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] SJTU

Project Page: https://capability-bench.github.io

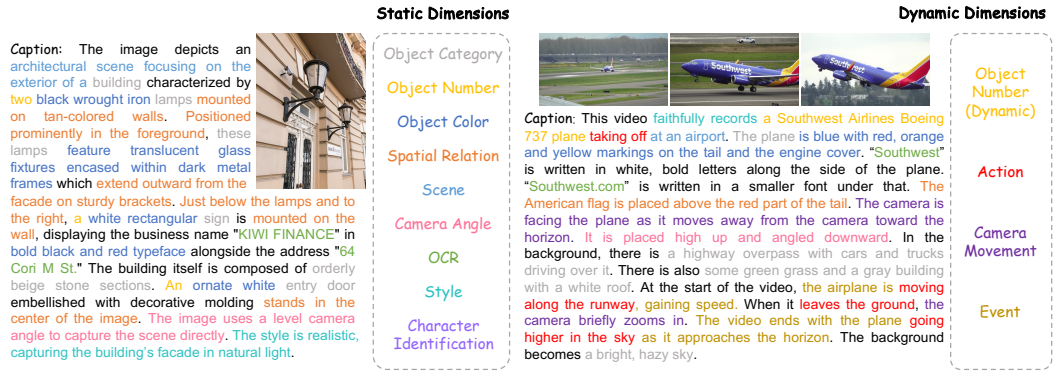

Figure 1: An example of image caption (left) and video caption (right) task. By analyzing the components of captions, we conclude 12 dimensions (9 static dimensions and 4 dynamic dimensions with object number shares on both static and dynamic), which all contribute to a detailed and comprehensive caption. The static dimensions are shared in both images and videos. Video data has additional dynamic dimensions that need to be judged with temporal relations.

Abstract

Visual captioning benchmarks have become outdated with the emergence of modern multimodal large language models (MLLMs), as the brief ground-truth sentences and traditional metrics fail to assess detailed captions effectively. While recent benchmarks attempt to address this by focusing on keyword extraction or object-centric evaluation, they remain limited to vague-view or object-view analyses and incomplete visual element coverage. In this paper, we introduce CAPability, a comprehensive multi-view benchmark for evaluating visual captioning across 12 dimensions spanning six critical views. We curate nearly 11K human-annotated images and videos with visual element annotations to evaluate the generated captions. CAPability stably assesses both the correctness and thoroughness of captions with *precision* and *hit* metrics. By converting annotations to QA pairs, we further introduce a heuristic metric, *know but cannot tell* ($K\bar{T}$), indicating a significant performance gap between QA and caption capabilities. Our work provides a holistic analysis of MLLMs' captioning abilities, as we identify their strengths and weaknesses across various dimensions, guiding future research to enhance specific aspects of their capabilities.

*Interns at Tongyi Lab, Alibaba Group

†Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

https://doi.org/10.52202/085713-2211

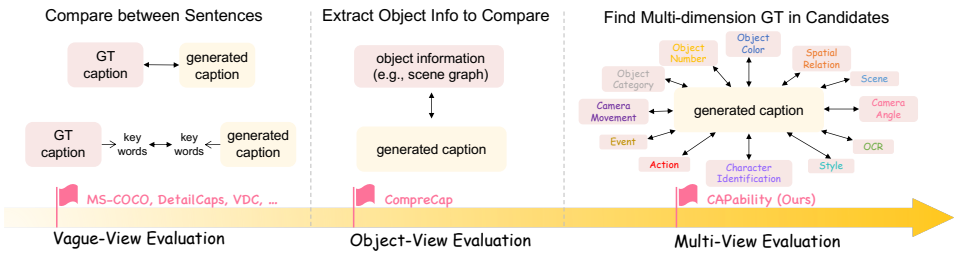

Figure 2: The development of visual caption benchmarks. Many works compare the ground-truth with generated sentences, which is vague. CompreCap [27] uses a scene graph to evaluate only object-related information. Our CAPability considers multiple views with a comprehensive evaluation.

1 Introduction

Visual captioning, which translates visual content into textual descriptions, is a fundamental task for both image and video understanding, affecting various downstream tasks [1, 2, 3]. The caption capability directly reflects the modality alignment ability [4, 5], forms a significant basis for image and video generation [6, 7, 8, 9], and provides the precondition of synthesizing a large-scale multi-modal understanding and reasoning dataset [10, 11]. To assess the capabilities of this task, researchers established several visual caption benchmarks in earlier years [12, 13, 14, 15].

With the rapid development of recent MLLMs [16, 17, 18, 19, 20, 21, 22, 23, 24], these traditional benchmarks have rapidly become outdated. This can be attributed to two main reasons: 1) The ground truths of traditional benchmarks often contain short sentences, missing many details. In contrast, recent MLLMs can produce much more detailed and fine-grained captions. 2) Traditional benchmarks use N-gram-based metrics (*e.g.*, BLEU [25], CIDER [26]) to directly compare the similarity between sentences, making evaluations unreliable due to their high sensitivity to sentence style.

Recently, new visual caption benchmarks have been introduced to update the outdated ones. As illustrated in Fig. 2, Dream-1K [28], DetailCaps [29] and VDC [30] extract keywords from both generated and reference captions, (*e.g.*, objects in DetailCaps, events in Dream-1K, objects, background, and camera information in VDC), and then compare the extracted information to score the caption, as opposed to earlier methods that directly compared sentences. We name all these methods vague-view evaluation as their evaluations still depend on the level of detail and accuracy of the ground-truth caption, which can suffer from human bias and cumulative errors from repeated extraction and comparison by LLMs. CompreCap [27] extracts object-related annotations from images (*e.g.*, scene graph) without ground-truth captions, thereby focusing on evaluating the object description capabilities of modern image MLLMs. We refer to this as object-view evaluation, as it drops entire sentences as ground truth, and evaluates captions based on object representation. Compared to traditional benchmarks, all these newly introduced approaches aim to provide more precise ground truths and evaluation methods, enhancing the reliability and interpretability of benchmarking.

Table 1: Our designed views and more detailed dimensions. We can treat a caption from the listed six views, and then split them into several dimensions.

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

However, the evaluation of these benchmarks remains incomplete as they focus on a single aspect of captions with limited visual elements, inadequately covering the full caption scope. For instance, they often overlook aspects like scene, text, and style. We argue that a multi-view evaluation for visual caption is essential. In this paper, we introduce a new comprehensive visual caption benchmark, CAPability. Our approach uses complete visual elements rather than caption sentences as annotations to evaluate both correctness and thoroughness for each dimension. The selection of dimensions is motivated by existing visual generation benchmarks [31, 32, 33], which is the inverse task of visual captioning. They usually treat and evaluate the generated visual content from different aspects (*e.g.*, objects, scene, style, camera, motion control). Similarly, we design 6 views and 12 dimensions for CAPability by analyzing several visual captions, as illustrated in Tab. 1 and Fig. 1. We believe these components contribute to a complete caption, as lacking any of them may align the caption

https://doi.org/10.52202/085713-2211

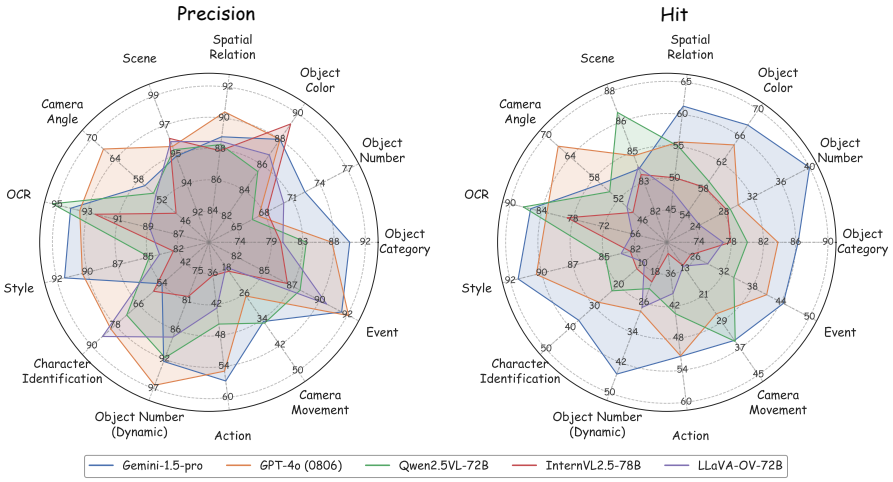

Figure 3: *Precision* and *hit* comparison of SOTA MLLMs on our CAPability. Models perform more variably on *hit* metric, which evaluates the thoroughness. GPT-4o performs the best on *precision*, and Gemini-1.5-pro [20] performs the best on *hit*.

with different visual content. There are 9 static and 4 dynamic dimensions, with *object number* encompassing both static and dynamic aspects. Static dimensions apply to both images and videos, while dynamic ones are exclusive to video. In our CAPability, we collect video data only for dynamic dimensions and image data for static dimensions for simplicity. We then manually annotate 11K images and videos for CAPability, providing sufficient samples.

In addition to multi-view annotation, we also focus on improving the evaluation quality of captions. While most methods assess only the correctness, we argue that considering both correctness and thoroughness of visual elements provides a more comprehensive evaluation for visual captioning. Therefore, we conduct comprehensive experiments using both *precision* and *hit* as our main metric, supplemented by another heuristic metric: the *know but cannot tell* ($K\bar{T}$). While *precision* only evaluates the accuracy of captions, *hit* considers both accuracy and coverage of captions compared with ground truth, and $K\bar{T}$ indicates the ratio when a model can answer related questions correctly but fails to convey the same information in the caption automatically. These metrics provide a robust framework for evaluating both correctness and thoroughness in MLLMs' detailed captions. To our knowledge, we are the first to heuristically highlight the gap in correctness and thoroughness capabilities of MLLMs across multiple views, providing deeper insight into specific capabilities or limitations of a model, thereby offering actionable guidance for further research and development, rather than just an overall score. Representative results are shown in Fig. 3, leading to the following conclusions: 1) Models perform more variably on *hit*, which evaluates the thoroughness and may be ignored by previous research. 2) GPT-4o seems the best on *precision*. For *hit*, Gemini-1.5-pro [20] has a leading advantage in many dimensions, followed by GPT-4o [34]. 3) Gemini-1.5-pro demonstrates strong object-counting abilities, while GPT-4o excels in identifying camera angles. 4) All models still struggle with dimensions like object numbers, camera angle, camera movement, character identification, and action. We hope our findings guide researchers to focus on improving these abilities in caption tasks. Our main contributions are listed as follows:

Table 2: 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.

Benchmark	Data Type		Anno-tations	Evaluation	
	Image	Video		Thoroughness	$K\bar{T}$
MS-COCO[12]	✓	-	Sentences	-	-
MSRVTT [14]	-	✓	Sentences	-	-
Dream-1K [28]	-	✓	Sentences	single dim	-
VDC [30]	-	✓	Sentences	-	-
DetailCaps [29]	✓	-	Sentences	-	-
CompreCap [27]	✓	-	Object Info	single dim	-
CAPability (Ours)	✓	✓	Multi-view Elements	✓	✓

- We introduce a new comprehensive visual caption benchmark, CAPability, featuring 6 views and 12 dimensions. By collecting and human-annotating nearly 11K images and videos, CAPability provides a novel and comprehensive methodology for caption benchmarking.

https://doi.org/10.52202/085713-2211

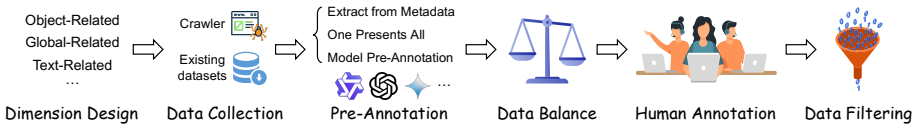

Figure 4: The pipeline of our data annotation for each dimension.

