OmniAlign-V: Towards Enhanced Alignment of MLLMs with Human Preference

Xiangyu Zhao[1,2*] **Shengyuan Ding**[2,3*] **Zicheng Zhang**[1,2]
Haian Huang[2] **Maosong Cao**[2] **Weiyun Wang**[2,4] **Jiaqi Wang**[2] **Xinyu Fang**[2,5]
Wenhai Wang[2] **Guangtao Zhai**[1,2] **Haodong Duan**[2†] **Hua Yang**[1,2†] **Kai Chen**[2†]
Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University[1]
Shanghai AI Laboratory[2] Nanjing University[3] Fudan University[4] Zhejiang University[5]

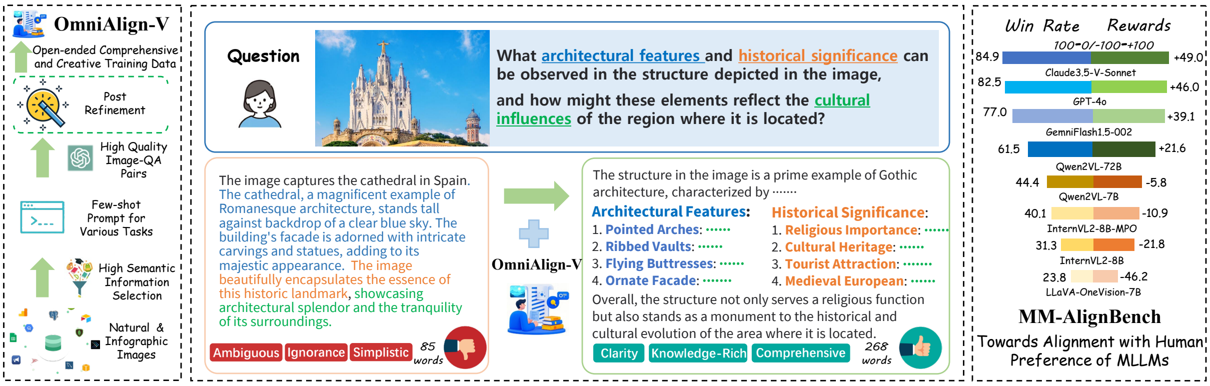


Figure 1: **OmniAlign-V** consists of curated images paired with open-ended, comprehensive question-answer pairs, significantly improving the alignment of MLLMs with human preferences. Additionally, we introduce **MM-AlignBench**, a human-annotated, high-quality benchmark specifically designed to evaluate the ability of MLLMs to align with human values.

Abstract

Recent advancements in open-source multi-modal large language models (MLLMs) have primarily focused on enhancing foundational capabilities, leaving a significant gap in human preference alignment. This paper introduces **OmniAlign-V**, a comprehensive dataset of 200K high-quality training samples featuring diverse images, complex questions, and varied response formats to improve MLLMs' alignment with human preferences. We also present **MM-AlignBench**, a human-annotated benchmark specifically designed to evaluate MLLMs' alignment with human values. Experimental results show that finetuning MLLMs with OmniAlign-V, using Supervised Fine-Tuning (SFT) or Direct Preference Optimization (DPO), significantly enhances human preference alignment while maintaining or enhancing performance on standard VQA benchmarks, preserving their fundamental capabilities.

1 Introduction

With the rapid advancement of large language models (LLMs) (OpenAI, 2023a; Touvron et al., 2023), multi-modal large language models (MLLMs) (OpenAI, 2023b; Team et al., 2023) have also seen significant improvements. Most open-source MLLMs (Liu et al., 2023b; Chen et al., 2023b) are developed by connecting a vision encoder (Dosovitskiy, 2020) to a pretrained LLM, followed by vision instruction tuning. Existing vision instruction tuning datasets (Liu et al., 2023b; Chen et al., 2023a, 2024a) and multi-modal evaluation benchmarks (Liu et al., 2023d; Yue et al., 2023; Lu et al., 2023) primarily focus on assessing fundamental capabilities (object recognition, OCR, *etc.*) of MLLMs, while paying little attention to human preference alignment. Consequently, while open-source MLLMs achieve comparable performance to proprietary counterparts on objective metrics of these foundational skills, they display a significant gap in the alignment with human preferences, which detrimentally impacts user experience in multi-modal conversational interactions, as demonstrated in Appx. A.

In this work, a preliminary investigation was conducted to quantitatively assess the degradation of human preference alignment in MLLMs(Sec. 3). Experimental results on text-only subjective evaluation benchmarks (Dubois et al., 2024; Li et al., 2024b) revealed that MLLMs exhibited a sub-

[*]Equal Contribution.
[†]Corresponding Author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18490–18515
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

stantial performance drop compared to their corresponding LLMs. A plausible hypothesis suggests that MLLMs suffer from catastrophic forgetting during vision instruction tuning. Following this hypothesis, a straightforward solution would be incorporating both multi-modal and text-only SFT data (Xu et al., 2024; Cao et al., 2025; Ding et al., 2023) for joint learning. However, our experiments (Tab. 2) indicate that this approach not only failed to yield improvements in human preference alignment under multi-modal contexts but also demonstrated negative impacts on the foundational multi-modal skills of MLLMs. These observations suggest that enhancing human alignment in multi-modal scenarios necessitates the development of specialized multi-modal instruction tuning datasets.

Existing multi-modal instruction tuning datasets predominantly focus on foundational capabilities, featuring simple language patterns and uniform response styles (Fig. 5 in appendix). This study argues that effective multi-modal datasets for enhancing human preference alignment should incorporate several critical characteristics: **open-ended questions**, **broad topic coverage**, **diverse response formats** (varying in length and style), and **strict adherence to instructions**. Based on these principles, we constructed **OmniAlign-V**. In terms of image sources, OmniAlign-V encompasses natural images and infographics such as posters and charts. Furthermore, a novel solution was developed to filter out semantically rich images from natural image collections. With respect to tasks, complex knowledge-based question answering, creative tasks, and reasoning tasks were designed for different image types. Each task category incorporates diverse subtasks, and state-of-the-art MLLMs were leveraged to obtain diverse, high-quality responses. Ultimately, OmniAlign-V comprises ∼200K multi-modal instruction tuning samples, exhibiting a significantly different overall data distribution compared to existing multi-modal SFT datasets.

Leveraging OmniAlign-V, comprehensive experiments were conducted to explore its full potential for enhancing human alignment in MLLMs. Integrating OmniAlign-V into the SFT stage of the LLaVA-NeXT (Liu et al., 2024a) structure with InternLM2.5-7B and Qwen2.5-32B yielded significant improvements in human preference alignment across various powerful LLM encoders, including InternLM2.5 (Cai et al., 2024) and Qwen2.5 (Bai et al., 2023a). Furthermore, on ground-truth-based VQA benchmarks like MMMU (Yue et al., 2023) or OCRBench (Liu et al., 2023e), MLLMs finetuned with OmniAlign-V displayed comparable or superior performance. Beyond its application in SFT, OmniAlign-V also demonstrated additional value when applied to Direct Preference Optimization (DPO) (Rafailov et al., 2024). Experimental results indicate that implementing OmniAlign-V for DPO yields further improvements in preference alignment, surpassing baseline models finetuned on the same dataset. After the SFT and DPO stage, **our LLaVA-Next baseline with Qwen2.5-32B surpasses the state-of-the-art model Qwen2VL-72B** (Chen et al., 2024b) finetuned on extensive proprietary datasets.

Throughout the exploration, we also observed that existing multi-modal human preference benchmarks (Lu et al., 2024; Qian et al., 2024) lack diversity in image sources, contain repetitive questions, and lack clarity. To address these limitations, we introduce **MM-AlignBench**, a high-quality benchmark comprising 252 carefully curated samples with diverse image sources and meticulously crafted questions by human annotators, enabling comprehensive evaluation of MLLMs' alignment with human preferences. MM-AlignBench, along with other MLLM human preference benchmarks, was employed throughout this study for evaluation.

The contributions of this study are as follows:

1. An in-depth investigation into the degradation of MLLM human alignment, analyzing the impact of both text-only and multi-modal tuning data.

2. The introduction of OmniAlign-V, a comprehensive open-ended multi-modal SFT dataset, complemented by OmniAlign-V-DPO. Extensive experiments demonstrate the effectiveness of these datasets in improving human preference alignment.

3. The development of MM-AlignBench, a carefully curated benchmark comprising high-quality, human-annotated samples specifically designed to evaluate MLLMs' human preference alignment.

2 Related Work

Alignment of LLMs. Alignment, encompassing the ability to follow human instructions and provide helpful assistance (Liu et al., 2024b), has long been a critical aspect of LLMs. Recent works (Wang et al., 2023; Xu et al., 2024; Cao et al., 2025) focused on generating high-quality SFT training data to enhance the alignment of LLMs. Besides, recent benchmarks (Liu et al., 2023c; Li et al., 2023, 2024b) have introduced open-ended and challenging questions to assess the alignment performance of LLMs. However, there is a notable lack of benchmarks designed to evaluate the alignment of MLLMs.

Visual Question Answering Datasets. In the early stages, Visual Question Answering (VQA) datasets were primarily used for basic semantic alignment between images and text (Radford et al., 2021; Li et al., 2022; Zhao et al., 2024). These datasets were relatively simple in structure. Popular VQA datasets (Hudson and Manning, 2019; Antol et al., 2015) predominantly featured straightforward questions that elicited single-word or short-sentence answers. Similarly, visual classification and detection datasets (Lin et al., 2014; Schuhmann et al., 2021; Sharma et al., 2018) typically consisted of brief descriptions that focused on the main object within the image.

