OMNICAPTIONER: One Captioner to Rule Them All

Yiting Lu[2,†], Jiakang Yuan[1,3,†], Zhen Li[4], Shitian Zhao[1], Qi Qin[1], Xinyue Li[1], Le Zhuo[1]
Licheng Wen[1], Dongyang Liu[1], Yuewen Cao[1], Xiangchao Yan[1], Xin Li[2], Tianshuo Peng[1,4]
Shufei Zhang[1], Botian Shi[1], Tao Chen[3], Zhibo Chen[2,✉], Lei Bai[1], Peng Gao[1], Bo Zhang[1,‡,✉]

[1] Shanghai Artificial Intelligence Laboratory, [2] University of Science and Technology of China,
[3] Fudan University, [4] The Chinese University of Hong Kong

https://alpha-innovator.github.io/OmniCaptioner-project-page
https://github.com/Alpha-Innovator/OmniCaptioner
https://huggingface.co/U4R/OmniCaptioner

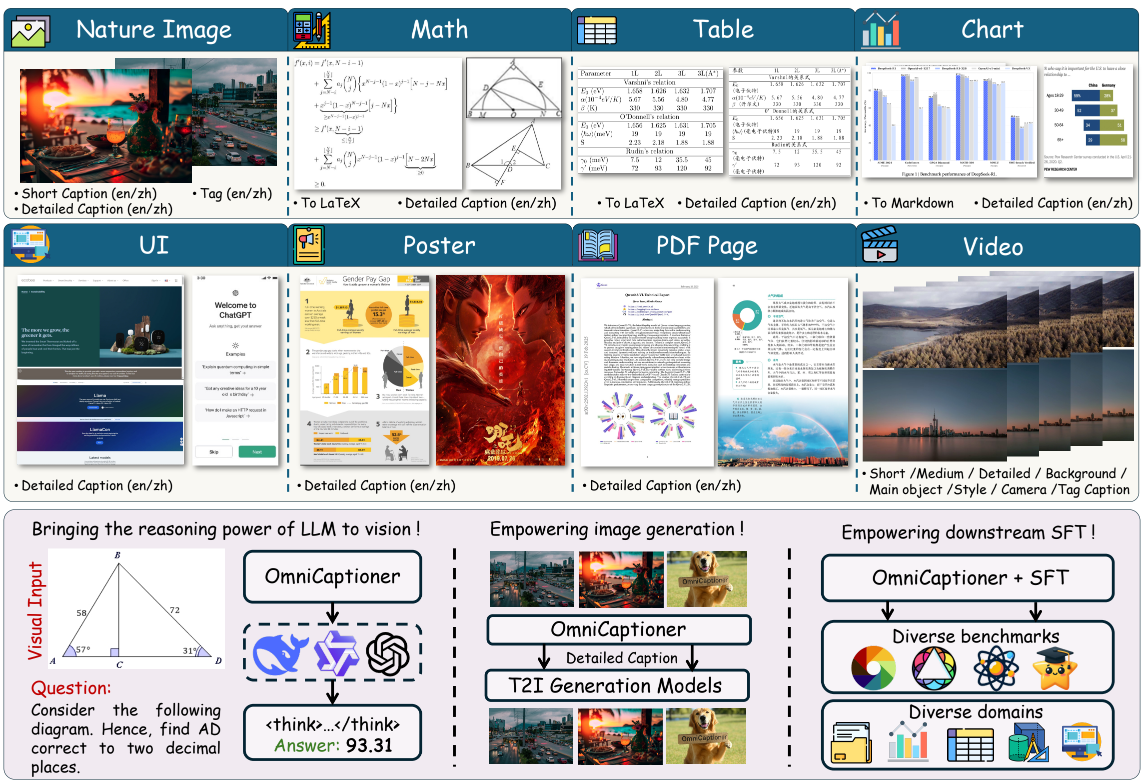


Figure 1: OMNICAPTIONER: the top section demonstrates its capability to process diverse visual domains. The bottom section highlights its applications in visual reasoning (associated with reasoning LLM), image generation (integrated with T2I generation models), and efficient downstream SFT tasks adaptation.

Abstract

† Equal contribution, ‡ Project Lead, ✉ Corresponding authors.

We propose OMNICAPTIONER, a versatile visual captioning framework for generating fine-grained textual descriptions across a wide variety of visual domains. Unlike prior methods limited to specific image types (*e.g.*, natural images or geometric visuals), our framework provides a unified solution for captioning natural images, visual text (*e.g.*, posters, UIs, textbooks), and structured visuals (*e.g.*, documents, tables, charts). By converting low-level pixel information into semantically rich textual representations, our framework bridges the gap between visual and textual modalities. Our results highlight three key advantages: (i) Enhanced Visual Reasoning with LLMs, where long-context captions of visual modalities empower LLMs, particularly the DeepSeek-R1 series, to reason effectively in multimodal scenarios; (ii) Improved Image Generation, where detailed captions improve tasks like text-to-image generation and image transformation; and (iii) Efficient Supervised Fine-Tuning (SFT), which enables faster convergence with less data. We believe the versatility and adaptability of OMNICAPTIONER can offer a new perspective for bridging the gap between language and visual modalities.

1 Introduction

Pretraining of Multimodal Large Language Models (MLLMs) (Liu et al., 2023; Li et al., 2024a; Chen et al., 2024d; Wang et al., 2024b; Bai et al., 2025), particularly in bridging the gap between visual and textual domains, has gained significant attention in recent years. Substantial progress has been achieved in image captioning and visual question answering, enabling models to serve as universal visual assistants through large-scale Supervised Fine-Tuning (SFT). However, MLLMs still face limitations in perceptual accuracy in the visual-text and structured image domains, particularly when handling synthesized images that exhibit a substantial domain gap from natural images, as illustrated in Fig. 3 (c).

Recent research has increasingly emphasized the role of image captioning in aligning modalities during multimodal pretraining, aiming to enhance both perception and reasoning across diverse domains through the SFT process. Meanwhile, domain-specific studies, such as those focusing on document understanding MLLMs (Luo et al., 2024a; Hu et al., 2024) and mathematical MLLMs (Peng et al., 2024; Zhang et al., 2025; Xia et al., 2024a), have leveraged domain-specific caption data to further improve modality alignment and advance multimodal pretraining. *These advancements highlight the need for a unified framework for multimodal pretraining centered on image captioning.* Also, despite progress in MLLMs, their multimodal reasoning capabilities still fall short of the textual reasoning abilities of LLMs. As shown in Fig. 2, when provided only with a question and no visual input on the MathVision and MathVerse benchmarks, DeepSeek-Distill-Qwen-7B (orange) significantly outperforms Qwen2-VL-Instruct (blue), demonstrating the strength of LLM-driven reasoning in multimodal tasks.

In this work, we bridge this gap by introducing the first OMNICAPTIONER framework, designed to generate fine-grained textual descriptions across diverse visual domains as shown in Fig. 1. Unlike prior approaches that focus on specific visual categories (*i.e.*, natural or geometry images), our approach enables a unified solution for diverse image types, paving the way for broader multimodal understanding. We focus on converting low-level pixel features into semantically rich textual representations, which preserve crucial visual details while bridging the modality gap between vision and language. OMNICAPTIONER has two characteristics: i) **Diverse Visual Domain Coverage**: We present a unified framework that supports diverse visual content, including natural images, visual text images (*e.g.*, poster, UI, textbook) and structured images (*e.g.*, geometry, equation, tables, charts). ii) **Pixel-to-Text Mapping**: By pairing these diverse image types with

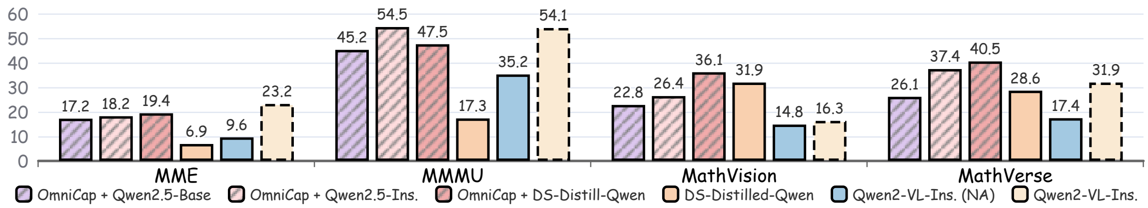


Figure 2: Performance comparison across different visual benchmarks for different LLMs/MLLMs (7B) with or without visual input. The bar with dashed borders denotes Qwen2-VL-Instruct, indicating it has pixel-level visual input, while others do not. Qwen2-VL-Ins.(NA) refers to a setting where only the question is provided as input. We divide the MME score by 100 to have the same scale as other benchmarks.

detailed captions, we convert low-level pixel information into semantically rich, fine-grained textual descriptions, enabling a deeper understanding of visual content, which effectively bridges the gap between visual and textual modalities.

To evaluate the effectiveness of OMNICAPTIONER, we conduct systematic assessments across both image understanding (*e.g.*, visual reasoning) and image generation tasks (*e.g.*, text-to-image generation). Our results reveal several key advantages: i) **Improved Visual Reasoning with LLMs**: Our detailed, long-context captions can be directly incorporated into powerful LLMs to address challenging visual reasoning questions, particularly for models like the DeepSeek-R1 (Guo et al., 2025) series. This approach enables LLMs to perform visual reasoning tasks in a training-free manner, leveraging rich textual descriptions without requiring additional fine-tuning. ii) **Enhanced Image Generation and Conversion**: The detailed captions produced by our framework significantly improve image generation tasks, such as image-to-text generation and image conversion, owing to their near-complete pixel-to-text mapping capability. iii) **Efficient SFT Process**: Leveraging the knowledge from pretraining on OMNICAPTIONER, the SFT process becomes more efficient, requiring less training data and achieving faster convergence.

Furthermore, the contributions of this paper are summarized below:

- **Unified Visual Captioning Framework**: We present OMNICAPTIONER, a unified framework for generating captions across diverse domains. Our approach seamlessly integrates captioning capabilities for natural images, visual text images (*e.g.*, posters, UI, and textbooks), and structured visual images (*e.g.*, tables, charts, equations, and geometric diagrams). OMNICAPTIONER sets a new standard for generalized visual captioning, enabling more effective and scalable vision-language understanding.
- **Comprehensive Pixel-to-Text Conversion**: Our framework leverages detailed captions to convert low-level pixel information into semantically rich, fine-grained textual descriptions, effectively bridging the gap between visual and textual modalities. Particularly, this enhances text-to-image generation by providing more precise and context-aware textual guidance, leading to improved visual fidelity and alignment with the intended semantics.
- **Improved Visual Reasoning with LLMs**: By incorporating detailed, long-context captions, our approach enables enhanced visual reasoning capabilities, especially when integrated into LLMs such as the DeepSeek-R1 series. Leveraging the perceptual information provided by OMNICAPTIONER, LLMs can infer and reason within the textual space to effectively solve visual reasoning tasks.

2 Related Works

Image Captioning. Image captioning tasks can be broadly classified into two categories. The first approach focuses on generating high-quality captions for natural images. Notably, ShareGPT4V (Chen et al., 2024a) improves vision-language alignment by collecting high-quality, attribute-specific captions through targeted prompts to GPT-4V for natural images, while models like Densefusion (Li et al., 2024b) leverage multiple expert models to synthesize captions for natural images. The second approach, exemplified by CompCap (Chen et al., 2024c), tackles the challenge of domain diversity during pretraining by incorporating synthetic images to enhance performance on underrepresented domains. However, the first approaches are often constrained by its focus on specific domains, while the second faces challenges due to the relatively small quantity of synthetic images used during training.

Multimodal Large Language Models. With the development of LLMs (Yang et al., 2024; Guo et al., 2025; Touvron et al., 2023; Yuan et al., 2025), integrating visual perception capability into LLMs (*i.e.*, MLLMs) has received increasing attention. To address the gap between different modalities, most of works (Wang et al., 2024b; Bai et al., 2025; Chen et al., 2024d; Xia et al., 2024b; Liu et al., 2023; Li et al., 2024a; Lin et al., 2023; Liu et al., 2024a) first pretrain on image captioning data to obtain a vision-language connector (*e.g.*, MLP-based or cross-attention based) and followed by SFT. To better integrate information from multiple modalities, several works (Lin et al., 2024; Luo et al., 2024b; Diao et al., 2024; Team, 2024) try to explore new architectures to process different modalities in a single Transformer model. In addition to model architecture, some works (Wang et al., 2024c) try to boost models' reasoning ability through post-training (*e.g.*, reinforcement learning) (Wang et al., 2024c) or test-time scaling (*e.g.*, monte-carlo tree search) (Yao et al., 2024a; Luo et al., 2025; Dong et al., 2024a). Furthermore, recent studies (Zhang et al., 2024a; McKinzie et al., 2024; Chen et al., 2024c; Deng et al., 2025) have systematically investigated the influence of data quality on on both the pretraining and SFT phases of MLLMs. MM1 (Zhang et al., 2024a) reveals that model capabilities induced through pretraining with high-quality data are effectively preserved after SFT. Most existing open-source MLLMs (Liu et al., 2023; Li et al., 2024a) primarily focus on pretraining with natural images, while domain-specific MLLMs (*e.g.*, math, chart) are trained exclusively on domain-specific caption data. In contrast, we propose a more unified pretraining approach that integrates diverse domain knowledge during pretraining. In addition, current MLLMs generally exhibit inferior reasoning capabilities compared to text-only LLMs, OMNICAPTIONER can generate detailed, long-context captions of different domains and use LLMs to address challenging visual reasoning tasks.

3 OMNICAPTIONER

To achieve a unified multimodal pretraining paradigm and handle diverse visual domains, we first construct a diverse caption dataset as shown in Sec. 4. We will provide the dataset description and then detail the dataset construction process in Sec. 3.1 and Sec. 3.2, respectively. And the pertaining process is described in Sec. 3.3.

3.1 Diverse Visual Caption Dataset

The diversity of our visual caption dataset is characterized by two dimensions: domain diversity (diverse data sources) and caption formula diversity. To achieve effective unified pretraining, the dataset needs to encompass a wider range of domains. For example, when acting as a documentation assistant, MLLMs need to comprehend tables and charts, while as a GUI agent, they are required to understand elements in web pages. As illustrated in the data distribution section

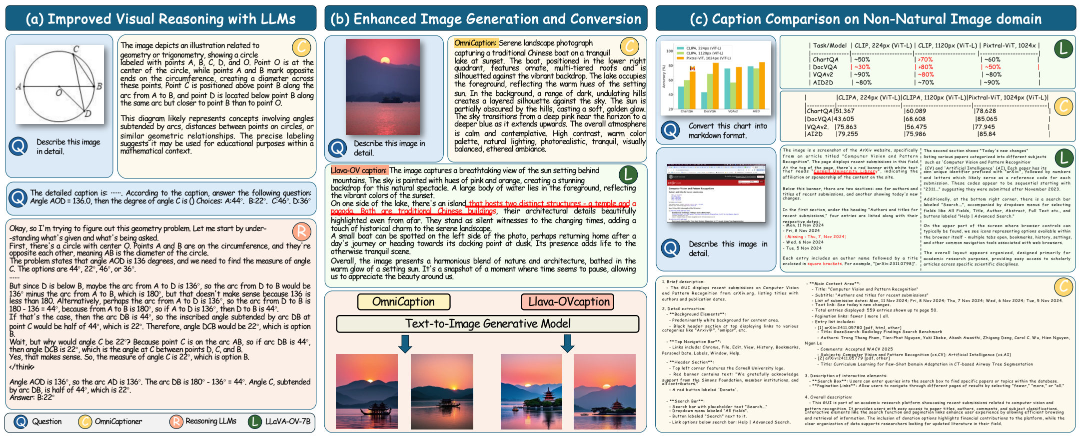


Figure 3: Illustration of OMNICAPTIONER's plug-and-play applications (Sub-figure a, b) and comparison between OMNICAPTIONER and LLava-OneVision-7B on non-natural image captioning (Sub-figure c). Sub-figure (a) shows that OMNICAPTIONER leverages LLMs' strong reasoning abilities to perform multimodal reasoning tasks. Sub-figure (b) highlights how hallucinated or inaccurate captions—like those from LLava-OneVision-7B can lead to inconsistent image conversion, revealing weakened alignment capabilities in text-to-image models when captions don't faithfully represent the original content. Sub-figure (c) highlights that LLaVA-OneVision-7B, due to limited exposure to non-natural images during pretraining, struggles with perception in such domains, often leading to hallucinations, whereas OMNICAPTIONER provides more accurate descriptions.

of Fig. 4, our caption dataset is composed of four major categories: natural images, structured images (including chart, table, and so on), visual text images (including UI images, posters, and so on), and video. This comprehensive data coverage enables our model to serve as a multi-domain assistant and further enhance the performance on downstream tasks. Furthermore, diverse types of captions may be necessary even for the same visual input. For instance, a chart image may require both structured tabular conversion and comprehensive analytical descriptions. To address this requirement, we define diverse caption formulas for each domain. This approach enables our model to generate diverse caption formats, including multilingual (Chinese and English) descriptions, varying granularity levels (from comprehensive to concise), and so on.

3.2 Dataset Construction

To generate high-quality captions for images across diverse domains, we propose a two-step caption generation pipeline. The design of our pipeline takes into account the need for accurate visual descriptions, the flexibility to support different stylistic outputs, the ability to perform reasoning and logic extrapolation, as well as bilingual captioning.

Seed Caption Generation. In the first stage, we focus on seed caption generation. The goal is to produce an initial caption that is as accurate as possible, with a comprehensive textual description of all relevant visual elements present in the visual signal. This stage leverages carefully designed prompts to guide the powerful closed-source multimodal model GPT-4o to describe all possible visual elements in natural images and visual-text images, ensuring an accurate pixel-to-word mapping. For structured images generated via code, the description is generated as accurately as

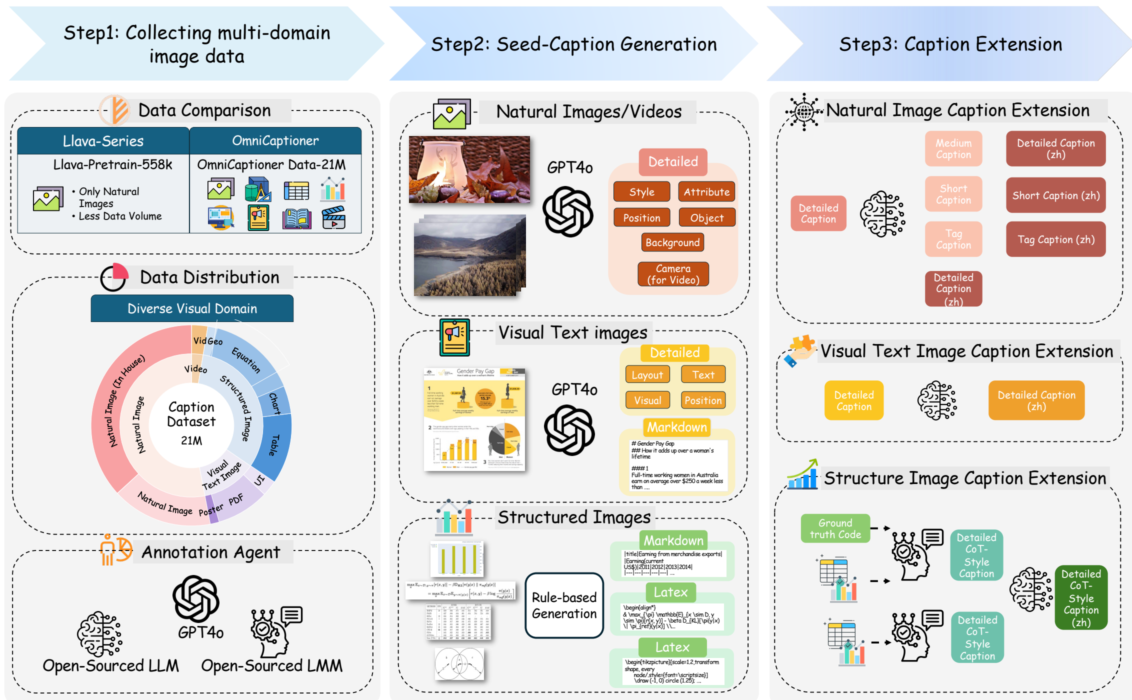


Figure 4: OMNICAPTIONER's diverse visual captioning pipeline. The pipeline consists of Seed-Caption Generation to ensure precise pixel-to-word mapping, and Caption Extension to enrich caption styles to support image generation and visual reasoning tasks. OMNICAPTIONER utilizes a **21M-caption dataset**, covering diverse domains beyond natural images, enabling more comprehensive captioning capabilities. For further details about dataset composition, please refer to Fig. 7 in Appendix A.

possible using predefined code rules. The generated seed caption serves as a reliable foundation for further refinement in the subsequent stage.