- We emphasize that captions should be evaluated for both correctness and thoroughness. Accordingly, we report *precision* and *hit* to combine correctness and thoroughness.
- We transform our annotations into a QA format to evaluate QA accuracy. Based on this approach, we assess an additional capability via the $K\bar{T}$ metric, which indicates the performance gap between QA and the captioning task.

2 Related Work

Multi-modal large language models. Based on the significant development of Large Language Models (LLMs) among various linguistic tasks [35, 36, 37, 38], many works try to extend the powerful capabilities into multi-modal understanding [39, 40, 41]. By integrating image content into LLMs, Multi-modal Large Language Models (MLLMs) also gain huge achievements [16, 42, 43, 17, 44, 45, 18]. Based on the pre-trained weights from image models, recent MLLMs also expand video understanding capabilities [46, 19, 10, 47, 21, 22, 23, 24]. With rapid development, MLLMs are powerful enough to describe both the image and video content in detail, which makes the traditional benchmarks with short captions outdated. More and more methods even try to produce re-captioned detailed descriptions by more powerful models rather than existing human-annotated short captions to train their model [19, 23]. Therefore, it is urgent to propose a new visual caption benchmark that adapts to modern MLLMs.

Visual caption benchmarks. Visual Caption is a fundamental task in computer vision. Early visual caption benchmarks, such as MS-COCO [12], NoCaps [13], MSR-VTT [14], and VATEX [15], usually contain a short sentence with limited visual information as the ground truth. They also use metrics like BLEU [25], CIDER [26], and METEOR [48] to calculate the matching score directly between two sentences, which is easily affected by the sentence style. Recently, from the annotation aspect, DetailCaps [29] extracts object-related information from the ground-truth caption, Dream-1K [28] splits the ground truth and candidates sentences into events. VDC [30] also extracts the object, background, and camera information from the video captions by question templates. However, they still rely on the ground-truth caption with human-bias, and require existing LLMs to extract and compare multiple times, thus increasing the cumulative error. CompreCap [27] explores directly annotating the object-related information in image captions, making the benchmarking more interpretable. On the contrary, we are the first time to propose a comprehensive visual caption benchmark covering both image and video data with 6 views and 12 dimensions. For evaluation, most methods only focus on correctness. Dream-1K [28] and CompreCap [27] begin to focus on thoroughness and calculate the recall of events or the object coverage in the segmentation map. However, they still remain incomplete as they only evaluate one dimension and limited metrics. We design metrics about both correctness and thoroughness, which may be ignored by previous work. We summarize the comparison with other visual caption benchmarks in Tab. 2, and we are the most holistic on all listed aspects.

3 CAPability

3.1 Multiple Dimension Data Annotation

The pipeline of our whole collection and annotation is shown in Fig. 4. We first design 6 views and split 12 dimensions, then collect nearly 1,000 images and videos for each dimension with low overlap among dimensions. For the collected data, we conduct pre-annotations by SOTA MLLMs and the following data balancing before the human annotation. After the human annotation of each dimension, we filter bad cases during the annotation and finally complete the data of CAPability.

Dimension design. Motivated by visual generation benchmarks [31, 32, 33], we conclude 6 views and split them into 12 dimensions based on the analysis of caption cases, as shown in Tab. 1. As

https://doi.org/10.52202/085713-2211

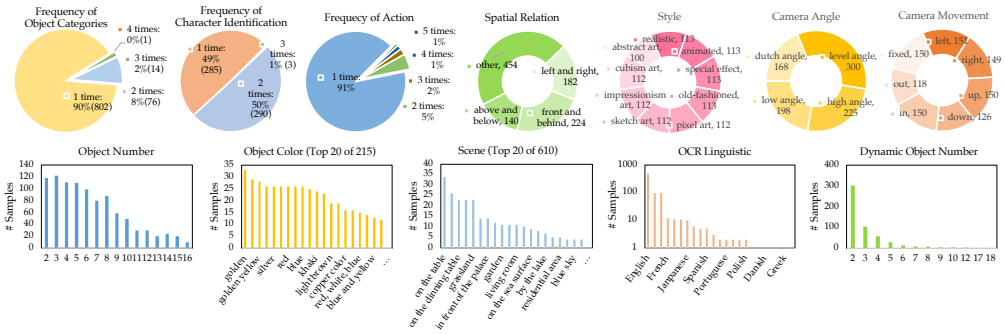

Figure 5: The annotation distribution of each dimension. We statistic different dimensions with different types. We count the frequency in object categories, character identification, and action as most of the descriptions only appear one time. For spatial relation, we summarize 4 categories and count their numbers. For style, camera angle, and camera movement, we count the samples of each category. For others, we plot bar charts to count and show the most frequent samples.

shown in Fig. 1, we design 9 static dimensions for both video and image, and 4 dynamic dimensions for video, covering most of what makes up a visual caption. We classify dimensions as dynamic or static based on the following principle: descriptions obtainable from a single frame are static, those requiring the entire video are dynamic, which are more related to temporal information. For the object number dimension, the number can be counted statically in an image, and also dynamically in a video, which is more challenging [49]. We also design the annotation type of each dimension as two types: open-ended, and specific categories. Specifically, we define 9 categories for style, 4 categories for camera angle, and 7 categories for camera movement. The specific categories of each dimension can be found in Fig. 5. See Appendix A.1 for details of dimension design.

Data collection. For convenience and problem simplification, we only collect image data for static dimensions and video data for dynamic dimensions. This is based on the common sense that the video understanding capabilities for MLLMs are usually built upon sufficient image understanding capabilities [46, 10, 23, 21]. Since an image or a video cannot cover all these dimensions of information, we directly collect data for each dimension independently and evaluate each dimension separately. For static dimensions, we mainly collect images from SA-1B [50], COYO-700M [51], Wukong [52], and Wikipaintings [53], and we also crawl a considerable amount of data from multiple public datasets and websites with CC0 license by ourselves. We also borrow parts of the image data and annotations from CompreCap [27] for the spatial relation dimension. For dynamic dimensions, we crawl and cut videos for camera movement dimension, borrow videos from Dream-1K [28] for action and event dimensions, and borrow videos from VSI-Bench [49] for the dynamic object number dimension. Fig. 6 shows our data sources for each dimension and their proportion, and Tab. A1 shows our data overlap among dimensions.

Pre-Annotation. The annotations may not be unique for different dimensions. For global-related, camera-related, and knowledge-related views, the annotations tend to be unique as an image only belongs to one kind of scene, style, *etc*. We directly pre-annotate them by extracting the metadata (*e.g.*, style for images in Wikipaintings [53]), or ask SOTA MLLMs to get a preliminary answer. For object-related, text-related, and temporal-related views, there could be multiple objects, texts, or actions in an image or video. However, it is extremely hard to annotate all objects or actions within an image or a video, as the categories of objects can be divided by almost infinite granularity [54, 55, 56]. Therefore, we do not pursue the most comprehensive annotation possible for each single sample, but randomly select only one object from the visual content, and the same for other dimensions, and reflect the accuracy and thoroughness through the evaluation of a large number of samples. We name this strategy as *One Represents All*. According to the law of large numbers, the distribution of randomly selection can approximate the expectation of covering different granularities of the entire visual content with a large amount of samples, thus ensuring the unbiased nature of the benchmark. Therefore, the key of this annotation strategy is to keep the selection as random as possible. To avoid humans' bias on selecting, we ask the three SOTA MLLMs, *i.e.*, GPT-4o [34], Gemini-1.5-pro [20], and Qwen-VL-Max [47] to list all objects and actions at the granularity they deem appropriate in an image or video, ask PaddleOCR [57] to list all texts in an image. We finally use Qwen2.5-Max [37] to merge the results together and randomly select one from the merged list to obtain the pre-annotated

https://doi.org/10.52202/085713-2211

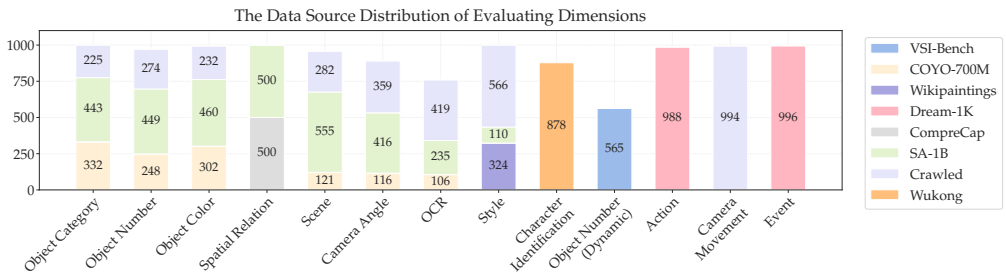

Figure 6: The data source count and distribution of each dimension. We collect nearly 1,000 images/videos for each dimension, crawl parts of data by ourselves, and sample some data from existing datasets to ensure diversity.

results. For further object-related dimensions, *e.g.*, object number, object color, and spatial relation, the object selection follows this strategy, then pre-annotate these attributes by MLLMs.

Data balance. Based on the pre-annotation results for each dimension, we conduct data balance strategy to control the difficulty and diversity. For dimensions with specific categories, we try to make the number of each category similar. For dimensions of open-ended descriptions, we count the frequency of descriptions, suppress the long-tail distribution, keep low-frequency words, ensuring the variety. See Appendix A.4 for examples. The final annotation distribution is shown as Fig. 5.

Human annotation. For different dimensions, we design different tasks for human annotators. For example, human annotators are asked to judge only right or wrong for object categories and actions rather than changing the annotated descriptions since we need to keep randomness. For dimensions with specific categories, we ask annotators to check the pre-annotated option and select the correct option. As for other dimensions with open-ended descriptions, we ask annotators to check the pre-annotations one by one and modify them based on pre-defined rules if there are any mistakes. We also conduct human-validation of all annotations to ensure the accuracy of annotations is above 97%.

Data filtering. We finally conduct data filtering to drop harmful visual content and re-balance the data since many of them are modified manually. The final distribution of each dimension is shown in Fig. 5. Some benchmark examples are shown in Appendix A.8.

3.2 Multiple Dimension Evaluation

Caption evaluation. As we collect and annotate the data of each dimension separately, we also evaluate each dimension independently. Different from matching the similarity between the caption and ground-truth sentences, our annotation drops the caption sentence, and we use GPT-4 Turbo [58] to take interpretable scores for each dimension. We use a similar prompt template for dimensions with specific categories (*i.e.*, style, camera angle, and camera movement), and use another similar prompt template for other open-set dimensions. See Appendix B.2 for the details of the prompts. Therefore, we can judge the caption into the following three situations: 1) **MIS**, which means the caption does not mention the corresponding content about the dimension. 2) **COR**, which means the caption mentions the corresponding content about the dimension, and describes it correctly compared with the annotations. 3) **INC**, which means the caption mentions the corresponding content about the dimension, but gives a wrong description compared with the annotations. As all data can be judged into these three situations, we can then calculate two metrics: 1) ***Precision***, which represents the accuracy on all samples that the model has mentioned, and thus only considers the correctness. 2) ***Hit***, which represents the accuracy on all annotated samples, no matter whether the dimension is described or missed in the caption, and thus also considers thoroughness. Our *Hit* is similar to recall, see Appendix A.7 for more discussions. Given the set of all samples as $S(\text{ALL})$, positive samples as $S(\text{COR})$, negative samples as $S(\text{INC})$, missed samples as $S(\text{MIS})$, the metrics can be calculated by:

$$\text{Precision} = \frac{|S(\text{COR})|}{|S(\text{COR}) \cup S(\text{INC})|}, \tag{1}$$

$$\text{Hit} = \frac{|S(\text{COR})|}{|S(\text{ALL})|}. \tag{2}$$

https://doi.org/10.52202/085713-2211

Table 3: The *precision* and *hit* of closed-source models and 72B open-source models on all dimensions. The *precision* represents the accuracy of what the models have described. The *hit* shows how many visual elements in the image can be described correctly. For video inputs, we send the whole video for Gemini, and uniformly sample 50 frames for GPT due to the API limitation of image number.