Instruction Tuning data. With the rapid development of MLLMs (Chen et al., 2024b; Bai et al., 2023b), the instruction tuning data has gained more attention. LLaVA (Liu et al., 2023b) leveraged advanced LLMs to generate instruction-following formatted data from traditional VQA datasets. Recent works have further expanded this approach by employing MLLM to generate captions (Chen et al., 2023a), complex question-answer pairs (Chen et al., 2024a; Gu et al., 2024), OCR data (Carter, 2024) and math-related data (Shi et al., 2024). Other efforts (Tong et al., 2024; Li et al., 2024a) aggregates multiple publicly available document image datasets into unified resources. However, these datasets primarily focus on fundamental visual capabilities with short dialogues and factual questions, resulting in suboptimal alignment with human preferences.

3 Human Alignment of MLLMs: The Preliminary Study

When handling open-ended and flexible questions involving images, state-of-the-art open-source MLLMs – despite excelling in certain recognition tasks – exhibit significantly weaker alignment with human preferences compared to GPT-4o (Fig. 6 in appendix). We hypothesize that this decline in multi-modal preference alignment stems from a reduction in the language model's proficiency after the multi-modal SFT stage. To test this hypothesis, we evaluated state-of-the-art MLLMs on text-only human preference alignment benchmarks (Liu et al., 2023c; Li et al., 2023, 2024b). Additionally, we constructed a LLaVA (Liu et al., 2023b) baseline using InternLM-2.5-7B, fine-tuned on the LLaVA-Next-778k (Liu et al., 2023a) SFT dataset. As shown in Tab. 1, the LLM's ability to handle text-only open-ended questions degrades significantly after multi-modal SFT. This degradation may be due to (1) insufficient quantity or quality of text-only samples during multi-modal SFT, or (2) the overly simplistic style of multi-modal fine-tuning data derived from traditional VQA datasets (Goyal et al., 2017; Hudson and Manning, 2019), often consisting of simple questions and short, factual answers.

Method	Type	AlignBench	AlpacaEval-V2	ArenaHard
InternLM2.5-7B	LLM	6.36	27.58	27.06
InternVL2-8B	MLLM	4.04(-36.4%)	3.35(-87.9%)	4.65(-82.8%)
LLaVA-Internlm	MLLM	4.66(-26.7%)	4.22(-84.7%)	4.93(-81.8%)
Qwen2-7B	LLM	6.02	24.47	32.84
Qwen2VL-7B	MLLM	4.92(-18.3%)	3.85(-83.4%)	6.46(-80.3%)
LLaMA3-8B	LLM	4.88	30.19	31.96
MiniCPM-V2.5	MLLM	3.84(-21.3%)	7.33(-75.7%)	8.05(-74.8%)
InternLM2-20B	LLM	5.49	43.35	33.75
InternVL2-26B	MLLM	4.39(-20.0%)	5.34(-87.7%)	11.25(-66.7%)
Hermes2-llama3-70b	LLM	5.72	46.09	57.40
InternVL2-76B	MLLM	4.33(-24.3%)	8.32(-81.9%)	16.17(-72.3%)

Table 1: **Language Alignment Benchmark Results.** After multi-modal SFT, MLLMs demonstrate significant decline in human preference alignment compared to their corresponding LLMs.

To assess the impact of the first factor on multi-modal alignment, we examined the role of text-only data in multi-modal SFT. The LLaVA-Next-778K dataset includes approximately 40K text-only samples from ShareGPT (Chiang et al., 2023), which are relatively outdated and of lower quality. To improve alignment, we experimented with replacing these samples with better-quality alternatives and increasing their proportion in the multi-modal fine-tuning data. Specifically, we sampled 40K / 80K instances from two high-quality language SFT datasets: Magpie-Llama3.3 (Xu et al., 2024) and Condor (Cao et al., 2025), using these to replace the original language data in LLaVA-Next. We then evaluated models fine-tuned on these different data mixtures using both text-only benchmarks and multi-modal benchmarks, including WildVision (Lu et al., 2024) for human preference alignment and various benchmarks from OpenVLM Leaderboard (Duan et al., 2024) for assessing various fundamental multi-modal capabilities.*

Results in Tab. 2 reveal several key findings. While MLLMs tuned with higher-quality text-only samples show significant improvements on text-

*MMVet is open-ended, while others are closed-ended.

Model	*Language Benchmarks*			*Multi-modal Benchmarks*				
	AlignBench	AlpacaEval-V2	ArenaHard	WildVision	MMVet	MMBench-V1.1	AI2D	OCRBench
LLaVA^I-LLaVANext778k	4.7	31.7	21.6	18.4/-55.1	41.2	73.7	74.2	39.7
LLaVA^I-LLaVANext_{mm}738k-Magpie40k	4.5 ↓	65.5 ↑	44.6 ↑	16.8/-58.9 ↓	37.7 ↓	73.1 ↓	73.4 ↓	38.7 ↓
LLaVA^I-LLaVANext_{mm}738k-Condor40k	5.6 ↑	72.9 ↑	55.7 ↑	16.8/-57.4 ↓	38.3 ↓	72.6 ↓	73.6 ↓	38.5 ↓
LLaVA^I-LLaVANext_{mm}738k-Condor80k	5.7 ↑	75.2 ↑	55.7 ↑	16.8/-57.0 ↓	38.3 ↓	72.6 ↓	74.0 ↓	37.6 ↓

Table 2: **Performance of Incorporating High-Quality Language Data**. LLaVA^I denotes LLaVA structure with InternLM2.5-7B as language model, and LLaVANext_{mm}738k refers to the multi-modal data in LLaVANext-778K. Integrating high-quality language data significantly improves alignment performance on language benchmarks. However, it leads to a decline in multi-modal alignment performance on benchmarks such as WildVision and MMVet. For AlpacaEval and ArenaHard, we present the winning rate against GPT-3.5.

only alignment benchmarks, they unexpectedly demonstrate degraded performance on both multi-modal alignment and common VQA benchmarks. This counter-intuitive phenomenon suggests that language alignment capability does not directly translate to multi-modal alignment. We therefore argue that high-quality multi-modal human-aligned data is crucial for improving MLLMs' human preference alignment in multi-modal contexts.

4 OmniAlign-V

Current MLLM instruction tuning datasets primarily focus on enhancing basic capabilities like perception, OCR, and mathematical reasoning. These datasets typically contain simple, brief question-answer pairs that inadequately capture human preferences and real-world interaction complexity, as shown in Fig. 5. We propose that multi-modal training data should also incorporate: (1) **Open-ended, Diverse, and Creative Questions requiring interdisciplinary knowledge**, (2) **Comprehensive and Knowledge-rich Responses**. To address these requirements, we develop a novel data synthesis pipeline to generate high-quality human-aligned multi-modal training data, resulting in the creation of **OmniAlign-V**.

4.1 Task Taxonomy: A Big Picture

Image content plays a crucial role in constructing multi-modal training data. To ensure comprehensive coverage, we classify images into two major categories: natural images and infographic images, as shown in Fig. 3(a). Our data synthesis pipeline first determines whether an input image belongs to natural images (captured from real-world scenes) or infographic images (human-created to convey information). Based on this classification, different vision-language tasks are assigned.

For natural images, we define three primary tasks: **Knowledge**, **Inferential**, and **Creation**, each requiring diverse and complex question formats with comprehensive, reasoned responses. These tasks enhance the model's ability to interpret real-world scenes effectively.

For infographic images, given their diverse content, we identify four key types that elicit intricate and challenging questions: **Arts**, **Charts**, **Diagrams**, and **Posters**. These categories necessitate a deep understanding of human-designed visuals, including both abstract and detailed elements.

4.2 Image Selection Strategy

For natural images, rich semantic content leads to more comprehensive and insightful QA pairs. To enhance training data quality, we developed a pipeline to select semantically rich images from diverse sources, as shown in Fig. 2(a). This pipeline consists of two key steps.

First, we use the Image Complexity (IC) recognition model IC9600 (Feng et al., 2022) to assign IC scores to images. Images with low semantic content —- characterized by few objects or simple, uniform backgrounds —- receive low IC scores and are then excluded. While IC9600 effectively filters out low-complexity images, high IC scores alone do not guarantee semantic richness. For instance, as shown in Fig. 7, an image filled with tents may have a high IC score but lacks sufficient semantic information for multi-modal training.

To refine selection, we employ the Recognize Anything Model (Zhang et al., 2024) to identify objects within images, filtering out those with high complexity but minimal meaningful content. This two-step approach ensures our pipeline accurately selects images that are both complex and semantically rich. Experiment results have validated the effectiveness of our proposed image selection strategy, as shown in Appx. E.

