

POINTS1.5: Building a Vision-Language Model towards Real World Applications

Yuan Liu, Le Tian, Xiao Zhou, Xinyu Gao, Kavio Yu, Yang Yu, Jie Zhou

Pattern Recognition Center, WeChat AI, Tencent Inc, China

{bensenliu}@tencent.com



<https://github.com/WePOINTS/WePOINTS>



<https://huggingface.co/WePOINTS/POINTS-1-5-Qwen-2-5-7B-Chat>

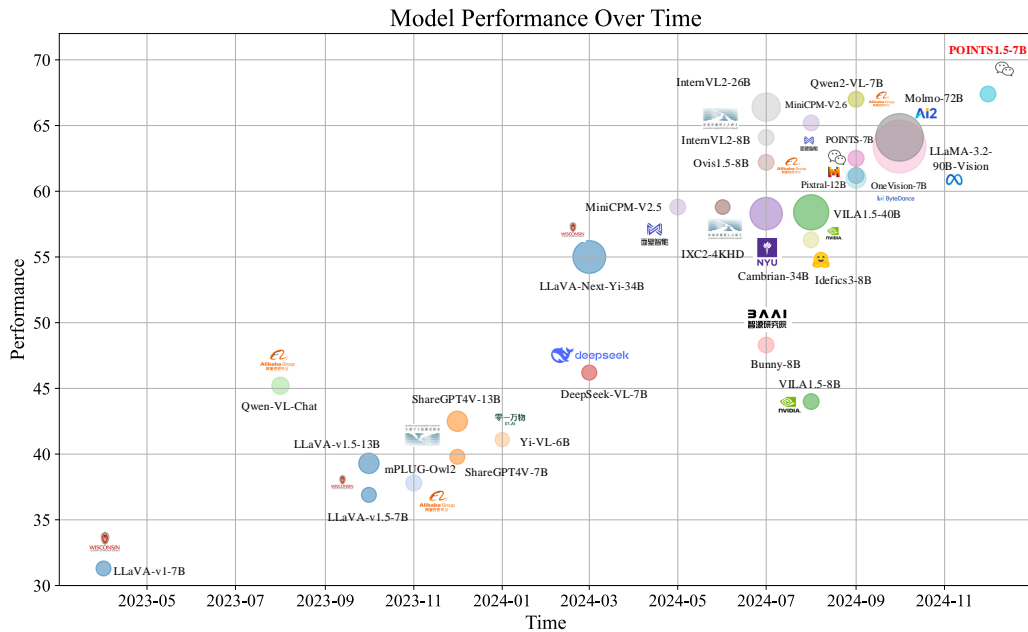


Figure 1: Performance of Open-Source Models on the OpenCompass Leaderboard[Contributors, 2023]. POINTS1.5 ranks first among all models under 10B in size, even outperforming models several times larger. The size of each bubble represents the model size.

Abstract

Vision-language models have made significant strides recently, demonstrating superior performance across a range of tasks, *e.g.* optical character recognition and complex diagram analysis. Building on this trend, we introduce a new vision-language model, **POINTS1.5**, designed to excel in various real-world applications. POINTS1.5 is an enhancement of POINTS1.0 and incorporates several key innovations: i) We replace the original CLIP vision encoder, which had a fixed image resolution, with a NaViT-style vision encoder that supports native dynamic high resolution. This allows POINTS1.5 to process images of any resolution without needing to split them into tiles. ii) We add bilingual support to POINTS1.5, significantly enhancing its capability in Chinese. Due to the scarcity of open-source Chinese datasets for vision-language models, we collect numerous images from the

Internet and annotate them using a combination of manual and automatic methods. iii) We propose a set of rigorous filtering methods for visual instruction tuning datasets. We comprehensively evaluate all these filtering methods, and choose the most effective ones to obtain the final visual instruction tuning set. Thanks to these innovations, POINTS1.5 significantly outperforms POINTS1.0 and demonstrates strong performance across a range of real-world applications. Notably, POINTS1.5-7B is trained on fewer than 4 billion tokens and ranks first on the OpenCompass leaderboard among models with fewer than 10 billion parameters¹.