	Methods	Obj. Cate.	Obj. Num.	Obj. Color	Spa. Rel.	Scene	Cam. Ang.	OCR	Style	Cha. Iden.	(D) Obj. Num.	Act.	Cam. Mov.	Event	Avg.
Precision	LLaVA-OV-72B	80.4	70.1	86.8	88.5	96.2	53.6	88.9	83.9	**84.5**	87.5	42.7	17.7	90.2	74.7
	Qwen2VL-72B	82.0	70.0	89.2	88.6	95.2	52.4	**95.9**	82.9	83.3	83.6	44.6	33.3	89.9	76.2
	InternVL2.5-78B	80.1	68.3	89.2	87.9	**96.4**	48.4	92.5	82.8	58.3	80.0	36.4	19.0	86.8	71.2
	Qwen2.5VL-72B	83.7	66.7	85.5	88.3	95.7	54.2	95.3	84.7	72.1	91.3	45.8	35.1	87.9	75.9
	GPT-4o (0806)	87.3	67.4	88.0	**90.4**	95.9	**67.0**	93.5	90.0	80.2	**96.4**	54.9	26.7	**92.1**	79.2
	Gemini-1.5-pro	**89.8**	72.4	88.0	88.8	95.3	56.4	94.1	**91.4**	54.0	92.0	**56.8**	34.6	91.5	77.3
	Gemini-2.0-flash	85.9	**78.6**	**90.4**	89.0	96.1	57.4	95.3	86.9	82.0	94.2	50.6	**35.4**	89.0	**79.3**
Hit	LLaVA-OV-72B	77.0	24.6	54.7	47.9	84.0	49.9	66.6	83.5	9.3	27.3	39.6	12.2	28.6	46.6
	Qwen2VL-72B	79.9	25.1	57.3	50.4	85.1	52.1	79.6	82.6	5.7	18.1	41.7	25.7	31.2	48.8
	InternVL2.5-78B	77.9	28.5	58.3	50.1	83.6	48.4	79.1	82.8	12.4	20.5	32.1	12.0	25.0	47.0
	Qwen2.5VL-72B	80.0	28.9	59.2	55.0	**86.9**	54.2	87.5	84.7	22.7	22.3	43.4	34.9	34.1	53.4
	GPT-4o (0806)	83.8	30.0	64.7	55.7	84.6	**67.0**	83.0	90.0	28.1	28.3	51.3	26.6	41.0	56.5
	Gemini-1.5-pro	**86.3**	**40.0**	**67.7**	**61.3**	83.9	56.4	86.1	**91.4**	36.5	**45.0**	**51.4**	34.6	**44.5**	**60.4**
	Gemini-2.0-flash	82.5	30.6	60.8	51.8	84.0	57.4	**88.8**	86.8	**37.9**	28.7	46.6	**35.2**	39.7	56.2

Question-answer pairs evaluation. As we annotate each descriptive element for each dimension rather than caption sentence, we can also convert our annotations to question-answer (QA) pair format to evaluate the MLLMs' general ability out of the horizon of caption only. See Appendix A.8 for examples of our CAPability-QA. Based on the QA accuracy, we introduce a new metric, *know but cannot tell* ($K\bar{T}$), which evaluates the situation when a model knows the answer (*i.e.*, can answer correctly when given it the related question), but cannot tell automatically in the caption without specific question as prompt. This evaluation is significant to the caption task of MLLMs, but is usually ignored by previous methods. Given the set of correct answers as $S_{qa}(\text{COR})$, $K\bar{T}$ can be calculated as:

$$K\bar{T} = \frac{|S_{qa}(\text{COR}) \cap [S(\text{INC}) \cup S(\text{MIS})]|}{|S_{qa}(\text{COR})|}. \tag{3}$$

4 Experiments

4.1 Experimental Setups

For comprehensively evaluating the state-of-the-art (SOTA) models, we both choose several popular open-source and closed-source MLLMs. For closed-source models, we evaluate GPT-4o (0806) [34], Gemini-1.5-pro [20], and Gemini-2.0-flash [59]. For open-source models, we evaluate InternVL2.5 [21], LLaVA-OneVision [19], NVILA [22], VideoLLaMA3 [23], Qwen2VL [47] and Qwen2.5VL [24] with their different LLM sizes. We use the same image prompt and video prompt to infer all models, see Appendix B.2 for the inference prompts. We use GPT-4 Turbo (1106-preview) [58] to take scores for all generated captions to complete our evaluation. See Appendix B.1 for more implementation details.

4.2 Main Evaluation Results

Precision and hit of closed-source API and 72B models. We report the *precision* and *hit* of closed-source and 72B models in Tab. 3. Gemini-2.0-flash and GPT-4o achieve the highest *precision* (79.3% and 79.2%), which represents their captions are truthful and accurate. When it comes to the *hit* metric, all results drop significantly as it is harder to cover as many visual elements as possible. Gemini-1.5-pro achieves the best with 3.9% higher than second place, *i.e.*, GPT-4o, which means it is better at identifying more elements correctly. Among 72B models, Qwen2.5VL achieves the best *hit*, with notable *precision*. We note that the high precision but moderate hit for models like GPT-4o suggests they tend to describe only what they are confident about and avoid parts they are less certain of (line 245). This behavior results in high accuracy for described elements, but at the cost of thoroughness and coverage. When we focus on each dimension, it is worth noting that these models behave differently in different dimensions. Gemini-1.5-pro has a huge advantage in object counting in both image and video, especially in thoroughness (10% better than the second place for image and 16.3% better than the second place for video in *hit* metric). GPT-4o excels at recognizing camera angle, as it is 9.6% higher than the second place on both *precision* and *hit*. Qwen2.5-VL performs

https://doi.org/10.52202/085713-2211

Table 4: The *precision* and *hit* of 7B open-source models on all dimensions. We keep their default settings for each model.

	Methods	Obj. Cate.	Obj. Num.	Obj. Color	Spa. Rel.	Scene	Cam. Ang.	OCR	Style	Cha. Iden.	(D) Obj. Num.	Act.	Cam. Mov.	Event	Avg.
Precision	LLaVA-OV-7B	79.5	67.8	87.3	88.7	95.3	41.9	87.7	**84.4**	**90.9**	**92.8**	38.9	20.2	87.3	74.0
	Qwen2VL-7B	80.3	68.3	**88.7**	89.0	95.4	39.9	**95.4**	77.1	83.3	73.5	42.5	24.2	86.7	72.6
	NVILA-8B	80.6	68.5	84.2	88.4	95.4	44.5	92.8	79.0	90.8	47.2	32.4	14.7	**92.1**	70.0
	InternVL2.5-8B	76.1	60.3	85.8	89.3	95.1	42.5	89.0	81.9	48.3	84.4	37.9	20.2	84.5	68.9
	VideoLLaMA3-7B	81.0	66.8	85.5	**90.6**	**97.0**	43.2	90.0	80.9	84.1	78.9	43.0	**30.2**	88.3	73.8
	Qwen2.5VL-7B	**82.0**	**73.5**	88.0	88.6	95.8	**47.7**	93.8	82.0	80.8	92.4	**43.7**	26.8	86.5	**75.5**
Hit	LLaVA-OV-7B	76.8	23.0	53.0	48.5	82.7	33.4	64.5	**83.4**	4.6	**32.0**	35.8	12.6	27.0	44.4
	Qwen2VL-7B	78.4	20.6	50.4	46.1	84.7	39.1	73.4	77.1	4.0	14.7	40.0	17.4	27.5	44.1
	NVILA-8B	78.2	**23.5**	54.6	46.6	81.3	37.9	69.1	77.5	6.7	10.4	26.1	7.2	19.8	41.5
	InternVL2.5-8B	73.8	23.0	52.2	49.3	83.0	42.5	75.3	81.9	**9.8**	19.1	34.6	19.2	27.8	45.5
	VideoLLaMA3-7B	77.0	22.7	53.4	**51.1**	83.0	40.2	67.2	79.6	4.2	7.3	**41.5**	25.1	**30.5**	44.8
	Qwen2.5VL-7B	**79.3**	19.7	**56.0**	49.0	**85.6**	**47.3**	**81.3**	82.0	9.1	19.3	40.4	**26.5**	30.3	**48.1**

Table 5: The accuracy↑ (higher is better) of CAPability-QA and the $K\bar{T}$ ↓ (lower is better) result.

	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

the best *hit* in the open-source models as it performs well on scene and camera movement. Object category, scene, OCR, and style seem simple for these powerful models, as they all achieve well on the F1-score. When it comes to the dimensions of object number, object color, spatial relation, style, character identification, and events, all of them show relatively high *precision* but low *hit*, which means they can describe these elements well when they are confident about them, but might miss some instances and ignore the thoroughness. As for the action and two camera-related dimensions, all models perform unsatisfactorily on both *precision* and *hit*. This phenomenon inspires researchers to focus more on these aspects of the model's capability. We acknowledge that this finding highlights well-known open challenges for current MLLMs. For example, accurate object counting requires fine-grained discrimination; some recent methods [60, 61] leverage segmentation or external tools to assist reasoning. Camera-related tasks demand a strong visual reasoning ability to infer subtle contextual cues. Character identification relies heavily on high-quality training data to build related knowledge. Action recognition is inherently challenging due to the temporal modeling.

Precision and hit of 7B models. The *precision* and *hit* of 7B models are shown in Tab. 4. Among these 6 models, Qwen2.5VL-7B achieves the best *precision* (75.5%) and *hit* (48.1%), demonstrating its awesome ability. Averagely, the 7B models perform a bit worse than 72B models, verifying the scaling law. Among the dimensions, they follow a similar pattern as 72B and closed-source models. Researchers should focus on the thoroughness of object number, object color, spatial relation, style, character identification, and event, and try to improve the ability of the action and the two camera-related dimensions.

QA-based evaluation and the $K\bar{T}$ metric. Due to the visual element annotation type, we can easily convert our annotations to QA format, evaluate the accuracy of closed-source APIs and 72B models, thus further calculating their $K\bar{T}$ metric, as shown in Tab. 5. We are surprised to find that the difference in their QA accuracy is not significant, which means they can have a similar level of understanding of the correct visual descriptions based on the specific questions. When it comes to the $K\bar{T}$ metric, these models vary a lot. Gemini-1.5-pro performs the best with the lowest $K\bar{T}$ (25.7%), but the $K\bar{T}$ of LLaVA-OneVision-72B, Qwen2VL-72B, and InternVL2.5-78B comes to more than 40%, which means they are more likely knowing the answer but cannot tell automatically.

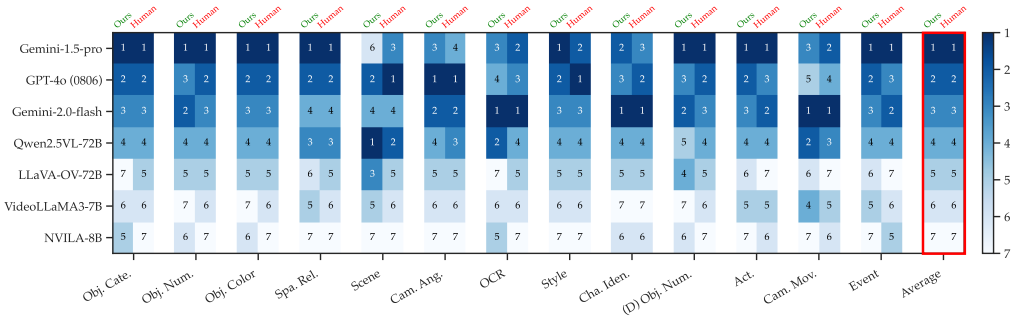

Figure 7: The metric ranking consistency comparison with human arena evaluation.

