-3.2-90B-Vision-Instruct, Qwen2-VL-72B-Instruct, Qwen2-VL-7B-Instruct, Qwen2-VL-2B-Instruct, LLaVA-v1.6-34B, LLaVA-v1.5-7B, InternVL2-26B, CogVLM2-llama3-chat-19B, and MiniCPM-V2.6-8B. By evaluating both commercial and open-source models, we aim to provide a comprehensive comparison of VLMs in the context of detailed captioning.

D GPT-4o vs Human Example

We observed that human captions occasionally overlook certain image details. For instance, in the upper image of Table 5, compared to humans, GPT-4o paid attention to the surrounding environment in greater detail. Additionally, its mention of "a red leash is attached to its collar, trailing across the grass" was more precise than simply stating that "a red leash leading to the bottom of the image". In



Figure 6: Detailed caption generated with different prompt.

the lower image, GPT-4o provided a more detailed description of the yacht and also noticed the poster on the window, a feature that was not mentioned by the human annotators.

E Quality Control

Throughout the annotation process, experts intermittently check the annotations of each annotator to ensure consistent high-quality results. When necessary, specific feedback was provided to guide the annotator and enhance the accuracy and consistency of subsequent annotations.

Additionally, the platform continuously monitored the annotators' activity. If an annotator's annotation time consistently fell below the average, the platform would trigger a feedback alert, prompting experts to pay closer attention to the quality of that annotator's work. Moreover, given the task's complexity, submissions made in under five seconds were considered as potential malicious attempts and were disregarded to maintain the integrity of the data. The rigorous annotator training procedures and stringent platform filtering mechanisms implemented throughout the annotation process ensured the collection of high-quality data.

F Protocol Example

We have included additional annotation examples in Table 6 to offer a clearer and more detailed illustration of our evaluation protocol. In the annotation process, annotators follow the established guidelines to carefully assess whether any issues exist within the pair of captions for a given image. They compare the two captions against each other, considering multiple aspects. Based on this comparison, annotators provide a annotation result, ensuring that the final result reflects the strengths and weaknesses of each caption. If it is difficult to distinguish between the captions, annotators are

also allowed to mark the result as tie.

G Failure Cases

In our analysis of VLM performance, we identified several common mistakes that persist even in models with strong overall performance. We provide four types of representative examples in Table 7. The last row of the table illustrates typical errors made by weaker models.

H Guidelines for Annotation

Our guidelines for judging which of the two image captions is better are shown in Figure 7.

While the guidelines provide a framework, they cannot account for all possible cases. Therefore, we encourage annotators to make informed judgments based on the specific circumstances and their own reasoning about which caption is more appropriate.

I Prompt Template

Building on [Chen et al. \(2024a\)](#); [Li et al. \(2024b\)](#), we designed the VLM-as-a-Judge prompt for pairwise judgments, as shown in Figure 8.

J CapArena-Auto Leaderboard

The current leaderboard of *CapArena-Auto* is provided in Table 8.