Caption Extension. The second stage, caption extension, is responsible for enhancing and diversifying the generated caption. Here, the focus shifts from purely accuracy to incorporating stylistic variation and domain-specific reasoning. The seed caption is extended by introducing bilingual outputs (Chinese and English), with variations ranging from detailed to medium-length, short, and tag-style captions. Additionally, we inject reasoning knowledge relevant to specific domains to enrich the semantic depth of the captions. This allows the captions to not only reflect the visual content but also accommodate nuanced understanding in different contexts. Specially, **for Natural Images**, we leverage the open-source LLM, Qwen2.5-32B, to adjust the caption length through different prompts, allowing captions to range from medium-length to short and tag-style. Additionally, these varied captions are translated into Chinese, facilitating the creation of bilingual prompts for image generation. The benefit of this approach is to enable more flexible and effective bilingual prompt extraction for image generation tasks. **For Visual Text Images**, we use open source LLM Qwen2.5-32B to translate the detailed subtitles generated by GPT-4o into the corresponding Chinese versions to ensure cross-language consistency. **For Structured Images**, which often relate to mathematical or document-based reasoning (*e.g.*, Chain-of-Thought (CoT) analysis), we prioritize the accuracy of the seed caption. After confirming the seed caption's accuracy, we input both the seed caption and the original image into the open-source multimodal model Qwen2-VL-76B for CoT-style caption generation. This approach allows us to condition the captioning process on both the seed caption's code (*e.g.*, Markdown, LaTeX) and the image content, reducing hallucinations and improving the reliability of the generated captions. Additionally, we collect structured images without seed captions and directly input them into the same multimodal model for CoT-style caption generation. By decoupling the caption generation process into these two stages, we ensure both high accuracy in representing visual content and flexibility in producing diverse, contextually appropriate captions.

3.3 Unified Pretraining Process

To effectively handle the multi-domain nature of the OMNICAPTIONER dataset, which spans a broad range of image types and captioning tasks, we propose a practical approach utilizing distinct system prompts. These prompts help minimize task conflicts and improve task coordination during training. By customizing system prompts for specific image categories and using a fixed set of question templates for various captioning styles, we differentiate between tasks and data types in the pretraining process. This approach facilitates efficient multi-domain training, ensuring robust model performance across diverse tasks and domains. To address the challenge of handling images with large variations in resolution and arbitrary aspect ratios, we leverage the powerful visual understanding capabilities of the Qwen2-VL-7B (Wang et al., 2024b) model. Given that the Qwen2-VL-Instruct model is inherently powerful in managing multi-domain image data, we initialize our model with the Qwen2-VL-Instruct weights. This initialization allows us to effectively fine-tune on the OMNICAPTIONER dataset, ensuring robust performance across a wide range of image resolutions and aspect ratios while benefiting from the model's ability to generalize across diverse domains.

4 One Captioner to Rule Them All

Improved Visual Reasoning Tasks with LLMs. Current MLLMs lag behind LLMs in reasoning capabilities. This discrepancy motivates us to investigate whether LLMs can directly perform visual reasoning without modality-alignment losses that may degrade their reasoning ability while still effectively handling diverse visual reasoning tasks. In this work, we integrate image captioning with large language models (LLMs) to enable seamless visual reasoning in textual space. As illustrated in Fig. 3 (a), firstly, our captioner converts input images (spanning natural images, charts, equations, and beyond) into linguistically dense descriptions that explicitly encode pixel-level structures (*e.g.*, spatial layouts, symbolic operators, tabular hierarchies) into textual space. These captions, acting as lossless semantic proxies, are then directly processed by powerful LLMs (*e.g.*, DeepSeek-R1 (Guo et al., 2025), Qwen2.5 series (Yang et al., 2024)) to perform task-agnostic visual reasoning, including geometric problem-solving and spatial analysis.

Just shown in Fig. 3, OMNICAPTIONER can transform geometric images into detailed and precise visual descriptions. OMNICAPTIONER accurately describes geometric images, such as a circle with a diameter and circumferential angles, detailing spatial relationships among points. This enables LLMs to perform logical inferences, like calculating angles, without direct pixel-level perception.

There are three key advantages to this approach: i) Decoupled Perception and Reasoning – By separating perception (handled by MLLMs) from reasoning (handled by LLMs), our method avoids conflicts between the two capabilities, leading to more effective and accurate visual reasoning. ii) Elimination of Modality-Alignment Training – Instead of requiring complex modality-alignment losses, our approach translates visual inputs into linguistic representations, allowing LLMs to process them naturally. This removes the need for additional multimodal training while preserving the reasoning strengths of LLMs. iii) Flexibility and Generalization – The plug-and-play design enables seamless integration of LLMs into diverse visual reasoning tasks without domain-specific tuning. This ensures broad applicability across different types of visual inputs, from geometric diagrams to complex tabular structures.

Enhanced Image Generation and Conversion. Detailed and accurate image captions play a pivotal role in both the training and inference stages of Text-to-Image (T2I) tasks. During training, such captions offer fine-grained supervision by explicitly aligning low-level/high-level visual patterns (*e.g.*, textures, spatial arrangements, object attributes) with precise linguistic semantics. At inference time, as shown in Fig. 3(b), detailed and precise captions substantially enhance image generation quality by guiding the model to follow instructions more faithfully—capturing spatial relationships, object interactions, and semantic details with higher fidelity. These benefits highlight the critical role of captions as a dense supervisory signal, enabling more precise instruction-following in T2I generation.

Efficient SFT Process. The training paradigm of MLLMs typically consists of two sequential phases: pretraining on image-caption data, followed by Supervised Fine-Tuning (SFT). Empirical studies (Chen et al., 2024c; Jiang et al., 2025; McKinzie et al., 2024) have demonstrated that diverse and high-quality image-caption data (*e.g.*, composite images) can significantly enhance image-language alignment and subsequently promote performance on downstream tasks, such as Visual Question Answering (VQA). OMNICAPTIONER leverages diverse and high-quality domain data (*e.g.*, table, chart, and so on) during the pretraining phase, enabling the model to acquire multi-domain knowledge. During the SFT phase, the multi-domain knowledge serves as a crucial foundation for rapid adaptation to downstream tasks across different domains.

5 Experiment

To evaluate OMNICAPTIONER, we conduct four primary experiments. The **first** experiment focuses on the caption evaluation from the perspective of objective metrics and subjective preference. The **second** experiment focuses on visual reasoning with a Caption-inserted Large Language Model. In this setup, detailed captions and corresponding questions are provided to the LLM, and its ability to answer the questions is evaluated. We use five benchmark datasets to assess the model's performance on this downstream task: MME (Fu et al., 2023), Mathverse (Zhang et al., 2024c), Mathvision (Wang et al., 2024a), MMMU (Yue et al., 2024) and Olympiad bench (He et al., 2024). For the LLMs, we select Qwen2.5-3B-Instruct (Yang et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), Qwen2.5-32B-Instruct (Yang et al., 2024), DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025), and DeepSeek-R1-Distill-LLaMA-70B (Guo et al., 2025), all chosen for their strong reasoning capabilities. The **third** experiment involves finetuning the text-to-image generation model (Qin et al., 2025; Gao et al., 2024; Xie et al., 2025) such as SANA-1.0-1.6B (Xie et al., 2025) with image-caption pairs generated by different captioners (*i.e.*, Qwen2-VL (Wang et al., 2024b), OMNICAPTIONER). The training setting uses a resolution of 1024 $\times$ 1024. The model's generative performance is then evaluated on the GenEval (Ghosh et al., 2023).The **fourth** experiment evaluates the efficiency of the SFT process. For this, we select the LLaVA-OneVision (Li et al., 2024a) data from the OV stage with chain-of-thought enhancement to assess the SFT version of OMNICAPTIONER across multiple commonly-used benchmarks (Fu et al., 2023; Yue et al., 2024; Masry et al., 2022; Mathew et al., 2021; Wang et al., 2024a; Zhang et al., 2024c; Lu et al., 2023).

Table 1: Performance comparison on various visual benchmarks between our OMNICAPTIONER-inserted LLMs and previous SOTA MLLMs. **We would like to emphasize** that by utilizing OMNICAPTIONER, LLMs can function as MLLMs without requiring additional training. Moreover, we have observed that, particularly in mathematical scenarios, caption-integrated LLMs surpasses MLLMs with comparable parameter sizes, where MLLMs have undergone rigorous data preparation and GPU-intensive training.

Model	MME	MMMU	MathVision	MathVerse	Olympiad
		Frontier Models			
GPT-4V	-	63.1	24.0	32.8	18.0
GPT-4o (2024-05)	-	69.1	30.4	50.2	25.9
Claude3.5-Sonnet	-	68.3	-	-	-
		3B-Level Models			
Qwen2-VL-2B (Wang et al., 2024b)	1872	41.1	12.4	21.0	-
InternVL2-2B (Chen et al., 2024d)	**1876**	36.3	12.1	**25.3**	0.4
MinniCPM-V2.0 (Yao et al., 2024c)	1808	38.2	-	-	-
OMNICAPTIONER + Qwen2.5-3B-Instruct	1599	**43.0**	**16.0**	22.2	**7.24**
		7B-Level Models			
Qwen2-VL-7B (Wang et al., 2024b)	2327	54.1	16.3	31.9	-
InternVL2-8B (Chen et al., 2024d)	2210	52.6	18.4	37.0	1.9
MiniCPM-Llama-V-2.5-8B (Yao et al., 2024b) 2024	45.8	-	-	-	-
Cambrain-1-8B (Tong et al., 2024)	-	42.7	-	-	-
LLava-Onevision-7B (Li et al., 2024a)	1998	48.8	-	26.2	-
MiniCPM-V2.6 (Yao et al., 2024b)	**2348**	49.8	18.3	25.7	-
OMNICAPTIONER + Qwen2.5-7B-Instruct	1824	**54.5**	26.4	37.4	**10.9**
OMNICAPTIONER + DS-R1-Distill-Qwen-7B	1942	47.5	**36.2**	**40.5**	7.8
		32B-Level Models			
InternVL-Chat-V1.5 (Chen et al., 2024d)	2194	46.8	15.0	28.4	0.6
InternVL2-26B (Chen et al., 2024d)	2260	51.2	17.0	31.1	3.5
Cambrian-34B (Tong et al., 2024)	-	49.7	-	-	-
VILA-1.5-40B	-	55.1	-	-	-
InternVL2-40B	**2307**	55.2	16.9	36.3	3.9
OMNICAPTIONER + Qwen2.5-32B-Instruct	1831	**59.7**	32.1	39.7	13.1
OMNICAPTIONER + DS-R1-Distill-Qwen-32B	2007	59.2	**43.3**	**43.7**	**13.2**
		72B-Level Models			
Qwen2-VL-72B (Wang et al., 2024b)	**2482**	64.5	25.9	-	11.2
InternVL2-76B (Chen et al., 2024d)	2414	62.7	23.6	**42.8**	5.5
LLaVA-OneVision-72B (Li et al., 2024a)	2261	56.8	-	39.1	-
OMNICAPTIONER + DS-R1-Distill-Llama-70B	2025	**64.6**	**42.9**	42.5	**13.7**

Table 2: Caption Metrics comparison across models. OMNICAPTIONER achieves the highest score in all metrics.

Metric	LLaVA-OV-7B	Qwen2-VL-7B	OmniCaptioner
BLEU	14.18	21.70	**22.35**
CLIPScore	30.12	32.71	**34.05**
CAPTURE	62.40	64.38	**64.88**

Table 3: User study results on caption preference (%). OMNICAPTIONER is preferred by human evaluators across both image types.

Domain	Qwen2-VL-7B	OMNICAPTIONER
Non-Natural Images	43.3	**56.7**
Natural Images	48.8	**51.2**

5.1 Main Results

Caption Quality Comparison. To evaluate the quality of generated captions, we conduct a comprehensive comparison using both objective metrics and subjective human evaluation. For the objective evaluation, we measure BLEU (Papineni et al., 2002), CLIPScore (Hessel et al., 2021), and CAPTURE scores (Dong et al., 2024b). As shown in Table 2, our method, OMNICAPTIONER, consistently outperforms strong baselines such as LLaVA-OneVision and Qwen2-VL-7B-Instruct, achieving the highest scores across all metrics: 22.35 on BLEU, 34.05 on CLIPScore, and 64.88 on CAPTURE, demonstrating superior alignment with both textual references and visual semantics. In addition to automated evaluation, we conduct a user study to assess human preference. We collect 90 images from three diverse sources—MMMU, ChartQA, and MME—and categorize them into natural and non-natural domains. Captions are generated by both OMNICAPTIONER and Qwen2-VL-7B-Instructand evaluated in a blind pairwise comparison setting by 10 human raters. Each rater selects the preferred caption based on relevance, informativeness, and fluency. As shown in Table 3, OMNICAPTIONER is preferred in 56.7% of non-natural image cases and 51.2% of natural image cases, indicating a consistent advantage in human-perceived quality across different image types.

Improved Visual Reasoning with LLMs. Our experimental results of Table 1 demonstrate that integrating captions into reasoning-enhanced Large Language Models (LLMs), without any additional fine-tuning, achieves state-of-the-art performance across multiple reasoning benchmarks, including MathVision (Wang et al., 2024a), MathVerse (Zhang et al., 2024c), MMMU (Yue et al., 2024), and Olympiad bench (He et al., 2024). This highlights the power of OMNICAPTIONER in boosting reasoning capabilities for multiple visual tasks. Specifically, OMNICAPTIONER-inserted LLMs significantly outperform existing models in MathVision across multiple model sizes, underscoring the enhancement of reasoning ability for complex visual and mathematical tasks. Notably, *OmniCaptioner + DS-R1-Distill-Qwen-7B* and *OmniCaptioner + DS-Distill-Qwen-32B* demonstrate exceptional performance on MathVerse benchmark, significantly outperforming previous models. These results further validate the efficacy of caption-based pretraining in bridging the LLM's comprehension of visual geometry content. In the MMMU benchmark, *OmniCaptioner + DS-R1-Distill-Qwen-72B* approaches the performance of Qwen2-VL-72B, with a minimal gap between them. This result serves as strong evidence that caption integration with reasoning-enhanced LLMs leads to significant visual understanding and reasoning for multidisciplinary content.

The successful integration of captions with LLMs across scales, from 3B to 72B, underscores that OMNICAPTIONER consistently enhances LLMs' reasoning abilities for visual tasks, yielding improvements irrespective of model size. These results highlight that our unified pretraining methodology, leveraging large-scale caption data, is a highly effective strategy for advancing visual reasoning across diverse tasks, outperforming existing approaches even when compared to large-scale fine-tuning methods.

Enhanced Image Generation. As illustrated in Tab. 4, to validate the importance of caption accuracy in T2I generation, our model demonstrates significant performance improvements over the Qwen2-VL-Instruct (Wang et al., 2024b) caption and original SANA, on GenEval benchmark. The original SANA model achieves a 64.61 overall score on GenEval, which is significantly improved to 65.27 with Qwen2-VL-Instruct and further to 67.58 withOMNICAPTIONER. This +2.97 absolute gain over the vanilla SANA model underscores the effectiveness of high-quality captions in guiding T2I generation. Also, our OMNICAPTIONER outperforms Qwen2-VL-Instruct across various aspects (except colors), showcasing the enhanced accuracy of our caption generation.

Efficient SFT. In Tab. 5, we compare the performance of several models on visual perception and reasoning tasks, including *LLaVA-OV-7B(SI)*, *LLaVA-OV-7B*, *Qwen2-VL-base+OV SFT*, and our

Table 4: Performance comparison of models trained with different captioners on GenEval (Ghosh et al., 2023) (Resolution: 1024 × 1024).

Methods	GenEval ↑					
	Color Attri.	Sin. Obj.	Pos.	Colors	Counting	Overall
SANA-1.0-1.6B (Xie et al., 2025)	38.50	98.75	21.25	86.70	**65.31**	64.61
SANA-1.0-1.6B + Qwen2-VL (Wang et al., 2024b)	44.29	98.44	26.64	**86.97**	57.81	65.27
SANA-1.0-1.6B + OMNICAPTIONER	**46.00**	**99.06**	**29.50**	84.57	64.06	**67.58**

Table 5: SFT performance comparison across diverse evaluation benchmarks. OmniCaptioner + OV SFT denotes the SFT model based on OMNICAPTIONER, while Qwen2-VL-base + OV SFT is based on Qwen2-VL-Base. LLaVA-OV-7B (SI) represents the model after the single-image training in LLaVA-OneVision (Li et al., 2024a).

SFT Model	Data	MME	MMMU	MathVerse	MathVista	DocVQA	ChartQA
LLaVA-OV-7B (SI) (Li et al., 2024a)	3.2M	2109	47.3	26.9	56.1	89.3/86.9	78.8
LLaVA-OV-7B (Li et al., 2024a)	4.8M	1998	48.8	26.2	63.2	90.2/87.5	80.0
Qwen2-VL-Base+OV SFT	1.6M	1905	44.4	24.9	53.8	84.2/-	53.5
OMNICAPTIONER+OV SFT	1.6M	2045	46.6	25.8	57.4	91.2/-	79.0

proposed *OmniCaptioner+OV SFT* model. While *LLaVA-OV-7B (SI)* and *LLaVA-OV-7B* use significantly larger datasets for SFT – 3.2M and 4.8M examples, respectively – our *OmniCaptioner+OV SFT* achieves comparable results with just 1.6M SFT examples used during the one-vision (OV) stage. A key difference lies in the unified pretraining phase of OMNICAPTIONER, which utilizes a diverse caption-based dataset prior to the SFT stage. This step equips the model with richer domain knowledge, enabling it to excel in visual instruction-following tasks despite fewer SFT examples. It also reveals that *Qwen2-VL-base + SFT* lags behind *OmniCaptioner + OV SFT*, indicating OMNICAPTIONER's superior visual perception capabilities.

5.2 Discussions and Findings

We consider conducting three important experiments when combining OMNICAPTIONER with reasoning-enhanced LLMs. First, we evaluate effectiveness using different Qwen versions. Second, we aim to explore the extent to which Qwen2-VL-Instruct (without image input) and mainstream reasoning-enhanced LLMs rely on visual modality information to solve visual reasoning tasks. Third, we compare OMNICAPTIONER to Qwen2-VL-Instruct by modifying the captions provided to the reasoning-enhanced LLMs. For more visualization results of image captioning, video captioning, and text-to-image generation task, please refer to Appendix E and Appendix F.

Effect of Different Qwen-Family Versions. Fig. 5 illustrates the performance progression of combining OMNICAPTIONER with different versions of Qwen on MMMU and MathVerse. As Qwen evolves from Qwen1-8B-chat to Qwen2.5-7B-Instruct, there is a steady improvement in visual reasoning capabilities, driven by the pixel-to-word captioning ability of OMNICAPTIONER. As illustrated in Fig. 2, the performance comparison between *OmniCaptioner + Qwen2.5-7B-Base*, *OmniCaptioner + Qwen2.5-7B-Instruct* and *OmniCaptioner + DS-R1-Distill-Qwen-7B* highlights the advantage of integrating the DeepSeek Distilled Qwen2.5, which excels in mathematical reasoning. The distilled variant (*DS-R1-Distill-Qwen-7B*) achieves the highest accuracy across ***MME (1942)***, ***MathVision (36.2)***, and ***MathVerse (40.5)***, emphasizing the benefits of distilled reasoning ability. In contrast, *Qwen2.5-7B-Instruct* is better suited for general world knowledge tasks, as reflected in its improved performance on the ***MMMU (54.5)***.

Table 6: Comparing different captioners through experiments with captioner-inserted LLM on multiple visual benchmarks.

Captioner Selection	LLM	MME	MMMU	MathVision	MathVerse
llava-onevision-qwen2-7b-ov	DS-R1-Distill-Qwen-7B	1646	22.4	31.7	36.6
InternVL2-8B	DS-R1-Distill-Qwen-7B	1789	23.1	34.4	39.9
Qwen2-VL-7B-Instruct	DS-R1-Distill-Qwen-7B	1914	42.4	31.6	33.0
OMNICAPTIONER (ours)	DS-R1-Distill-Qwen-7B	**1942**	**47.5**	**36.2**	**40.5**

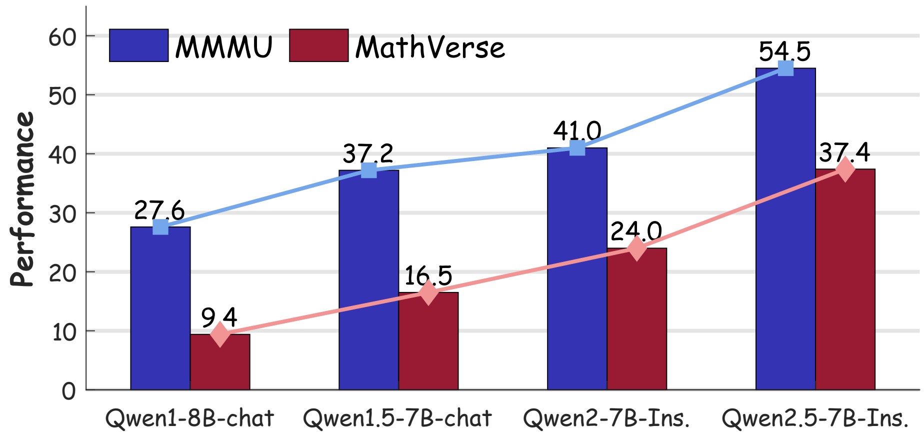


Figure 5: Integrate OMNICAPTIONER into different versions of LLMs, enabling them to handle tasks in multimodal scenarios.

Impact of Visual Modality on Reasoning-Enhanced LLMs. From Fig. 2, the performance of *Qwen2-VL-7B (NA)* and *DeepSeek-Distill-Qwen-7B* suggests that the absence of image input significantly restricts their ability to solve visual reasoning tasks. In contrast, *OmniCaptioner + DS-R1-Distill-Qwen-7B*, which retains visual modality, achieves substantially higher accuracy than its non-visual input LLM, highlighting the critical role of visual information in enhancing reasoning capabilities. Furthermore, non-visual input LLM *DS-R1-Distilled-Qwen-7B* significantly outperforms the no-image MLLM (*i.e.*, *Qwen2-VL-Instruct-7B*) on MathVision and MathVerse, demonstrating the superior reasoning ability of R1 Serious model.