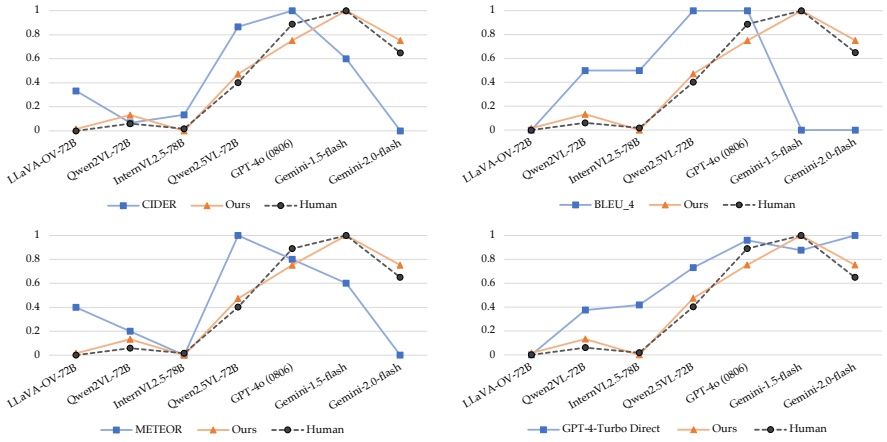

Figure 8: Consistency comparison with other metrics, we calculate CIDER, BLEU_4, METEOR metrics, and conduct direct GPT-4-Turbo evaluation. Our evaluation shows the highest consistency.

This phenomenon shows the performance gap between the strong instruction (QA) task and the weak instruction (caption) task, which may be ignored by previous work.

4.3 Metric Analysis

Consistency with human evaluation. To demonstrate the reliability of our annotations and evaluation pipeline, we build a human arena to manually compare the caption performance of closed-source models and 72B open-sourced models for each dimension. Specifically, we randomly select 5K pair-wise caption samples for each dimension and require human annotators to select a better one each time. The judging criteria include both the correctness and thoroughness for each dimension. We count the model ranking of each dimension, and compare it with the harmonic mean of *precision* and *hit*. The results are shown in Fig. 7. On the scene dimension, scores of all models are high and close in our evaluation, which may cause slight inconsistencies. But for most dimensions, our ranking shows high consistency with human results, leading to high reliability of final evaluation results.

Comparison with other metrics. We compare our metrics with other traditional metrics to demonstrate the credibility of our CAPability. We randomly select 200 images or videos for each dimension, and ask human annotators to annotate truthful and detailed captions as the ground truth. Then we calculate traditional metrics (*i.e.*, CIDER [26], BLEU_4 [25], and METEOR [48]) between the human-annotated and model-generated captions. Besides, we also use GPT-4-Turbo [58] to directly score the captions from 0 to 5 based on the ground truth. We further randomly select 200 inference samples for each dimension and each model, and manually score them to get human evaluation results as a reference. We report the harmonic mean for CAPability and linearly scale all results to 0-1 for easier comparison. The results are shown in Fig. 8. As our result shows high consistency with human evaluation, all traditional metrics fail to align. The phenomenon shows that existing traditional metrics cannot reliably and reasonably evaluate the detailed visual captions. The direct GPT4-Turbo evaluation shows better consistency than traditional metrics, but still has gaps with CAPability, lacks

https://doi.org/10.52202/085713-2211

reasonable interpretability, and is too vague for evaluation. Our CAPability can give scores from various aspects with clear judging criteria, thus obtaining an accurate and comprehensive evaluation.

5 Conclusion

In this work, we present CAPability, the first comprehensive visual caption benchmark through 6-view and 12-dimensional analysis. Unlike existing benchmarks that rely on oversimplified metrics or limited visual elements, CAPability introduces a correctness-thoroughness dual evaluation framework based on *precision*, *hit*, and $K\bar{T}$. Through this meticulous evaluation process, we uncover specific areas needing improvement across leading models, such as their gap between *precision* and *hit*, the challenges in aspects like camera angle detection, character identification, and action recognition. We also indicate the "know but cannot tell" phenomenon of MLLMs, which may be ignored by previous work. We believe that CAPability will play a pivotal role in advancing the field of visual captioning by encouraging the development of models that holistically understand and describe visual content. We open-source all our annotated data to facilitate future research.

Acknowledgment

This work is supported by the National Nature Science Foundation of China (62425114, 62121002, U23B2028, 62232006). We thank Alibaba Group for providing human annotation resources, closed-source model APIs support, and GPU computing resources. We thank all outsourcing annotators' efforts, and acknowledge the support of the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC, and USTC super-computing center for providing computational resources for this project.

References

[1] Kaixun Jiang, Zhaoyu Chen, Hao Huang, Jiafeng Wang, Dingkang Yang, Bo Li, Yan Wang, and Wenqiang Zhang. Efficient decision-based black-box patch attacks on video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4379–4389, 2023.

[2] Zhihang Liu, Jun Li, Hongtao Xie, Pandeng Li, Jiannan Ge, Sun-Ao Liu, and Guoqing Jin. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 3855–3863, 2024.

[3] Kaixun Jiang, Zhaoyu Chen, Jiyuan Fu, Lingyi Hong, Jinglun Li, and Wenqiang Zhang. Videopure: Diffusion-based adversarial purification for video recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[4] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

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

[6] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[7] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[8] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024.

[9] Yujie Wei, Shiwei Zhang, Hangjie Yuan, Biao Gong, Longxiang Tang, Xiang Wang, Haonan Qiu, Hengjia Li, Shuai Tan, Yingya Zhang, et al. Dreamrelation: Relation-centric video customization. *arXiv preprint arXiv:2503.07602*, 2025.

[10] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[11] Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal chain of thought: Long-video understanding by thinking in frames. *arXiv preprint arXiv:2507.02001*, 2025.

[12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[13] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

[14] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[15] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591, 2019.

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2023.

[17] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[18] OpenAI. Gpt-4v(ision) system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`, 2023.

[19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[20] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[21] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[22] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024.

[23] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[24] Qwen Team. Qwen2.5-vl. `https://qwenlm.github.io/blog/qwen2.5-vl/`, January 2025.

[25] K Papinesi. Bleu: A method for automatic evaluation of machine translation. In *Proc. 40th Actual Meeting of the Association for Computational Linguistics (ACL), 2002*, pages 311–318, 2002.

[26] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[27] Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. *arXiv preprint arXiv:2412.08614*, 2024.

[28] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024.

[29] Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024.

[30] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jeng-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024.

 https://doi.org/10.52202/085713-2211

[31] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

[32] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[33] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024.

[34] OpenAI. Gpt-4o(mini) system card. `https://openai.com/index/hello-gpt-4o/`, 2024.

[35] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[36] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. `https://lmsys.org/blog/2023-03-30-vicuna/`, March 2023.

[37] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[38] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[39] Zhihang Liu, Chen-Wei Xie, Pandeng Li, Liming Zhao, Longxiang Tang, Yun Zheng, Chuanbin Liu, and Hongtao Xie. Hybrid-level instruction injection for video token compression in multi-modal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8568–8578, 2025.

[40] Xiaoyi Bao, Siyang Sun, Shuailei Ma, Kecheng Zheng, Yuxin Guo, Guosheng Zhao, Yun Zheng, and Xingang Wang. Cores: Orchestrating the dance of reasoning and segmentation. In *European Conference on Computer Vision*, pages 187–204. Springer, 2024.

[41] Xiaoyi Bao, Chenwei Xie, Hao Tang, Tingyu Weng, Xiaofeng Wang, Yun Zheng, and Xingang Wang. Dynimg: Key frames with visual prompts are good representation for multi-modal video understanding. *arXiv preprint arXiv:2507.15569*, 2025.

[42] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024.

[43] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[44] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`, January 2024.

[45] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[46] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`, April 2024.

[47] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[48] Satanjeev Banerjee and Alon Lavie. 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, 2005.

[49] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024.

[50] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[51] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

[52] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022.

[53] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013.

[54] Chufeng Tang, Lingxi Xie, Xiaopeng Zhang, Xiaolin Hu, and Qi Tian. Visual recognition by request. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15265–15274, 2023.

[55] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023.

[56] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. In *European Conference on Computer Vision*, pages 471–490. Springer, 2024.

[57] Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, et al. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*, 2022.

[58] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[59] Google Deepmind. Gemini 2.0 is now available to everyone. `https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/`, February 2025.

[60] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

[61] Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Vlm-r3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *arXiv preprint arXiv:2505.16192*, 2025.

[62] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[63] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[64] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[65] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

 https://doi.org/10.52202/085713-2211

[66] Kent Allen, Madeline M Berry, Fred U Luehrs Jr, and James W Perry. Machine literature searching viii. operational criteria for designing information retrieval systems. *American Documentation (pre-1986)*, 6(2):93, 1955.

[67] Nancy Chinchor. Appendix b: Muc-7 test scores introduction. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*, 1998.

[68] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

[69] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 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, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims presented in the abstract and introduction accurately reflect the contributions and scope of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations in the appendix performed by the authors.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/085713-2211

Justification: During the inference process, we set the temperature parameter of the MLLMs to 0 to eliminate randomness and ensure the validity of the theoretical results derived. The proposed metrics are discussed with proper reference in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We introduce the detailed steps to obtain experiment results in the paper and explain the implementation details in the appendix. We provide the code and benchmark data to ensure the result can be reproducible. We also provide the inference and evaluation JSON files in the code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have provided open-accessed code and data link in the submission.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed specifics are provided in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the results of repeating the experiment several times in the appendix to verify the evaluation stability.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

 https://doi.org/10.52202/085713-2211

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the computing resource we use in the appendix, though it is not the matter we care about for proposing a benchmark.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper adheres to the NeurIPS Code of Ethics. All experimental procedures, data handling, and reporting practices were conducted with full compliance to ethical standards, ensuring transparency, integrity, and respect for all stakeholders involved

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We highlight some findings based on our evaluation results, which inspire the researchers. We also discuss the potential positive and negative societal impacts in the appendix.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: As mentioned in the paper, we conduct human filtering when annotating every image and video, ensuring there are no unsafe images or videos released.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper provides appropriate credits and references to the creators or original owners of the assets used. We centrally state the licenses and copyright in the appendix.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

https://doi.org/10.52202/085713-2211

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have made our benchmark data publicly available, and the link and croissant file are provided in the submission.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We ask outsourcing annotators to complete the benchmark construction, but it does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use LLMs for editing typos and formatting, and use LLMs for pre-annotation, followed by human annotation. In our evaluation, we also employ LLM-as-judge evaluation. However, the core motivation, method, and development do not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-2211

Appendix

Overview

A More Details of CAPability

A.1 Details of Dimension Design

We argue that multi-dimensional evaluation is significant to visual caption evaluation and is more comprehensive than previous work. So how to choose proper dimensions? We refer to existing VQA benchmarks [62, 63, 64, 65] and visual generation benchmarks [31, 32, 33]. VQA benchmarks usually design various types of questions to include multi-dimensional evaluation and analysis of MLLMs. For instance, MMBench [64] defines 20 ability dimensions, including attribute recognition, attribute comparison, action recognition, spatial relationship, physical property, OCR, object localization, image style, image scene, identity reasoning, *etc*. MVBench [64] covers 20 challenging video tasks including action, object, position, count, scene, pose, attribute, character, cognition, *etc*. Due to the flexible design of questions, VQA benchmarks can be naturally built with comprehensive dimensions. Different from the VQA task, the visual caption task does not require specific questions, but inspects the alignment of visual and textual information. Visual generation is the inverse task of visual captioning, as it requires models to generate specific visual content based on detailed textual descriptions. GenEval [31] designs 6 different tasks to evaluate text-to-image alignment, including single object, two object, counting, colors, position, and attribute binding. VBench [32] comprises 16 dimensions, including subject consistency, background consistency, object class, human action, color, spatial relationship, scene, style, *etc*. We follow their explored dimensions to design proper dimensions for visual captioning. Finally, we design 6 views, covering object, global, text, camera, temporal, and knowledge. The object-related view includes object category, object color, object

https://doi.org/10.52202/085713-2211

number, and spatial relation, the global-related view includes scene and style, the text-related view evaluates the OCR capability of captions, the camera-related view covers the camera angle and movement, the temporal-related view contains action and event, and we also design a view to evaluate the knowledge of MLLMs, *i.e.*, character identification.