Image	Detailed Caption
	<p>human: A dog laying on it's belly looking at the camera, The dog appears to be by a tree on a lawn of grass with dead leaves around the area. The dog has white and curly fur, It has a bone held by its paws, and has a red leash leading to the bottom of the image. The dog is looking directly at the camera and has its body shifted to the right.</p> <p>GPT-4o-0806: A fluffy, cream-colored puppy is resting against the trunk of a large tree in a grassy area. The puppy has a soft, curly coat and is holding a chew stick in its front paws. A red leash is attached to its collar, trailing across the grass. The ground is scattered with dry leaves, and the background shows part of a house with some garden elements. The puppy looks content and relaxed in its outdoor setting.</p>
	<p>human: A medium three quarters frontal view of a white yacht on a dock near a body of water that is partially visible at the bottom of the view. Behind the yacht is a large blue building with a blue toned color on the right side of the view, and on its left side the color is gray. The building extends from the middle left of the view to the right side. At the top of the view the blue sky is visible behind large metal constructs and poles that stand above the yacht and nearby building. Between the metal columns and poles are large nets that extend across the upper portion of the view. Reflections of the white yahct are visible on the water at the bottom of the view. The yacht has blue accents around the bottom and has two different deck sections.</p> <p>GPT-4o-0806: The image depicts a large yacht docked at a marina. The yacht has a sleek, modern design with a white hull and a blue stripe near the waterline. The upper structure features a combination of white and wooden elements, giving it a classic and elegant appearance. The main deck has wooden railings and a covered area with blue fabric, likely for protection against the elements. The yacht has multiple levels, with large windows on the lower deck and a spacious upper deck area, possibly for navigation and leisure. There are antennas and equipment on the top, indicating communication or navigation systems. In the background, there is a large building with a blue exterior and multiple windows, some of which display images or posters. The building is adjacent to the marina, with a walkway and railings visible alongside the water. Above the building, there are tall metal structures, possibly part of a nearby facility or infrastructure. The water is calm, reflecting parts of the yacht and the surrounding structures. The sky is clear, suggesting a sunny day.</p>

Table 5: Examples of GPT-4o perform better than human





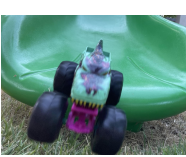



Image	Pairwise Caption Battle	Preference
	<p>Caption1 (Qwen2-VL-7B): ... The background shows more cars parked outside, indicating that the dealership has a large inventory. The showroom also features a high ceiling with large windows, allowing for ample natural light. ...</p> <p>Caption2 (Qwen2-VL-72B): ... Outside the large windows, the view shows a parking lot with several cars parked, including a mix of sedans and SUVs. The parking lot is surrounded by greenery, with trees and bushes visible in the distance. The sky is partly cloudy, with patches of blue sky and white clouds. ...</p>	Caption2 is better (Caption1 has baseless speculation).
	<p>Caption1 (Gemini-2.0-flash-exp): ... The most prominent feature is a large, reversed text that reads "BLACK SASH" when viewed correctly (likely through a window). The letters are black, three-dimensional, and seemingly made of a hard, glossy material. ...</p> <p>Caption2 (MiniCPM-V2.6-8B): The image shows a wall with the word "BLACK ASH" written in large, black, capital letters. Below the text, there are several colorful ribbons or banners attached to the wall. ...</p>	Caption1 is better (correctly point out "reversed" and accurately identify it).