Effect of Different Captioners. Tab. 6 presents a comparative analysis of different captioners on multiple perception and reasoning benchmarks. Our model, incorporating DeepSeek-Distill-Qwen2.5-7B, achieves superior performance across all evaluated tasks, significantly outperforming previous approaches. These results highlight the effectiveness of OMNICAPTIONER, whose captions provide more precise and contextually accurate descriptions than those generated by Qwen2-VL-7B-Instruct. The enhanced caption quality contributes to improved visual reasoning tasks, particularly in tasks requiring multi-step inference and detailed visual understanding.

6 Conclusion

We have introduced OMNICAPTIONER, a unified framework that bridges visual and textual modalities through fine-grained pixel-to-text mapping across diverse domains, including natural images, visual-text images and structured images. By converting low-level visual patterns into semantically rich captions, our approach empowers reasoning-enhanced LLMs (*e.g.*, DeepSeek-R1) to achieve enhanced visual reasoning, and enables precise text-to-image generation through comprehensive semantic preservation. This work pioneers a scalable paradigm for multimodal alignment and reasoning, achieving seamless visual-language interoperability without costly label-supervised fine-tuning.

Acknowledgement

The research was supported by Shanghai Artificial Intelligence Laboratory, a locally commissioned task from the Shanghai Municipal Government, the Shanghai Municipal Science and Technology Major Project, and Shanghai Rising Star Program (Grant No. 23QD1401000).

References

Anas Awadalla, Le Xue, Manli Shu, An Yan, Jun Wang, Senthil Purushwalkam, Sheng Shen, Hannah Lee, Oscar Lo, Jae Sung Park, et al. Blip3-kale: Knowledge augmented large-scale dense captions. *arXiv preprint arXiv:2411.07461*, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024a.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024b.

Xiaohui Chen, Satya Narayan Shukla, Mahmoud Azab, Aashu Singh, Qifan Wang, David Yang, ShengYun Peng, Hanchao Yu, Shen Yan, Xuewen Zhang, et al. Compcap: Improving multimodal large language models with composite captions. *arXiv preprint arXiv:2412.05243*, 2024c.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024d.

Xueqing Deng, Qihang Yu, Ali Athar, Chenglin Yang, Linjie Yang, Xiaojie Jin, Xiaohui Shen, and Liang-Chieh Chen. Coconut-pancap: Joint panoptic segmentation and grounded captions for fine-grained understanding and generation. *arXiv preprint arXiv:2502.02589*, 2025.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*, 2024.

Guanting Dong, Chenghao Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. Progressive multimodal reasoning via active retrieval. *arXiv preprint arXiv:2412.14835*, 2024a.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Peng Gao, Le Zhuo, Dongyang Liu, Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, Chen Lin, Rongjie Huang, Shijie Geng, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

Zihan Huang, Tao Wu, Wang Lin, Shengyu Zhang, Jingyuan Chen, and Fei Wu. Autogeo: Automating geometric image dataset creation for enhanced geometry understanding. *arXiv preprint arXiv:2409.09039*, 2024.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303*, 2024b.

Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*, 2024.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Junpeng Liu, Tianyue Ou, Yifan Song, Yuxiao Qu, Wai Lam, Chenyan Xiong, Wenhu Chen, Graham Neubig, and Xiang Yue. Harnessing webpage uis for text-rich visual understanding. *arXiv preprint arXiv:2410.13824*, 2024b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15630–15640, 2024a.

Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *arXiv preprint arXiv:2410.08202*, 2024b.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*, 2025.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pp. 304–323. Springer, 2024.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Tianshuo Peng, Mingsheng Li, Hongbin Zhou, Renqiu Xia, Renrui Zhang, Lei Bai, Song Mao, Bin Wang, Conghui He, Aojun Zhou, et al. Chimera: Improving generalist model with domain-specific experts. *arXiv preprint arXiv:2412.05983*, 2024.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, et al. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024a.

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*, 2024b.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024c.

Renqiu Xia, Mingsheng Li, Hancheng Ye, Wenjie Wu, Hongbin Zhou, Jiakang Yuan, Tianshuo Peng, Xinyu Cai, Xiangchao Yan, Bin Wang, et al. Geox: Geometric problem solving through unified formalized vision-language pre-training. *arXiv preprint arXiv:2412.11863*, 2024a.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, et al. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*, 2024b.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

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.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024a.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024b.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024c.

Jiakang Yuan, Xiangchao Yan, Botian Shi, Tao Chen, Wanli Ouyang, Bo Zhang, Lei Bai, Yu Qiao, and Bowen Zhou. Dolphin: Closed-loop open-ended auto-research through thinking, practice, and feedback. *arXiv preprint arXiv:2501.03916*, 2025.

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*, pp. 9556–9567, 2024.

Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024a.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning. *arXiv preprint arXiv:2404.16635*, 2024b.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024c.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. *arXiv preprint arXiv:2407.08739*, 2024d.

Shan Zhang, Aotian Chen, Yanpeng Sun, Jindong Gu, Yi-Yu Zheng, Piotr Koniusz, Kai Zou, Anton van den Hengel, and Yuan Xue. Open eyes, then reason: Fine-grained visual mathematical understanding in mllms. *arXiv preprint arXiv:2501.06430*, 2025.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. Multimodal table understanding. *arXiv preprint arXiv:2406.08100*, 2024.

A OmniCaptioner Dataset Composition

As shown in Fig. 7, the OMNICAPTIONER dataset is a large-scale multimodal benchmark comprising images, tables, charts, mathematical geometry/equations, posters, PDFs, UI elements, and videos, with captions available in both English and Chinese. The dataset includes natural images sourced from in-house collections, BLIP3Kale (Awadalla et al., 2024), and DenseFusion (Li et al., 2024b). Tabular data are collected from the arXiv website and the open-source MMTab dataset (Zheng et al., 2024), while chart data originate from arXiv website and TinyChart (Zhang et al., 2024b). Mathematical content, including equations and geometric structures, is sourced from arXiv and generated from datasets such as MAVIS (Zhang et al., 2024d) and AutoGeo (Huang et al., 2024). UI data are obtained from the MultiUI dataset (Liu et al., 2024b), while poster images feature OCR-based captions. Video captions are derived from OpenVid (Nan et al., 2024) and Panda (Chen et al., 2024b), covering multiple attributes such as detailed descriptions, style, background, tags, camera angles, and object information. Fig. 6 illustrates the token length distribution for different caption types associated with natural images, categorized into detailed, medium, short, and tag captions.

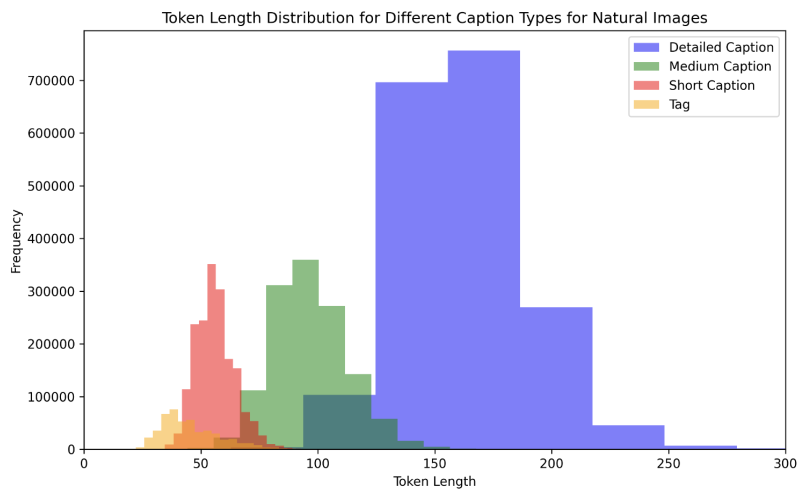


Figure 6: Token length distribution for natural images.

B Experimental Setup

We fine-tune the Qwen2-VL-7B-Instruct model on a large-scale captioning dataset using 64 A100 GPUs. The training process is distributed using torchrun with the DeepSpeed ZeRO-3 optimization strategy.

Hyperparameters:

- **Batch Size:** 256 (1 per device, with gradient accumulation of 8)
- **Learning Rate:** 1e-5 (base model), 1e-5 (merger module), 2e-6 (vision tower)
- **Weight Decay:** 0.0
- **Warmup Ratio:** 3%

- **Scheduler:** Cosine decay
- **Precision:** BF16 enabled,
- **Gradient Checkpointing:** Enabled

Training Details:

- **Image Resolution:** From **2**×**28**×**28** to **6400**×**28**×**28** pixels
- **Epochs:** 1

C Prompt for Caption Annotation

Natural Image Annotation Prompt for GPT-4o

You are an expert in image captioning, segmentation labeling, and stylistic descriptions at the level of an Oscar-winning cinematographer, photographer, or illustrator. Your task is to give me an extremely information-dense description of each image I send you.
Remember that you may need to caption images from all visual domains imaginable: Photography, Movie Stills, Animated Pixar movies, Sketches, IKEA assembly instruction diagrams, Screenshots, UIs, Cave Paintings, Abstract Art, Product Photography, all forms of illustrations, and many more genres. Therefore, you need to be quite descriptive and effective at identifying the artistic medium and render technique utilized. Avoid unnecessary repetitions and redundancy. Only write what you feel reasonably confident about. Occasional mistakes are okay, but do not hallucinate what you do not actually see in each image.
Your response should for the most part answer three questions:

1. How would you describe this image and its environment overall?
   (e.g., "Photo portrait of a white, middle aged man in front of a white background looking to the right.")
2. What are all objects you see in this image and where exactly are they placed?
   (e.g., "A yellow Taxi driving forwards in the left foreground. Pedestrian crossing and cracked asphalt street in the center. Many cars in the background. New York buildings skyline in the background.")
3. What are all purely stylistic properties that this image shows?
   (e.g., "Underexposed, dark, moody, Photorealistic, Shallow depth of field, natural lighting, golden hour, warm color palette, high contrast, automotive photography, tack sharp, glossy texture, muted brown earth tones, low angle perspective, rustic urban landscape.")

Style and Formatting Instructions:

- Fuse all answers into a single, coherent string.
- Do not use semantic labels such as "Stylistic properties:".
- Use full sentences for overall descriptions, and tightly punctuated keywords for visual style.
- Do not start with "The image depicts" or "This image has". Go straight to the content.
- Avoid phrases like "suggesting that", "potentially", "might be". Be visually confident.

- Describe all identifiable objects: position, size, color, material, orientation, relation to others.
- Describe stylistic traits: lighting, color grading, rendering method, medium, realism level.
- Format output as a single dense caption. Use periods or commas, but no lists or line breaks.

Output Format: One highly descriptive string. No list. No section labels. No bullets. No line breaks.
Examples:
Example 1: Photo of a sleek, grey Ferrari F8 parked in a narrow cobblestone alleyway of an old Italian village. The car is positioned in the center of the frame, facing slightly to the right. The background features rustic buildings with weathered, beige plaster walls and wooden shutters. A building to the left has a wooden door and an arched entrance partially covered by ivy. A vibrant red bougainvillea climbs up the left building, and a green bush with yellow flowers is visible next to it. In the distant background, there's a hillside with dense greenery. The foreground includes out-of-focus branches with yellow leaves, framing the image. Photo, underexposed, dark mood, medium depth of field, soft natural lighting, golden hour, warm color palette, photorealistic, high contrast, automotive photography, tack sharp, glossy texture, nostalgic, serene, visually balanced.
Example 2: Ethereal 3D image of a character playing on a piano suspended in a blue dreamlike environment. Scene from Pixar's animated movie "Soul". The character, a man wearing a hat and glasses, is seated on a stool to the left of the piano. The piano is a glossy black grand piano, centrally positioned, with its lid open. The scene is bathed in vibrant purple and blue lighting, creating an ethereal and otherworldly atmosphere. The background is filled with abstract light patterns and gradients, enhancing the surreal feel of the image. Pixar animation, medium depth of field, soft diffused lighting, neon color palette, photorealistic textures, high contrast, ethereal, whimsical, visually balanced, dynamic composition, dramatic lighting effects, digital animation.

Video Annotation Prompt for GPT-4o

You are describing a video represented by frames extracted at a rate of one frame per second. Based on these frames, provide detailed captions in English from the following aspects:
1. Short Caption: Summarize the video in one detailed sentence, capturing key actions and the overall mood.
2. Background Caption: Describe the background, including objects, location, weather, time, and dynamic elements like movements.
3. Main Object Caption: Describe the main subject's actions, attributes, interactions, and movements across the frames, including changes in posture, expression, or speed.
4. Reference Caption: Provide a detailed, dense caption (around 250 words) describing all visible actions, environmental details, and emotional atmosphere. Use a structured approach covering:

- Subject
- Subject actions
- Environment and background
- Visual language (style, composition, lighting)

- Camera language (movement, angles, focal length, shot sizes)

Highlight the mood and tone, and create a vivid narrative rich enough for AI to recreate the video.
5. Standard Summary: Provide a concise, approximately 100-word summary that highlights the main actions, key subjects, and important environmental details.
6. Style Tags: Provide a single, comma-separated string of tags (at least 5) that includes video types, video style, and any relevant attributes.
7. Key Tags: Provide a single, comma-separated string of tags including key objects, people, or entities (3–5), location, time, environment (2–3), notable qualities (2–4), video style, and camera techniques (1–2).

Important Camera Work Requirement:
For all sections, descriptions related to the camera work (including shot types, camera angles, and camera movements) should primarily reference the following camera caption provided by the user:
`{camera_caption_only}`
This includes all mentions of how the video is framed, how the camera moves, and any stylistic elements related to the camera language. If the provided camera caption contains significant errors or inconsistencies, you may adapt the descriptions as needed, ensuring they remain accurate and cohesive. Avoid forcing unnecessary mentions of camera techniques.

Important Guidelines:

- Avoid describing each frame individually or using phrases like “first frame”.
- Do not start with “The scene...”, “In this video...”, etc. Write in vivid, flowing narrative form.
- For *Reference Caption* and *Standard Summary*, avoid any reference to "the video".
- Be cohesive and immersive. Avoid short descriptive fragments; instead use continuous, vivid narration.
- Strictly follow the format with all 7 sections labeled.
- Never decline a description. If objects or individuals are unidentifiable, describe their visual features or behavior.
- Only include camera work details if they align with the provided camera caption or visual evidence.

Poster Annotation Prompt for GPT-4o

You are an AI assistant specialized in analyzing Poster images and converting them into a structured markdown format. You need to provide a detailed caption for an English poster. The main content of the poster includes text and non-text elements. Based on these elements, provide a concise English poster description in the order in which they appear in the poster, and describe them according to the following requirements:
Instructions:

1. Describe the textual and visual elements in the order they appear in the poster:

- For textual content: specify the font type (e.g., Heiti, Songti, Kaiti, Dengxian, handwriting), font color, font size, position (e.g., top, bottom, top-left, bottom-right, center), alignment (e.g., centered, left-aligned, right-aligned), whether it is obstructed, and layout characteristics (e.g., vertical, horizontal).

- For visual elements: describe their properties (e.g., color, shape, size, dynamic or static, texture), position (e.g., centered, dispersed), layering with respect to text or other elements, and any decorative effects (e.g., border, shadow, gradient, texture).

2. Describe the layout and interaction between text and visual elements, including their spatial relationships (e.g., overlap, separation, symmetry).
3. Provide an overall assessment of the poster's style (e.g., bright, minimalistic, vintage, modern, tech-oriented, natural, artistic).
4. Avoid speculating on the poster's topic, narrative, or intention—focus solely on visual and structural features.
5. Keep the description concise and accurate, focusing on visual aspects. Do not include unrelated content.

GUI Annontation Prompt for GPT-4o

You are an AI assistant specialized in analyzing Graphical User Interface (GUI) images and converting them into structured markdown format. GUI images often contain background, navigation, interaction, visual and text information, layout, and icons. Your task is to: Provide detailed annotation of the Graphical User Interface (GUI) image. Based on the GUI's visual, text elements and layout, provide detailed descriptions in the following aspects:
1. **Brief description**: Summarize the GUI's main purpose and content in one concise but specific sentence.
2. **Detailed extraction**:

- If the GUI image contains background elements, describe background elements (e.g., colors, images, dynamic elements, and so on).
- Extract all elements from right to left and from top to bottom.
- Extract the content of the GUI image in detail and completely, without missing any part.
- Don't miss any text that appears on the image.
- Use markdown format.

3. **Description of interactive elements**: If the GUI image includes interactive elements (e.g., search boxes, buttons, and so on), describe them and their functionality and usage.
4. **Overall description**: Provide a summary of about 100 words, summarizing the main functions and usage scenarios of the GUI display page.
Instructions:

- Please structure your response as follows: 1. Brief description, 2. Detail extraction, 3. Description of interactive elements, 4. Overall description.
- Ensure that the layout and visual language mentioned in all sections are consistent with this description.
- If you find major errors or inconsistencies in the description, you can adjust it as needed, but you must ensure accuracy and consistency.
- Please provide content strictly in the specified format, ensuring that all 4 sections are covered.

- Do not refuse any description request, even if the specific content cannot be identified, describe elements of the GUI image by inferring the characteristics.
- Ensure that the text description and visual language are consistent, but do not over-emphasize certain details or repeat content.

Caption Extension Prompt for Medium Caption Using Qwen2.5-32B-Instruct

Task: Summarize the following long caption into a medium-length caption.
The medium caption should be shorter than the long caption. It should retain the key information from the long caption while improving clarity and brevity.
Input: Long Caption: {caption}

Caption Extension Prompt for Short Caption Using Qwen2.5-32B-Instruct

Task: Summarize the following medium caption into a short-length caption.
The short caption should be shorter than the long caption. It should retain the key information from the short caption while improving clarity and brevity.
Input: Medium Caption: {caption}

Caption Extension Prompt for Key Tags Using Qwen2.5-32B-Instruct

You are given a detailed caption in English. Your task is to extract key tags (keywords) from the caption that capture the main concepts or themes.
Summarize the key tags. Each set of tags should be concise and represent the core ideas of the caption.
Use the following JSON format for your output:
`{"tag1, tag2, tag3,..."}`
The provided caption: {caption}

Caption Translation Prompt Using Qwen2.5-32B-Instruct (English to Chinese)

You are given a detailed caption in English. Your task is to translate this detailed English caption to a Chinese caption that preserves the meaning and visual richness of the original.
The provided caption: {caption}

Caption Extension Prompt for Detailed Analysis of Table (Qwen2-VL-76B-Instruct)

Please help me analyze the table image and the corresponding LaTeX code. The provided LaTeX code represents the structure of the table.
Provided LaTeX Code: {Latex}

Caption Extension Prompt for Detailed Analysis of Equation (Qwen2-VL-76B-Instruct)

Please help me analyze the equation image and the corresponding LaTeX code. The provided LaTeX code represents the structure of the equation.
Provided LaTeX Code: {Latex}

Caption Extension Prompt for Detailed Analysis of Chart (Qwen2-VL-76B-Instruct)

Please help me analyze this image of chart and corresponding markdown in detail.
Provided markdown format of this chart image: {markdown}

D System Prompt Example for OMNICAPTIONER

Fig. 8 presents different system prompts used in OMNICAPTIONER for various image types. It categorizes prompts into three sections: natural images, visual text images, and structured images. These prompts guide the model's captioning style and task-specific adaptations.

E Caption Visualization

As illustrated in Fig. 9 to Fig. 15, we present a comprehensive visualization of captioning results across multiple tasks using OMNICAPTIONER, including natural images, table images, chart images, math images, poster images, and videos. For natural images, we demonstrate the impact of different system prompts on caption generation, showcasing how specific prompts can elicit world knowledge in the model's responses in Fig. 16. In the case of structured images from Fig. 17, different system prompts lead to distinct stylistic variations in captioning, reflecting the adaptability of the model to various formatting requirements. Additionally, we visualize how OmniCaptioner-generated captions can enhance DeepSeek-R1-Distill-LLaMA-70B in Fig. 18, Fig. 19 and Fig. 20, enabling it to tackle visual tasks more effectively. These visualizations highlight the versatility and robustness of OMNICAPTIONER in handling diverse multimodal data, demonstrating its potential for improving vision-language understanding.

F Text-to-Image Generation

The visualization from Fig. 21 demonstrates that OMNICAPTIONER's detailed captions significantly enhance the text-to-image (T2I) alignment in models like SANA 1.0 (Xie et al., 2025). By providing precise and richly descriptive textual caption, the generated images exhibit improved fidelity to the original prompts. We also present some image conversion examples in Fig. 22 to illustrate the pixel-to-word ability of our OMNICAPTIONER. All the generated images shown above are produced by the generation model trained on image data labeled by OMNICAPTIONER, fine-tuned using SANA 1.0 with 1.6B parameters.