We believe these dimensions contribute to a comprehensive visual caption benchmarking. However, it is undeniable that there may still be other dimensions that also contribute to caption evaluation. This phenomenon exists in all multi-dimensional benchmarks, but the purpose of our design is to find dimensions that are as comprehensive as possible and sufficient to differentiate model capabilities. As we design 12 dimensions, the evaluation is strong enough to evaluate models from various aspects. We also welcome the proposal of more constructive dimensions.

We explain each dimension in detail about what it represents here.

- **Object category.** This dimension measures the ability of whether models can give a correct description of a specific object in the image. The object is randomly selected from the image.
- **Object number.** Given a kind of randomly selected object existing in several numbers in an image or a video, this dimension measures the ability of whether models can count the object correctly. For videos, models should watch the whole video and dynamically count the number based on the camera.
- **Object color.** Given a kind of randomly selected object in an image, this dimension measures the ability of whether models can correctly describe the color.
- **Spatial relation.** Given two nearby objects in an image, this dimension measures the ability of whether models can correctly describe the spatial relationship of the two objects. We sample 500 images from our collected data, and sample another 500 images from CompreCap [27], with their spatial relationship descriptions.
- **Scene.** Given an image, this dimension measures the ability of whether models can obtain and tell the global scene of the image correctly.
- **Camera angle.** Given an image, this dimension measures the ability of whether models can obtain and tell the camera angle correctly when shooting the image.
- **OCR.** Given an image, this dimension measures the ability of whether models can recognize and tell the text appearing in the image correctly.
- **Style.** Given an image, this dimension measures the ability of whether models can obtain and tell the global style of the image correctly.
- **Character identification.** Given an image, this dimension measures the ability of whether models can recognize the character or the person in the image.
- **Action.** Given a video, this dimension measures the ability of whether models can recognize the action in the video. We use the video data of Dream-1K [28] and re-annotate the action from their annotations.
- **Camera movement.** Given a video, this dimension measures the ability of whether models can obtain and tell the camera angle correctly when recording the video. We search videos by ourselves and cut them into short clips, filtering complex movement composition. We only have simple camera movement in our data, but existing models still perform unsatisfactorily.
- **Event.** Given a video, this dimension measures the ability of whether models can tell a complete event in the video. We refer Dream-1K [28] to design this dimension, and we extract the events from their annotations. Different from other dimensions with atom-level elements, the event is usually composed of subjects and actions, which measures the temporal summarization ability of the model.

A.2 Explanation for One Represents All Strategy

"One represents all" is designed for object selection for object-related dimensions, text for OCR dimension, and action. We further build other dimensions related to attributes and relationships of objects (object color, object number, spatial relation) based on the selected objects. By aggregating results over a large number of samples with random pairing, we achieve statistical coverage across a

 https://doi.org/10.52202/085713-2211

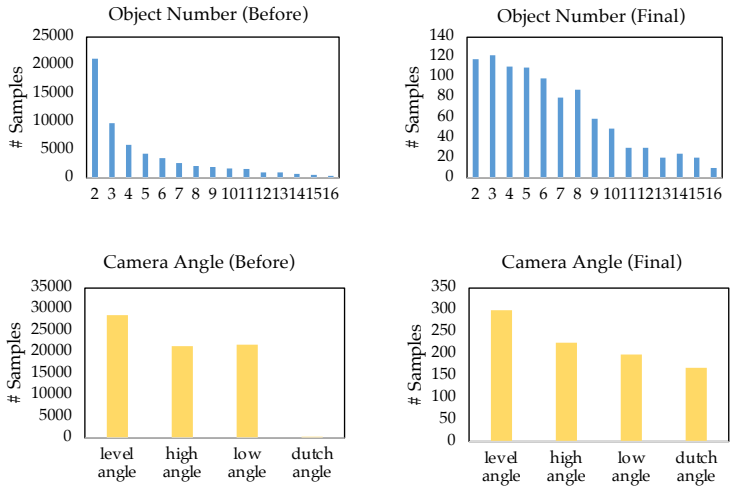

Figure A1: Two examples (object number and camera angle) of data distribution before data balance (pre-annotated) and after data-balanced selection (final human-annotated).

broad spectrum of relationships, granularities, and contexts. Therefore, our design ensures that both single-object properties and inter-object relationships are robustly evaluated at the benchmark level, while providing a practical and scalable way to balance annotation cost and comprehensive coverage.

We focus on keeping the randomness of element selection, thus covering the whole visual content in a statistical sense, based on the law of large numbers. Therefore, we can get the ability to evaluate the thoroughness of the generated captions by calculating the hit. The details about our efforts to ensure the randomness are as follows. 1) To minimize bias and keep randomness in the pre-annotation stage, we utilize multiple SOTA models (GPT-4o, Gemini-1.5-pro, Qwen-VL-Max) to generate candidate objects and take the union of their outputs. We then use code (Python's random library) to randomly select one object from this union. 2) During manual annotation, as humans tend to select the most obvious objects in the image, annotators are not allowed to reselect or change the target object if the pre-annotation is incorrect, thus mitigating the bias; they are only permitted to judge correctness and filter out incorrect samples. This ensures that neither model nor human bias influences which element is selected for annotation, thus keeping randomness.

A.3 Details about Human Subjectivity Controlling

We acknowledge that differences in human annotator judgment can introduce subjectivity, particularly in interpreting synonyms or resolving ambiguous cases. To reduce this, annotators are required to flag any samples for which they have low confidence. All low-confidence samples are then independently annotated by two additional annotators, and the final decision is determined by majority vote among the three annotations. We also have a spot check and verification process. For each dimension, we ask other annotators (or ourselves) to randomly select 20% of the samples for verification. If the annotation accuracy falls below 97%, we hold a meeting to highlight the incorrect annotations, revise all annotations for that dimension, and repeat this process until the annotation accuracy meets our requirements.

A.4 Details of Data Balance and Final Distribution

The purpose of data balance is to suppress the long-tail distribution, thus ensuring there are a certain number of samples of different difficulties in the benchmark. Fig. A1 shows two examples of the comparison of pre-annotated data distribution and final human-annotated data distribution. For object number and camera angle dimension, we first randomly sample nearly 75K samples and conduct model pre-annotation. The 75K samples consist of approximately 1/3 SA-1B [50], 1/3 COYO-100M [51], and 1/3 crawled from multiple public datasets and websites. For object number dimension, we select all samples with counting from 2 to 16 same objects within an image. As shown in Fig. A1 (left upper), the counting follows the long-tail distribution, there are fewer images

Table A1: The overlap among each dimension.

Obj. Cate.	Obj. Num.	Obj. Color	Spa. Rel.	Scene	Cam. Ang.	OCR	Style	Cha. Iden.	(D) Obj. Num.	Cam. Mov.	Act./Event
11.3%	10.8%	10.5%	2.5%	4.6%	4.6%	2.6%	0.5%	0%	0%	0%	0%

with more objects within an image. Therefore, we conduct data-balanced sampling. Specifically, we separately sample images for different counts, thus forcing the number of each count to be more balanced after the human correcting and filtering process. For camera angle dimension, the dutch angle data is rare, therefore we keep all dutch angle data, and sample the same number of the other three categories. After the human correcting and filtering, the number of these categories varies slightly. The situation in other dimensions is also similar.

As we consider the counts, categories, *etc*. For each dimension to conduct the data balance and not consider the data source (*i.e.*, from SA-1B, COYO-700M, or crawled by ourselves) during this process, the final data source distribution for dimensions of object category, object number, object color, scene, camera angle, OCR varies. For data of object-related dimensions (object category, object number, object color, spatial relation), global-related dimensions (scene and style), and a small part of camera angle and OCR dimensions, we collect the base data in a hybrid way. Our approach involves first performing pre-annotation for these dimensions on a large pool P_{all} of images (100K). We then filter out those samples that are not suitable for the corresponding dimension independently, donated as P_i^{pre} (nearly 40K - 80K), where i represents each dimension. After that, we conduct balanced sampling and manual annotation, resulting in approximately 1K samples per dimension, donated as P_i^{final}. For spatial relation, we directly choose 1/2 CompreCap [27] and 1/2 SA-1B images as SA-1B is more likely to contain high-resolution images with complex object relationship scenes. For style, we choose all realistic images from SA-1B, and crawl animated, special effect, and old-fashioned images by ourselves, all art-related images are from Wikipaintings [53]. For character identification, we use all images from the public dataset, *i.e.*, Wukong-100M [52] rather than crawling to ensure proper copyright. For dynamic object number, we directly use data from VSI-Bench [49]. For action and event, we directly use videos from Dream-1K [28]. We crawl all videos for camera movement dimension by ourselves, as there is little camera movement data in existing datasets. While there may be some overlap among pre-annotated P_i^{pre}, the average of overlapping samples in the final P_i^{final} is 3.95%, as shown in Tab. A1. We thank all public datasets and benchmarks, their excellent images, videos, and annotations provide much convenience for building our CAPability.

A.5 Details about Annotators

We employed 24 outsourcing annotators to complete the annotation task, including 14 female annotators and 10 male annotators. Their ages range from 24 to 30 years old, and all are from mainland China. Each annotator holds at least an associate's or bachelor's degree, demonstrating their expertise level. In total, the annotators labeled 23,824 samples, averaging approximately 993 samples per person. Among these, 3,323 samples were annotated by three different annotators, and the final result was determined through voting.

We specifically instructed annotators on how to handle such cases during the annotation process for each dimension. For example, in the color dimension, if the main object is primarily one color with minor secondary colors (such as a white airplane with some painted markings), only the primary color should be annotated. If the object exhibits multiple primary colors, annotators were asked to label up to three colors if the object can be clearly described within this limit; otherwise, the sample was filtered out. As a result, each object in our dataset is annotated with at most three colors.

A.6 Ethical Impact

The geographic and cultural backgrounds of the data are diverse. For data obtained from other public datasets, the distribution of potential bias in our benchmark largely follows that of the original datasets, as we performed random sampling. For the data we collected ourselves, there is a certain degree of cultural bias towards mainland China and East Asian cultures, although mainstream Western

https://doi.org/10.52202/085713-2211

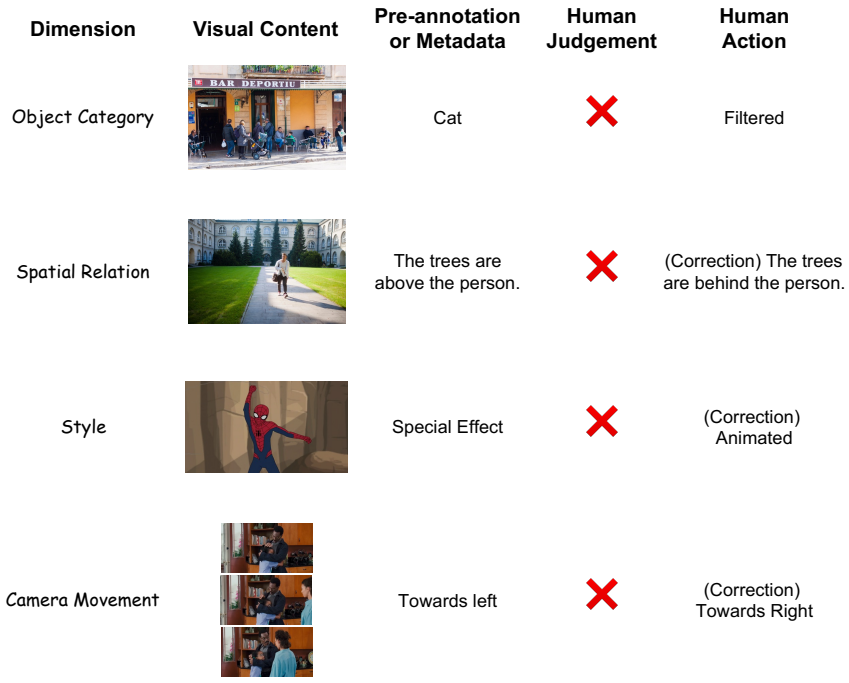

Figure A2: Bad case examples of pre-annotation or metadata, human annotators filter or correct the wrong pre-annotation or metadata one by one.

cultures are also represented to some extent. However, coverage of cultures from other regions is relatively limited.