4.3 Data Generation Pipeline

SFT QA Generation. In Fig. 2(b), we outline the generation process of SFT QA pairs. For vision-language tasks involving natural images

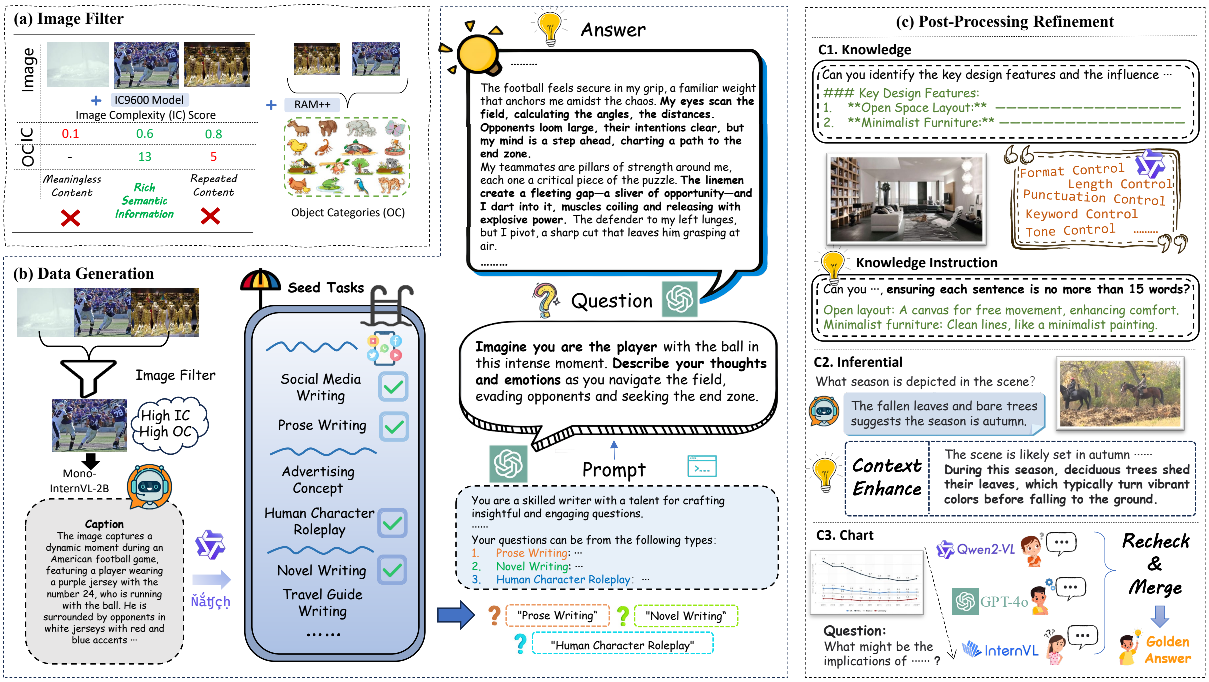


Figure 2: **Overall pipeline of OmniAlign-V**. By utilizing an image filter and employing a customized pipeline for distinct tasks, we curate semantically rich images paired with high-quality open-ended question-answer sets. Post-refinement further enhances both the variety and quality of our dataset.

(knowledge, creative, inferential), we first apply our image selection strategy to filter semantically rich images from CC-3M(Sharma et al., 2018), Flickr30k (Plummer et al., 2015), and GQA (Hudson and Manning, 2019). For infographic image tasks, images are collected from various existing sources (Masry et al., 2022; Li et al., 2018; Kembhavi et al., 2016). More details are provided in Appx. B.

Knowledge & Inferential Tasks. In preliminary experiments, we found that GPT-4o can generate diverse, content-relevant questions for the two tasks when provided with well-designed few-shot prompts. Consequently, we carefully designed a single prompt for each task category, incorporating comprehensive instructions and selected few-shot examples. These prompts were then used with GPT-4o across diverse images to generate QA pairs.

Creative Tasks. We noticed that a single prompt cannot generate sufficiently diverse, content-relevant creative questions. Therefore, we developed a more sophisticated pipeline inspired by Condor (Cao et al., 2025). First, we created a set of seed creative questions:

$$\mathcal{Q}_s = \{Q_1, Q_2, ...Q_N\} \tag{1}$$

where each seed question corresponds to a distinct creative task. Since directly using all seed questions as few-shot examples leads to repetition and lack of diversity, we employ a light-weight MLLM (Luo et al., 2024) to generate detailed captions C for each image. An LLM $\mathcal{M}$ then selects a relevant subset of seed questions according to the caption:

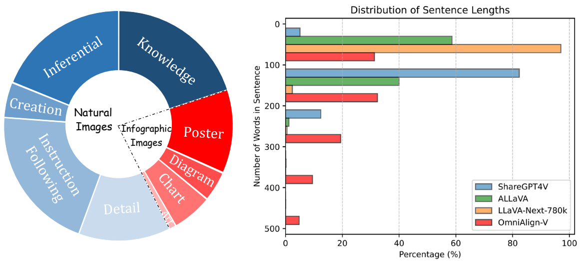
$$\mathcal{Q}_s' = \mathcal{M}(C, \mathcal{Q}_s), \|Q_s'\| \ll \|Q_s\| \tag{2}$$

Finally, we randomly select three question types from Q_s' as few-shot examples for GPT-4o, preserving both quality and diversity of synthesized data.

Infographic Tasks. For infographic image tasks (*Charts*, *Diagrams*, *etc.*), questions and answers are closely tied to specific image content. Unlike natural images, these visuals convey information primarily through text, colors, lines, and symbolic elements, making evaluation based on image complexity or object categories unsuitable. Therefore, instead of applying image selection strategies, we carefully select image sources containing rich, detailed information. We then design specialized prompts for GPT-4o to generate questions that require comprehensive background knowledge understanding. The difference is shown below:

InfographicVQA: What is the respiratory disease death rate for individuals aged 70+?

OmniAlign-V: How does the respiratory disease death rate for individuals aged 70+ compare to the other age groups, and what might this suggest?



(a) Distribution of tasks. (b) Comparison with other datasets.

Figure 3: **Data distribution of OmniAlign-V.** Our dataset includes a diverse range of tasks, characterized by a more balanced distribution of answer lengths compared to those observed in ALLaVA and ShareGPT4V.

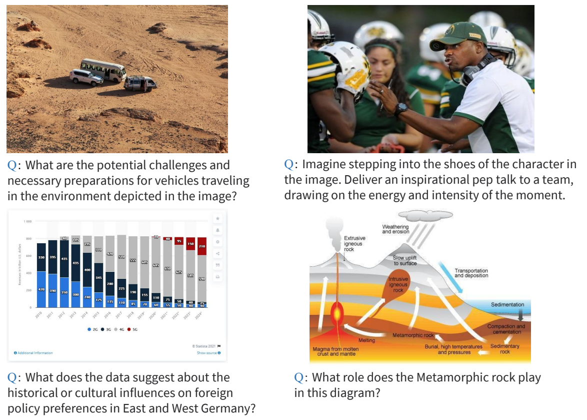


Figure 4: **Samples in MM-AlignBench**.

Post Refinement. To further improve the quality of the synthesized data, we implemented a series of post-processing methods for refinement, as illustrated in Fig. 2(c).

Instruction Augmented Knowledge QAs. Instruction following is a crucial capability significantly impacting human preference. To enhance this, we incorporate complex instructions and restrictions into our knowledge QA pairs and reformulate responses accordingly. As shown in Fig. 2(c1), for each knowledge QA, we use a powerful LLM to select an appropriate instruction type that can be integrated into the question without depending on visual content. The instruction is then incorporated into the existing question, and an LLM adjusts the corresponding answer to ensure alignment with both the modified question and original context, resulting in instruction-augmented knowledge QAs.

Enriched Inferential QAs. For many inferential QAs, the answers lack sufficient detail to fully explain underlying logic and background knowledge. To address this, we employ a knowledge-rich LLM to enrich responses with detailed explanations, relevant background information, and logical reasoning (Fig. 2(c2)). This refinement enhances alignment with user preferences.

Quality Improved Infographic QAs. For infographic tasks, particularly those involving *Chart* data, we observed that even SOTA MLLMs struggle with complex charts and detailed questions. While GPT-4o excels at explaining background knowledge but often produces inaccurate OCR results, SOTA open-source MLLMs (Chen et al., 2024b; Bai et al., 2023b) show superior OCR accuracy but lack detailed explanations. Therefore, we developed a refinement pipeline to generate responses combining rich background knowledge and accurate OCR results. We filter out questions where GPT-4o and SOTA open-source MLLM responses show significant discrepancies in trends. For remaining questions, the responses from different MLLMs are merged to produce a final response that is both precise and comprehensive. The merged answers are further reviewed by human experts to ensure their quality and consistency. More details are demonstrated in Appx. C.

OmniAlign-V comprises 39K Knowledge QAs, 37K Inferential QAs, 10K Creative QAs, 38K Instruction-Following (Knowledge) QAs, and 44K Infographic QAs (2K Art, 8K Diagram, 11K Chart, 23K Poster). Additionally, we prompt GPT-4o to generate 35K QAs focusing on image details, resulting in a total of 205K high-quality SFT training samples. Examples are shown in Figs. 9 to 12.

DPO Data Generation. OmniAlign-V's high-quality, human-aligned QA pairs can serve as positive samples for DPO training. Inspired by Reject Sampling (Casella et al., 2004), we generate negative samples by prompting a LLaVA-Next baseline (generator G) trained on LLaVA-Next-778k. For each question Q_i, the generator produces N responses with high temperature, $\mathbf{R} = \{r_1^i, r_2^i, ...r_N^i\}$. An LLM Judger J then evaluates these responses to select the one that most deviates from the original question's intent and context as the negative sample r_{Neg}^i, ensuring clear contrast between positive and negative samples.

5 MM-AlignBench

Current benchmarks for assessing multi-modal alignment capabilities are limited. While WildVision (Lu et al., 2024) aims to evaluate human preferences in real-world interactions, it employs repetitive and simplistic question formats that inadequately assess response quality, as shown in Fig. 8.

Model	Data	LLM	MM-AlignBench	WildVision	MIA-Bench	MMVet	MMMU	MMBenchV1.1	AI2D	OCRBench
LLaVA	LLaVANext-778k	InternLM2.5-7B	3.6 / -82.1	18.4 / -55.1	75.4	41.2	42.6	73.6	74.1	39.7
LLaVA	OmniAlign-V$_{mix}$	InternLM2.5-7B	50.0 / +3.8	28.2 / -34.6	85.4	43.5	43.3	73.7	74.7	41.3
			+ 46.4 / 85.9	+ 9.8 / 20.5	+ 10.0	+ 2.3	+ 0.7	+ 0.1	+ 0.6	+ 1.6
LLaVANext	LLaVANext-778k	InternLM2.5-7B	20.6 / -42.7	23.4 / -45.0	76.9	41.8	44.1	75.1	74.7	56.2
LLaVANext	OmniAlign-V$_{mix}$	InternLM2.5-7B	57.1 / +11.1	29.6 / -31.3	86.7	47.7	46.8	74.9	77.5	58.9
			+ 36.5 / 53.8	+ 6.2 / 13.7	+ 9.8	+ 5.9	+ 2.7	- 0.2	+ 2.8	+ 2.7
LLaVANext	LLaVANext-778k	Qwen2.5-32B	26.6 / -29.0	25.2 / -41.3	86.0	47.7	55.2	79.3	79.6	55.9
LLaVANext	OmniAlign-V$_{mix}$	Qwen2.5-32B	62.3 / +19.4	40.2 / -14.9	89.6	56.9	60.7	80.6	81.7	55.9
			+ 35.7 / 48.4	+ 15.0/26.4	+ 3.6	+ 9.2	+ 5.5	+ 1.3	+ 2.1	+ 0.0

Table 3: **Evaluation Results of Integrating OmniAlign-V into the SFT stage**. By integrating OmniAlign-V, the multi-modal preference alignment of MLLMs significantly improved. Additionally, we also observe comparable or better performance on common VQA benchmarks. In 'Model' column, LLaVA and LLaVANeXT denote the model structure and training strategy. For MM-AlignBench and WildVision, notation *A/B* denotes *Winning Rate/Rewards*.