Natural Image (Data Source: In-house)		
Detailed_en (1.9M)	Detailed_zh (1.9M)	Medium_en (650K)
Medium_zh (642K)	Short_en (350K)	Short_zh (324K)
Tag_en (200K)	Tag_zh (169K)	
Nature Image (Data Source: BLIP3Kale)		
Detailed_en (55K)	Detailed_zh (132K)	Medium_en (483K)
Medium_zh (465K)	Short_en (1.6M)	Short_zh (1.3M)
Nature Image (Data Source: DenseFusion)		
Detailed_en (432K)	Detailed_zh (348K)	Medium_en (19K)
Medium_zh (13K)	Short_en (1K)	Short_zh (2k)
Table (Data Source: arXiv, MMtab)		
To LaTeX(2.8M)	Detailed_en (168K)	Detailed_zh (34K)
Chart (Data Source: arXiv, Tinychart)		
To Markdown (767K)	Detailed_en (253K)	Detailed_zh (79K)
Math-Equation (Data Source: arXiv)		
To LaTeX (3M)	Detailed_en (382K)	Detailed_zh (141K)
Math-Geometry (Data Source: Self-generated, Mavis, Autogeo)		
To LaTeX (102K)	Detailed_en (300K)	
Poster (Data Source: In-house)		
OCR (82K)	Detailed_en (134K)	Detailed_zh (98K)
PDF and UI (Data Source: MultiUI, arXiv)		
UI Caption (709K)	PDF OCR (47K)	Pure Text OCR (2M)
Video (Data Source: Openvid, Pandas)		
Detailed_en (600K)	Tag (600K)	Main Object (600K)
Style (600K)	Medium (600K)	Short (600K)
Background (600K)	Camera (600K)	

Figure 7: Dataset composition for pretraining OMNICAPTIONER.

System Prompt For Natural Images:

"Detailed" :
You are a helpful assistant focused on providing detailed descriptions and background information for the generated images. Analyze the given image and generate a comprehensive caption that includes the visual style, spatial relationships between elements, texture details, descriptions of the main objects, and relevant world knowledge to enhance understanding.

"Medium" :
You are a helpful assistant specialized in creating medium-length captions for the generated images. Analyze the provided image and generate a caption that captures the key visual elements, while maintaining clarity and coherence.

"Short" :
You are a helpful assistant focused on creating short captions for the generated images. Analyze the provided image and generate a concise caption that highlights the main subject.

"Tag":
You are a helpful assistant specialized in generating key tags for the generated images. Analyze the provided image and create a list of relevant tags that capture the main subjects, themes, and notable elements;

"Detailed_Zh" :
你是一位专注于提供详细描述和背景信息的助手。分析给定的生成图像，生成一个全面的描述，包含视觉风格、元素之间的空间关系、纹理细节、主要对象的描述，以及增强理解的相关背景知识。

"Medium_Zh" :
你是一位专注于创建中等长度图像描述的助手。分析所提供的生成图像，生成一个描述，捕捉关键视觉元素，保持清晰和连贯。

"Short_Zh" :
你是一位专注于创建简短图像描述的助手。分析提供的生成图像，生成一个简洁的描述，突出主要主体。

"Tag_Zh":
你是一位专注于为图像生成关键词标签的助手。分析提供的生成图像，创建一个相关标签列表，捕捉主要主题、元素和显著特点

"Detailed_Natural" :
You are a helpful natural image captioner. Provide a comprehensive description of the natural image, including the main subject, background elements, lighting conditions, color distribution, textures, spatial arrangement, and any potential dynamic context.

"Medium _Natural" :
You are a helpful natural image captioner. Describe the main content, background in the medium-length text.

"Short _Natural" :
You are a helpful natural image captioner. Describe the main content, background in the short-length text.

"Detailed _Natural _Zh" :
您是一位乐于助人的自然图片分析助手。请提供自然图像的全面描述，包括主要主题、背景元素、光照条件、颜色分布、纹理、空间排列以及任何潜在的背景。

"Medium _Natural _Zh" :
您是一位乐于助人的自然图片分析助手。请用中等长度的文本描述主要内容和背景。

"Short _Natural _Zh" :
您是一位乐于助人的自然图片分析助手。请用短文本描述主要内容和背景。

System Prompt For Visual Text Images:

"Visual_Text" :
You are an advanced model designed to accurately analyze the image with text items. You can describe the text information and visual information in the image, including font style, size, color, background, text layout and other visual objects in detail.

"UI" :
You are analyzing a UI webpage layout. Provide a detailed caption describing the layout's structure, including the arrangement, style, and functionality of key components such as buttons, navigation bars, input fields, and visual elements.

"OCR" :
You are an advanced OCR model designed to accurately extract text from images. Your task is to analyze the provided image and return the text in a clear, readable format.

"Visual_Text_Zh" :
你是一个精确分析文本内容图像的先进助手。你可以详细描述图像中的文本信息和视觉信息，包括字体样式、大小、颜色、背景、文本布局和其他视觉对象

System Prompt For Structured Images:

"Chart" :
You are a professional data visualization analyst. Given a chart image, first accurately perform OCR on any textual and numeric content (including titles, legends, axes, labels, and annotations), and you can convert it into Markdown format, then structure and analyze the extracted data to identify key trends and insights.

"Table" :
You are a data conversion and extraction expert. Given a table image, you can convert it into CSV, HTML, Markdown or LaTeX formats, then extract and summarize the key relationships or insights from the data.

"Equation" :
You are an equation analysis expert. Given an equation image, you can convert it into proper LaTeX format, then summarize any key mathematical properties, patterns, or insights it conveys.

"Geometry":
You are a geometry analysis expert. Given a geometric figure, you can convert it into a corresponding LaTeX representation, then provide insights or interpretations about the structure or properties.

"Structured_zh":
你是一个高级模型，旨在分析和解释图表、图形、表格、数学公式、数学几何图和数据可视化。你的任务是从提供的图像中提取相关信息，并以机器可读格式或结构化格式返回文本。

Figure 8: Different system prompts used for OMNICAPTIONER.

System Prompt Type: "Detailed"

Aerial photograph of a traditional Chinese temple complex perched on a rugged mountain landscape, likely during a festival. The central focus is a large, ornate pagoda with a golden roof, surrounded by smaller pagodas and pagoda-style structures, all intricately detailed with red and gold accents. Crowds of people are visible throughout the scene, moving along pathways and gathered around the buildings, indicating a bustling atmosphere. The terrain is rocky and partially covered with snow, with steep cliffs and narrow pathways leading to and around the structures. In the background, a series of distant mountains create a dramatic backdrop, with a soft blue hue from the sky. The lighting is warm and golden, suggesting a sunrise or sunset, casting long shadows and highlighting the textures of the rocks and the architectural details. High contrast, vibrant colors, dynamic composition, cultural richness, serene yet lively, visually stunning, panoramic perspective.

System Prompt Type: " Medium"

Aerial view of a traditional Chinese temple complex on a snowy mountain, featuring pagoda-style buildings with golden roofs and red accents. Crowds fill the temple grounds and surrounding pathways, with a large pagoda at the center. The scene includes rugged, snow-covered mountains and distant peaks under a clear blue sky, captured in vibrant colors with high contrast and natural lighting, creating a serene yet dynamic atmosphere.

System Prompt Type: " Short"

An aerial view of a bustling temple complex with pagoda-style buildings perched on rugged mountain peaks, surrounded by a snow-dusted landscape and bathed in warm, golden hour light, creating a serene yet dynamic scene.

System Prompt Type: " Tag"

Aerial view, mountainous landscape, temple complex, golden pagodas, rocky terrain, vibrant colors, cultural significance, dramatic perspective, natural lighting, high contrast, serene atmosphere, expansive scale, architectural photography.

System Prompt Type: "Detailed_zh"

空中俯瞰一座历史悠久的寺庙群，它坐落在崎岖的悬崖边上，四周是白雪皑皑的山脉。寺庙的建筑风格传统，屋顶铺着黄色瓦片，屋檐装饰华丽，主要建筑位于悬崖中央，两侧是通往山顶的楼梯。寺庙周围聚集了许多人，显示出这是一个热闹的场景。悬崖上覆盖着雪斑，与寺庙的暖色调形成鲜明对比。远处的山脉在柔和的蓝色天空中逐渐消失，暗示着这是清晨或黄昏的光线。高角度视角，丰富的色彩搭配，自然与建筑元素的完美融合，宁静而又充满活力的氛围，清晰的细节，自然与人类建筑的和谐共存。

System Prompt Type: "Medium_zh"

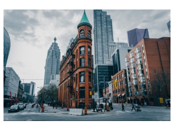
空中俯瞰一座历史悠久的寺庙群，寺庙群坐落在崎岖的悬崖上，四周被群山环绕。寺庙的屋顶是金色的，墙壁是红色的，寺庙通过一座桥与悬崖相连。一条蜿蜒的小径通向山顶的两个金色塔楼，而一群身着传统服饰的人们正朝着寺庙前进。白雪覆盖了地面，与岩石的土色调形成对比。这一幕捕捉了黄昏时分的宁静而壮丽的氛围，突出了寺庙的宏伟和自然景观的美丽。

System Prompt Type: "Short_zh"

空中俯瞰一座历史悠久的寺庙群，寺庙屋顶铺着金色瓦片，建筑色彩斑斓，坐落在崎岖的悬崖上，四周环绕着白雪皑皑的山脉，一条蜿蜒的小径通向山顶，整个场景在金色时刻的光线下显得格外迷人。

System Prompt Type: "Tag_zh"

寺庙, 悬崖, 山脉, 黄昏, 空中视角, 传统建筑, 人群, 鲜艳色彩, 自然美景, 宁静



System Prompt Type: "Detailed"

Urban street scene featuring the iconic Flatiron Building in Toronto, positioned centrally with its distinctive red brick facade and green roof. The building is flanked by modern skyscrapers, including a tall, glass skyscraper to the left and a dark, reflective building to the right. In the foreground, a few pedestrians cross the street, with one pushing a stroller on the right. The street is lined with orange traffic cones and signs, indicating construction or roadwork. The background includes a cloudy sky and a mix of architectural styles, from modern to classic. The scene is captured in natural lighting with a cool color palette, emphasizing the contrast between the historic and contemporary urban elements. Wide-angle perspective, high contrast, sharp focus, architectural photography, dynamic composition, urban exploration, visually balanced.

System Prompt Type: " Medium"

A wide-angle shot of Toronto's Flatiron Building, a historic red-brick structure with a green roof, stands out against modern skyscrapers. Pedestrians cross the street in front, while a black car and construction barriers are visible in the foreground. The overcast sky and cool color palette create a balanced, dynamic urban scene with a mix of historic and contemporary architecture.

System Prompt Type: " Short"

A wide-angle urban photograph captures Toronto's Flatiron Building, a historic red-brick structure with a green spire, surrounded by modern skyscrapers, with a few pedestrians and a black car in the foreground, set against an overcast sky.\n", "Tag": "Flatiron Building, Toronto, urban landscape, historic architecture, modern skyscrapers, city street, overcast sky, construction, pedestrians, traffic cones, vibrant reds, cool tones, dynamic composition, architectural photography.

System Prompt Type: " Tag"

Flatiron Building, Toronto, urban landscape, historic architecture, modern skyscrapers, city street, overcast sky, construction, pedestrians, traffic cones, vibrant reds, cool tones, dynamic composition, architectural photography

System Prompt Type: "Detailed_zh"

城市街道场景，以一座历史悠久的红砖建筑为中心，建筑顶部有一个绿色的圆锥形塔楼，位于画面的右侧。这座建筑拥有许多窗户和黑色的消防逃生梯。在画面的左侧，一座现代的玻璃建筑与一座具有独特绿色尖顶的摩天大楼形成对比。在画面的右侧，一座现代的砖砌建筑，窗户很大，与历史建筑相得益彰。街道上人来人往，有些人正在过马路，有些人站在人行道上。交通锥和路标被放置在街道交叉口。天空阴沉，给场景带来了一种沉闷的灰色调。城市摄影，中等景深，自然漫射光，冷色调，历史与现代建筑的对比，动态构图，平衡的框架，轻微的饱和度，城市氛围。

System Prompt Type: "Medium_zh"

城市景观，以一座历史悠久的红砖平房建筑为中心，建筑顶部为绿色圆锥形，背景是现代摩天大楼。场景中包括行人、推着婴儿车的人、停放的汽车和施工围栏，天空阴沉。照片捕捉到了新旧建筑风格的融合，以及一种忧郁、高对比度的氛围，细节清晰。

System Prompt Type: "Short_zh"

多伦多的标志性建筑——熨斗大厦，矗立在现代摩天大楼之中，街道场景热闹非凡，行人络绎不绝，施工围挡随处可见，这一切都映衬在阴云密布的天空下，构成了一幅充满活力且对比强烈的都市画卷。

System Prompt Type: "Tag_zh"

"城市景观, 平铁大楼, 现代摩天大楼, 建筑对比, 阴天, 城市摄影, 历史与现代建筑

Figure 9: Natural image captioning.

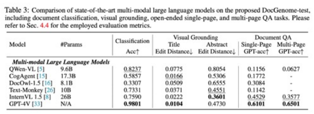


Table 3: Comparison of state-of-the-art multi-modal large language models on the proposed DocGenome-test, including document classification, visual grounding, open-ended single-page, and multi-page QA tasks. Please refer to Sec. 4.4 for the employed evaluation metrics.

Model	#Params	Classification	Visual Grounding		Document QA	
			Title	Abstract	Single-Page	Multi-Page
		Acc↑	Edit Distance↓	Edit Distance↓	GPT-acc↑	GPT-acc↑
Multi-modal Large Language Models						
QWen-VL [5]	9.6B	0.8237	0.0775	0.8054	0.1156	0.0627
CogAgent [15]	17.3B	0.5857	0.0166	0.5306	0.1772	-
DocOwl-1.5 [16]	8.1B	0.3307	0.0509	0.6555	0.3084	-
Text-Monkey [26]	10B	0.7331	0.0371	0.4551	0.1142	-
InternVL 1.5 [8]	26B	0.7590	0.0222	**0.3601**	0.4529	0.3577
GPT-4V [33]	N/A	**0.9801**	**0.0104**	0.4730	**0.6101**	**0.6501**

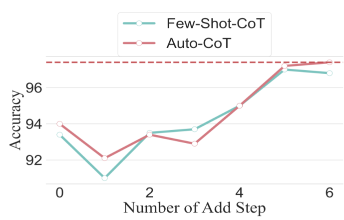
System Prompt Type: "Table"

```
\begin{table}[t]
\centering
\captionsetup{width=1.0\linewidth}
\caption{Table 3: Comparison of state-of-the-art multi-modal large language models on the proposed DocGenome-test, including document classification, visual grounding, open-ended single-page, and multi-page QA tasks. Please refer to Sec.~\textcolor{red}{4.4} for the employed evaluation metrics.}
\vspace{-0.2cm}
\setlength{\tabcolsep}{1.8mm}{
\begin{tabular}{ll|c|cc|cc}
\toprule
\multirow{3}{*}{\textbf{Model}} & \multirow{3}{*}{\textbf{\#Params}} & \textbf{Classification} & \multicolumn{2}{c}{\textbf{Visual Grounding}} &
\multicolumn{2}{c}{\textbf{Document QA}} \\
& & \textbf{Acc$\uparrow$} & \textbf{Title} & \textbf{Abstract} & \textbf{Single-Page} & \textbf{Multi-Page} \\
& & & \textbf{Edit Distance$\downarrow$} & \textbf{Edit Distance$\downarrow$} & \textbf{GPT-acc$\uparrow$} & \textbf{GPT-acc$\uparrow$} \\
\midrule
\multicolumn{7}{c}{\textbf{\textit{Multi-modal Large Language Models}}} \\
QWen-VL~[\textcolor{cyan}{5}] & 9.6B & \underline{0.8237} & 0.0775 & 0.8054 & 0.1156 & 0.0627 \\
CogAgent~[\textcolor{cyan}{15}] & 17.3B & 0.5857 & \underline{0.0166} & 0.5306 & 0.1772 & - \\
DocOwl-1.5~[\textcolor{cyan}{16}] & 8.1B & 0.3307 & 0.0509 & 0.6555 & 0.3084 & - \\
Text-Monkey~[\textcolor{cyan}{26}] & 10B & 0.7331 & 0.0371 & \underline{0.4551} & 0.1142 & - \\
InternVL 1.5~[\textcolor{cyan}{8}] & 26B & 0.7590 & 0.0222 & \textbf{0.3601} & \underline{0.4529} & \underline{0.3577} \\
GPT-4V~[\textcolor{cyan}{33}] & N/A & \textbf{0.9801} & \textbf{0.0104} & 0.4730 & \textbf{0.6101} & \textbf{0.6501} \\
\bottomrule
\end{tabular}
}

\vspace{-0.3cm}
\end{table}
```

Team	Chassis	Engine	Tires	No	Drivers
Chip Ganassi Racing	Reynard 98i	Honda	Firestone	1	Alex Zanardi
				12	Jimmy Vasser
Marlboro Team Penske	Penske PC27-98	Mercedes	Goodyear	2	Al Unser, Jr.
				3	André Ribeiro
Walker Racing	Reynard 98i	Honda	Goodyear	5	Gil de Ferran
Newman-Haas Racing	Swift 009.c	Ford XB	Goodyear	6	Michael Andretti
				11	Christian Fittipaldi Roberto Moreno
Team Rahal	Reynard 98i	Ford XB	Firestone	7	Bobby Rahal
				8	Bryan Herta
Hogan Racing	Reynard 98i	Mercedes	Firestone	9	JJ Lehto
Della Penna Motorsports	Swift 009.c	Ford XB	Firestone	10	Richie Hearn
				43	Hideshi Matsuda
Project Indy	Reynard 97i	Ford XB	Goodyear	15	Roberto Moreno Domenico Schiattarella
Bettenhausen Racing	Reynard 98i	Mercedes	Goodyear	16	Hélio Castroneves
PacWest Racing Group	Reynard 98i	Mercedes	Firestone	17	Maurício Gugelmin
				18	Mark Blundell
Payton/Coyne Racing	Reynard 98i	Ford XB	Firestone	19	Michel Jourdain, Jr.
				34	Dennis Vitolo Gualter Salles
Patrick Racing	Reynard 98i	Ford XB	Firestone	20	Scott Pruett
				40	Adrián Fernández
Tasman Motorsports Group	Reynard 98i	Honda	Firestone	21	Tony Kanaan
Arciero-Wells Racing	Reynard 98i	Toyota	Firestone	24	Hiro Matsushita Robby Gordon
				25	Max Papis
Team KOOL Green	Reynard 98i	Honda	Firestone	26	Paul Tracy
				27	Dario Franchitti
Forsythe Racing	Reynard 98i	Mercedes	Firestone	33	Patrick Carpentier
				99	Greg Moore
All American Racing	Reynard 98i Eagle 987	Toyota	Goodyear	36	Alex Barron
				98	P. J. Jones Vincenzo Sospiri
Davis Racing	Lola T98/00	Ford XB	Goodyear	77	Arnd Meier

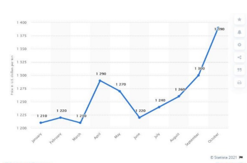
System Prompt Type: "Table"

```
当然！让我们逐步分析这个表格。

#### 表格概述
该表格列出了各种赛车队及其相应的车架、引擎、轮胎品牌以及他们的驾驶员名单。以下是各列的详细介绍：

1. **车队**：赛车队的名字。
2. **车架**：赛车使用的特定车型或底盘。
3. **引擎**：赛车所搭载的引擎类型。
4. **轮胎**：赛车所使用轮胎的品牌。
5. **编号（No）**：赛车号码。
6. **驾驶员**：与每个车队相关联的驾驶员名字，包括他们的国籍国旗。

#### 详细分析

##### 车队名称和车架
- 大多数车队使用Reynard 98i车架，有少数车队使用Penske PC27-98、Swift 009.c等其他车架。
- Reynard 98i是大多数车队中最常用的车架。

##### 引擎类型
- 大多数车队使用Ford XB引擎，其次是Mercedes和Toyota。

##### 轮胎品牌
- Firestone是最常见的轮胎品牌，其次是Goodyear。

##### 驾驶员信息
- 每个驾驶员都与一个国家相关联，这在“编号”列中的国旗上有所体现。
- 国籍从美国到巴西不等，表明了国际化的赛车运动社区。

#### 特定观察
- Chip Ganassi Racing使用Reynard 98i车架搭配Honda引擎和Firestone轮胎，驾驶员为Alex Zanardi和Jimmy Vasser。
- Marlboro Team Penske使用Penske PC27-98车架搭配Mercedes引擎和Goodyear轮胎，驾驶员为Al Unser, Jr.和André Ribeiro。
- Della Penna Motorsports使用Swift 009.c车架搭配Ford XB引擎和Firestone轮胎，驾驶员为Richie Hearn和Hideshi Matsuda。

#### 结论
此表提供了关于不同赛车队、他们使用的设备和驾驶员的信息快照。它突出了某些车架和引擎的普遍性，同时也展示了来自不同国家的驾驶员多样性。这些数据可用于分析赛车运动中使用的车辆和驾驶员的分布情况。
```



System Prompt Type: "Chart"

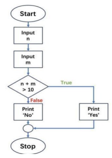
```
The title of the chart is not explicitly mentioned, but it appears to be related to "Few-Shot-CoT" and "Auto-CoT."

#### Axes:
- **X-Axis (Horizontal)**: Labeled as "Number of Add Step," ranging from 0 to 6.
- **Y-Axis (Vertical)**: Labeled as "Accuracy," with values ranging from 92 to 98.

#### Data Series:
There are two lines representing different methods or models:
1. **Few-Shot-CoT**: Represented by a teal line.
2. **Auto-CoT**: Represented by a red line.

#### Key Observations:

##### Few-Shot-CoT Line:
- At \( x = 0 \), the accuracy starts at around 93.
- There is a significant drop in accuracy between \( x = 0 \) and \( x = 2 \), where the accuracy drops sharply to just below 92.
- From \( x = 2 \) onwards, there is a steady increase in accuracy, reaching close to 97 by \( x = 6 \).

##### Auto-CoT Line:
- At \( x = 0 \), the accuracy starts slightly higher than Few-Shot-CoT, around 94.
- The accuracy fluctuates less compared to Few-Shot-CoT.
- By \( x = 6 \), the accuracy reaches its highest point, close to 98.

#### Analysis:
1. **Initial Performance**:
   - Both methods start relatively high, indicating good initial performance before any additions were made.

2. **Impact of Add Steps**:
   - For Few-Shot-CoT, the sharp decline after adding one step suggests that the addition of steps significantly impacts its performance negatively.
   - Auto-CoT shows less impact on its performance when additional steps are added, maintaining a relatively stable trend despite fluctuations.

3. **Final Performance**:
   - By \( x = 6 \), both methods show high accuracy, with Auto-CoT slightly outperforming Few-Shot-CoT.

#### Conclusion:
The chart illustrates how the number of add steps affects the accuracy of two different methods, Few-Shot-CoT and Auto-CoT. While Few-Shot-CoT experiences a dramatic decrease followed by recovery, Auto-CoT maintains a more consistent performance across all steps.
```



System Prompt Type: "Chart"

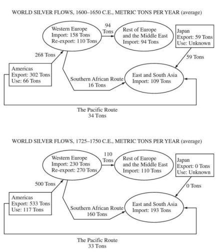
```
| Month | Price (U.S. Dollars/ton) |
|-------|------------------------|
| Jan | 1210 |
| Feb | 1220 |
| Mar | 1210 |
| Apr | 1290 |
| May | 1270 |
| Jun | 1220 |
| Jul | 1240 |
| Aug | 1260 |
| Sep | 1300 |
| Oct | 1390 |
```

Figure 10: Table/Chart image captioning.