A.7 Discussion about the Designed Metrics

We define two metrics, *precision* and *hit*, to evaluate both the correctness and thoroughness. The *hit* is similar to the meaning of *recall*, which measures how many visual elements in the image/video can be described correctly. However, the *recall* is usually related to the TP/FP/FN/TN system [66], which is inconsistent with our evaluation situations. To avoid ambiguity and misunderstanding, we name our metric as *hit* rather than *recall*.

We give the analysis from the TP/FP/FN/TN perspective here. The inconsistency comes from the conflict of the situation definition. In our pipeline, there are three situations when evaluating: 1) MIS, 2) COR, 3) INC, but the TP/FP/FN/TN system does not define the situation of MIS. In the TP/FP/FN/TN system, the *precision* and *recall* are defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{A1}$$

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{A2}$$

The existence of MIS leads to the ambiguity of the definition of FN and FP. If there is no existing MIS, we can only calculate the accuracy without considering *precision* or *recall*. Now we try to analyze with MIS. Based on the definition of TP, we can map our COR to TP with no doubt, as it correctly predicts the answer. If we want to calculate the *precision*, we can map our INC to FP, as it wrongly predicts the answer. Therefore, the precision can be calculated by Eq. 1 or Eq. A1, which is consistent. When we consider recall, $TP+FN$ should be the number of all ground truths, as there are no negative samples (TN) in our annotation, which means $TP+FN = S(\text{COR}) \cup S(\text{INC}) \cup S(\text{MIS})$. This leads to S(INC) being included in both FP and FN. As the TP/FP/FN/TN system does not define the MIS, the TP/FP/FN/TN-based *precision* and *recall* cause contradiction. Therefore, we name our metric of $S(\text{COR})/(S(\text{COR}) \cup S(\text{INC}) \cup S(\text{MIS}))$ as *hit* rather than *recall* to avoid ambiguity as it does not fit the TP/FP/FN/TN-based definition.

https://doi.org/10.52202/085713-2211

Figure A3: Examples of visual content and annotations for each dimension. We outline some visual elements by the red box in the image or video to make them easier to identify.

In addition to the TP/FP/FN/TN system, there are also other ways to define *precision* and *recall*. MUC-7 [67] defines the *precision* and *recall* with the situation of MIS. Apart from COR, INC, MIS, which own the same meaning as ours, MUC-7 also defines SPU, which represents the number of spurious, and equals 0 in our situation. MUC-7 defines the *precision* and *recall* as follows:

$$\text{Precision} = \frac{S(\text{COR})}{S(\text{COR}) + S(\text{INC}) + S(\text{SPU})}, \tag{A3}$$

$$\text{Recall} = \frac{S(\text{COR})}{S(\text{COR}) + S(\text{INC}) + S(\text{MIS})}. \tag{A4}$$

Based on this kind of definition, our *hit* equals the "*recall*".

However, as the TP/FP/FN/TN system is too famous and standard to define the *precision* and *recall*, we finally decide to use the "*hit*" rather than "*recall*" to avoid misunderstanding.

A.8 Benchmark Examples

Examples of human annotation process. We show some visual cases of the human annotation process in Fig. A2. All of examples are with wrong pre-annotation or metadata. Human annotators check the pre-annotation/metadata one by one and filter/correct the mistakes for each dimension.

https://doi.org/10.52202/085713-2211

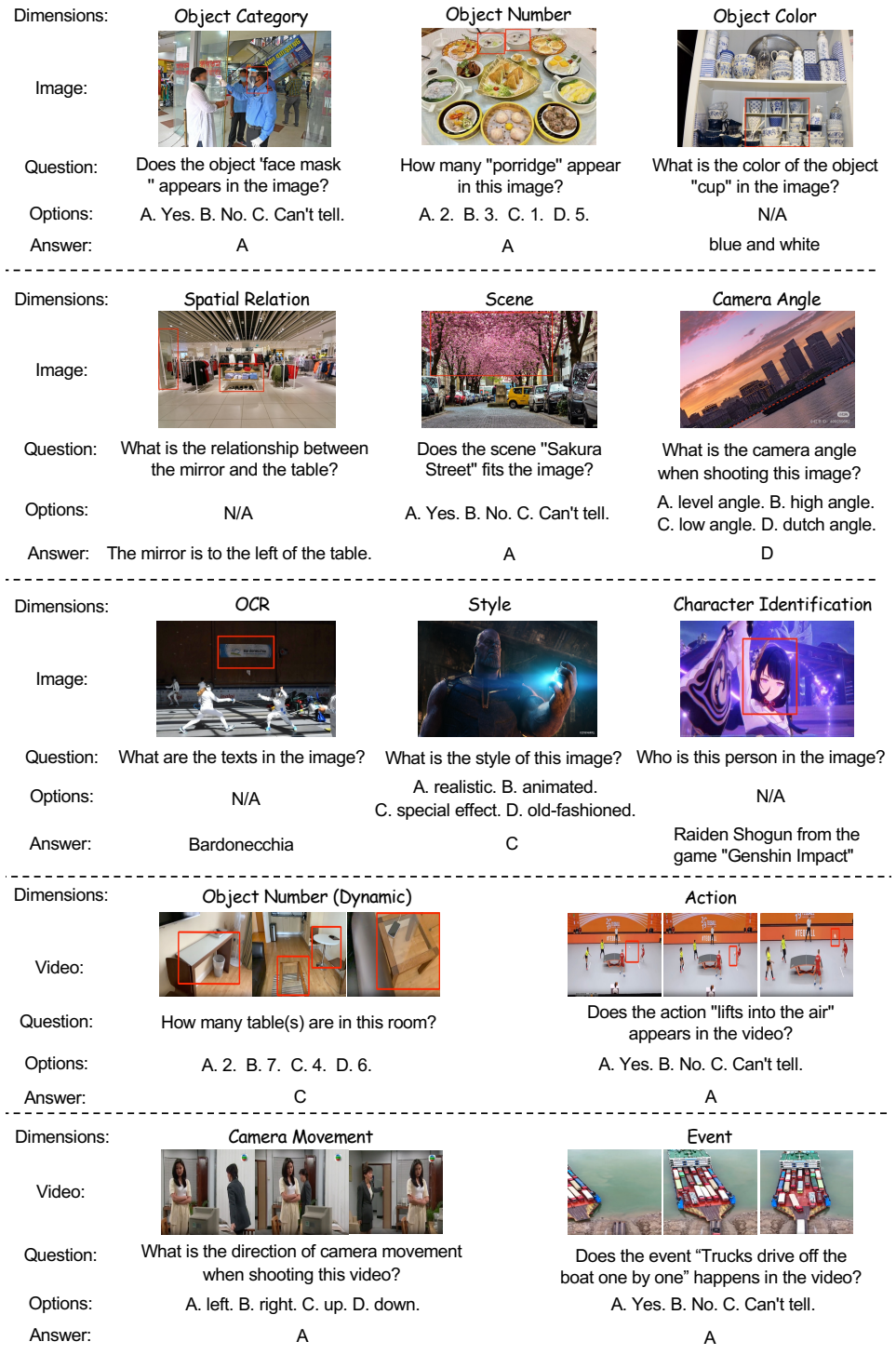

Figure A4: Examples of visual content and converted QA annotations for each dimension. The visual content is the same as Fig. A3. We outline some visual elements by the red box in the image or video to make them easier to identify.

Examples of annotations. We show some visual cases with our annotations in Fig. A3. We outline some visual elements by the red box in the image or video to make them easier to identify. We collect our data from various sources, and we crawled some visual content from the Internet by ourselves, ensuring diversity.

Image

```
IMAGE_PROMPT = "Please describe the image in detail. Your description should follow these rules:\n"\
"a) You should describe each object in the image in detail, including its name, number, color, and spatial relationship between objects.\n"\
"b) You should describe the scene of the image.\n"\
"c) You should describe the camera angle when shooting this image, such as level angle, high angle, low angle, or dutch angle.\n"\
"d) You should describe the style of the image, such as realistic, animated, special-effect, old-fashioned and so on.\n"\
"e) If there are any texts in the image, you should describe the text content.\n"\
"f) If you know the character in the image, you should tell his or her name.\n"\
"Directly output your detailed description in a elaborate paragraph, instead of itemizing them in list form. Your description: "
```

Video

```
VIDEO_PROMPT = "Please describe the video in detail. Your description should follow these rules:\n"\
"a) You should describe each events in the video in order, especially focusing on the behavior and action of characters, including people, animals.\n"\
"b) You should describe each object in the video in detail, including its name, number, color, and spatial relationship between objects.\n"\
"c) You should describe the scene of the video.\n"\
"d) You should describe the camera movement when shooting this video, especially the direction, such as pan left, track right, tilt up, boom down, zoom in, dolly out, and so on.\n"\
"e) You should describe the style of the video, such as realistic, animated, special-effect, old-fashioned and so on.\n"\
"f) If there are any texts in the video, you should describe the text content.\n"\
"g) If you know the character in the video, you should tell his or her name.\n"\
"Directly output your detailed description in a elaborate paragraph, instead of itemizing them in list form. Your description: "
```

Figure A5: The image prompt and video prompt for all models when inferring captions.

Examples of converted QA pairs. As we directly annotate the visual elements in the image or video rather than the caption sentence, we can easily convert our annotation into the format of question-answer (QA) pairs, and we name it as CAPability-QA. We use CAPability-QA to evaluate the QA accuracy and the *know but cannot tell* ($K\bar{T}$) metric. In Fig. A4, we also show the same visual cases as Fig. A3 for each dimension with converted QA format. Most of the dimensions are converted to the format of a multiple-choice QA task with several options, and the object color, OCR, and character identification dimensions are designed as open-ended QA tasks.

B More Experimental Analysis

B.1 Implementation Details

We use 4x80G GPUs to run all open-sourced model inference. We use `transformers` to deploy LLaVA-OneVision, InternVL2.5, VideoLLaMA3, and NVILA, use `vLLM` to deploy Qwen2VL and Qwen2.5VL, as their official repositories suggested. For all our evaluated model, we follow their official configurations to run the inference. We set the temperature of all open-source models to 0, while keeping the default for closed-source APIs. All maximum output token length is set to 8192. We list the configurations as follows.

LLaVA-OneVision. We set `anyres-max-9` for image, and uniformly sample 32 frames for video.

Qwen2VL and Qwen2.5VL. We keep the default minimum and maximum image pixels in package `qwen_vl_utils`, which is $4 * 28 * 28$, and $16384 * 28 * 28$, respectively. We also keep default video settings, the fps is set to 2.0, the maximum frames are 768, the minimum video pixel is $128 * 28 * 28$, and the maximum video pixel is $768 * 28 * 28$.