Model	Type	AlpacaEvalv2	ArenaHard
		v.s. GPT-3.5/GPT-4	
LLaVANext^{I}	MLLM	29.8 / 3.8	21.4 / 4.93
LLaVANext^{I}-OA	MLLM	50.1 / 7.8	30.4 / 9.33
InternLM2.5-7B	LLM	78.3 / 26.2	47.5 / 19.1
LLaVANext^{Q}	MLLM	50.6 / 7.0	54.5 / 18.0
LLaVANext^{Q}-OA	MLLM	77.6 / 18.0	87.2 / 58.1
Qwen2.5-32B	LLM	92.1 / 37.1	94.8 / 75.9

Table 4: **Performance on text-only alignment benchmarks.** OA denotes models trained with OmniAlign-V. We present the winning rate against GPT-3.5 and GPT-4 (original setting). OmniAlign-V can also enhance MLLM's performance on text-only alignment benchmarks.

To address this, we developed MM-AlignBench, featuring human-curated images and questions for more nuanced evaluation.

We selected high-quality images from SAM-1B (Kirillov et al., 2023), CC-3M-Test (Sharma et al., 2018), AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), and InfographicVQA (Mathew et al., 2022). For natural images, we applied our image selection strategy from Sec. 4.2 to identify 2,000 semantically rich images and combine them with 1,000 carefully selected infographics images. GPT-4o was then used to generate diverse questions for these images. Human experts reviewed and refined the image-question pairs, filtering out low-quality, repetitive, or contextually weak samples. This process resulted in 252 high-quality question-image pairs, featuring diverse question types and semantically rich images. Several examples are shown in Fig. 4.

For evaluation, we follow WildVision's approach, using GPT-4o as the judge model to compare model responses with reference responses generated by Claude3V-Sonnet (Anthropic, 2024), reporting winning rates and reward scores.

6 Evaluation Results

6.1 SFT with OmniAlign-V

We conduct extensive experiments to demonstrate OmniAlign-V's effectiveness. We combine OmniAlign-V with LLaVA-Next-778k (excluding text-only samples), creating OmniAlign-V$_{mix}$ with 946K training samples. We evaluate various MLLMs tuned on OmniAlign-V against their counterparts tuned on LLaVA-Next-778k.

Our evaluation spans multiple multi-modal benchmarks, including standard VQA benchmarks (Yu et al., 2023; Liu et al., 2024c, 2023e; Yue et al., 2023; Kembhavi et al., 2016) and human-preference alignment benchmarks: MM-AlignBench, WildVision (Lu et al., 2024), and MIA-Bench (Qian et al., 2024). Results in Tab. 3 show that OmniAlign-V significantly improves human alignment across all benchmarks. Moreover, our training data improves general multi-modal capabilities, particularly on MMVet and MMMU, demonstrating a trend distinct from text-only data.

Notably, despite excluding language samples from training data, models maintain stronger language alignment than those trained on LLaVA-Next-778k, as shown in Table 4. This suggests that while high-quality language data alone may not significantly impact multi-modal capabilities, enhancing multi-modal data quality can improve both language and multi-modal performance, highlighting the crucial role of high-quality, human-aligned multi-modal training data.

6.2 DPO with OmniAlign-V-DPO

We conduct DPO post-training on three models: LLaVA-Next trained with LLaVA-Next-778k, LLaVA-Next trained with OmniAlign-V$_{mix}$, and InternVL2-8B. Results in Tab. 5 show that DPO tuning significantly improves performance on real-world questions (WildVision) across all models.

Model	Stage	MM-AlignBench	WildVision
LLaVANext[I]	SFT	9.5 / -69.2	30.4 / -34.2
LLaVANext[I]	SFT+DPO	11.1 / -64.5	35.5 / -23.4
LLaVANext[I]-OA	SFT	57.1 / +11.1	29.6 / -31.3
LLaVANext[I]-OA	SFT+DPO	64.3 / +22.4	41.8 / -10.1
InternVL2-8B	SFT	31.4 / -21.8	48.6 / +1.4
InternVL2-8B	SFT+DPO	64.7 / +19.4	51.4 / +1.9

Table 5: **Performance of applying DPO with OmniAlign-V-DPO**. For models finetuned with human-aligned data, by employing DPO training, the model's alignment with human perference further improved.

Model	Win Rate ↑	Reward↑	B+	B	T	W	W+
Claude3.5V-Sonnet	84.9	+51.4	70	144	13	25	0
GPT-4o	81.3	+49.0	81	124	12	31	4
GPT-4V	82.5	+46.0	57	151	12	31	1
GeminiFlash1.5-002	77.0	+39.1	56	138	14	35	9
LLaVANext-OA-32B-DPO	74.2	+36.9	49	138	20	40	5
Qwen2VL-72B	61.5	+21.6	43	112	15	75	7
LLaVANext-OA-32B	62.3	+19.4	31	126	19	62	14
Claude-3V-Sonnet	50	0	-	-	-	-	-
Qwen2VL-7B	44.4	-5.8	28	84	5	101	34
InternVL2-72B	44.4	-6.9	19	93	8	98	34
InternVL2-8B-MPO	40.1	-10.9	26	75	10	100	41
InternVL2-8B	31.3	-21.8	18	61	15	109	49
LLaMA3.2-Vision-11B	27.8	-33.7	18	52	4	98	80
LLaVANext-Qwen32B	26.6	-29.0	16	51	10	121	54
LLaVA-OneVision-7B	23.8	-46.2	14	46	1	75	116
MiniCPM-V-2.5	12.7	-53.0	9	23	8	116	96
Xcomposer2.5-7B	7.5	-74.0	5	14	3	63	167
Idefics3-8B	2.7	-92.3	3	4	0	15	230

Table 6: **Performance of existing MLLMs on MM-AlignBench**. B+/B/T/W/W+ denotes MuchBetter/Better/Tie/Worse/MuchWorse. Our LLaVA-Next-OmniAlign(OA)-32B-DPO, trained with OmniAlign-V and applied DPO with OmniAlign-V-DPO, demonstrates outstanding performance, surpassing a wide range of strong MLLMs, even Qwen2VL-72B.

While the baseline trained solely on LLaVA-Next-778k shows minimal improvement on MM-AlignBench, models incorporating OmniAlign-V during SFT demonstrate substantial gains after DPO. Similarly, InternVL2-8B, a state-of-the-art MLLM partially trained on proprietary multi-modal corpora, shows significant improvement on MM-AlignBench post-DPO. This indicates that if a model has been trained on data aligned with human preferences, such as open-ended or long-context data during SFT phase, the subsequent DPO training on high-quality human-aligned data can significantly activate the model's ability, leading to a considerable improvement in alignment performance. In contrast, if the model has not been exposed to such alignment-focused datasets during SFT, training with open-ended data alone via DPO will not significantly improve its capabilities of alignment. These findings demonstrate the value of OmniAlign-V in both SFT and DPO stages for enhancing human preference alignment.

Model	MM-AlignBench	WildVision	MMVet
LLaVANext-778k	20.6 / -42.7	23.4 / -45.0	41.7
+Know / Infer / Detail	23.4 / -42.1	23.2 / -43.2	42.8
+Instruction Following	36.5 / -17.3	26.2 / -37.5	44.6
+Creation	43.7 / -5.0	26.6 / -37.7	44.4
+Chart / Diagram / Poster	57.1 / +11.1	29.6 / -31.3	47.7

Table 7: **Ablation of Subsets in OmniAlign-V.** By progressively incorporating tasks within OmniAlign-V, the model's alignment performance gradually improves.

6.3 MM-AlignBench

We evaluate various state-of-the-art MLLMs (OpenAI, 2023a; Team et al., 2023; Anthropic, 2024; Bai et al., 2023b; Chen et al., 2024b; Li et al., 2024a; OpenBMB, 2024; Dong et al., 2024; Meta, 2024; Laurençon et al., 2024; Wang et al., 2024) on MM-Alignbench, with results shown in Tab. 6. Closed-source models like GPT, Claude, and Gemini demonstrate strong alignment in responding to open-ended questions. In contrast, while Qwen2-VL and InternVL2 excel at common VQA tasks, they show relatively lower human preference alignment. This highlights the importance of prioritizing MLLM alignment for improved daily human interactions. Our LLaVA-OA-32B, trained with OmniAlign-V, achieves exceptional performance, outperforming numerous strong MLLMs and nearly matching Qwen2VL-72B. After applying DPO with OmniAlign-V-DPO, **LLaVA-OA-32B-DPO achieves winning rate of 72.6 with an average reward of +33.5, surpassing the performance of Qwen2VL-72B**. These results highlight the high quality and effectiveness of the OmniAlign-V dataset in improving model alignment with human preferences.