System Prompt Type: "Detailed"

Flowchart illustrating a simple conditional process. The flowchart begins with an oval labeled "Start" at the top. Two rectangular boxes follow, labeled "Input n" and "Input m," connected by arrows pointing downwards. A diamond-shaped decision box is positioned below, reading "n + m > 10." This decision box has two outcomes: "True" and "False." If "True," the flowchart moves to a rectangle labeled "Print 'Yes'" on the right side. If "False," it moves to a rectangle labeled "Print 'No'" on the left side. The "False" outcome is connected to a circle labeled "Stop" at the bottom. The "True" outcome is connected to another rectangle labeled "Print 'Yes'" on the right side. The flowchart uses a combination of blue and green text for clarity. Clean, minimalist design, clear labeling, structured layout, functional, educational, technical illustration.



System Prompt Type: "OCR"

WORLD SILVER FLOWS, 1600–1650 C.E., METRIC TONS PER YEAR (average)

- **Western Europe**
  - Import: 158 Tons
  - Re-export: 110 Tons

- **Rest of Europe**
  - Import: 94 Tons

- **Japan**
  - Export: 59 Tons
  - Use: Unknown

- **Americas**
  - Export: 302 Tons
  - Use: 66 Tons

- **Southern African Route**
  - 16 Tons

- **East and South Asia**
  - Import: 109 Tons

- **The Pacific Route**
  - 34 Tons

WORLD SILVER FLOWS, 1725–1750 C.E., METRIC TONS PER YEAR (average)

- **Western Europe**
  - Import: 230 Tons
  - Re export: 270 Tons

- **Rest of Europe**
  - Import: 110 Tons

- **Japan**
  - Export: 0 Tons
  - Use: Unknown

- **Americas**
  - Export: 533 Tons
  - Use: 117 Tons

- **Southern African Route**
  - 160 Tons

- **East and South Asia**
  - Import: 193 Tons

- **The Pacific Route**
  - 33 Tons

System Prompt Type: "Visual_Text_zh"

海报描述：
1. 顶部左上角为手写字体文字"taopii"，白色，横排，居中对齐。
2. 背景整体为红色主色调，呈现简约风格。
3. 中间部分有三个拱形窗框，从左至右依次为蓝色、橙色和绿色背景，每个窗框内都有一个剪影人物正在阅读书籍，分别为蓝白配色、红黄配色和绿白配色，静态设计，具有装饰性效果。
4. 窗框下方中央位置为主标题文字："戀愛"二字采用较大的黑体字，白色，竖直排列；右侧是较小字号的副标题" 愛的浪漫周代製造"，字体类型与主标题一致但字号更小，同样为白色，竖直排列。
5. 主标题左侧有一列竖排文字："自由恋愛/日本統治時代にけるロマンス"，字体为细黑体，白色，竖排，位于中间偏下位置。
6. 左侧中部有时间信息，"2019年10月16日"和"2020年5月17日"，字体为细黑体，白色，竖排，分两行排列，靠近页面边缘。
7. 底部为一行英文文字"Romantic love in the modern"，字体为无衬线字体，白色，横排，居中对齐。
8. 最底部为多行中文及标识信息，字体为细黑体，白色，横排，居中对齐，包含多个品牌或合作方标志，以及"@辛未设计"的字样，字体略大，位于最右边。

整体风格：现代感十足，色彩鲜明，以红色为主基调，搭配简洁的几何元素和插画形式，营造出活泼且艺术化的氛围。

System Prompt Type: "Detailed"

Monochrome poster design featuring bold text with an abstract backdrop. The central focus is the phrase "ISLAND TALES: TAIWAN AND AUSTRALIA" prominently displayed across the middle in large serif font. Above it, smaller text reads "ISL," partially obscured by torn paper effects revealing underlying layers. To the left, vertically stacked text states "ISLAND TALES," followed by "TAIWAN AND AUSTRALIA." On the right side, Chinese characters are aligned vertically within a black rectangle labeled "TAIPEI - PERTH." Below these texts, dates "2019 11.16" and "2020 3.1" appear alongside location information "GALLERY 2A, TAIPEI FINE ARTS MUSEUM." The bottom right corner features small branding logos. The background consists of layered textures resembling aged paper or fabric, creating depth through varying shades of gray and white. Graphic design, high contrast, minimalist aesthetic, vintage-inspired typography, textured overlay, balanced composition.

System Prompt Type: " Visual_Text_zh"

海报描述：
1. 顶部左侧文字"ISLAND TALES"，使用黑体大写字母，白色字体，黑色矩形底框衬托，横排居左对齐。下方为两行小字"TAIWAN AND AUSTRALIA"，同样采用黑体大写字母，白色字体，黑色矩形底框衬托，横排居左对齐。

2. 右侧竖直排列的文字"小说·双城计画"，使用手写风格字体，白色字体，黑色矩形底框衬托，竖排右对齐。其上方有较小字号文字"TAIPEI DERTH"，使用细线字体，白色字体，与主标题形成对比。

3. 中间部分以撕裂纸张效果呈现不规则形状区域，内含褪色文字"ISLAND TALES"、"TAIWAN AND"及"AUSTRALIA"，字体为较淡的灰色，半透明效果，与背景融合，增加层次感。

4. 底部左侧文字"2019 11.16"至"2020 3.1"，使用黑体数字和日期格式，白色字体，黑色矩形底框衬托，横排左对齐。下方是更小字号文字"GALLERY 2A, TAIWEN FINE ARTS MUSEUM"，同样为黑体字体，白色字体，黑色矩形底框衬托，横排左对齐。

5. 底部右侧标志"台北市立艺术博物馆"，使用简约设计，白色字体，位于底部边缘，横排居中对齐；下方附带社交媒体账号"@辛未设计"，字体极小，白色字体，横排居中对齐。

整体风格：复古风，黑白配色为主，搭配手写风格文字和撕裂纸张纹理，营造出一种质感强烈且富有艺术气息的设计。

Figure 11: Visual-Text image captioning.

$$
\begin{aligned}
&[1+R(\eta)]^{1/4}(\Delta^{(2)}_{00}-4\Psi^{(2)})=A\,\cos[kr_s(\eta)]+B\,\sin[kr_s(\eta)]\\
&-4\frac{k}{\sqrt{3}}\int_0^\eta d\eta'[1+R(\eta')]^{3/4}\left(\Phi^{(2)}(\eta')+\frac{\Psi^{(2)}(\eta')}{1+R}\right)\sin_k[\eta,\eta']\\
&+\frac{\sqrt{3}}{k}\int_0^\eta d\eta'[1+R(\eta')]^{3/4}\left(\mathcal{S}'_\Delta+\frac{\mathcal{H}R}{1+R}\mathcal{S}_\Delta-\frac{4}{3}ik_i\,\mathcal{S}^i_V\right)\sin_k[\eta,\eta']\,,
\end{aligned}
$$

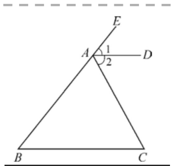
System Prompt Type: "Equation"

```
图像包含一个涉及多个变量和函数的复杂数学方程。方程以清晰易读的形式呈现，每个项都明确分离并组织。以下是方程的详细描述：

\[ [1 + R(\eta)]^{1/4} (\Delta_{00}^{(2)} - 4 \Psi^{(2)}) = A \cos[k r_s(\eta)] + B \sin[k r_s(\eta)] \]

\[ -4 \frac{k}{\sqrt{3}} \int_0^{\eta} d\eta' [1 + R(\eta')]^{3/4} \left( \Phi^{(2)}(\eta') + \frac{\Psi^{(2)}(\eta')}{1 + R} \right) \sin_k[\eta, \eta'] \]

\[ + \frac{\sqrt{3}}{k} \int_0^{\eta} d\eta' [1 + R(\eta')]^{3/4} \left( S'_\Delta + \frac{H R}{1 + R} S_\Delta - \frac{4}{3} i k_i S_V^i \right) \sin_k[\eta, \eta'] \]

### 方程分解：

1. **左端（LHS）：**
 - 左边的第一个项是 \( (1 + R(\eta))^{1/4} \)，其中 \( R(\eta) \) 是 \(\eta\) 的函数。
 - 第二个项是 \( (\Delta_{00}^{(2)} - 4 \Psi^{(2)}) \)，涉及到两个函数 \( \Delta_{00}^{(2)} \) 和 \( \Psi^{(2)} \)。

2. **右端（RHS）：**
 - 右边的第一个部分是三角函数：
 \[ A \cos[k r_s(\eta)] + B \sin[k r_s(\eta)] \]
 这里，\( A \) 和 \( B \) 是常数，而 \( k r_s(\eta) \) 是依赖于 \(\eta\) 的函数。

- 第二个部分是积分项：
 \[ -4 \frac{k}{\sqrt{3}} \int_0^{\eta} d\eta' [1 + R(\eta')]^{3/4} \left( \Phi^{(2)}(\eta') + \frac{\Psi^{(2)}(\eta')}{1 + R} \right) \sin_k[\eta, \eta'] \]
 这个积分涉及到几个组成部分：
 - \( \Phi^{(2)}(\eta') \) 和 \( \Psi^{(2)}(\eta') \) 是 \(\eta'\) 的函数。
 - 积分限从 0 到 \(\eta\)。
 - 涉及到正弦函数 \( \sin_k[\eta, \eta'] \)。

 - 第三个部分也是另一个积分项：
 \[ + \frac{\sqrt{3}}{k} \int_0^{\eta} d\eta' [1 + R(\eta')]^{3/4} \left( S'_\Delta + \frac{H R}{1 + R} S_\Delta - \frac{4}{3} i k_i S_V^i \right) \sin_k[\eta, \eta'] \]
 这个积分与第二个类似，但函数不同：
 - \( S'_\Delta \)，\( S_\Delta \)，和 \( S_V^i \) 是额外的函数或算符。
 - 积分限从 0 到 \(\eta\)。

### 分析：

这个方程看起来是一个复杂的微分或积分方程，可能来自量子力学、场论或其他理论物理领域。\(\sin_k[\eta, \eta']\) 和其他三角函数的存在表明它可能与波动现象有关。\(\int_0^{\eta}\) 表示这些积分是在某个范围内的特定点进行评估的。

### 结论：

这个方程代表了各种函数及其导数之间的关系，很可能用于模拟某些物理系统或理论模型。对这一方程的具体解释将取决于其使用的上下文。
```

$$
\begin{aligned}
&[1+R(\eta)]^{1/4}(\Delta^{(2)}_{00}-4\Psi^{(2)})=A\,\cos[kr_s(\eta)]+B\,\sin[kr_s(\eta)]\\
&-4\frac{k}{\sqrt{3}}\int_0^\eta d\eta'[1+R(\eta')]^{3/4}\left(\Phi^{(2)}(\eta')+\frac{\Psi^{(2)}(\eta')}{1+R}\right)\sin_k[\eta,\eta']\\
&+\frac{\sqrt{3}}{k}\int_0^\eta d\eta'[1+R(\eta')]^{3/4}\left(\mathcal{S}'_\Delta+\frac{\mathcal{H}R}{1+R}\mathcal{S}_\Delta-\frac{4}{3}ik_i\,\mathcal{S}^i_V\right)\sin_k[\eta,\eta']\,,
\end{aligned}
$$

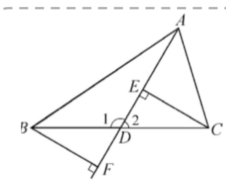
System Prompt Type: "Equation"

```
\begin{eqnarray*}
&&[1+R(\eta)]^{1/4}(\Delta^{(2)}_{00}-4\Psi^{(2)})=A\,\cos[kr_s(\eta)]+B\,\sin[kr_s(\eta)]\nonumber\\
&&-4\frac{k}{\sqrt{3}}\int_0^\eta
d\eta'[1+R(\eta')]^{3/4}\left(\Phi^{(2)}(\eta')+\frac{\Psi^{(2)}(\eta')}{1+R}\right)\sin_k[\eta,\eta']\nonumber\\
&&+\frac{\sqrt{3}}{k}\int_0^\eta
d\eta'[1+R(\eta')]^{3/4}\left(\mathcal{S}'_\Delta+\frac{\mathcal{H}R}{1+R}\mathcal{S}_\Delta-
\frac{4}{3}ik_i\,\mathcal{S}^i_V\right)\sin_k[\eta,\eta']\,,
\end{eqnarray*}
```



System Prompt Type: "Geometry"

The image depicts a geometric diagram featuring a triangle labeled with vertices A, B, and C, and an additional vertex E. The triangle is oriented with its base along the horizontal axis and its sides forming the sides of the triangle. The angle at vertex A is marked with two angles, labeled as angle 1 and angle 2, indicating the measurement of the angles at that vertex. The diagram is simple and uses minimalistic lines to illustrate the geometric shapes and angles, providing a clear visual representation of the triangle's structure.

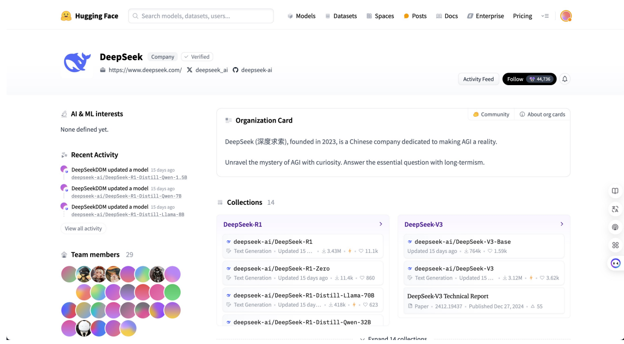


System Prompt Type: "Geometry"

The image depicts a geometric diagram featuring a triangle labeled with vertices A, B, and C, along with additional points D, E, and F. The triangle is oriented with its sides and angles clearly marked. Point D is positioned at the intersection of two lines, labeled with angles 1 and 2, indicating the angles at which these lines meet the sides of the triangle. Point E is located on the side of the triangle opposite vertex A, while point F is positioned on the side opposite vertex B. This diagram likely serves as a visual aid for understanding geometric concepts, such as angles and intersections within triangles.

Figure 12: Math image captioning.

System Prompt Type: "UI"



1. Brief description:
  - The GUI image is a profile page on the Hugging Face platform for the DeepSeek organization, showcasing their AI & ML interests, recent activity, collections of models, and team members.

2. Detail extraction:
  - **Background Elements**: The background is predominantly white with light gray sections for content organization.
  - **Top Navigation Bar**:
    - Logo: "Hugging Face" with a smiling face emoji.
    - Search bar: Placeholder text "Search models, datasets, users..."
    - Navigation links: "Models", "Datasets", "Spaces", "Posts", "Docs", "Enterprise", "Pricing", and a dropdown menu icon.
    - User icon with a notification badge indicating "44,736".
  - **Profile Section**:
    - Logo: A blue whale icon with "DeepSeek" text.
    - verified badge.
    - URL: "https://www.deepseek.com/ deepseek_ai deepseek-ai".
    - Buttons: "Activity Feed", "Follow", notification icon.
  - **Content Sections**:
    - **AI & ML Interests**: "None defined yet."
    - **Recent Activity**:
      - "DeepSeekDDM updated a model 15 days ago deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B".
      - "DeepSeekDDM updated a model 15 days ago deepseek-ai/DeepSeek-R1-Distill-Qwen-7B".
      - "DeepSeekDDM updated a model 15 days ago deepseek-ai/DeepSeek-R1-Distill-Llama-8B".
      - "View all activity" link.
    - **Team members**: Grid of circular avatars with a total count of "29".
    - **Organization Card**:
      - Title: "Organization Card".
      - Description: "DeepSeek (深度求索), founded in 2023, is a Chinese company dedicated to making AGI a reality. Unravel the mystery of AGI with curiosity. Answer the essential question with long-termism."
    - **Collections**:
      - Title: "Collections 14".
      - **DeepSeek-R1**:
        - "deepseek-ai/DeepSeek-R1" with details: "Text Generation", "Updated 15 days ago", "3.43M", "11.1k".
        - "deepseek-ai/DeepSeek-R1-Zero" with details: "Text Generation", "Updated 15 days ago", "11.4k", "860".
        - "deepseek-ai/DeepSeek-R1-Distill-Llama-70B" with details: "Text Generation", "Updated 15 days ago", "418k", "623".
        - "deepseek-ai/DeepSeek-R1-Distill-Qwen-32B".
      - **DeepSeek-V3**:
        - "deepseek-ai/DeepSeek-V3-Base" with details: "Updated 15 days ago", "764k", "1.59k".
        - "deepseek-ai/DeepSeek-V3" with details: "Text Generation", "Updated 15 days ago", "3.12M", "3.62k".
        - "DeepSeek-V3 Technical Report" with details: "Paper", "2412.19437", "Published Dec 27, 2024", "55".

3. Description of interactive elements:
  - Search bar: Allows users to search for models, datasets, or users.
  - "Activity Feed" and "Follow" buttons: Redirect to the user's activity feed or follow DeepSeek.
  - "View all activity" link: Redirects to a detailed list of recent activities.
  - Collection links: Expandable sections for additional details about each collection.
  - Team member avatars: Likely clickable to view more information about team members.

4. Overall description:
  - This GUI page serves as a profile overview for DeepSeek on the Hugging Face platform, showcasing their focus on AI and machine learning. It highlights their recent updates, model collections, and team members. The page is designed to provide users with a comprehensive overview of DeepSeek's activities, including their recent model updates and technical reports. The interface is user-friendly, with clear navigation options and interactive elements to explore further details about DeepSeek's activities and offerings. This page is likely used by researchers, developers, and enthusiasts interested in AI technology and DeepSeek's contributions to the field.

Figure 13: UI captioning.