InternVL2.5. We use the official video and image process function and uniformly sample 32 frames for video.

VideoLLaMA3. We use image model for image dimensions and video model for video dimensions. The fps is set to 1, and the maximum frames are 180 for videos.

NVILA. We use the official image and video process function in `VILA` repository, and uniformly sample 8 frames for videos, as suggested in the official config.

GPT-4o. Due to the maximum frame number limits of GPT API, we uniformly sample 50 frames for videos, and keep the original spatial size of both images and videos, sending them to the API server.

Gemini-1.5-pro and Gemini-2.0-flash. As Gemini API supports video, we directly send the original image and video to the API server. For very few videos with too large file size, we downsample the fps to 3, and send the downsampled video to the API server for connection stability.

https://doi.org/10.52202/085713-2211

```
Object Number
object_number_user_prompt = "Given an image caption and the number of an object with format {object: number} as follows:\n"\
f"Image Caption: {caption}\n"\
f"Object Number: {{{object_category}: {object_number}}}\n"\
f"Please analyze the image caption. Determine whether the provided object number is correctly described in the caption, and explain why. You may need to count in the caption to
determine how many the provided objects it describes.\n"\
"Give score of 0 if the caption does not mention the specific number of provided object (including the use of words such as 'some' and 'various' in the caption rather than giving specific
numbers) or not mention the provided object. Give score of 1 if the caption describes the object number correctly. Give score only of -1 if the caption gives the wrong number.\n"\
"Output a JSON formed as:\n"\
"{\"object_number\": \"copy the provided {object: number} here\", \"score\": \"put your score here\", \"reason\": \"give your reason here\"}\n"\
"DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only output the JSON. Do not add Markdown syntax. Output:"

Camera Movement
camera_movement_category_explains = [
"left: the camera angle swings left (pan left), or the camera moves left (track left)",
"right: the camera angle swings right (pan right), or the camera moves right (track right)",
"up: the camera angle swings up (tilt up), or the camera moves up (boom up)",
"down: the camera angle swings down (tilt down), or the camera moves down (boom down)",
"in: camera pushes toward the subject (dolly in), or enlarges the frame (zoom in)",
"out: camera moves away the subject (dolly out), or expands the visible area, makeing the subject appear smaller (zoom out)",
"fixed: camera is almost fixed and does not change",
]
camera_movement_categories = [c.split(":")[0] for c in camera_movement_category_explains]
camera_movement_user_prompt = "Given a video caption, your task is to determine which kind of camera movement is included in the caption.\n"\
f"Video Caption: {caption}\n"\
f"Please analyze the video caption and classify the descriptions of camera movement into the following categories: {camera_movement_categories}\n"\
f"Here are the explanations of each category: " + '\n'.join(camera_movement_category_explains) + "\n"\
"If the caption explicitly mentions one of the above camera movement categories, write the result of the category into the 'pred' value of the json string. Note do not infer the camera
movement categories from the whole caption. You should only search the descriptions about the camera movement. If there is no description of the camera movement in the video
caption or the description does not belong to any of the above categories, write 'N/A' into the 'pred' value of the json string.\n"\
"Output a JSON formed as:\n"\
"{\"pred\": \"put your predicted category here\", \"reason\": \"give your reason here\"}\n"\
"DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only output the JSON. Do not add Markdown syntax. Output:"
```

Figure A6: Two prompt examples for different types of evaluation sub-tasks. The example of object number represents dimensions with open-ended descriptions, and the example of camera movement represents the dimensions with specific categories.

Table A2: The referring ratio of all models, which only reflects the referring ratio of each dimension without considering the accuracy.

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-7B	96.6	33.9	60.8	54.7	86.7	79.7	73.6	98.8	5.0	34.5	92.2	62.4	30.9	62.3
Qwen2VL-7B	97.6	30.2	56.8	51.8	88.7	98.0	77.0	99.9	4.8	20.0	94.0	72.0	31.7	63.3
NVILA-8B	97.0	34.3	64.9	52.7	85.3	85.1	74.5	98.1	7.4	22.1	80.6	48.6	21.5	59.4
InternVL2.5-8B	97.0	38.1	60.9	55.2	87.3	100.0	84.6	100.0	20.3	22.7	91.4	94.8	32.9	68.1
VideoLLaMA3-7B	95.1	34.0	62.5	56.4	85.6	93.1	74.7	98.5	5.0	9.2	96.5	83.1	34.5	63.7
Qwen2.5VL-7B	96.7	26.8	63.7	55.3	89.4	99.2	86.7	100.0	11.3	20.9	92.5	98.6	35.0	67.4
LLaVA-OV-72B	95.8	35.2	63.1	54.1	87.4	93.3	74.9	99.5	11.0	31.2	92.6	69.0	31.7	64.5
Qwen2VL-72B	97.4	35.8	64.3	56.9	89.4	99.6	83.0	100.0	6.8	21.6	93.5	77.1	34.7	66.2
InternVL2.5-78B	97.2	41.7	65.3	57.0	86.7	100.0	85.5	100.0	21.3	25.7	88.2	63.3	28.8	66.2
Qwen2.5VL-72B	95.6	43.3	69.2	62.3	90.7	100.0	91.8	100.0	31.4	24.4	94.8	99.4	38.9	72.5
GPT-4o (0806)	96.0	44.5	73.5	61.6	88.2	100.0	88.8	100.0	35.1	29.4	93.4	99.4	44.5	73.4
Gemini-1.5-pro	96.1	55.3	77.0	69.0	88.1	99.9	91.4	100.0	67.5	48.9	90.5	100.0	48.6	79.4
Gemini-2.0-flash	96.1	39.0	67.2	58.2	87.5	100.0	93.2	99.9	46.2	30.4	92.0	99.6	44.6	73.4

B.2 Prompts of Inference and Evaluation

Inference prompt. We use the same prompts for all models to produce the visual captions. The image prompt and video prompt are shown in Fig. A5. To decrease the inference difficulty, we prompt the models to output the information of all our designed dimensions with a detailed caption. Despite this, the models still show a huge difference in the hit rate of each dimension, which may be due to the variety of training data related to the caption.

Evaluation prompt. As we divide the evaluation of dimensions into two types: 1) dimensions with specific categories (*i.e.*, style, camera angle, and camera movement), 2) dimensions with open-ended descriptions. Therefore, we design two kinds of templates for evaluating, and fine-tune them within each dimension. In Fig. A6, we take the object number dimension and camera movement dimension as examples, to show our prompts for evaluation. For dimensions with specific categories, we ask GPT-4-Turbo to extract the key information and classify the caption into our pre-defined categories or the 'N/A' class. The correct classification is considered positive, the wrong one as negative, and the 'N/A' result is considered a miss. For dimensions with open-ended descriptions, we ask GPT-4-Turbo to directly compare the annotations and the caption, and give out the result of positive, negative, or miss with reasons.

B.3 More Experimental Results

Referring ratio among all models. Apart from the *precision* and *hit*, we can also report another metric, *referring ratio*, which represents the referring ratio about the dimension in visual caption and

Table A3: The average metric of image dimensions and video dimensions.

Models	LLaVA-OV-72B	Qwen2VL-72B	InternVL2.5-78B	Qwen2.5VL-72B	GPT-4o (0806)	Gemini-1.5-pro	Gemini-2.0-flash
image precision	81.4	82.2	78.2	80.7	84.4	81.1	84.6
video precision	59.5	62.9	55.6	65.0	67.6	68.7	67.3
image hit	55.3	57.5	57.9	62.1	65.2	67.7	64.5
video hit	26.9	29.2	22.4	33.7	36.8	43.9	37.6

Table A4: The PPL results of each generated caption by different models.

PPL models	LLaVA-OV-72B	Qwen2VL-72B	InternVL2.5-78B	Qwen2.5VL-72B	GPT-4o (0806)	Gemini-1.5-pro	Gemini-2.0-flash
Qwen3 PPL	3.56	4.60	5.13	5.01	8.64	8.45	8.16
LLaMA3.1 PPL	4.89	6.70	7.75	7.54	11.29	10.76	10.68

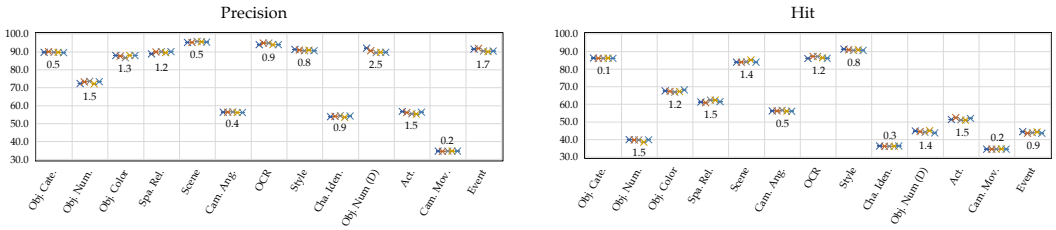

Figure A7: The evaluation of repeating 5 times for Gemini-1.5-pro captions. We tag the fluctuation range beside the data point.

can be calculated as:

$$\text{Referring Ratio} = \frac{|S(\text{COR}) \cup S(\text{INC})|}{|S(\text{ALL})|}. \tag{A5}$$

This metric only considers the pure thoroughness of the caption in each dimension without considering the accuracy.

We report the *referring ratio* in Tab. A2. For example, it is considered a reference if the caption mentions any object for the object category dimension, or mentions any angle information for the camera angle dimension, but for the object number or color dimension, it is only considered a reference if the caption mentions any number or color information of the correct object. We find the referring ratio seems to increase as the size of models increases, which may be due to more knowledge and stronger instruction following ability for larger models. Among all dimensions, the referring ratio of character identification performs the worst, the existing models prefer to keep silent as they usually cannot recognize the person and character well. The closed-source models would be more likely to reveal the names of characters, and we guess this may be due to stronger knowledge and more diverse training data.

Metric analysis between different modalities. Tab. A3 shows the metric analysis between image dimensions and video dimensions. Across all models, performance on image dimensions is substantially higher than on video dimensions, for both precision and hit metrics. Even the best-performing models (GPT-4o, Gemini series) show nearly 20% or greater drop in precision/hit from image to video dimensions. For example, all models perform well on object category, scene, OCR, and style for both precision and hit, but all models cannot achieve a satisfactory level on all video dimensions for hit. This shows that time series modeling and multi-frame information integration are still the main challenges of current MLLMs. It is true that some dimensions may be inherently more difficult to represent in video than in pictures. This may be due to the following two reasons: 1) There is more redundant content with more visual tokens input into MLLMs, handling the long sequence is likely more difficult than a shorter one. 2) Temporal change and movement should be considered for video dimensions, which may lead to confusion for models to recognize. Nevertheless, images can also represent the strength of videos in this dimension to a certain extent (such as OCR), because video understanding also requires a good spatial understanding ability as a prerequisite. The video dimensions we designed are more focused on the challenge of temporality. In the future, we will consider introducing more video data to fully evaluate the video's ability in the spatial dimensions.