1 Introduction

Vision-language models [Liu et al., 2024b, Li et al., 2024, Bai et al., 2023, Liu et al., 2024d, Dong et al., 2024a, Chen et al., 2024c] have made remarkable strides in recent years, showcasing their potential to tackle complex tasks such as geometry math problems and optical character recognition (OCR). Despite these advancements, open-source models still lag behind closed commercial models like GPT-4o [OpenAI, 2023] and Claude-3.5-Sonnet in addressing certain real-world challenges. To bridge this gap, the open-source community has made significant efforts, exemplified by the Qwen2-VL series [Wang et al., 2024b], which has demonstrated performance comparable to, or even surpassing, these commercial models. In line with this trend, we introduce POINTS1.5, a more robust model than its predecessor, POINTS1.0 [Liu et al., 2024d], which currently holds the top position on the OpenCompass leaderboard among models with fewer than 10 billion parameters.

The development of vision-language models generally follows two distinct paths: i) The LLaVA-style architecture, which integrates a pre-trained vision encoder, a randomly initialized projector, and a pre-trained large language model; and ii) Models where the large language model is randomly initialized, and both visual and text tokens are jointly used to train the language model, as seen in works like Emu3 [Wang et al., 2024c]. The LLaVA-style architecture has shown superior performance in visual understanding tasks, and POINTS1.5 continues to follow this approach. This architecture involves continual post-training of the large language model to enhance its ability to interpret visual information. The pre-training stage primarily serves to align the projection layer with the space of vision and text tokens [Li et al., 2023b]. We identify two critical factors for developing a superior LLaVA-style vision-language model: i) A high-performance vision encoder that can accurately and uniquely represent an image, and ii) High-quality visual instruction tuning datasets that enable the model to understand image content and exhibit strong instruction-following capabilities. Based on this analysis, POINTS1.5 introduces the following innovations.

Native Dynamic High Resolution. Enabling a vision-language model to process images of any resolution without down-sampling offers numerous benefits, such as reducing hallucinations and enhancing performance on text-intensive tasks. Historically, many vision encoders, such as Vision Transformer [Dosovitskiy, 2020] and ConvNext [Liu et al., 2022], could only handle fixed-resolution images. Previous works [Liu et al., 2024a, Dong et al., 2024b, Chen et al., 2024c, Liu et al., 2024d] often split large images into tiles to accommodate the vision encoder, disrupting the spatial structure of the original image. In contrast, POINTS1.5 employs a NaViT-style architecture, following the approaches of Qwen2-VL [Wang et al., 2024b] and Idefics2 [Laurençon et al., 2024], allowing it to process arbitrary-resolution images without splitting them, resulting in significant improvements over the dual CLIP vision encoders used in POINTS1.0.

Bilingual Support. In POINTS1.0, the English corpus comprised over 95% of the total data. In this version, we have increased the amount of Chinese data for both the pre-training and visual instruction tuning stages. Due to the limited availability of open-source Chinese datasets, gathering a large quantity of Chinese corpus is challenging. For the pre-training stage, we followed the strategy used to obtain the 1 million pre-training data in POINTS1.0, creating an additional 1 million Chinese pre-training data from LAION-5B-cn [Schuhmann et al., 2022] using CapFusion [Yu et al., 2024] and perplexity filtering [Liu et al., 2024d]. This was combined with the original English data to form a final dataset of 2 million for pre-training. For the visual instruction tuning stage, we employed two strategies: (i) translating existing conversation datasets into Chinese using a large language model (LLM), and (ii) for Chinese OCR datasets, collecting related images from the internet and using an