6.4 Ablation Study

We conduct an ablation study to evaluate each subset of OmniAlign-V, reporting results on MM-Alignbench, WildVision, and MMVet in Tab. 7. Performance improves progressively as different tasks from OmniAlign-V are incorporated. Notably, Instruction-Following data significantly enhances performance across all three benchmarks, demonstrating its crucial role. The creation data subset uniquely improves performance on MM-Alignbench but not on WildVision and MMVet. This discrepancy can be attributed to the absence of multi-modal creative question types in these two benchmarks, suggesting their incompleteness in capturing full spectrum of alignment challenges.

7 Conclusion

In this paper, we introduce **OmniAlign-V**, a dataset designed to enhance the alignment of MLLMs with human preferences, as well as **MM-AlignBench**, a high-quality, specific-purpose benchmark for evaluating such alignment. We investigate the impact of both language and multi-modal training data, emphasizing the critical role of multi-modal open-ended training data. By incorporating OmniAlign-V into SFT and DPO stages, we achieve significant improvements in the alignment of MLLMs.

8 Limitation

Although the OmniAlign-V pipeline can be easily scaled to support larger datasets, the scale of the dataset used in this paper may be insufficient for large-scale training due to cost limitations. Deeper exploration into the alignment of MLLMs is still needed to address these limitations and further advance the field.

9 Acknowledgements

This research was partly supported by grants of National Natural Science Foundation of China (NSFC, Grant No. 62171281), Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 20DZ1200203, 2021SHZDZX0102), National Key R&D Program of China (No.2022ZD0161600), the Shanghai Postdoctoral Excellence Program (No.2023023), China Postdoctoral Science Fund (No.2024M751559), and Shanghai Artificial intelligence Laboratory.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Maosong Cao, Taolin Zhang, Mo Li, Chuyu Zhang, Yunxin Liu, Haodong Duan, Songyang Zhang, and Kai Chen. 2025. Condor: Enhance llm alignment with knowledge-driven data synthesis and refinement. *arXiv preprint arXiv:2501.12273*.

Jimmy Carter. 2024. Textocr-gpt4v. `https://huggingface.co/datasets/jimmycarter/textocr-gpt4v`.

George Casella, Christian P Robert, and Martin T Wells. 2004. Generalized accept-reject sampling schemes. *Lecture notes-monograph series*, pages 342–347.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *Preprint*, arXiv:2402.11684.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi

Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Preprint*, arXiv:2404.16821.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Tinglei Feng, Yingjie Zhai, Jufeng Yang, Jie Liang, Deng-Ping Fan, Jing Zhang, Ling Shao, and Dacheng Tao. 2022. Ic9600: a benchmark dataset for automatic image complexity assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8577–8593.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. 2024. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

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

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *Preprint*, arXiv:2408.12637.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Juzheng Li, Hang Su, Jun Zhu, Siyu Wang, and Bo Zhang. 2018. Textbook question answering under instructor guidance with memory networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3655–3663.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024b. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and

Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, et al. 2023c. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2024b. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *Preprint*, arXiv:2308.05374.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023d. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

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

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023e. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

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

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*.

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

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

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Meta. 2024. Llama3.2-vision.

OpenAI. 2023a. Chatgpt. https://openai.com/blog/chatgpt.

OpenAI. 2023b. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenBMB. 2024. Minicpm: Unveiling the potential of end-side large language models.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. 2024. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

WikiArt Volunteer Team. 2025. Wikiart.

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

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

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

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

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2024. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732.

Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. 2024. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*.

A Alignment of MLLMs with Human Preference

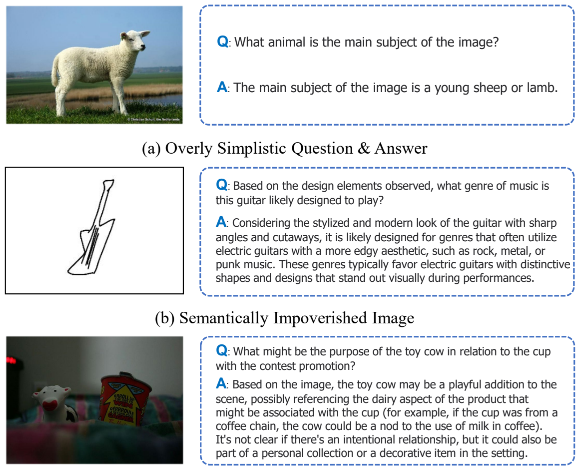
Although open-source MLLMs have already matched or even surpassed proprietary models like the GPT and Claude series in common VQA tasks like OCR and Visual Perception, a significant gap remains in their alignment with human preferences. When posed with open-ended questions that require knowledge-rich responses, even the most advanced open-source MLLM, InternVL2-76B, struggles to provide comprehensive answers with high readability, as illustrated in Fig. 6. In contrast, GPT-4o not only accurately identifies the main objects relevant to the question but also provides well-structured responses enriched with comprehensive contextual knowledge, achieving a high level of alignment with human preferences.

B Image sources

We carefully select image sources for *Arts*, *Charts*, *Diagrams*, and *Posters* tasks. For *Arts*, WikiArt (Team, 2025) is selected as the image source, offering a diverse range of painting styles. We uniformly sample 2000 images across all painting styles to ensure diversity in the dataset. For *Charts*, we select ChartQA (Masry et al., 2022), a dataset featuring charts that contain substantial statistical information. ChartQA includes several subcategories, from which we filter out simplistic charts with only two columns and retain those that contain charts with rich contextual information and diverse types. For *Diagrams*, we choose images from TextbookVQA (Team, 2025), which provides diagrams rich in natural content and detailed explanations. We exclusively utilize the question images and teaching images from the image sources, as they meet our specific requirements. For *Posters*, we utilize InfographicsVQA (Mathew et al., 2022), a dataset containing high-quality posters with intricate designs and informative content.

C Chart Post-Refine Pipeline

After curating images along with high-quality question, we first employ the current most powerful models(GPT-4o, InternVL2-72B, Qwen2VL-72B) to generate answers separately. A powerful LLM(Qwen2.5-72B) is then utilized to extract objective facts from the chart within each generated answer. The facts extracted from the answers of multiple models are compared for consistency. If



(a) Overly Simplistic Question & Answer

(b) Semantically Impoverished Image

(c) Ambiguous Question

Figure 5: Examples of limitation with current multi-modal instruction tuning dataset.

the facts differ significantly and lead to entirely different conclusions, such responses are flagged for further review or discarded to avoid misinformation. For cases where the facts exhibit only minor differences, we merge the detailed factual content from Qwen2VL-72B into the comprehensive explanations provided by GPT-4o. The merged answers are further reviewed and supervised by human experts to ensure their quality and consistency.

D Training Details

Our training strategy largely follows the approaches adopted by LLaVA and LLaVANext. CLIP-Large-336-Patch14 is employed as the visual encoder. In line with the LLaVA training strategy, we first conduct a pretraining stage where both the visual encoder and the LLM are frozen. We utilize the LLaVA-pretrain-558k and ALLaVA-pretrain-728k datasets for pretraining. The batch size is uniformly set to 256 and learning rate is set to 1e-3. During this phase, images are resized to 336×336, and no image-splitting method is applied.

For SFT stage, we unfreeze the LLM for LLaVA and further unfreeze the visual encoder for LLaVANext. In the case of LLaVANext, we apply the image-splitting method, setting the maximum split size to 3×3. The batch size is uniformly set to 128 and learning rate is set to 2e-5.. LLaVA-InternLM2.5-7B is trained using 8×A800 GPUs for 12 hours. LLaVANext-InternLM2.5-7B is trained using 16×A800 GPUs for 13 hours. LLaVANext-Qwen2.5-32B is trained using 32×A800 GPUs for 24 hours.

Model	MM-AlignBench	WildVision	MMVet
LLaVANext-77k	15.1/-52.6	13.6 / -63.1	37.7
+ 33k w.o. Imager Filter	31.4 / -42.3	22.0 / -42.3	42.0
+ 33k w. Imager Filter	35.3 / -41.0	22.6 / -37.5	44.4

Table 8: Ablation study on the impact of utilizing image filter.

E Ablation on Image Filter

We conduct ablation studies to evaluate the impact of using our image filter. We randomly sample 77k images from the LLaVA-Next-778K SFT dataset as a baseline and separately select images from CC3M both with and without applying our image filter. Subsequently, the selected images are used to generate Knowledge and Inferential Question-Answer pairs following the pipeline described in Sec. 4.3. We then assess the effect of these datasets on the performance of LLaVA-Next-InternLM2.5-7B, with the results presented in Tab. 8. It can be observed that by utilizing the image filter, the selected images contain richer semantic information, leading to the generation of higher-quality image-question pairs. This, in turn, enhances the model's performance in terms of alignment with human preferences.

F Human Experts

In this study, two authors (both PhD students in computer science) serve as human experts. They are responsible for reviewing and refining the questions in MM-AlignBench, as well as evaluating and filtering incorrect merged cases following the Chart Post-Processing Refinement stage.

G License

The InternLM and Qwen models are licensed under the Apache-2.0 license. The ChartQA dataset is distributed under the GNU General Public License v3.0. The remaining datasets are licensed under CC BY-NC 4.0, which permits only non-commercial use.

Text Prompt

How does the **loading and unloading process** depicted in the image **contribute to international shipping and trade**?
What are the **potential logistical challenges** faced at such ports?

InternVL2-76B: Ambiguous | Ignorance | Simplistic

The loading and unloading process shown in the image is a critical component of international shipping and trade, as it facilitates the movement of goods between different countries and regions. The use of cranes and trucks allows for the efficient transfer of cargo from ships to land transport, and vice versa. This process is essential for the global supply chain, enabling the distribution of goods to various markets.
Potential logistical challenges at such ports include coordinating the arrival and departure of ships, managing the flow of trucks and other vehicles, ensuring the safety and security of cargo, and dealing with weather conditions that can impact operations. Additionally, there may be challenges related to customs clearance, documentation, and compliance with international trade regulations. The image suggests a well-organized operation, but these challenges are inherent to the complex and dynamic environment of international shipping and trade.