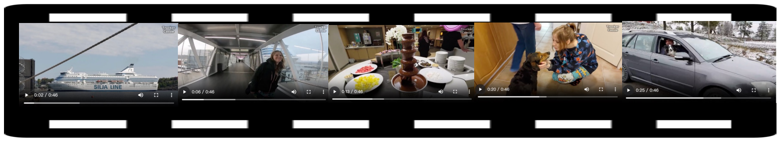
System Prompt Type: "OCR"

```
# 大气的组成

现代大气成分是地球漫长演化的结果，在短时间内不会发生明显变化。近地面的大气是由干洁空气、水汽以及微小颗粒组成的混合物。

## 干洁空气

通常将不包含水汽的纯净大气称为干洁空气，它是大气的主体，平均约占低层大气体积的99.97%。干洁空气中比重最大的是氮气，其次是氧气。氮元素是地球生物体内蛋白质的重要组成部分，是许多生物过程的基本元素。

此外，干洁空气中还有氩气、二氧化碳和一些微量气体，它们虽然比重较小，但仍然影响着地球的自然环境和人类活动。例如，二氧化碳和甲烷都是能产生温室效应的气体，它们比重的变化会在一定程度上引起全球气候变化，进而影响人类活动。

### 思考

- 现代大气中干洁空气的各组成成分比重是否会发生变化？试举例说明。
- 大气中的二氧化碳有什么作用？

### 水汽

水汽是大气中最重要的成分之一，它主要来自海水的蒸发，还有一部分来自地表水体的蒸发以及植物的蒸腾作用。大气中的水汽与云、雾、雨、雪以及虹等自然现象有着密切的关系。

在近地面大气中，水汽含量因地区和季节不同而存在差异。在低纬度的温暖洋面上，水汽含量大；而干旱的沙漠和极地地区，水汽含量极少。一般情况下，同一地区夏季水汽含量较大。

> 虹是太阳光照射在大气中水汽上发生折射和反射作用形成的彩色圆弧，由外到内呈红、橙、黄、绿、蓝、靛、紫七种颜色。图上为上海天空出现的虹。
```

System Prompt Type: "OCR"

```
the image and the exclusion of undesired segments by identifying them as "negative" points, thereby improving the precision of path selection and
the overall accuracy of path extraction. After isolating the paths segment as an image mask, we apply skeletonization and thinning techniques
with OpenCV  and skimage  to refine the mask to a $1$-pixel size, effectively removing any extraneous artifacts and noise. The result is a clear
set of points defining the intended path.
The user can select multiple paths this way and assign them to multiple objects (extracted by SAM) in the diagram for the animation execution,
we utilize the GSAP.js animation library\footnote{https://gsap.com/}, animating the object along the determined path and integrating additional
animation parameters, such as speed and direction, editable by the user.

\textit{\hspace{-0.5cm} Optics Simulation}. We developed a custom optics simulator utilizing P5.js visual graphics library\footnotemark[4]. Our
simulator currently supports convex lenses, conical lenses and mirrors. It calculates the positions of two representative light rays based on the
object and focal point positions, emulating the common practice of manual circuit drawing.

\textit{\hspace{-0.5cm} Circuit Simulation}. We developed a custom circuit simulator designed to operate within a web browser, incorporating
principles of circuit theory, such as Kirchhoff's laws. Utilizing the Gemini Multimodal Vision Model (\textit{gemini-ls-pro}), our system identifies
and segments resistor, capacitor, and battery symbols within circuit diagrams by detecting bounding boxes. Contour detection is then applied on
the image, which isolates lines and discriminates them based on their orientation. By identifying junctions within the diagram, the system
automatically links the bounding boxes of detected resistors or voltage sources to lines, symbolizing wires. The circuit is represented using a
simple array structure, which is transmitted to the web interface through Firebase real-time database and subsequently visualized.

\### Technical Evaluation

\#### Method.
We evaluated the accuracy and versatility of our pipeline through technical evaluations. We first gathered six different physics textbooks
covering topics such as kinematics, optics, circuit theory, and magnetism. From each textbook, we randomly selected ten pages containing
diagrams relevant to each simulation category (kinematics, circuits, optics, and animation), resulting in a total of 200 diagrams for our sample
dataset. We applied our detection pipeline across these diagrams for each simulation category. For object segmentation, we simply select objects
via mouse interaction. For line segmentation (for animated diagrams), we employ four points, two positive and two negative prompts, to segment
the line after that, multiple authors manually review the results by looking at the generated outcomes due to the absence of a standardized and
automated way to check the results, guided by a rubric described below. The complete list of pages and figures evaluated with our system will be
provided in the supplementary materials. Our analysis focused on measuring the error rate in various components of the pipeline.

\#### Results.
Table~1 presents a summary of our technical evaluation results. The success rates for the different components of the simulation are as follows:
kinematics at 64\%, optics at 44\%, circuits at 40\% ($62$\% with minor edits), and animation at 66\%.

Kinematics, Optics, and Animation work through semi-automatic segmentation. Notably, in kinematics, the success rates for polygon generation
and placement are \textbf{$72$\%} and $\mathbf{70}$\%, respectively, indicating effective conversion into physics-simulatable bodies with
proper segmentation. However, challenges arise in kinematics simulations due to limitations in supporting certain features ($6$\%), such as
rotational motion, body specific gravity, un-supported objects like ropes, and issues with simulating curved surfaces smoothly. Additionally, we
noted that $74$\% of the diagrams just required minor adjustments, like modifications to simulation parameters, to achieve accurate simulation
results. The success rate without any authoring and modification process was at $40$\%. Animation and optics were also consistent with the
number, but we observed that the line segmentation success rate was lower, despite using the same Segment Anything technique. For optics in
particular, simulation failures comes from diagrams our simulator does not support, such as those with multiple lenses (detecting two as one),
prisms, new lens types (like an eye), etc.

Our Circuit simulation pipeline utilizes a line detection method to localize and identify inversion in conjunction with the Gemini model to detect
symbols. The line detection success rate is on the lower side at $45$\% leading to an overall success rate of simulation at $62$\%. Without any
edits the pipeline is successful $40$\% of the time. The main reason for this is because of overlapping or crossing wires or artifacts in the
diagrams. On the other hand, symbol recognition is reasonably accurate at $72$\%. The symbols missed by the Gemini model were mainly due to
the error in the bounding boxes returned, which did not align with the symbol.
\begin{table}[!ht]
  \centering
  {
    \caption{Technical evaluation results}


     \scriptsize
       \begin{tabular}{llcc}
         & Kinematics \\
         Object Segmentation    &  86\%\\
         Polygon Generation     &  72\%\\
         Polygon Placement      &  70\%\\
         Simulation           &  64\%\\

         & Animation \\
         Object Segmentation     &  86\%\\
         Line Segmentation      &  70\%\\
         Simulation           &  64\%\\


         & Circuits \\
         Object Segmentation     &  86\%\\
         Simulation           &  44\%\\
         & Optics \\
         Line Detection         &  45\%\\
         Symbol Recognition       &  72\%\\
         Simulation           &  40\%\\
         Simulation with Minor Edits&  62\%
       \end{tabular}}
\end{table}
\section*{USER STUDY}
We evaluated our system through two user studies: 1) a usability study with twelve participants, and 2) expert interviews with experts from the
field of computer vision and human-computer interaction (HCI). In addition, we conducted a small survey among students about their experience
working with diagrams and how they would improve them. Finally, we compared our approach against existing tools to provide a baseline comparison.
```

Figure 14: PDF captioning.



System Prompt Type: "Video"

Short_Caption:
A serene cruise ship journey transitions into lively indoor activities onboard as passengers enjoy leisurely activities amidst vibrant surroundings.

Background_Caption:
The video begins at sea under soft daylight conditions, showcasing an expansive ocean view framed by distant landmasses. The ship's deck displays clean white structures against blue skies. Indoors, spacious rooms feature bright lighting, colorful decor, and modern furnishings, while communal areas host dynamic activities such as dining, music-making, and socializing beneath warm ambient lights.

Main_Object_Caption:
Passengers interact actively throughout various segments; one woman explores her cabin, adjusting items like bedding and storage spaces before transitioning to a lively lounge area where she engages in musical instruments. Her movements vary between deliberate organization indoors and expressive gestures during recreational activities. Her expressions shift subtly—calmness when tidying up and joy or focus during playing or conversing.

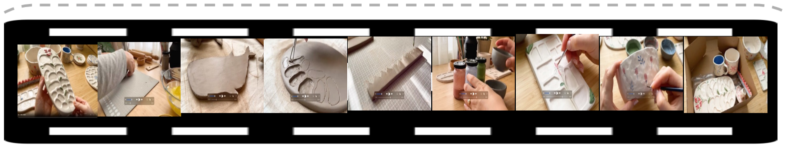
Reference_Caption:
The sequence opens outdoors with static medium shots capturing the sleek design of a massive cruise ship sailing across calm waters bathed in natural light. Transitioning inside, handheld close-ups follow individuals navigating their cabins adorned with personal touches like photos and art pieces. A woman organizes her space methodically, adjusting furniture and ensuring everything is in place. The camera shifts fluidly to showcase communal areas filled with activity. Vibrant reds and yellows dominate the interior tones, enhancing the festive mood as guests interact through music, dining, and conversation. Close-ups capture intricate details—the clinking of glasses over a table, fingers pressing drum pads rhythmically, and animated facial expressions among friends enjoying the evening's vibrancy. Static wide shots frame the entire setting for clarity and balance within compositions that highlight both individual actions and group dynamics. Subtle dolly-like motions enhance intimacy without disrupting the scene's relaxed atmosphere. Lighting plays a critical role here: soft natural hues outside contrast sharply with warm artificial lighting indoors, creating visual harmony reflective of the blend between tranquil exploration and energetic camaraderie onboard.

Standard_Summary:
The video captures moments aboard a large cruise ship beginning with views of the vessel gliding across peaceful waters surrounded by distant landscapes. Inside, passengers engage in various leisure activities. One woman tidies her room attentively, organizing her belongings amid personal decorations. Later scenes depict lively communal areas with music-making, dining, and cheerful conversations unfolding dynamically. Warm lighting enhances these interactions, contrasting beautifully with the openness of outdoor views seen earlier. The smooth interplay between handheld close-ups, wide angles, and subtle tracking creates a balanced narrative of relaxation paired with spirited engagement onboard.

Style_Tags:
Travel Vlog, Cruise Ship Experience, Leisure Activities, Documentary-style Video, Lifestyle Clip

Key_Tags:
cruise ship, passengers, communal areas, daytime, relaxing environment, vibrant interiors, travel documentary style, handheld camera techniques, dynamic lighting contrasts, scenic ocean backdrop

Camera_Caption:
The frames exhibit varied shot types including Medium Shots, Close-Ups, and Wide Shots captured predominantly using eye-level perspectives. The progression suggests a combination of static framing and slight handheld motion, particularly evident in close-up sequences, likely due to minimal positional adjustments. Some frames display slow lateral movement resembling minor pans or tracking.



System Prompt Type: "Video"

Short_Caption:

A step-by-step DIY tutorial for creating decorative ceramic plates using paint pens on clay, followed by painting and assembling them into unique vases.

Background_Caption:

The setting is an indoor crafting area with wooden flooring, white walls adorned with shelves holding various tools and materials such as brushes, jars of paint, and containers filled with supplies. Natural light streams through large windows that reveal glimpses of greenery outside, suggesting it's daytime. The workspace remains static throughout but features scattered objects like trays containing clay pieces, jars labeled "Black," "White," or "Blue," and small bowls used to mix paints.

Main_Object_Caption:

The main focus shifts between two hands actively engaging in creative activities involving clay and painted ceramics. Initially, hands press a textured rolling pin over clay to create oval-shaped bases, which are later dried under sunlight before being painted. With steady movements, they use fine-tipped paint markers to draw intricate patterns resembling animal prints onto the clay surfaces while maintaining precision. After drying again, the painted pieces are carefully assembled together to form vases, secured firmly within their final structure. Each action emphasizes care and attention to detail during both preparation and decoration processes.

Reference_Caption:

Two hands diligently transform clay into artistic ceramic pieces starting at home-based workstations. Using a textured rolling pin, flat clay sheets are flattened evenly against a surface. The dried clay pieces are then cut into oval shapes suitable for decorating. Paint markers, pre-labeled in black, white, blue tones, are employed skillfully—creating elaborate patterns mimicking animal fur across several clay slabs. These decorated elements undergo further steps where individual pieces are aligned and attached to each other, forming functional yet stylish vases. In natural daylight pouring through nearby windows, the vibrant greens visible beyond add warmth and liveliness to the scene. The camera captures these transitions via close-up shots emphasizing texture details—the grainy rolling pin, smooth paint strokes, and matte finish after baking—while medium shots provide context for the broader crafting setup. Static framing ensures consistency without dynamic movement, enhancing clarity and instructional quality. This immersive process culminates when the finished products are displayed side by side alongside unpainted pieces, showcasing creativity blended with utility.

Standard_Summary:

The sequence showcases how handcrafted ceramic vases are made, beginning with shaping clay oval bases using a textured rolling pin. Once dried, detailed designs featuring animal-like patterns are meticulously drawn on top utilizing paint pens. The vases are later assembled by connecting multiple clay pieces. Set indoors amidst organized supplies and bathed in soft natural lighting, every stage—from rolling pins to intricate patterns—is captured clearly. The result is aesthetically pleasing, blending artistry with functionality as the completed vases take shape.

Style_Tags:

DIY Tutorial, Educational Video, Handcrafting Process, Step-by-Step Guide, Artistic Creation, Indoor Setting, Creative Demonstration, Instructional Content.
Key_Tags: Ceramic plates, paint pens, clay rolls, textured rolling pin, vases, crafting room, natural lighting, stationary camera angles, artistic patterns, assembly techniques.

Camera_Caption:

Predominantly handheld shots alternating between Medium Shots and Close-Ups focusing on specific objects held up towards the camera. Minimal camera motion suggests mostly static positioning with slight adjustments to maintain alignment.

Figure 15: Video captioning.

System Prompt Type: "Detailed"

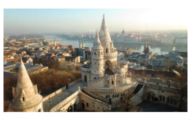
Iconic photo of the Hollywood sign against a clear blue sky backdrop. The large white letters spelling "HOLLYWOOD" dominate the lower third of the frame, positioned centrally atop a brown hillside with sparse greenery and brush. Several communication towers rise behind the sign, adding structural complexity to the scene. In the upper left corner, there's a line of shrubs and trees marking the horizon. Bright natural lighting casts minimal shadows, creating sharp contrasts between the bright letters and the earthy tones of the hillside. Photorealistic, high contrast, wide depth of field, midday sun, vibrant color palette.

System Prompt Type: "Detailed_Natural"

The Hollywood sign, an iconic landmark located on Mount Lee in Los Angeles County, California, is depicted in this image as it sits atop a hillside with a clear blue sky above. The large white letters that spell out 'HOLLYWOOD' are prominently displayed against the backdrop of the terrain below, which consists of patches of greenery and some structures.

System Prompt Type: "UI"

This triptych painting is titled "The Lamentation over Christ," created by an unknown artist from around 1480-1526. It consists of three panels, each depicting different scenes with rich religious symbolism.
Left Panel:
- The scene shows a group of women in dark robes standing solemnly before a landscape backdrop.
- They appear to be weeping or mourning, possibly representing the five Stations of the Cross, where the Virgin Mary stands at the foot of the cross, surrounded by sorrowful figures like St. John and other women.
Middle Panel:
- This panel illustrates a central figure lying on a platform amidst a crowd of mourners dressed in red and green robes.
- Some individuals hold religious objects while others show various expressions of grief. - In the background, there's a depiction of the crucifixion with Jesus on the cross, enhancing the emotional depth of the piece.
Right Panel:
- The scene features several figures in traditional religious attire.
- A prominent figure wearing white robes appears to be leading a procession towards the left side of the panel. - Other individuals hold religious items or stand solemnly beside him. - The background includes additional figures observing the main action. Overall, the artwork showcases intricate details typical of late Gothic art, focusing on deep emotions and spiritual themes associated with the Christian faith.

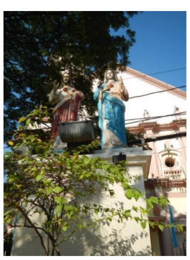


System Prompt Type: "UI"

The image showcases an aerial view of Fisherman's Castle (halácskás) located on Buda Castle Hill in Budapest, Hungary. The historic structure features distinctive conical towers that rise majestically against a clear sky, with sunlight casting warm tones over its stone facade. In the background, the Danube River flows through the cityscape, surrounded by various architectural styles that include domes and spires. This iconic landmark is a significant part of Hungarian history and serves as a popular tourist attraction.

System Prompt Type: "Detailed"

In the heart of Budapest, Hungary, stands the historic Fisherman's Bastion, a grand architectural structure boasting multiple towers and turrets. This castle-like edifice, with its intricate details and medieval design, is situated on a hill overlooking the city. The cityscape below is a vibrant mix of old and new buildings, with a river running through its center. The riverbank is dotted with bridges, and the city extends into the distance, revealing more buildings and structures. The sky above is clear, and the sun casts a warm glow over the scene, creating an enchanting tableau. Although the image is visually appealing, it is essential to note that the Fisherman's Bastion is a real historical site in Budapest.



System Prompt Type: "OCR_textqa"

The Church of Our Lady of Czestochowa in Iloilo showcases two statues on a pedestal: one male figure with outstretched hands, wearing a red robe and white sash, symbolizing blessings orTEAS; the other female figure holding a blue book or scroll, dressed in blue robes, signifying knowledge orTEAS. The setting includes trees casting shadows on the pedestal, a building background with architectural details like arches and decorative elements.

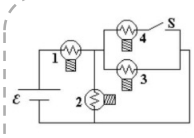
System Prompt Type: "Detailed"

Statue scene featuring two religious figures prominently positioned atop a white pedestal under a clear blue sky. The statue on the left depicts a male figure wearing red robes with an outstretched hand holding a heart symbolizing divinity, while the right figure shows a female in a blue robe holding a baby, both adorned in traditional attire. They stand amidst lush green foliage partially obscuring their lower bodies. In the midground is a large black planter. Behind them rises part of a pinkish-white church building showcasing intricate architectural detailing including arches and a clock tower visible at the top right corner. Overhead utility wires stretch across the frame against dense tree branches creating dappled shadows over the scene. Daytime photo, natural lighting casting soft shadows, vibrant color palette emphasizing blues and pinks contrasted by earthy tones from leaves, high depth of field capturing details throughout, serene yet slightly chaotic composition due to overlapping elements.

System Prompt Type: " Detailed_Natural"

In the heart of Iloilo City lies an outdoor setting adorned with religious statues on pedestals. The left statue depicts Jesus Christ, recognizable by his crown of thorns and red robe draped over one shoulder. To the right stands a female figure, likely the Virgin Mary or another revered saint, holding a book and dressed in blue robes. A pot sits between them. In the foreground, lush green plants flourish, while the background reveals part of a building boasting intricate architectural details.

Figure 16: Natural image captioning with different system prompts.



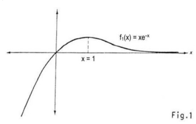
System Prompt Type: "UI"

The image depicts an electrical circuit diagram illustrating various components connected by wires. Key elements include:
1. **Bulbs (Lamps)**:
- Labeled as "1", "2", "3", and "4".
- Each bulb is represented with zigzag patterns indicating their light emission.
2. **Wires**:
- Represented as straight lines connecting different parts of the circuit.
3. **Power Source**:
- denoted as "E" on the left side of the diagram, representing the voltage source providing power to the circuit.
4. **开关（Switches）**:
- Indicated at certain points where the flow of current can be switched between two paths.
5. **Source (S)**: -
Positioned at the top right corner, suggesting it might represent the starting point or reference for the circuit.

This schematic illustrates fundamental concepts related to electrical circuits, specifically showing how different components interact when connected through conductive paths like wires. The design helps visualize the basic principles of electrical connections and energy flow within a circuit system.

System Prompt Type: "Detailed"

Technical schematic diagram of a simple electrical circuit featuring four light bulbs labeled 1, 2, 3, and 4. The circuit includes a power source labeled $\\mathcal{E}$ on the left, connected to a horizontal line. The first bulb, labeled 1, is positioned at the top left, connected to a vertical line. The second bulb, labeled 2, is at the bottom center, connected to a vertical line. The third bulb, labeled 3, is to the right of the second bulb, connected to a horizontal line. The fourth bulb, labeled 4, is at the top right, connected to a horizontal line. The schematic uses simple line art with minimal shading, monochrome color scheme, technical illustration style, precise and functional, clear labeling, and a straightforward layout.



System Prompt Type: "Table"

The image displays a mathematical function graphed on a coordinate plane. The function is labeled as "f(x) = x e^x," where "f(x)" represents the function value at a given point "x," and "e^x" is the exponential function of "x." The graph shows the curve of this function, which starts at the origin (0,0) and increases as "x" increases. Key features of the graph include:

1. **X-Axis (Horizontal Axis)**:
Labeled with "x," it represents the independent variable.

2. **Y-Axis (Vertical Axis)**: Labeled with "y," it represents the dependent variable.

3. **Function Curve**: The curve starts at the origin (0,0) and rises steeply as "x" increases, then levels off as "x" continues to increase.

4. **Label "x = 1"**: This indicates a specific point on the x-axis where the function value is evaluated.

5. **Label "f(x) = x e^x"**: This is the equation of the function being plotted.

6. **Label "Fig.1"**: This suggests that this is the first figure in a series of graphs. The graph is a simple yet informative representation of the function "f(x) = x e^x," which is a common function in mathematics and physics, often used in various applications such as modeling growth rates, decay processes, and more.

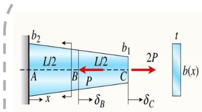
System Prompt Type: "UI"

he image displays a mathematical curve representing the function \(f(x) = x e^{-x} \). The graph shows how the function changes with respect to \(x \), illustrating its behavior at different values of \(x \).

- **X-Axis**: Labeled "x", indicating the independent variable.
- **Y-Axis**: Labeled "f(x)", indicating the dependent variable which is equal to \(x e^{-x} \).
- **Curve**: A smooth curve that starts high on the y-axis when \(x = 0 \) and gradually decreases as \(x \) increases.

Noteable points:
1. At \(x = 0 \), \(f(x) = 0 \).
2. As \(x \) approaches infinity, \(f(x) \) also approaches zero but never reaches it.
3. There is an inflection point where the rate of change of \(f(x) \) equals zero, typically around \(x = 1 \).

A dashed line labeled \(x = 1 \) highlights the specific value of interest along the x-axis for reference within the context of the curve. This figure likely serves as part of a larger discussion or analysis involving exponential decay functions, possibly related to probability distributions like the Poisson distribution mentioned elsewhere in the document.



System Prompt Type: "Detailed"

Technical illustration of a mechanical system featuring a trapezoidal cross-section with labeled components. The trapezoid is divided into sections labeled A, B, and C, with $L/2$ marked on each section. The left side of the trapezoid is labeled b_2, and the right side is labeled b_1. A red arrow labeled $2P$ extends from the right side of the trapezoid towards the right side of the image. The system includes a vertical line labeled t on the far right, indicating a thickness or a parameter related to the trapezoid. The diagram includes annotations such as δ_B and δ_C, suggesting displacement or stress components. The background is white, and the trapezoid is shaded in a light blue gradient, enhancing the technical and schematic appearance. Technical drawing, schematic style, precise line work, minimalistic color palette, functional, educational, engineering-focused.