The naturalness and coherence analysis. In the current landscape of MLLMs, the naturalness and coherence of generated captions are generally no longer a sufficient or even primary differentiator of

https://doi.org/10.52202/085713-2211

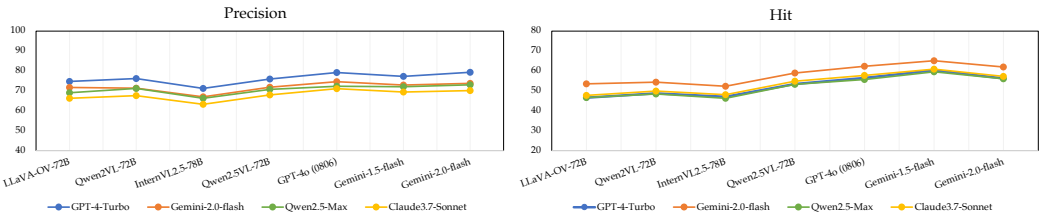

Figure A8: The stability analysis with three different evaluation models on 7 MLLMs' captions. The results on all metrics show a high degree of consistency.

model performance. Linguistic quality is more considered for NLP language models, and modern LLMs consistently generate grammatically correct and fluent sentences [68]. As MLLMs are built from LLMs, the main challenge and research focus of MLLMs has shifted toward evaluating the factual accuracy and relevance of the generated captions, rather than their linguistic quality. For example, the papers of MLLMs do not report any metrics about the naturalness and coherence, such as PPL [19, 24], modern benchmarks for MLLMs do not measure the naturalness and coherence either [69, 65]. To support the view that the naturalness and coherence of generated captions are not the primary challenges, we also further calculate perplexity (PPL) of the generated captions of each model. PPL is a fundamental metric for language models that gauges how "surprised" the model is by a given text, thus reflecting their naturalness and fluency, the lower PPL means better coherence. Specifically, we conduct qwen3-32B and LLaMA-3.1-8B-Instruct to calculate the PPL of generated captions from these models, as shown in Tab. A4. (There seems to be some gap between the closed-source APIs and open-source models. This may be because the training data distribution among open-source models is more similar to Qwen3 and LLaMA3.1 than closed-source APIs.) Among closed-source APIs, the PPLs of them are similar to or lower than GPT-4o. Among open-source models, the PPLs of them are similar to or lower than Qwen2.5VL-72B. Based on the common sense of GPT-4o and Qwen2.5VL-72B can generate coherent and human-like sentences, we can draw a conclusion that the naturalness and coherence are not the main challenge for all these models.

Evaluation stability. To validate the stability and robustness of our GPT-4-Turbo-based evaluation method, we take the inferred caption of Gemini-1.5-pro as the example, run our evaluation 5 times, and the result is shown in Fig. A7. We tag the fluctuation range, *i.e.*, the difference between the maximum and minimum scores, besides the data point. Fig. A7 shows our strong stability, and our average range of precision and recall are 1.1% and 1.0%, respectively. This demonstrates the reliability and interpretability of our evaluation method, which matches annotated elements in the generated captions. Moreover, we introduce three other models, Gemini-2.0-flash, Qwen2.5-Max, and Claude3.7-Sonnet to replace the GPT-4-Turbo in our evaluation pipeline. All of them are the most popular and powerful SOTA language models. We re-run the evaluation process with new evaluation models with the same evaluation prompts. The average result of evaluations for closed-source models and open-source 72B models is shown in Fig. A8. Though the judgment criteria for each model may be slightly different, leading to only a slight difference in the scores, the evaluation results show high consistency as they keep the same rank of these MLLMs as GPT-4-Turbo. This further demonstrates the high reliability, interpretability, and stability of our evaluation methods.

B.4 Visualization of Inference and Evaluation.

In Fig. A9 and Fig. A10, we visualize the inferred caption of Gemini-1.5-pro, GPT-4o (0806), and Qwen2.5VL-72B in object number dimension and camera angle dimension. In Fig. A9, the annotation of the given image is 7 pig trotters. Gemini-1.5-pro refers to the correct number of pig trotters, and we thus give it a positive. GPT-4o recognizes the trotters in the image, but counts with a wrong number, 6, and we thus give it a negative. As for Qwen2.5VL-72B, it says there are 7 chicken thighs in the image, recognizing the wrong object category. However, this is the dimension of the object number, and therefore we only evaluate the correctness and thoroughness of the number, without considering the categories of objects. As Qwen2.5VL-72B does not mention the pig trotters, we give it a miss. In Fig. A10, the annotation of the given image is dutch angle, which means the lens has a certain angle of deflection along the central axis, making the horizon crooked. Gemini-1.5-pro says the image is taken from a medium-high angle, and we classify it into the high angle category, thus

https://doi.org/10.52202/085713-2211

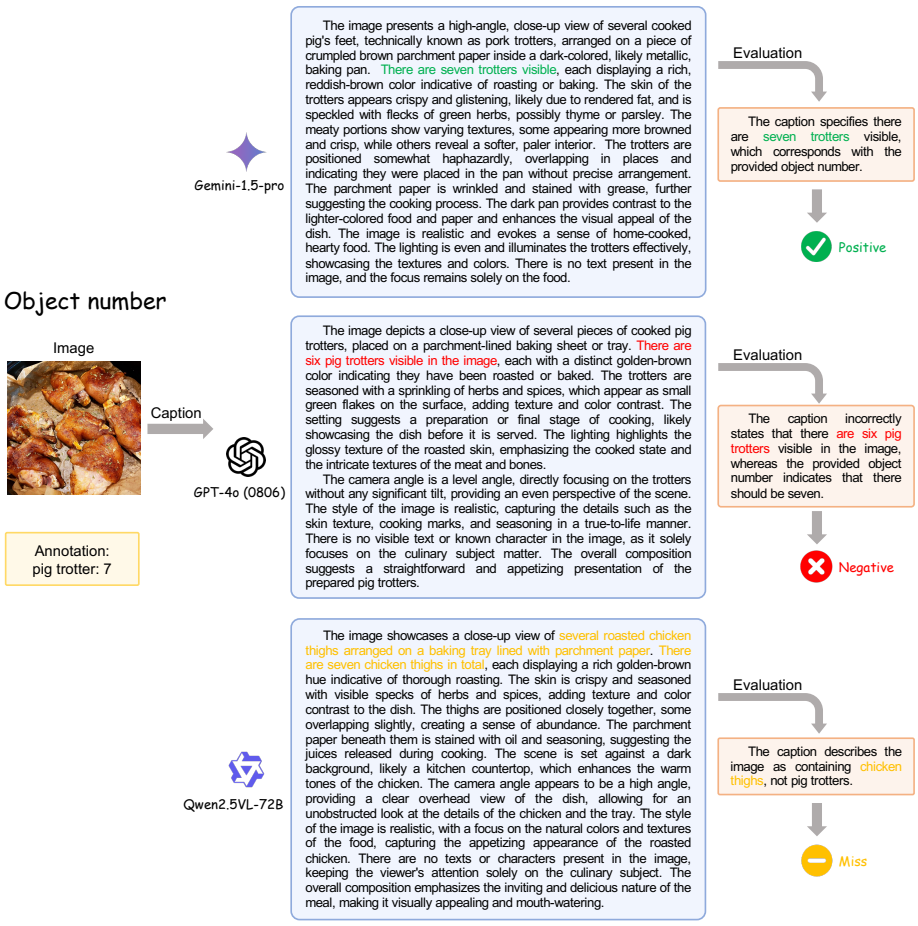

Figure A9: Examples of inference and evaluation on object number dimension. We select the inferred caption from Gemini-1.5-pro, GPT-4o, and Qwen2.5VL-72B as instances.

negative. GPT-4o explicitly points it out as a subtle dutch angle, thus is classified into the dutch angle category, which is positive. Qwen2.5VL-72B describes the image shot from a slightly elevated angle, and it appears to be a level angle, which is also negative. These two figures show our evaluation pipeline, which is precise and reliable.

C Copyright and License

CAPability comprises data from SA-1B, COYO-700M, Wukong-100M, Wikipaintings, VSI-Bench, CompreCap, Dream-1K, and uses some craweled data from inner retrieval system, each governed by its own licensing terms:

- **SA-1B**: SA-1B Dataset Research License[1]
- **COYO-100M**: CC-BY-4.0 License
- **Wukong-100M**: CC-BY-NC-SA-4.0 License
- **Wikipaintings**: BSD 2-clause License
- **VSI-Bench**: Apache-2.0 License
- **CompreCap**: CC-BY-4.0 License
- **Dream-1K**: Apache-2.0 License

[1]https://ai.meta.com/datasets/segment-anything-downloads/

https://doi.org/10.52202/085713-2211

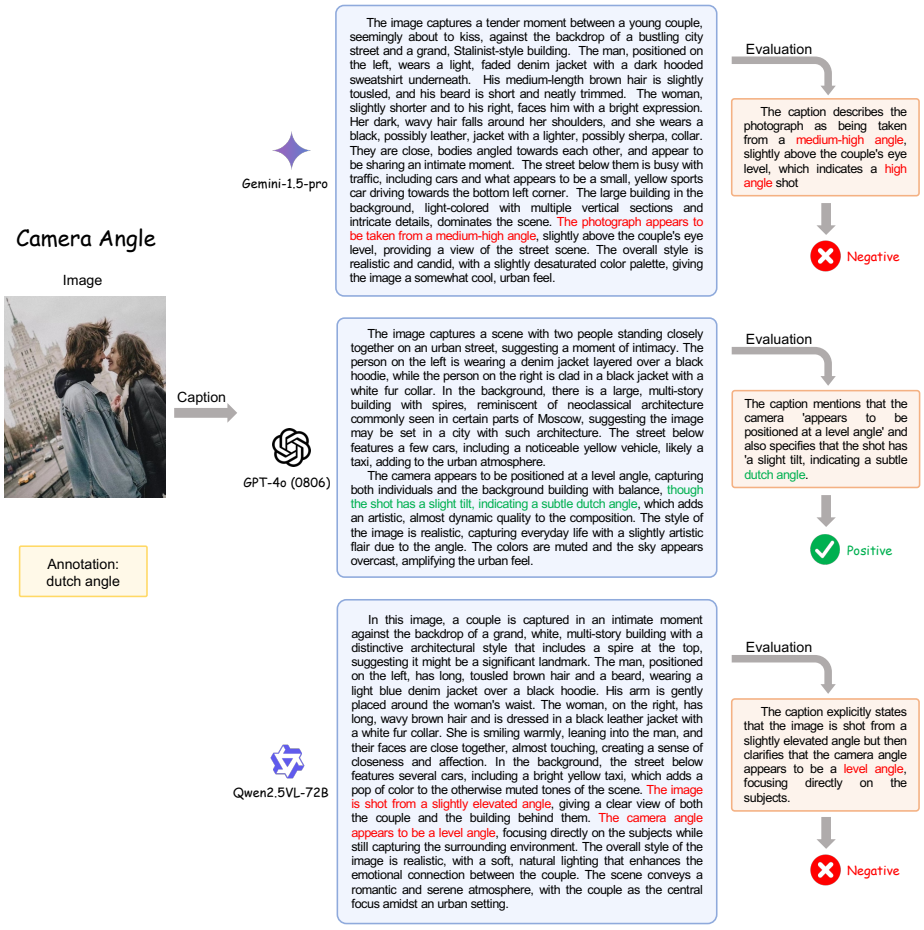

Figure A10: Examples of inference and evaluation on camera angle dimension. We select the inferred caption from Gemini-1.5-pro, GPT-4o, and Qwen2.5VL-72B as instances.

- **Craweled data**: Our craweled data are all retrieved from inner multi-modal retrieval system, which contains various public datasets, and visual contents from websites with CC0 license.

D Limitations

Different from multi-choice VQA benchmarks which evaluate models by definite and explicit choice accuracy, our CAPability is a visual caption benchmark, which still depends on LLMs for evaluation. Therefore, it is still hard to ensure a completely correct evaluation. We try to split the evaluation into several dimensions, thus makes the evaluation as simple and clear as possible. Therefore, the LLM-based evaluation can be more accurate. Due to the LLM language capability limitation, it can still make wrong judgments and requires a carefully designed prompt for constraint.

E Societal Impacts

As our proposed CAPability can perform comprehensive caption evaluations of MLLMs, this work can help LLM users make informed choice, and leads the community to build more and more strong MLLMs. The potential negative impacts are similar to other LLM-related works, The development of MLLMs and MLLMs' benchmarks pose societal risks like the perpetuation of biases, the potential for misinformation, job displacement, *etc*.