GPT-4o: Clarity | Knowledge-Rich | Comprehensive

The loading and unloading process depicted in the image is a critical component of international shipping and trade. This process involves transferring goods between ships and other forms of transportation, such as trucks, which facilitates the movement of goods across different regions and countries. Here's how it contributes to international shipping and trade, along with potential logistical challenges:

Contribution to International Shipping and Trade:

1.Efficient Movement of Goods:
The process ensures that goods are efficiently transferred from ships to trucks or vice versa, enabling the seamless flow of products from manufacturers to consumers across the globe.
2.Intermodal Transportation:
Ports serve as hubs for intermodal transportation, where goods are transferred between different modes of transport (e.g., sea to land). This interconnectivity is essential for global supply chains.
3.Economic Growth:
Ports facilitate international trade, which is a significant driver of economic growth. Efficient port operations can enhance a country's trade capabilities and competitiveness.
4.Supply Chain Management:
Effective loading and unloading processes are crucial for maintaining the integrity and timing of supply chains, ensuring that goods reach their destinations on schedule.

Potential Logistical Challenges:

1.Congestion:
Ports can become congested, especially during peak times, leading to delays in loading and unloading. This can disrupt supply chains and increase costs.
2.Infrastructure Limitations:
Inadequate infrastructure, such as insufficient cranes, storage facilities, or transportation links, can hinder the efficiency of port operations.
3.Customs and Regulatory Compliance:
Navigating customs and regulatory requirements can be complex and time-consuming, potentially causing delays in the movement of goods.
4.Labor Issues:
Labor shortages, strikes, or disputes can impact port operations, leading to delays and increased costs.
5.Weather Conditions:
Adverse weather conditions can disrupt port activities, affecting the loading and unloading process and causing delays.
6.Security Concerns:
Ensuring the security of goods and preventing theft or damage is a significant challenge. Ports must implement robust security measures to protect cargo.
7.Technological Integration:
Integrating advanced technologies for tracking, managing, and automating port operations can be challenging but is necessary for improving efficiency and reducing errors.
8.Environmental Regulations:
Ports must comply with environmental regulations, which can impact operations and require investments in cleaner technologies and practices.
Addressing these challenges requires coordinated efforts between port authorities, shipping companies, logistics providers, and government agencies to ensure smooth and efficient port operations, which are vital for sustaining international trade and economic growth.

Figure 6: GPT-4o shows superior alignment with human preference than InternVL2-76B.

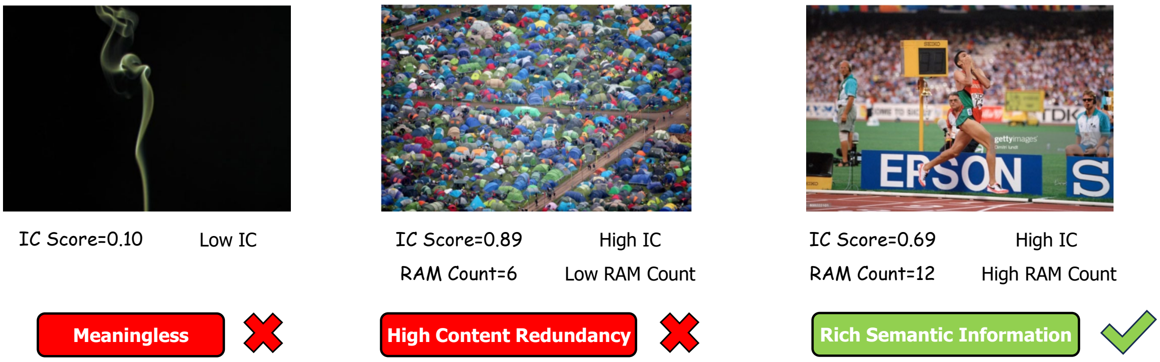


Figure 7: Demonstration examples of our image filter.

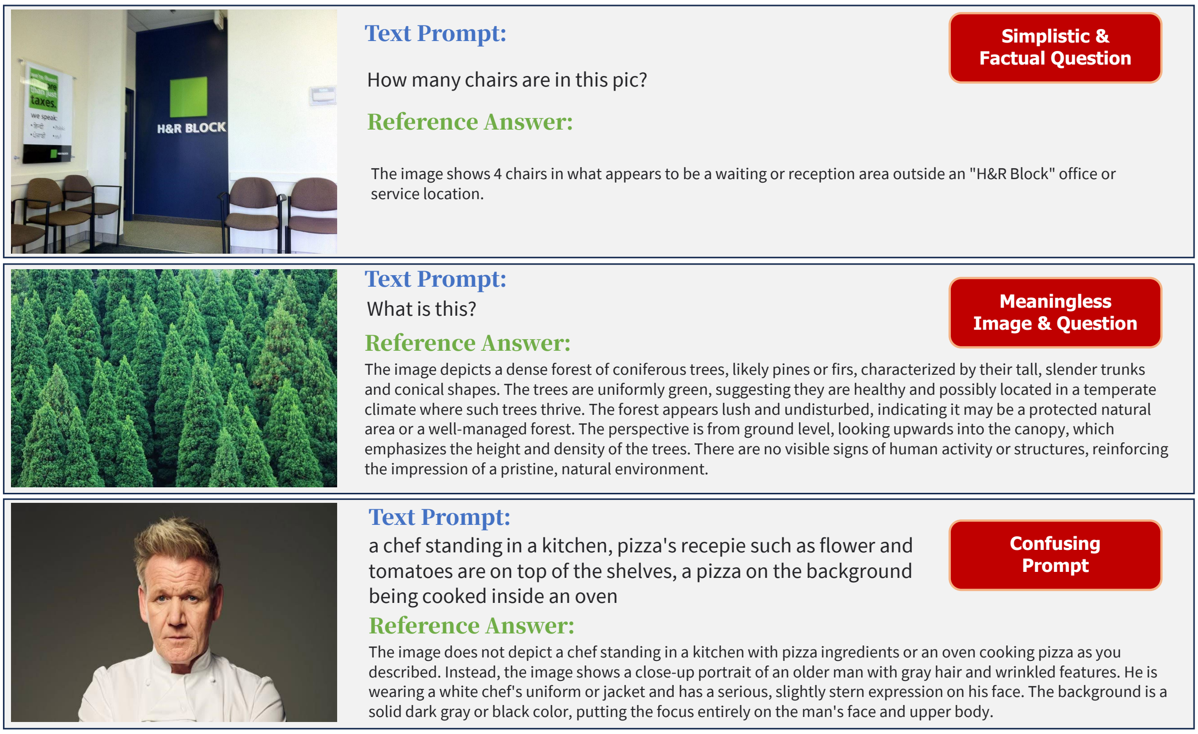
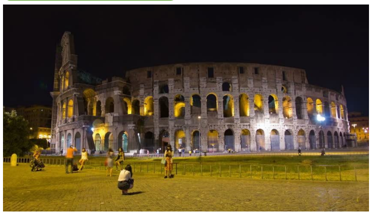


Figure 8: Examples of limitation within current multi-modal benchmark for alignment.

Knowledge



Please introduce the history of the landmark in the image, and suggest a culturally enriching leisure activity plan for visitors.

History of the Landmark
The landmark in the image is the Colosseum, an iconic symbol of ancient Rome located in the heart of the city. Constructed between 70-80 AD under the reign of Emperor Vespasian and completed by his successor Titus, the Colosseum was the largest amphitheater ever built at the time and could hold between 50,000 to 80,000 spectators. It was primarily used for gladiatorial contests and public spectacles such as animal hunts, mock sea battles, and executions. Despite suffering damage from natural disasters and stone robbers, the Colosseum remains a testament to Roman engineering and architectural prowess.
Culturally Enriching Leisure Activity Plan
Morning: Guided Historical Tour
Start your day with a guided tour of the Colosseum. A knowledgeable guide will provide insights into the architectural techniques used in its construction and recount tales of the gladiators who fought within its walls. This tour will help visitors appreciate the historical significance and the cultural heritage of the Colosseum.
Afternoon: Roman Forum Exploration
After the Colosseum, head to the nearby Roman Forum, once the center of Roman public life. Walk through the ruins and imagine the political, religious, and economic activities that took place here. Consider using an audio guide or a historian-led tour for deeper understanding.
Evening: Traditional Italian Dining Experience
Conclude your day with a dinner at a nearby trattoria, indulging in traditional Italian cuisine. Opt for dishes such as pasta alla carbonara or cacio e pepe, paired with a glass of local wine. End your meal with a classic tiramisu or gelato, savoring the flavors of Rome.
This plan not only immerses visitors in the rich history of the Colosseum and its surroundings but also offers a taste of authentic Roman culture.

Inferential

What occasion are the people in the image likely celebrating?

The people in the image are likely celebrating a holiday, most probably Christmas. This conclusion is drawn from several key elements visible in the scene:
- Wrapped Gifts: The presence of wrapped gifts suggests a gift-giving occasion, which is a common tradition during Christmas.
- Stockings on the Fireplace: Stockings hung by the fireplace are a traditional Christmas decoration, often used for Santa Claus to fill with small gifts and treats.
- Festive Decorations: The overall festive atmosphere, including other possible decorations like a Christmas tree or lights, further supports the idea that this is a Christmas celebration.
These elements together create a clear picture of a family or group gathered to enjoy the warmth and joy of the holiday season.