System Prompt Type: "UI"

1. **Brief description**:
The diagram illustrates an engineering or physics scenario involving forces, displacements, and dimensions on two interconnected structures.

2. **Detailed extraction**:
- **Background Elements**:
  - A white background with no additional decorative elements.
- **Visual Components (Right to Left)**:
  1. **Left Structure**:
    - A vertical rectangular shape labeled "b_2" at its top edge.
    - A horizontal dashed line across it indicating measurement or division.
    - A blue gradient-filled rectangle extending from left to right, marked by labels:
      - "L/2" near the middle section.
    - Two black arrows pointing outward from both sides of the structure, representing displacement or force directions.
    - An arrow pointing downward labeled "x", indicating movement along the horizontal axis.
    - A label "A" is placed below the center of the structure.
    - A label "B" is positioned next to the left end of the structure.
  2. **Middle Section**:
    - A red double-headed arrow labeled "P" extends horizontally between points B and C, suggesting a force vector.
    - Dashed lines connecting B and C indicate measurements or divisions.
    - Labels "L/2" appear twice, once above each side of the central region.
  3. **Right Structure**:
    - A tall, narrow blue gradient-filled rectangle labeled "t".
    - A label "b(x)" is written vertically beside the rectangle.
    - Two black arrows point outward from either side of the right structure, likely representing displacement or force directions.
    - A label "C" is located directly beneath the right structure.
    - A label "δ_C" is displayed below the left end of the right structure.

3. **Description of interactive elements**:
- No visible interactive elements like buttons or sliders; all components are static graphical representations for educational purposes.

4. **Overall description**:
This diagram provides a clear visualization of structural mechanics concepts, specifically focusing on forces, displacements, and dimensions involved in a mechanical system. It shows how forces act upon different parts of a structure, using directional arrows and labels to represent specific parameters such as lengths ("L/2"), distances ("x," "δ_B," "δ_C"), and widths ("b_2," "b_1"). The use of gradients adds depth and clarity while maintaining simplicity for easy comprehension. Such diagrams are typically used in educational settings to illustrate complex physical principles effectively.

Figure 17: Structured image captioning with different system prompts.

System Prompt Type: " Detailed"

A man in a red shirt and black pants stands in the left foreground, observing a modified white pickup truck with extended wheel arches and large tires, positioned diagonally in the center. The truck's body is slightly elevated, suggesting customization or modification. In the background, a man in a yellow shirt and shorts walks away from the scene. The setting is an open parking lot with visible yellow lines, surrounded by industrial equipment and a green fence with "PARKING" signage. The lighting is natural, casting soft shadows on the ground. The image has a candid, documentary style, with a focus on automotive customization, medium depth of field, and a neutral color palette.

Question

Is there a man wearing a red shirt in the image? Please answer yes or no.

Thinking Process from DeepSeek-R1-Distill-Qwen-70B

Okay, so I need to figure out if there's a man wearing a red shirt in the given image based on the description provided. Let me start by reading through the detailed caption again carefully.

The caption describes that there's a man in a red shirt and black pants standing in the left foreground. He's observing a modified white pickup truck. So right there, it mentions a man wearing a red shirt. That seems pretty straightforward.

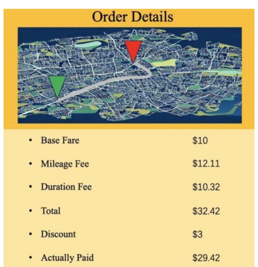
I should check if there are any other people mentioned and what they're wearing. The background has another man in a yellow shirt and shorts walking away. So, only two men are present, one in red and one in yellow.

The question is specifically asking about a man in a red shirt. Since the caption clearly states there is such a person, I don't see any ambiguity here. There's no mention of any other colors or conflicting information.

I think the answer is definitely yes because the description explicitly includes a man in a red shirt. I don't see any reason to doubt that part of the caption. It's also the first person described, so it's prominent in the image.

Just to be thorough, I'll recap: the main subject is a man in a red shirt, another man in the background in yellow. Setting is a parking lot with trucks and industrial stuff. No other people or clothing colors mentioned. So yes, the man in red is definitely there.
</think>

Yes



System Prompt Type: " Detailed"

Map-based order details interface with a detailed city map in the upper half, featuring a network of streets and green areas. Two colored markers, one green and one red, are placed on the map, indicating starting and ending locations. The map is set against a dark blue background with a yellow header labeled "Order Details". Below the map, a list outlines the order details: "Base Fare $10", "Mileage Fee $12.11", "Duration Fee $10.32", "Total $32.42", "Discount $3", "Actually Paid $29.42". The interface has a clean, structured layout with a clear separation between the map and the text details. Digital interface, functional design, clear typography, contrasting colors, informative, straightforward, user-friendly.

Question

Here are the order details for my taxi ride. Should I actually pay $32.42? Please answer yes or no.

Thinking Process from DeepSeek-R1-Distill-LLama-70B

"Okay, so I need to figure out whether I should actually pay $32.42 for my taxi ride based on the information provided. Let me break this down step by step.

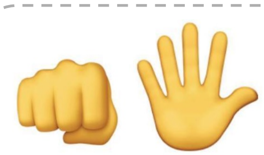
First, looking at the order details: the total comes out to $32.42 before any discounts. Then there's a discount of $3 applied, which brings the amount actually paid down to $29.42. So, the question is asking if I should pay $32.42, but according to the details, after the discount, I only need to pay $29.42.

Wait, maybe I'm misunderstanding. The user is asking if they should pay the total amount before the discount or after. Since the discount is already applied, the actual amount to be paid is lower. So, paying $32.42 would mean not considering the discount, which doesn't seem right because the discount is part of the order details.

I should also consider why there's a discount. Maybe it's a promotional offer, a coupon, or some kind of loyalty program. Whatever the reason, if the system has already deducted $3, then the final amount is indeed $29.42. Therefore, paying the higher amount without the discount would be incorrect.

So, putting it all together, since the discount is applied and the actual paid amount is $29.42, the user shouldn't pay $32.42. They should only pay the discounted amount. Hence, the answer is no, they don't need to pay $32.42; instead, they pay $29.42.
</think>

No, you should not pay $32.42. After applying the $3 discount, the amount you actually need to pay is $29.42."



System Prompt Type: " Detailed"

Digital illustration of two emoji icons against a plain white background. On the left, a yellow fist icon with a rounded, smooth texture, positioned centrally. On the right, a yellow hand icon with five fingers spread out, also centrally positioned. Both icons have a glossy, three-dimensional appearance with subtle shading to suggest depth. Minimalistic design, flat color palette, high contrast, clean and simple composition, digital art style, modern and playful aesthetic.

Question

May I ask if in the game of finger-guessing game, did the right side of the picture win? Please answer yes or no.

Thinking Process from DeepSeek-R1-Distill-LLama-70B

"Okay, so I need to figure out whether the right side of the picture wins in the finger-guessing game. Let me start by understanding what the finger-guessing game is. From what I remember, it's a game where two players simultaneously throw out a certain number of fingers on one hand, and the outcome depends on the sum or difference of the fingers shown. Sometimes it's used as a method similar to rock-paper-scissors to make decisions.

Looking at the image description, there are two emoji icons. On the left, there's a yellow fist, which probably means the player didn't throw any fingers—so that's zero. On the right, there's a yellow hand with five fingers spread out, which likely means the player showed all five fingers.

In some versions of the game, each player's number of fingers is compared. If one player shows more fingers than the other, they win. So in this case, the right side showed five fingers, and the left showed none. Five is greater than zero, so the right side would win.

But wait, I should consider different rules because variations exist. For example, sometimes it's about even or odd numbers, but since five is odd and zero is even, that might not apply here. Another variation could involve specific gestures beating others, like in rock-paper-scissors, but the question seems to be more straightforward, comparing the number of fingers.

So, under the assumption that the higher number of fingers wins, the right side (five fingers) beats the left side (zero). Therefore, the answer should be yes, the right side won.

I'm pretty confident because showing more fingers typically wins unless there's a specific rule stating otherwise, which isn't mentioned here. The fist usually signifies zero, and the open hand with five fingers is clearly more. So yeah, the right side wins.
</think>

Yes."

Figure 18: Visualization of thinking process with OMNICAPTIONER for natural images.

Project	Beta	IRR
W	.80	9.4%
X	.95	10.9
Y	1.15	13.0
Z	1.45	14.2

Question

An all-equity firm is considering the following projects:
The T-bill rate is 3.5 percent, and the expected return on the market is 11 percent.Which projects would be incorrectly accepted or rejected if the firm's overall cost of capital was used as a hurdle rate?

A. Project X would be incorrectly rejected; Project Z would be incorrectly accepted.
B. Project Y would be incorrectly rejected; Project X would be incorrectly accepted.
C. Project Z would be incorrectly rejected; Project X would be incorrectly accepted.
D. No project would be incorrectly rejected; Project Z would be incorrectly accepted.

System Prompt Type: " Table"

Certainly! Let's analyze the table in the image step by step.

Table Analysis

Columns:
1. **Project**: This column lists the names of different projects.
2. **Beta**: This column shows the beta values associated with each project. Beta is a statistical measure that indicates the volatility of an investment relative to the overall market.
3. **IRR (Internal Rate of Return)**: This column displays the internal rate of return for each project. IRR is the expected return of an investment over a specific period.

Rows:
- **W**:
  - Beta: 0.80
  - IRR: 9.4%
- **X**:
  - Beta: 0.95
  - IRR: 10.9%
- **Y**:
  - Beta: 1.15
  - IRR: 13.0%
- **Z**:
  - Beta: 1.45
  - IRR: 14.2%

Observations:
1. **Beta Values**:
   - The beta values increase from project W to project Z.
   - Project W has the lowest beta (0.80), indicating less volatility compared to the market.
   - Project Z has the highest beta (1.45), indicating higher volatility compared to the market.

2. **Internal Rate of Return (IRR)**:
   - The IRR also increases from project W to project Z.
   - Project W has the lowest IRR (9.4%).
   - Project Z has the highest IRR (14.2%).

Conclusion:
- As the beta value increases, the IRR also increases. This suggests that projects with higher beta values are expected to have higher returns, but also face greater volatility.
- Project Z, with the highest beta and IRR, is expected to have the highest returns but also the highest volatility.
- Project W, with the lowest beta and IRR, is expected to have the lowest returns but also the lowest volatility.

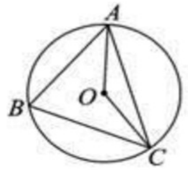
This analysis can help in understanding the relationship between volatility (beta) and expected returns (IRR) for different projects.

Thinking Process from DeepSeek-R1-Distill-Qwen-70B

Alright, let me try to work through this problem step by step. So, the question is about whether certain projects would be incorrectly accepted or rejected when using the firm's overall cost of capital as a hurdle rate. The firm is all-equity, which might simplify things because we don't have to worry about debt in the cost of capital calculation.

First, I remember that the cost of capital for an all-equity firm is essentially the required rate of return on equity. Since the company isn't using any debt, the weighted average cost of capital (WACC) would just be the cost of equity.
The given data includes the T-bill rate (risk-free rate) at 3.5%, and the expected return on the market is 11%. So, I think I'll need to use the Capital Asset Pricing Model (CAPM) to calculate the required rate of return for each project. CAPM formula is:
Required Return = Risk-free Rate + Beta * (Market Return - Risk-free Rate)
So, for each project (W, X, Y, Z), I can compute their required return by plugging in their respective betas.
Let me jot down the betas and IRRs:
- Project W: Beta 0.80, IRR 9.4%
- Project X: Beta 0.95, IRR 10.9%
- Project Y: Beta 1.15, IRR 13.0%
- Project Z: Beta 1.45, IRR 14.2%
Now, calculating the required return for each:
Starting with Project W:
Required Return = 3.5% + 0.80*(11% - 3.5%)
= 3.5% + 0.80*7.5%
= 3.5% + 6%
= 9.5%
So, the required return is 9.5%, and the IRR is 9.4%. Hmm, that's very close. If the IRR is slightly below the required return, the project would be rejected because it doesn't meet the hurdle rate.
Next, Project X:
Required Return = 3.5% + 0.95*(7.5%)
= 3.5% + 7.125%
= 10.625%
The IRR is 10.9%, which is higher than 10.625%, so this project should be accepted.
Project Y:
Required Return = 3.5% + 1.15*7.5%
= 3.5% + 8.625%
= 12.125%
But the IRR is 13.0%, which is higher, so accept.
Project Z:
Required Return = 3.5% + 1.45*7.5%
= 3.5% + 10.875%
= 14.375%
The IRR is 14.2%, which is lower than the required return. So, this project should be rejected because it doesn't meet the hurdle rate.
Wait, but looking back at the calculations, Project Z's required return is 14.375%, and its IRR is 14.2%, so it's actually below. That means if the company uses the cost of capital as a hurdle rate, they would reject Project Z even though its IRR is positive but just barely below the required return.
But wait, the question is about which projects would be incorrectly accepted or rejected. So, if the firm's cost of capital is the hurdle rate, then any project with IRR above the required return should be accepted, and those below should be rejected.
But the question is phrased as which projects would be "incorrectly" accepted or rejected. I'm a bit confused here because I thought the required return is the correct hurdle rate. Maybe the issue is that the firm is using the overall cost of capital as the hurdle rate for all projects, regardless of their individual risks. But since each project has different betas, using the same hurdle rate for all might lead to incorrect decisions.
Wait, perhaps the firm is using the company's overall cost of capital, which is the same for all projects, instead of adjusting for each project's specific risk (beta). So, maybe they're using a single hurdle rate for all projects, which could lead to errors.
Let me check the options again:
A. Project X would be incorrectly rejected; Project Z would be incorrectly accepted.
B. Project Y would be incorrectly rejected; Project X would be incorrectly accepted.
C. Project Z would be incorrectly rejected; Project X would be incorrectly accepted.
D. No project would be incorrectly rejected; Project Z would be incorrectly accepted.
From my earlier calculations, if each project's required return is calculated using their respective betas, then:
- Project W: Required 9.5% vs IRR 9.4% → Reject
- Project X: Required 10.625% vs IRR 10.9% → Accept
- Project Y: Required 12.125% vs IRR 13.0% → Accept
- Project Z: Required 14.375% vs IRR 14.2% → Reject
So, correctly, W and Z are rejected, X and Y are accepted.
But the question is about using the firm's overall cost of capital as a hurdle rate. Wait, the overall cost of capital for the firm (since it's all-equity) is the cost of equity, which is the required return on the firm's assets. But each project has a different beta, so perhaps the firm is mistakenly using the same cost of capital for all projects.
Alternatively, maybe the firm's overall cost of capital is calculated based on the company's average beta, which might not match the individual project betas.
Wait, but the firm is all-equity, so the cost of capital is the cost of equity, which is the required return on the entire firm. However, each project has a different beta, so the required return for each project should be different.
If the firm uses the same hurdle rate (the company's cost of equity) for all projects, regardless of their beta, that would be incorrect. For example, if the company's cost of equity is, say, 10%, then they might accept projects with IRR above 10% and reject those below. But in reality, each project's required return should be based on its own beta.
Wait, but in the CAPM, the required return is specific to each project based on its beta. So if the firm uses the company's overall cost of capital (which is based on the company's beta) as the hurdle rate for all projects, it might incorrectly accept or reject some projects.
Wait, perhaps I'm overcomplicating. Let me think differently. The question is asking which projects would be incorrectly accepted or rejected if the firm uses its overall cost of capital as the hurdle rate instead of the project-specific required return.
So, first, calculate the firm's overall cost of capital, which is the cost of equity. Since the firm is all-equity, its cost of capital is the same as its cost of equity. The cost of equity can be calculated using CAPM based on the firm's beta. But the problem is, we don't know the firm's beta. Hmm, that's a problem. Maybe the firm's beta is not given, so perhaps we have to assume that the firm's beta is the same as one of the projects? Or perhaps the firm's beta is such that the cost of capital is somewhere in between.
Wait, maybe the question is implying that the firm's overall cost of capital is being used as a single hurdle rate for all projects, ignoring their individual betas. So, for example, if the firm's cost of capital is, say, 10%, then any project with IRR above 10% would be accepted, and below would be rejected, regardless of their actual risk.
But without knowing the exact cost of capital, I'm stuck. Wait, perhaps the firm's cost of capital is calculated based on the market's expected return and the T-bill rate, but without a specific beta, how?
Wait, perhaps the question is simplifying things. Let me re-examine the problem statement.
The question says: "An all-equity firm is considering the following projects... The T-bill rate is 3.5 percent, and the expected return on the market is 11 percent. Which projects would be incorrectly accepted or rejected if the firm's overall cost of capital was used as a hurdle rate?"
So, perhaps the firm's overall cost of capital is calculated using its own beta, but since the firm is all-equity, and we don't have its beta, maybe we're supposed to infer that the cost of capital is based on an average or something else.
Alternatively, maybe the question is considering that the firm's cost of capital is the same as the market's expected return, but that doesn't make sense because the market return is 11%, which is the expected return on the market portfolio, not necessarily the firm's required return.
Wait, perhaps the confusion is that the firm's cost of capital is the same as the market's required return, which is 11%, so they use 11% as the hurdle rate for all projects. Then, any project with IRR above 11% would be accepted, others rejected. But looking at the IRRs:
- W: 9.4% → Reject
- X:10.9% → Reject
- Y:13.0% → Accept
- Z:14.2% → Accept
But according to the correct approach using CAPM, the required returns are:
- W:9.5% → Reject (since 9.4<9.5)
- X:10.625% → Accept (10.9>10.625)
- Y:12.125% → Accept (13>12.125)
- Z:14.375% → Reject (14.2<14.375)
So, if the firm uses 11% as the hurdle rate, then:
- W:9.4 <11 → Reject (correct decision)
- X:10.9 <11 → Reject (but according to CAPM, X should be accepted because 10.9>10.625)
- Y:13>11 → Accept (correct)
- Z:14.2>11 → Accept (but according to CAPM, Z should be rejected because 14.2<14.375)
So, using 11% as the hurdle rate, the firm would incorrectly reject Project X (since its IRR is 10.9% which is above its required return of 10.625%) and incorrectly accept Project Z (since its IRR is below its required return).
Looking at the options, option A says: Project X would be incorrectly rejected; Project Z would be incorrectly accepted. That matches what I just found.
Wait, but earlier I thought that the required return for the firm (cost of capital) is 11%, which is the market return. But actually, the market return is 11%, but the firm's cost of equity is based on its beta. However, since the firm is all-equity, and assuming the firm's beta is 1 (since it's an average company), then the required return would be 3.5 + 1*(11-3.5)=11%, which matches the market return. So yes, the firm's cost of capital is 11%.
Therefore, using 11% as the hurdle rate, Projects Y and Z are accepted, W and X are rejected. But according to the correct project-specific required returns, X should be accepted and Z should be rejected.
Thus, the firm would incorrectly reject X (because using 11%, X's IRR is 10.9% <11, so reject, but actually X's required return is 10.625%, so it should be accepted). And incorrectly accept Z (because Z's IRR is 14.2% >11, but its required return is 14.375%, so it should be rejected).
Therefore, the answer should be option A: Project X would be incorrectly rejected; Project Z would be incorrectly accepted.
</think>
The correct answer is:
A. Project X would be incorrectly rejected; Project Z would be incorrectly accepted.
Step-by-Step Explanation:
1. **Calculate Required Returns Using CAPM:**
- For each project, compute the required return using the CAPM formula:
\[
\text{Required Return} = R_f + \beta \times (R_m - R_f)
\]
Where \(R_f = 3.5\% \) and \(R_m = 11\% \).
- **Project W:** \(3.5\% + 0.80 \times 7.5\% = 9.5\% \)
- **Project X:** \(3.5\% + 0.95 \times 7.5\% = 10.625\% \)
- **Project Y:** \(3.5\% + 1.15 \times 7.5\% = 12.125\% \)
- **Project Z:** \(3.5\% + 1.45 \times 7.5\% = 14.375\% \)
2. **Determine Correct Accept/Reject Decisions:**
- Accept if IRR > Required Return; otherwise, Reject.
- **W:** 9.4% < 9.5% → Reject
- **X:** 10.9% > 10.625% → Accept
- **Y:** 13% > 12.125% → Accept
- **Z:** 14.2% < 14.375% → Reject
3. **Firm's Cost of Capital as Hurdle Rate:**
- Assume the firm uses its overall cost of capital (11%) as the hurdle rate for all projects.
- Accept if IRR > 11%; Reject otherwise.
- **W:** 9.4% < 11% → Reject (Correct)
- **X:** 10.9% < 11% → Reject (Incorrect, should be Accepted)
- **Y:** 13% > 11% → Accept (Correct)
- **Z:** 14.2% > 11% → Accept (Incorrect, should be Rejected)
4. **Conclusion:**
- **Project X** is incorrectly rejected because its IRR (10.9%) exceeds its required return (10.625%) but is below the hurdle rate of 11%.
- **Project Z** is incorrectly accepted because its IRR (14.2%) is below its required return (14.375%) but above the hurdle rate of 11%.
Thus, the correct answer is **A**.

Figure 19: Visualization of thinking process with OMNICAPTIONER.