	<p>Caption1 (GPT-4o-0806): ... It shows a person sitting on a wooden chair by a lake, but instead of a human head, the person has the head of a dog. The dog's head is large and fluffy, with a friendly expression and perked ears. The person is wearing a gray shirt, white pants, and brown slip-on shoes, and their arms are resting on the armrests of the chair. ...</p> <p>Caption2 (Claude-3.5-Sonnet): ... The scene shows a large, fluffy dog sitting on someone's lap in a wooden chair on what looks like a dock or pier by a body of water. The dog, which seems to be a mixed breed with long fur, is front and center, taking up most of the person's torso and obscuring their upper body. ...</p>	Caption2 is better (correctly identify the spatial relationship).
	<p>Caption1 (Claude-3.5-Sonnet): ... There are at least three turtles visible, with one larger turtle perched on top and two smaller ones below. The turtles' shells have distinct patterns and colors, ranging from dark green to brown. ...</p> <p>Caption2 (human): ... There is one big turtle and a little small one on top of its shell and in front of the big turtle there is a small one on the log looking up. ...</p>	Caption2 is better (accurately identify the characteristics of the three turtles).
	<p>Caption1 (Gemini-1.5-pro-002): A light green and purple monster truck toy is seen in motion, just about to drive off the bottom of a bright green plastic slide. The toy has large black tires, a pink chassis, and a textured, mottled design on its body that incorporates darker green and black markings, resembling scales or flames. ...</p> <p>Caption2 (Qwen2-VL-72B):...The truck is positioned on a grassy surface, and behind it is a large, green, plastic slide. The slide has a smooth, curved surface and is likely part of a playground set. ...</p>	Caption1 is better (precisely capture the motion state).
	<p>Caption1 (GPT-4o-mini-0718): ... The car is notable for its body damage—specifically, a significant scrape or dent running along the lower part of its driver's side, suggesting that it may have been in a minor collision or scraped against a post or another vehicle. ...</p> <p>Caption2 (InternVL2-26B): ... The car is positioned with its rear facing the camera, and the license plate is visible. The car appears to be in good condition, with no apparent damage. ...</p>	Caption1 is better (correctly describe the condition of the car).
	<p>Caption1 (human): ... The shadow depicts a person standing up straight and suspended in the air while jumping. The person can be seen wearing shorts, and the edge of a backpack can be seen. ...</p> <p>Caption2 (Gemini-2.0-flash-exp): ...Shadow Details:...**Pose:** The arms are positioned slightly away from the body and the legs are separated at the feet. This indicates a casual, natural standing position. ...</p>	Caption1 is better (accurately describe the jumping state).
	<p>Caption1 (CogVLM2-llama3-chat-19B): This image depicts a pink toy car with a yellow roof, parked on a paved surface. The toy car has a simplistic design, with a steering wheel on the left side and a single seat. Atop the toy car is a blue helmet, which is not typically associated with toys. The background shows a wooden fence and a gravel ground, suggesting an outdoor setting, possibly a backyard or a park.</p> <p>Caption2 (MiniCPM-V2.6-8B): The image shows a child's toy car, which is pink with a beige roof. The car is equipped with a steering wheel and has a blue helmet placed on the roof. The car is situated on a paved area with a wooden fence and gravel in the background. The toy car appears to be designed for young children, providing a safe and playful environment for imaginative play.</p>	Tie (Both captions accurately describe the appearance of the toy car, the surrounding objects and environment).