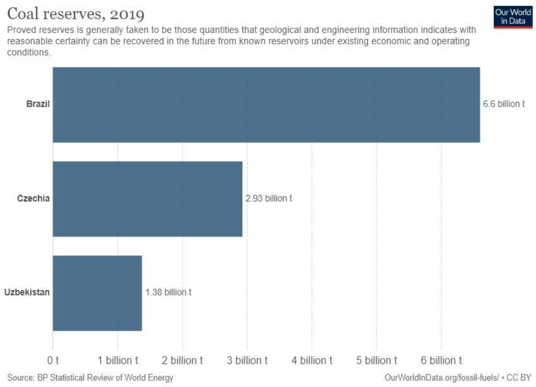
Creation

Imagine you are the athlete in this image. Describe your thoughts and emotions as you prepare for an important game. What motivates you to give your best performance on the ice?

Standing by the rink, I feel the electric buzz of anticipation coursing through my veins. The cool air of the arena sharpens my focus, while the roar of the crowd fuels the fire within me. Wearing the Vancouver jersey, I am more than just another athlete on the ice; I am part of something greater, a legacy I am determined to uphold.
As I grip my stick, memories of early mornings on frozen ponds flash through my mind. Each moment, each sacrifice, has led me to this pivotal game. My heart beats in sync with the rhythm of skates slicing through ice, a reminder that this is where I belong.
The motivation to perform doesn't just come from a desire to win, but from the faces of my teammates who have become family, the unwavering support from fans, and the pride in representing my city. I visualize each play, every pass, every shot, letting the anticipation transform into confidence.
Tonight, as the whistle blows and the puck drops, I am ready to give everything I have. This is my stage, my moment, and I intend to seize it with every ounce of passion in me.

Figure 9: Examples of each task in OmniAlign-V.

Chart

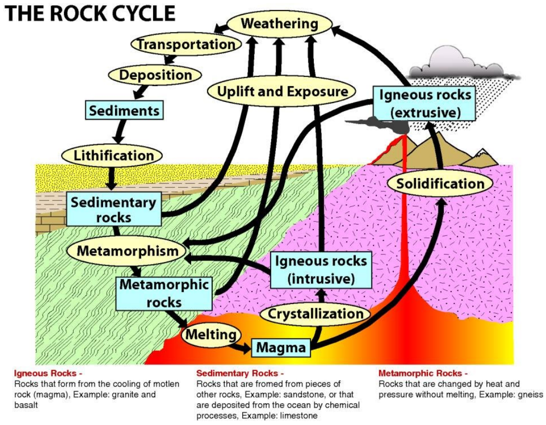


What is the relationship between the coal reserves of Brazil, Czechia, and Uzbekistan, and how might this affect their energy policies?

The coal reserves of Brazil, Czechia, and Uzbekistan differ significantly, which may influence their energy policies and strategies:
1. Coal Reserves Overview
- Brazil: 6.6 billion tons (largest reserves among the three).
- Czechia: 2.93 billion tons (moderate reserves).
- Uzbekistan: 1.38 billion tons (smallest reserves).
2. Implications for Energy Policies
- Brazil:
- Its large coal reserves might provide a reliable domestic energy source for power generation and industry.
- However, Brazil has a strong focus on renewable energy (e.g., hydropower) and may prefer to limit coal use to reduce environmental impacts.
- Coal could be used as a backup or exported for economic gains.
- Czechia:
- With moderate reserves, coal has traditionally played a key role in Czechia's energy mix, particularly for electricity generation.
- Czechia faces pressure to transition away from coal due to European Union climate targets, potentially accelerating the shift toward renewables and nuclear energy.
- Uzbekistan:
- Limited coal reserves may constrain its reliance on coal for energy.
- Uzbekistan has been investing in natural gas and renewable energy, and its smaller coal reserves may push it to focus more on diversifying its energy sources.
3. Global Transition Trends
- All three countries may face international pressure to reduce coal use in favor of cleaner energy sources due to global climate change goals.
- The extent of their reliance on coal will depend on domestic policies, economic priorities, and access to alternative energy resources.
4. Export Potential
- Brazil, with the largest reserves, has more potential for coal exports, which could shape its energy trade policies.
- **Czechia and Uzbekistan are less likely to focus on coal exports due to smaller reserves and domestic needs.**

In summary, Brazil's larger reserves provide more flexibility for domestic use and export, while Czechia and Uzbekistan's smaller reserves may push them to prioritize energy diversification and transition to cleaner energy.

Diagram



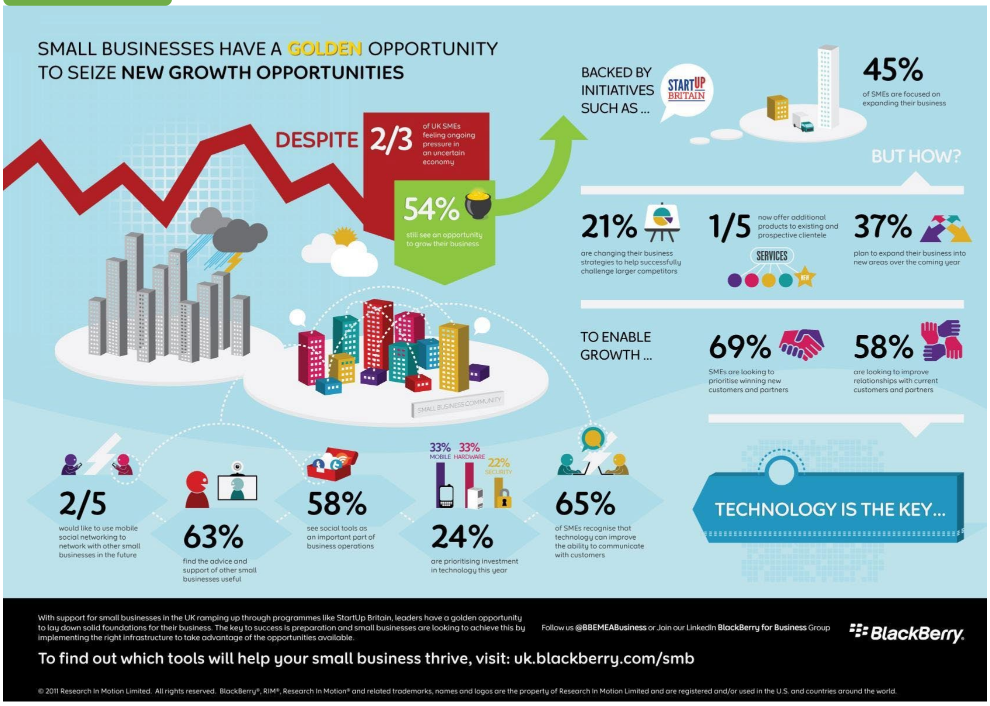
How does the relationship between weathering and sediment formation affect the rock cycle?

The relationship between weathering and sediment formation is a vital component of the rock cycle, as it provides the materials necessary for the formation of sedimentary rocks. Here's how they interact and affect the rock cycle:
1. Weathering Breaks Down Rocks:
- Weathering is the process in which rocks are broken down into smaller particles, such as sediments, through physical, chemical, or biological means.
- Rocks exposed at the Earth's surface, including igneous, metamorphic, or even sedimentary rocks, are subjected to weathering.
2. Sediment Formation and Transportation:
- The weathered material becomes sediments, which are transported by wind, water, or ice to new locations. This transportation step is essential for distributing sediments across different environments.
3. Deposition and Lithification:
- Sediments are eventually deposited in layers, often in water bodies like rivers, lakes, or oceans. Over time, these sediments are compacted and cemented through lithification to form sedimentary rocks.
4. Sedimentary Rocks and the Rock Cycle:
- Sedimentary rocks can undergo further changes in the rock cycle:
- They may be subjected to heat and pressure to become metamorphic rocks.
- They can also be subducted into Earth's mantle, where they melt and form magma, which solidifies into igneous rocks.
5. Continuous Cycling:
- **Weathering and sediment formation ensure that materials from older rocks re-enter the rock cycle, highlighting the interconnectedness of Earth's processes.**

In summary, weathering and sediment formation are essential steps in recycling materials within the rock cycle, connecting the processes that form sedimentary, metamorphic, and igneous rocks.

Figure 10: Examples of each task in OmniAlign-V.

Poster



In what ways does the infographic suggest small businesses can differentiate themselves from larger competitors, and what strategies are highlighted as crucial?

The infographic highlights several ways small businesses can differentiate themselves from larger competitors and underscores specific strategies as crucial for growth:
Ways to Differentiate:
1. Agility in Business Strategy:
- 21% of small businesses are changing their business strategies to effectively compete with larger companies.
- The focus is on flexibility and responsiveness to market changes.
2. Expanding Product and Service Offerings:
- 1 in 5 small businesses now offer additional products or services to attract and retain clients.
3. Geographical Expansion:
- 37% of SMEs plan to expand into new areas over the coming year to tap into untapped markets.
4. Customer Relationships:
- 69% prioritize winning new customers and partners.
- 58% focus on strengthening relationships with existing customers and partners.
Crucial Strategies:
1. Leveraging Technology:
- 65% of SMEs recognize that technology improves communication with customers.
- 24% are prioritizing investment in technology (e.g., mobile hardware and security systems).
2. Social Media and Networking:
- 58% see social tools as essential for operations.
- 2 in 5 SMEs plan to use mobile social networking to connect with other small businesses.
3. Community and Support:
- 63% find advice and support from other small businesses helpful, emphasizing the importance of collaboration and shared insights.
Key Takeaway:
The infographic emphasizes that investing in technology, adapting strategies, and building strong customer relationships are critical for small businesses to differentiate themselves and thrive in competitive markets.

Figure 11: Examples of each task in OmniAlign-V.

Art

What are the prominent features of the setting and figures depicted in the painting, and how do they reflect the themes and techniques commonly used by the artist?