Question:

Please directly answer the question and provide the correct option letter, e.g., A, B, C, D. Question: As shown in the figure, triangle ABC is the inscribed triangle of circle O. If angle ABC = 70.0, then the degree of angle AOC is equal to () Choices:
A:140°
B:130°
C:120°
D:110°

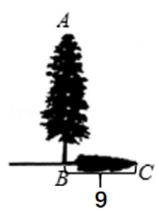
System Prompt Type: " Geometry"

The image depicts an illustration related to the concept of vectors or geometry within physics contexts. It features three points labeled A, B, and C arranged on a circle with an additional point marked O at its center. The lines connecting these points form triangles, suggesting potential applications in vector addition or trigonometry problems. This diagram is commonly used in educational materials to illustrate fundamental principles about vector operations or spatial relationships in physics.

Thinking Process from DeepSeek-R1-Distill-Qwen-70B

Okay, so I've got this geometry problem here. Let me try to figure it out step by step. The problem says that triangle ABC is inscribed in circle O, which means all three vertices of the triangle lie on the circumference of the circle. That makes triangle ABC a circumscribed triangle around the circle, but wait, actually, when a triangle is inscribed in a circle, the circle is called the circumcircle, and the triangle's vertices lie on the circle's perimeter. So, point O must be the center of this circle.
The question gives me that angle ABC is 70 degrees, and it's asking for the measure of angle AOC. The options are A:140°, B:130°, C:120°, D:110°.
Hmm, angle ABC is at point B, so it's one of the angles inside triangle ABC. Since the triangle is inscribed in the circle, angle ABC is an inscribed angle. I remember that inscribed angles have a relationship with central angles. Specifically, an inscribed angle is equal to half the measure of the central angle that subtends the same arc.
So, angle ABC subtends arc AC. That means the central angle AOC, which also subtends arc AC, should be twice as large as angle ABC. Because the central angle is twice the inscribed angle for the same arc.
Let me write that down:
Angle ABC = 70° (inscribed angle)
Arc AC = 2 * angle ABC = 2 * 70° = 140°
But wait, angle AOC is the central angle that subtends arc AC. So angle AOC should be equal to the measure of arc AC, right? Because central angles and their arcs have the same measure.
Wait, no, that's not quite right. Actually, the central angle is equal to the measure of its subtended arc. So if arc AC is 140°, then angle AOC is also 140°.
Wait, but let me double-check. Inscribed angle theorem states that an inscribed angle is half the measure of the central angle subtending the same arc. So if angle ABC is 70°, then the central angle AOC would be 140°, because 70° * 2 = 140°.
Yes, that makes sense. So angle AOC is 140 degrees, which corresponds to option A.
I think that's the correct approach. Let me visualize it: points A, B, and C on the circle, with O at the center. Triangle ABC has angle at B of 70°, which is an inscribed angle. The central angle from A to O to C would cover the same arc AC but from the center, making it twice as big. So 70° times 2 is 140°, so angle AOC is 140°, which is option A.
</think>
The angle at the center (AOC) is twice the inscribed angle (ABC). Therefore, angle AOC is 140°.
Answer: A:140°

A certain mathematics learning interest group measured the shadow length of Xiaoliang in the sun as 1.5. Knowing that Xiaoliang's height is 1.8. As shown in the figure, then the height of tree AB is ()
Choices: A:10.8m B:9m C:7.5m D:0.3m



System Prompt Type: " UI"

1. **Brief description**:
The image presents an educational question about tree height calculation using given measurements for shadow length and human height.

2. **Detail extraction**:
- **Background Elements**:
- White background.
- **Text Content (from top to bottom)**:
1. "A certain mathematics learning interest group measured the shadow length of Xiaoliang in the sun as 1.5. Knowing that Xiaoliang's height is 1.8. As shown in the figure, then the height of tree AB is () Choices: A:10.8m B:9m C:7.5m D:0.3m"

- **Visual Elements**:
- An illustration shows a silhouette of a tree labeled with 'A' at its top and 'B', 'C'.
- A horizontal line represents ground level marked with numbers from left to right as 'B', '9', and 'C'. The number '9' is highlighted below the line, possibly indicating a reference point or value.

3. **Description of interactive elements**:
- There appears to be no interactive elements like buttons or dropdowns present in the interface based on the provided static image; however, it seems likely that this might be part of a larger problem-solving application where users would select one of four options (A:10.8m, B:9m, C:7.5m, D:0.3m) based on the calculated height of Tree AB.

4. **Overall description**:

This image displays a mathematical problem centered around calculating the height of a tree named Tree AB by comparing shadows cast under sunlight with known measurements of another object (Xiaoliang). The user must use the information provided, including the shadow length of 1.5 meters and the height of Xiaoliang at 1.8 meters, along with given choices among three possible heights for Tree AB. This setup is typical for educational tools aimed at teaching geometric calculations related to light rays and their interactions with surfaces. The presence of a simple graphical representation suggests a focus on applying basic principles of geometry to solve practical problems.

Thinking Process from DeepSeek-R1-Distill-Qwen-70B

Okay, so I've got this math problem here about calculating the height of a tree using some shadow measurements. Let me see if I can figure it out step by step.
First, the problem says that Xiaoliang's shadow length is 1.5 meters, and his height is 1.8 meters. Then there's a tree, AB, whose height we need to find. The options are A:10.8m, B:9m, C:7.5m, D:0.3m.
I remember something about similar triangles from geometry class. When two objects are casting shadows under the same sunlight, their heights and shadow lengths are proportional because the sun's rays are parallel. So, the ratio of Xiaoliang's height to his shadow should be the same as the ratio of the tree's height to its shadow.
Let me write that down:
Xiaoliang's height / Xiaoliang's shadow = Tree's height / Tree's shadow
So plugging in the numbers:
1.8 m / 1.5 m = Tree's height / Tree's shadow
Wait, but the problem doesn't directly give the tree's shadow length. Looking back at the image description, there's a horizontal line with markers B, 9, and C, with 9 highlighted. Maybe that 9 is the shadow length of the tree? Or perhaps it's the distance between B and C?
Hmm, the image shows a tree with labels A, B, C, D. The ground has B, 9, C. Since B is probably the base of the tree, and C is the end of the shadow, the distance from B to C would be the tree's shadow length. So if the number 9 is highlighted, maybe that's the shadow length, which is 9 meters.
If that's the case, then the tree's shadow is 9 meters. Now, using the proportion:
1.8 / 1.5 = Tree's height / 9
Let me solve for the tree's height. Cross-multiplying:
1.8 * 9 = 1.5 * Tree's height
16.2 = 1.5 * Tree's height
Divide both sides by 1.5:
Tree's height = 16.2 / 1.5 = 10.8 meters.
Wait, but looking at the options, A is 10.8m. But wait, let me double-check because sometimes these problems can trick you. Is the shadow length really 9 meters? Because sometimes the numbers might represent something else, like the distance from the tree to another point.
Alternatively, maybe the 9 is the height of the tree, but that doesn't make sense because the options include 9m as an option. If the tree's shadow is 9m, then our calculation makes sense. But if the 9 represents something else, like the distance from Xiaoliang to the tree, then we'd have to consider similar triangles in a different way.
Wait, another approach: if Xiaoliang's shadow is 1.5m and his height is 1.8m, the ratio is 1.8/1.5 = 1.2. So for every meter of shadow, the object is 1.2 meters tall. If the tree's shadow is 9 meters, then the height would be 1.2 * 9 = 10.8m, which is option A.
But looking at the image again, the ground has B, 9, and C. Maybe B is the base of the tree, and C is where the shadow ends, making the shadow length BC, which is 9 meters. That fits with the earlier assumption.
Alternatively, if the 9 is the height of the tree, then the answer would be B, but that wouldn't require any calculation. Since the problem asks us to calculate, it's more likely that 9 is the shadow length, leading us to 10.8m, which is option A.
Wait, but I'm a bit confused because sometimes in these problems, the shadow length might be from the top of the tree to the tip of the shadow, which could involve more steps. But given the information, I think the straightforward approach is correct.
Another thought: maybe the 9 is the distance between B and C, which is the shadow length, so yes, 9 meters. Therefore, the tree's height would be 1.8/1.5 * 9 = 10.8m, which is option A.
Wait, but the options include B:9m. Maybe I made a mistake. Let me check the ratios again.

If Xiaoliang's height is 1.8m and his shadow is 1.5m, the ratio is 1.8:1.5, which simplifies to 6:5. So, for every 5 units of shadow, the height is 6 units. If the tree's shadow is 9m, then 9 divided by 5 is 1.8, multiplied by 6 gives 10.8m. So that's correct.
Alternatively, if the tree's shadow is 7.5m, then 7.5/1.5=5, so the height would be 1.8*5=9m, which is option B. Wait, that's another possibility. Did I misinterpret the shadow length?
Looking back, the image shows the ground with B, 9, and C. It's possible that the 9 is the distance from B to C, which is the shadow length. But if the tree's shadow is actually 7.5m, then the height would be 9m. But how do I know which one it is?
Wait, maybe the image shows the ground with B, then 9, then C, meaning the distance from B to C is 9 meters. So the shadow length is 9m. Therefore, the height is 10.8m, option A.
Alternatively, maybe the 9 is the height, but that doesn't make sense because it's on the ground. So I think the correct answer is A:10.8m.
Wait, but the initial problem statement says "as shown in the figure," so without seeing the figure, it's a bit tricky. But based on the description, the ground has B, 9, C, with 9 highlighted. So I think 9 is the shadow length, making the tree's height 10.8m, option A.
Wait, but I'm second-guessing because sometimes the shadow length might be from the top of the tree to the tip, which could involve more steps. Alternatively, if Xiaoliang is standing next to the tree, his shadow and the tree's shadow form similar triangles.
Wait, let me try another approach. If Xiaoliang's height is 1.8m and his shadow is 1.5m, the ratio is 1.8/1.5 = 1.2. So for the tree, if its shadow is S meters, then its height H = 1.2S.
If the shadow length S is 9m, then H=10.8m, option A. But if the shadow length is 7.5m, then H=9m, option B.
But according to the image, the ground has B, 9, C, so I think S=9m, so H=10.8m, option A.
Wait, but maybe the 9 is not the shadow length but the distance from Xiaoliang to the tree. Then we'd have to set up similar triangles considering both distances. Let me think.
If Xiaoliang's shadow is 1.5m, and the tree's shadow extends beyond him, then the total shadow length would be Xiaoliang's shadow plus the distance from him to the tree. But the problem doesn't specify that. Hmm.
Alternatively, maybe the 9 is the distance from the tree to the tip of its shadow, which would be the shadow length. So S=9m, H=1.2*9=10.8m, option A.
I think that's the most straightforward interpretation. So the answer should be A:10.8m.
</think>
A
Step-by-step explanation:
1. **Identify the given values**: Xiaoliang's height (1.8m) and shadow length (1.5m).
2. **Set up the proportion using similar triangles**:

```
\[
\frac{\text{Xiaoliang's height}}{\text{Xiaoliang's shadow}} = \frac{\text{Tree's height}}{\text{Tree's shadow}}
\]
```

Substituting the known values:

```
\[
\frac{1.8}{1.5} = \frac{H}{9}
\]
```

3. **Solve for the tree's height (H)**:

```
\[
H = \frac{1.8}{1.5} \times 9 = 1.2 \times 9 = 10.8 \text{ meters}
\]
```

Answer: A:10.8m

Figure 20: Visualization of thinking process with OMNICAPTIONER for math images.

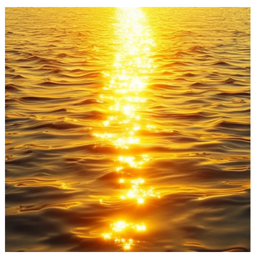
Minimalist still life photograph of a floral arrangement in a rustic ceramic vase, centrally positioned on a dark wooden surface. The vase is earthy brown with a textured finish. The arrangement features large, white chrysanthemums on the left, accompanied by glossy green leaves. On the right, there are delicate pine branches with needle-like leaves and a sprig of soft, feathery foliage. The background is a smooth, neutral beige wall, creating a serene and uncluttered atmosphere. Natural lighting, soft shadows, muted color palette, balanced composition, organic textures, tranquil, understated elegance.

Dramatic seascape photograph featuring a large, illuminated cumulus cloud dominating the center of the image against a deep blue sky. The cloud is brightly lit, creating a stark contrast with the dark ocean waves crashing onto the shore in the midground. The shoreline curves from the bottom left to the center, with wet sand reflecting the cloud's light. The horizon line is slightly above the center, with the ocean stretching out to the right. The lighting is intense and focused on the cloud, casting a soft glow on the wet sand. The scene is serene yet dramatic, with a moody atmosphere. Photorealistic, high contrast, vivid blue and white color palette, natural lighting, wide depth of field, tranquil yet powerful, visually balanced.

Close-up portrait of a young woman with long dark hair, looking directly at the camera. She is wearing a colorful, patterned sweater with shades of purple, green, and red. Her expression is neutral, with subtle makeup highlighting her eyes and lips. The background features a softly lit bedroom with a bed partially visible on the left, a lamp casting warm light, and a blurred piece of furniture or clothing in the background. The lighting is soft and natural, creating a warm and intimate atmosphere. The image has a shallow depth of field, focusing sharply on the woman's face while the background remains softly blurred. Photorealistic, warm color palette, intimate, serene, visually balanced.

Illustration of a rusted, partially submerged military structure with a large gun turret in the right foreground, surrounded by lush greenery on top. A person stands on the structure's edge, gazing into the distance. The background features towering, fluffy white clouds against a clear blue sky. The sea occupies the lower half of the image, with gentle waves and a few small figures swimming in the distance. The structure's weathered orange and brown tones contrast with the vibrant blue and white of the sky and clouds. Stylized, semi-realistic, high contrast, vivid color palette, dynamic composition, serene atmosphere, illustrative art style, sense of isolation and exploration.

A serene photograph capturing the golden reflection of the sun on a vast expanse of water. The sun is positioned at the top center, casting a brilliant, shimmering trail of light across the rippling surface. The water is textured with gentle waves, creating a rhythmic pattern that leads the eye towards the horizon. The entire scene is bathed in warm, golden hues, enhancing the tranquil and meditative atmosphere. High contrast, natural lighting, golden hour, photorealistic, expansive composition, reflective surface, peaceful, visually harmonious.

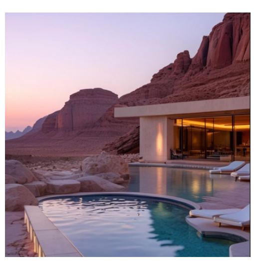


Serene landscape photograph of a dense forest reflected in a still lake. The forest occupies the upper half of the image, with a variety of trees displaying autumn foliage in shades of green, yellow, and orange. The reflection in the lake below mirrors the forest perfectly, creating a symmetrical visual effect. The left side of the image shows a slightly denser cluster of trees, while the right side is more open, revealing more of the water's surface. The water is dark and still, enhancing the reflection. Natural lighting, high contrast, vibrant autumn colors, photorealistic, tranquil atmosphere, balanced composition, reflective symmetry, crisp detail, peaceful and contemplative mood.

Close-up captures raindrops splashing onto the wet, textured surface of stone, creating concentric ripples. Two green leaves fall onto the water, one in the upper left corner and the other in the bottom center. Small patches of green moss and grass sprouts peek out from cracks in the stone, adding natural detail to the image. The stone takes on a dark, earthy hue with traces of rust and moss, contrasting with the bright green leaves. Natural light, high contrast, shallow depth of field, tranquil atmosphere, organic textures, dynamic water patterns, earthy color palette, tranquil, contemplative mood.

A close-up photograph of an otter standing in shallow, icy water, with a thick layer of snow resting on its head, resembling a hat. The otter is positioned slightly to the right of the center, facing forward with a curious expression. Its fur is a mix of dark brown and white, with the white fur more prominent on its face and chest. The background is a blurred expanse of icy water with patches of snow, creating a serene winter scene. Natural lighting, high contrast, photorealistic, sharp focus on the otter, soft focus on the background, cool color palette, tranquil, whimsical, visually balanced.

Modern architectural photography showing a luxury desert resort at dusk. In the foreground, a tranquil swimming pool is surrounded by rocky landforms, reflecting soft ambient light. Lounge chairs are placed around the pool, facing the water. In the midground, a modern-style building with large windows and warm interior lighting contrasts with the rugged desert landscape. In the background, towering rocky landforms appear particularly spectacular under the pastel sky. The entire scene is serene and luxurious, with natural and artificial elements blending harmoniously. Photography style: Medium Depth of Field, Soft Natural Light, Warm Color Pairing, High Contrast, Architectural Photography, Tranquility, Visual Balance, Harmonious Blend of Natural and Artificial Elements.



This is an illustration of a young man sitting in a traditional Japanese room. He is in the center of the frame, wearing a white shirt and green pants, holding a book with a red cover. The room is filled with various objects, including a vending machine filled with colorful drinks in the foreground and a green and orange fish-shaped ornament hanging above his head. The walls are decorated with posters and art, including a yellow bird above his head. There are several wooden chairs scattered around the room, one of which is draped with a blue towel. Sunlight shines through the windows, illuminating the scene and casting soft shadows. This illustration is colorful and clean-lined, with a manga style, warm color combinations, dynamic composition, and a warm, inviting atmosphere.

Figure 21: The detailed caption from OMNICAPTIONER enhances the alignment capability of text-to-image generation by providing precise descriptions, ensuring that the generated image accurately reflects the intended concepts, attributes, and relationships. The generation model here is fine-tuned on images labeled by OMNICAPTIONER, using the SANA 1.0 model with 1.6B parameters.

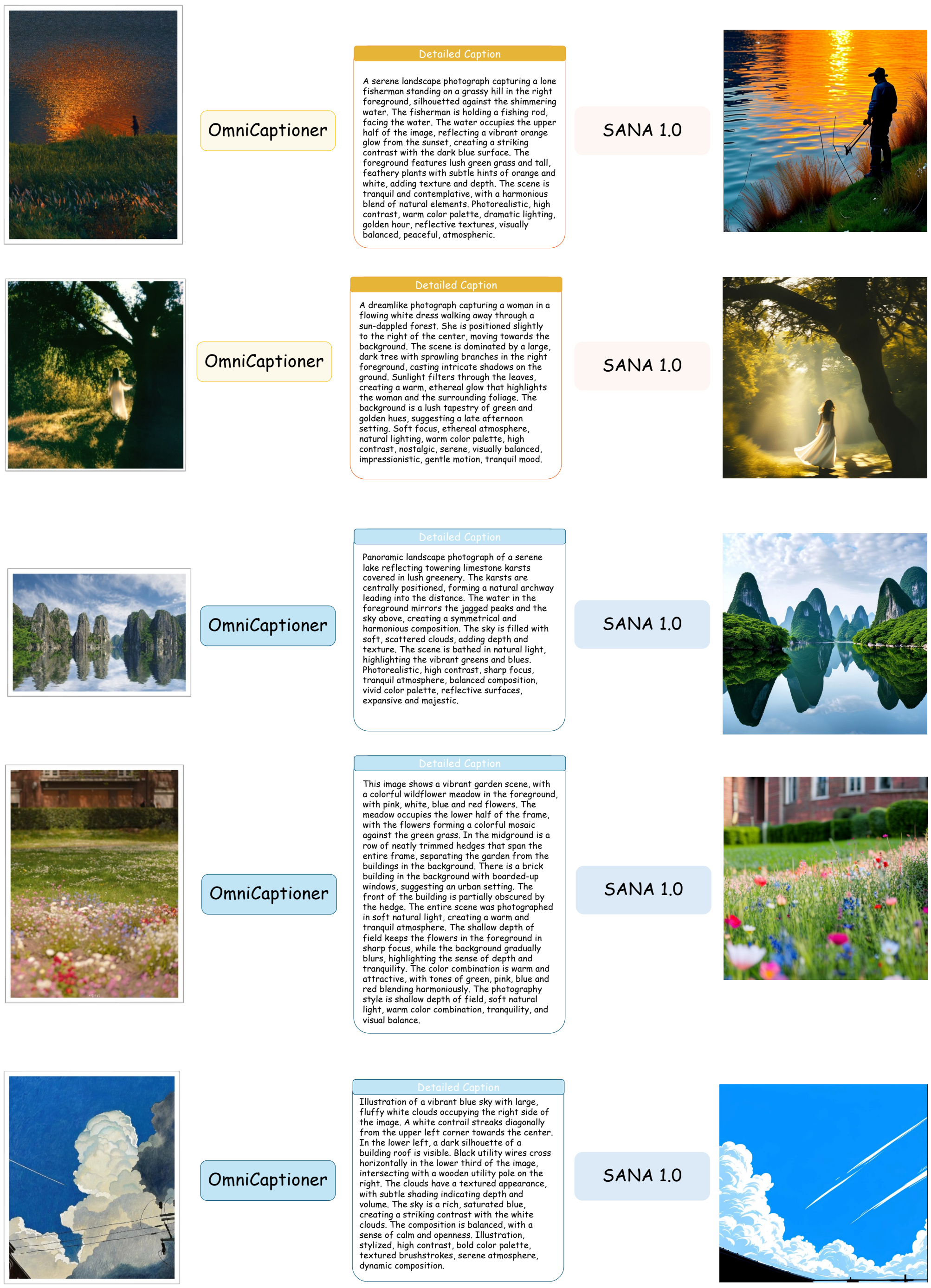


Figure 22: Image Conversion through OMNICAPTIONER and SANA-1.0. The generation model, SANA-1.0, is fine-tuned on images annotated by OMNICAPTIONER, enabling more accurate and semantically aligned image generation.