Table 6: More example of the evaluation protocol.


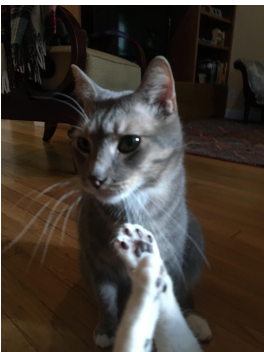



Category	Image	Failure Caption
Unusual Scene		<p>GPT-4o-0806: In this image, a flatbed tow truck is transporting a white GMC pickup truck. Unusually, a white Cadillac sedan is partially lodged underneath the rear of the pickup, with its front end lifted off the ground. The scene appears to be on a highway, with a shipping container visible in the background, indicating a busy roadway. The Cadillac's position suggests it may have collided with the pickup, resulting in this precarious arrangement. The surrounding environment includes concrete barriers and other vehicles, emphasizing the urban setting.</p> <p>Comment: A misinterpretation of the scene, where it should be the rear end being lifted, not the front, as commonly assumed.</p>
Important Details		<p>Claude-3.5-Sonnet-0620: This image shows a close-up of a gray cat sitting on a wooden floor. The cat has striking green eyes and is looking directly at the camera. Its fur appears to be a mix of light and dark gray, giving it a silvery appearance. The cat's whiskers are prominently visible. In the foreground, we can see the cat's paw raised, showing its pink paw pads. This gives the impression that the cat might have been in the middle of grooming or was about to make a playful gesture. The background of the image shows parts of what seems to be a living room. There's a glimpse of furniture, possibly a chair or sofa, and what looks like a bookshelf or cabinet in the distance. The room appears to be dimly lit, making the cat the main focus of the photograph. The wooden floor the cat is sitting on looks to be hardwood, and there's a hint of a rug or carpet visible at the edge of the frame. Overall, this image captures a candid and endearing moment of a domestic cat in its home environment, showcasing its curiosity and the close relationship between the photographer and the feline subject.</p> <p>Comment: There should be two cats in the scene, and the paw in the foreground actually belongs to the other cat, not the one in the close-up.</p>
Require Association		<p>GPT-4o-mini-0718: A series of five artistic shields are displayed on a patch of dirt next to a sidewalk. The shields, painted in vibrant reds and adorned with intricate designs that resemble leaves or vines, vary in size and orientation. They are positioned in a line, suggesting movement, with the largest shield at the far left and gradually decreasing in size towards the right. A grassy area and a white wooden fence can be seen in the background, completing the scene with a touch of nature and structure. The setting appears to be an outdoor space, likely a park or community area, illuminated by bright sunlight.</p> <p>Comment: It should be pointed out that this is a set of tomato slices, which is important for accurately describing the scene.</p>
Identify Clock Time		<p>Gemini-1.5-pro-002: The image shows a close-up of an analog clock on a textured, light blue wall. The clock has a black plastic or metal frame and a white face with black Arabic numerals. The time displayed is approximately 1:53. The word "QUARTZ" is printed in small black letters below the center of the clock face. The wall shows some signs of wear and tear. There's a noticeable crack running diagonally across the top left portion of the frame, and some smaller imperfections and possible chipped areas scattered around the wall's surface. In the top right corner, partially obstructing the view of the wall, is the corner of a metal fence or gate. A galvanized metal pole is attached to the wall with a metal bracket. The wall appears damaged around the area where the pole is attached, with some of the surface material missing, revealing what looks like the underlying concrete or plaster. A small section of chain-link fencing is also visible behind the pole. This suggests the clock might be located in a more utilitarian or perhaps outdoor setting.</p> <p>Comment: The correct time is about 1:07.</p>
General Mistake		<p>LLaVA-1.6-34B: The image captures a simple yet intriguing scene of a stack of six rectangular concrete blocks arranged in a pyramid-like structure on a sandy ground. The blocks, varying in shades of gray and brown, are stacked in such a way that the smallest block is at the top, and the largest block is at the bottom. The pyramid casts a shadow on the ground, adding depth to the scene. The sandy ground beneath the pyramid is scattered with small rocks and pebbles, providing a natural contrast to the man-made structure. The image is taken from a slightly elevated angle, giving a clear view of the pyramid and its surroundings. The overall composition of the image suggests a sense of balance and stability, as the pyramid stands firm on the sandy ground.</p> <p>Comment: The description of the arrangement of the bricks in terms of quantity and size is incorrect. There are more than six equal-sized bricks.</p>

Table 7: Examples of failure cases.

Guidelines

As an annotator, your task is to judge which of the two image captions is better. Below are some guidelines for your reference:

- **Precision:** The caption should accurately correspond to the content of the image, providing precise information about it. Common examples of imprecision include errors in color, quantity, spatial relationships, or the posture of people.
- **Informativeness:** Salient information in the image should be reflected in the caption. Since it is impossible to include every detail, you will need to subjectively judge which aspects of the image are important. For instance, describing an otter as "a small animal" is precise, but it is less informative than specifying "an otter".
- **Hallucination:** Captions that include descriptions of objects or elements that are clearly absent from the image should be significantly penalized.
- **Caption length:** Longer captions are not inherently better. For captions with equivalent informativeness, shorter ones are either better or at least not worse. Additionally, overly lengthy, verbose, or redundant expressions should be penalized.
- **Attention to detail:** Annotators should pay close attention to the details in the image to distinguish the quality of the descriptions.
- **Assistive description:** Imagine a visually impaired person asking you to describe the image for them. How would you convey the image to them?
- **Reverse thinking:** What image does the caption lead us to imagine? Does the caption effectively lead you to imagine the intended image?
- **Ties are acceptable:** If you find it genuinely difficult to determine which caption is better (e.g., both captions are excellent), marking a tie is acceptable.