The painting 'The Charmes of Life' by Antoine Watteau showcases a vivid and lively scene set in a pastoral landscape, typical of Watteau's work. Watteau is known for his 'fêtes galantes' style, which depicts scenes of aristocratic outdoor entertainment.
Features and Themes:
- Setting: The painting features a lush, picturesque countryside, suggesting a leisurely and idyllic lifestyle. The use of soft brushwork and a subtle color palette creates an ethereal, dreamlike quality.
- Figures: The elegantly dressed figures are engaged in music and conversation, portraying a sense of enjoyment and harmony. The attire and gestures reflect the social elegance of the Rococo period.
- Architectural Elements: The grand columns and structured terrace highlight the blend of nature and refined culture, a common theme in Watteau's works.
Techniques:
- Color and Light: Watteau's use of light captures the luminous quality of the outdoors, enhancing the serene and carefree atmosphere.
- Composition: The arrangement of figures in a relaxed, informal manner typifies the fluidity and spontaneity characteristic of Rococo art.
Overall, the painting reflects Watteau's mastery in capturing the fleeting pleasures of life, combining sophisticated elegance with a deep appreciation for nature and human interaction.

Instruction Following

What elements in the image reflect urban fashion trends, and how do these elements align with current fashion industry predictions for seasonal attire? Please describe each element using a metaphor and ensure your response includes at least three distinct metaphors.

The image showcases several elements reflective of urban fashion trends, each of which can be described using vivid metaphors to highlight their significance and alignment with current fashion industry predictions for seasonal attire.
1. Floral Dress: The central figure's floral dress is like a garden in full bloom, with its bright, bold patterns that capture the essence of spring and summer. This aligns with fashion industry predictions that emphasize vibrant colors and nature-inspired designs as key trends for these seasons. Just as a garden bursts with life and color, the dress brings a sense of freshness and vitality to the urban landscape.
2. Bright Pink Shoes: The bright pink shoes are like a neon sign in a bustling city, drawing attention and adding a pop of color to the outfit. This technique of using bold, eye-catching colors is a common strategy in urban fashion to make outfits more dynamic and memorable. The shoes serve as a beacon of style, guiding the eye and adding a modern, urban flair to the ensemble.
3. Bright Orange Handbag: The bright orange handbag carried by the person on the left is like a vibrant sunset, blending practicality with style. It complements the idea of combining functional fashion items with bold, statement pieces, a growing trend in urban fashion. Just as a sunset transforms the sky with its warm, radiant hues, the handbag adds a touch of warmth and energy to the overall look, making it both practical and visually striking.
These elements collectively reflect the broader industry move towards integrating bold, statement pieces with functional fashion items that cater to an urban lifestyle, ensuring that the wearer stands out while remaining practical and stylish.

Figure 12: Examples of each task in OmniAlign-V.

Prompt for Knowledge Task

Examine the image provided and generate a knowledge-based and exploratory question based on the content of the image and supply corresponding detailed answers.
Question Guidelines:
- Your question should invite insightful discussion on the types of elements in the image, such as:
- **Objects**: For example, animals, plants, food, or products.
- E.g., "What breed of dog is in the picture, and what are their characteristics?", "Can you give me a recipe for the food in image?", "Write a Product Description for the product in the image.".
- **Locations and Features**: Relevant to countries, landmarks, famous people, or scenic spots.
- E.g., "Please introduce the history of the landmark in the picture.", "How did the states in image get their names?", "Who is the person in the image and what is him famous for?".
- **Activities and Technologies**: Related to sports, machines, technology, and environmental details.
- E.g., "Can you explain how the game in image is played?", "How is the machine shown in the image operated?"
- **Events, Literature, and Media**: Concerning books, movies, or series in the image.
- E.g., "Write a short description about the movie or series in the image.", "Think of books that would be enjoyable for someone who liked the books in the image."

Answer Guidelines:
- Ensure your answers are factual and comprehensive.
- Please use Markdown formatting in your text to enhance the content, making it visually appealing and easy to read. Include appropriate headings, subheadings, lists, code blocks, and other Markdown elements to optimize your answers.

Output Format:
Your response should strictly follow this format:
“‘json
{
"question": "Question text",
"answer": "Answer text"
}
‘‘‘

Figure 13: **An Example of the prompt for Knowledge task generation.**

Prompt for Creative tasks

You are a skilled writer with a talent for crafting insightful and engaging questions based on the content of a given image.
Task Guidelines:
- Analyze the content of the provided image and select one of the question types listed below. Use it to create engaging questions that lead the viewer to explore and interpret the image in different level of complexity.
- Be closely tied to the content of the image, emphasizing its primary visual or thematic elements.
- Your questions should **avoid directly referencing specific details** in the image. Instead, they should encourage deeper reflection, ensuring the question cannot be answered without seeing the image.
- Avoid overly rigid or direct phrasing, focusing instead on open-ended exploration.
Question Types
Your questions can be from the following types, each followed by an example with a different level of complexity.
- Simple (basic observation or initial reflection)
- Moderate (more thought-provoking, requiring a deeper understanding and more structured response)
- Difficult (complex or abstract, requiring analysis and strict formatting)
{Match Types}
Output Format:
Your response should strictly follow this format:
```json
{
"question": "Question text",
"type": "Question type",
"level": "Question level"
} ```

Figure 14: **An Example of the prompt for Creative task generation.**
```

Prompt for Inferential tasks

You are an image analysis expert skilled in posing high-quality **inferential questions**. Please provide 2-5 of the most insightful questions you can think of, following these guidelines:
For Questions:
- **Focus on image-based questions:** Ensure that your question cannot be answered without analyzing the image. **You should not directly provide image's data or details in your generated questions.** For example, "What might be the impact on the radio industry due to 34 stations having their licenses revoked?" includes the data in the image and can be answered without analyzing the image, so it is a bad question.
- Ensure that questions are natural, not overly rigid. **You should be quite certain and confident about the questions you pose and their answers, and avoid using words like "possibly", "maybe" or "might be" in both questions and answers.**
- Your questions must require reasoning beyond the direct content of the image, making reasonable inferences based on the information presented. - The scope of the questions should not be overly broad or delve into political, philosophical, speculative, sensitive, or controversial topics. Stay within the context of the scene and elements inferred from it.
For Answers:
- You should provide a clear and concise answer to the question.
Good Examples:
- What precautions are the people on the boat taking to stay comfortable during the trip? - Is there anything else on the table other than the pizza? - Why do these people choose to dress in this style?
- What decorative element is present in this public restroom that is not typical? Bad Examples:
- What might indicate

Figure 15: **An Example of the prompt for Inferential task generation.**

Prompt for Chart tasks

You're a great image analyst. You need to analyze the image provided and generate some insightful questions based on the content of the image.
Question generation guidelines:
- Ensure that your questions require the image to be answered and do not include explicit information from the image. Instead, pose questions that prompt the respondent to analyze the image to find the answer.
- Your question should be explainable and require some reasoning to answer.
- Your question could contain different analytical perspectives, such as trends, comparisons, causal inference, etc.
- - Your questions should be insightful but also clear and straightforward. Avoid overly complex or niche questions.
Bad question examples:
- "What might be some factors contributing to the significantly higher private health expenditure per person in Argentina compared to Fiji and Benin?" (This includes specific details from the chart.) Its correct clarification should be "Is there any difference in private health expenditure per person between Argentina, Fiji, and Benin? If so, what might cause the difference?"
- "What trends can be observed in private health expenditures per person among the three countries shown?" (This question is unclear because 'trends between countries' is not a standard analytical concept. Trends typically refer to patterns over time or categories, not direct cross-entity comparisons.)
Output format:
Your response should strictly follow this format:
{
"questions": [
{
"question": "Question text"
},
{
"question": "Another question text"
}
]
}

Figure 16: **An Example of the prompt for Chart task generation.**

Prompt for Poster tasks

You're an excellent image analyst and good at generating insightful questions about the image. You need to analyze the image provided and generate some insightful questions based on the content of the image, and you should answer the questions you generated.
Possible Categories Reference:
1. **Cultural and Social Context**
2. **Analysis and Inference**.
3. **Visual Elements and Design Techniques**.
Question Guidelines:
- **Focus on image-based questions:** Ensure that your question cannot be answered without analyzing the image. **You should not directly provide image's data or details in your generated questions.** For example, "What might be the impact on the radio industry due to 34 stations having their licenses revoked?" includes the data in the image and can be answered without analyzing the image, so it is a bad question.
- **Encourage thoughtful, structured responses:** Your question should be explainable and need some reasoning to answer. You should not generate questions that are just extracting information from the image. For example, "What percentage of organizations verify the past employment records of new employees according to the image?" is a bad question.
- **Ensure diversity in the questions:** Cover different aspects of the image, encouraging multiple perspectives. You can choose some appropriate categories from the possible categories reference. **For the same category, you can generate multiple questions.**
- **Generate high-quality questions:** You can choose to generate challenging questions, but their answers should be able to clearly explain.

Output Format:
Your response should strictly follow this format:

```
"questions": [
{
"question": "Question text"
}
]
```

Figure 17: **An Example of the prompt for Diagram task generation.**

Prompt for Diagram tasks

You're a great diagram analyst. You need to analyze the diagram provided and generate 2-4 insightful questions based on the content of the diagram.
Question Guidelines:
- Your questions should be guiding and should not directly point to the content. For example: "How does acid rain affect water bodies, soil, and plant life?" should be changed to "How does the process in image affect water bodies, soil, and plant life?"
- Your question should invite insightful discussion, such as:
- **Interpretation**: Symbol Interpretation, Data Extraction
- **Examples**:
- "What is the role of the cytokine-producing cell in the process shown?"
- "Enumerate the steps outlined in the flowchart."
- **Relation Analysis**:
- **Examples**:
- "How does variable A affect variable B in the diagram?"
- "How many ways can A to B be achieved in the diagram?"
- **Inference**:
- **Examples**:
- "What can be inferred about the system's stability from the diagram?"
- "What does the bacterium do once it has the hybrid plasmid?"
Output Format:
Your response should strictly follow this format:
```json
{
"question":["Question text 1", "Question text 2", ...],
}
```

Figure 18: **An Example of the prompt for Diagram task generation.**