Figure 7: Guidelines for Annotation























Model	Score_Avg↑	Score_GPT	Score_Cog	Score_CPM	Length_Avg
 Gemini-1.5-pro-002	56.17	29.0	61.0	78.5	168.6
 GPT-4o-0806	44.00	0	55.5	76.5	115.8
 Qwen2.5-VL-72B-Instruct	35.33	-1.0	49.0	58.0	163.7
 Gemini-2.0-flash-exp	30.83	-2.0	39.5	55.0	417.0
 Ovis-2-34b	27.00	-15.0	33.5	62.5	120.2
 Claude-3.5-Sonnet-0620	21.50	-14.0	30.0	48.5	147.9
 InternVL2-26B	13.00	-38.5	20.0	57.5	236.3
 GPT-4o-mini-0718	9.33	-36.0	17.0	47.0	139.8
 Ovis-1.6-27b	3.00	-49.5	14.5	44.0	94.2
 GLM-4V-Plus	-0.17	-51.5	13.0	38.0	109.3
 CogVLM2-llama3-chat-19B	-8.50	-56.5	0	31.0	115.9
 Qwen2-VL-72B-Instruct	-9.00	-50.5	-4.5	28.0	114.5
 LLaVA-Onevision-72B-sft	-12.33	-57.5	-6.0	26.5	200.9
 LLaMA-3.2-vision-90B-Instruct	-25.67	-72.0	-13.0	8.0	160.3
 Hunyuan-standard-vision	-26.00	-63.0	-19.0	4.0	354.1
 InternVL2-5-8B	-29.83	-71.0	-29.0	10.5	117.8
 MiniCPM-V2.6-8B	-38.00	-80.0	-34.0	0	106.7
 Qwen2-VL-2B-Instruct	-48.67	-86.0	-49.5	-10.5	116.8
 Qwen2-VL-7B-Instruct	-49.00	-78.0	-59.0	-10.0	97.8
 LLaVA-1.6-34B	-67.50	-92.0	-53.5	-57.0	124.8
 cambrian-34b	-75.00	-93.0	-76.0	-56.0	120.2
 LLaVA-1.5-7B	-94.00	-99.5	-92.0	-90.5	74.4

Table 8: CapArena-Auto Leaderboard.

Template prompt of VLM-as-a-Judge

(System Prompt)

You are a highly capable multimodal AI assistant tasked with evaluating image captions.

(Instruction)

Given an image and two candidate captions, you are required to determine which of the two captions is better.

(Noticement)

Below are some guidelines for your reference:

1. **Precision**: The caption should accurately correspond to the content of the image, providing precise information about it. Common examples of imprecision include errors in color, quantity, spatial relationships, or the posture of people.
2. **Informativeness**: Salient information in the image should be reflected in the caption. Since it is impossible to include every detail, you will need to subjectively judge which aspects of the image are important. For instance, describing an otter as “a small animal” is precise, but it is less informative than specifying “an otter”.
3. **Hallucination**: Captions that include descriptions of objects or elements that are clearly absent from the image should be significantly penalized.
4. **Attention to detail**: Annotators should pay close attention to the details in the image to distinguish the quality of the descriptions.
5. **Assistive description**: Imagine a visually impaired person asking you to describe the image for them. How would you convey the image to them?
6. **Reverse thinking**: What image does the caption lead us to imagine? Does the caption effectively lead you to imagine the intended image?
7. **Ties are acceptable**: If you find it genuinely difficult to determine which caption is better (e.g., both captions are excellent), marking a tie is acceptable.

While the above guidelines provide a framework, they cannot cover all possible cases. Therefore, we encourage you to make **subjective judgments** based on the specific circumstances and your own reasoning about which caption is better.

(Response Format)

Format your response into two lines as shown below: Reason: <your thoughts and reasoning process for the judgment>

Judgment: <Caption 1 is better>/<Caption 2 is better>/<Tie>

Figure 8: Template prompt of VLM-as-a-Judge.