¹Result is obtained on December 8th, 2024

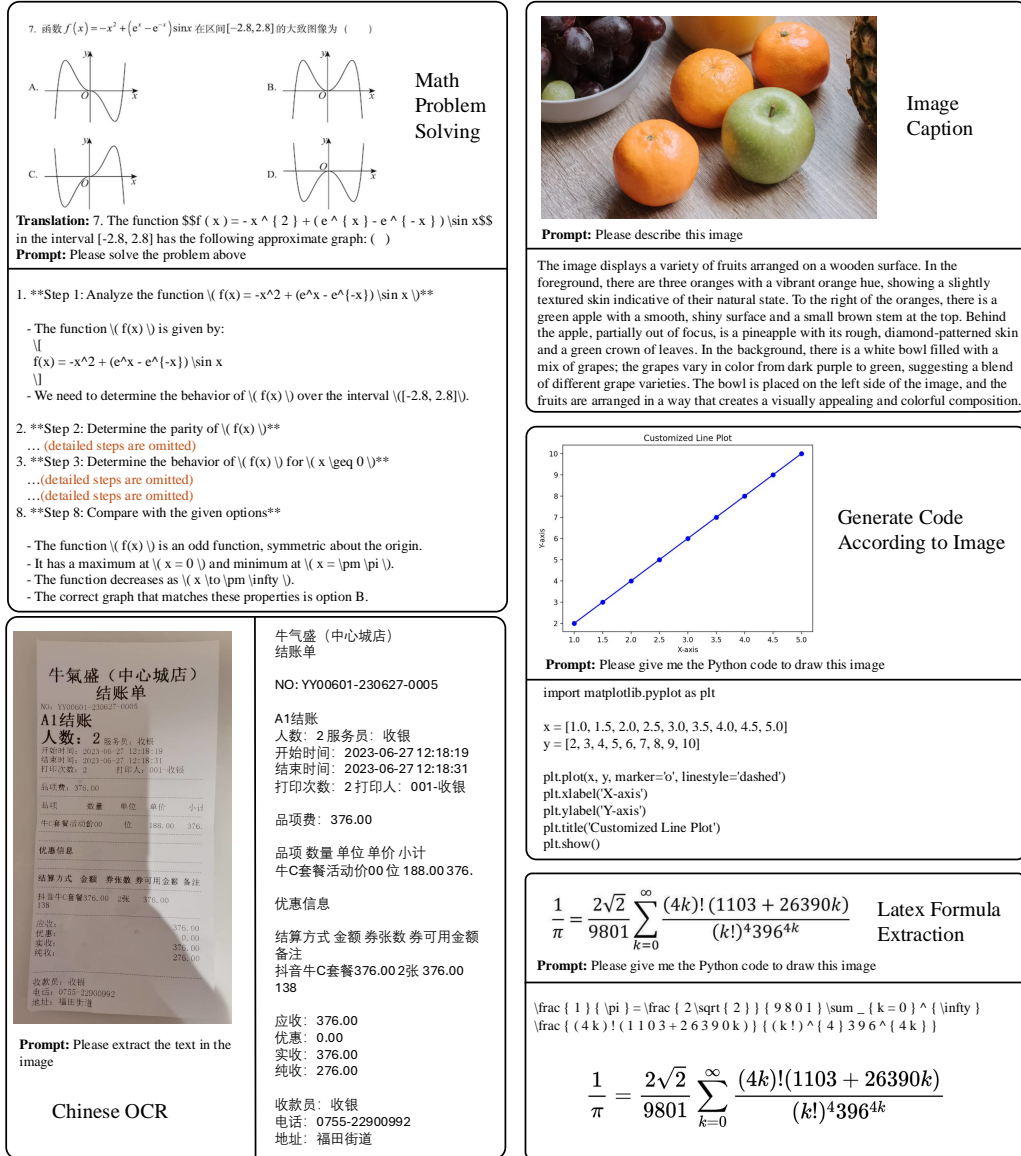


Figure 2: POINTS1.5 shows great potential to solve challenging real world problems.

existing vision-language model, such as Qwen2-VL-72B, to extract text from these images. Human labelers then verified these annotations, correcting minor errors or discarding them if the errors were significant.

Visual Instruction Tuning Set Filtering. We manually reviewed each dataset used in POINTS1.0 and identified two significant issues: i) A large number of grammatical errors in some datasets, and ii) Some questions could be answered without referring to the image. To address the first issue, we employed a large language model (LLM), such as Qwen2.5-72B [Yang et al., 2024], to detect grammatical errors in the existing data samples. We then either discarded these erroneous samples or corrected them and reintegrated them into the original dataset. For the second issue, we used an LLM to answer the questions without the image. If the LLM provided the correct answer, the corresponding data sample was labeled accordingly.

Combining these innovations, POINTS1.5 brings significant improvements compared to POINTS1.0 and performs well across a range of real-world applications. Notably, POINTS1.5-7B ranks first on the OpenCompass leaderboard among models with fewer than 10 billion parameters.

2 Model Architecture

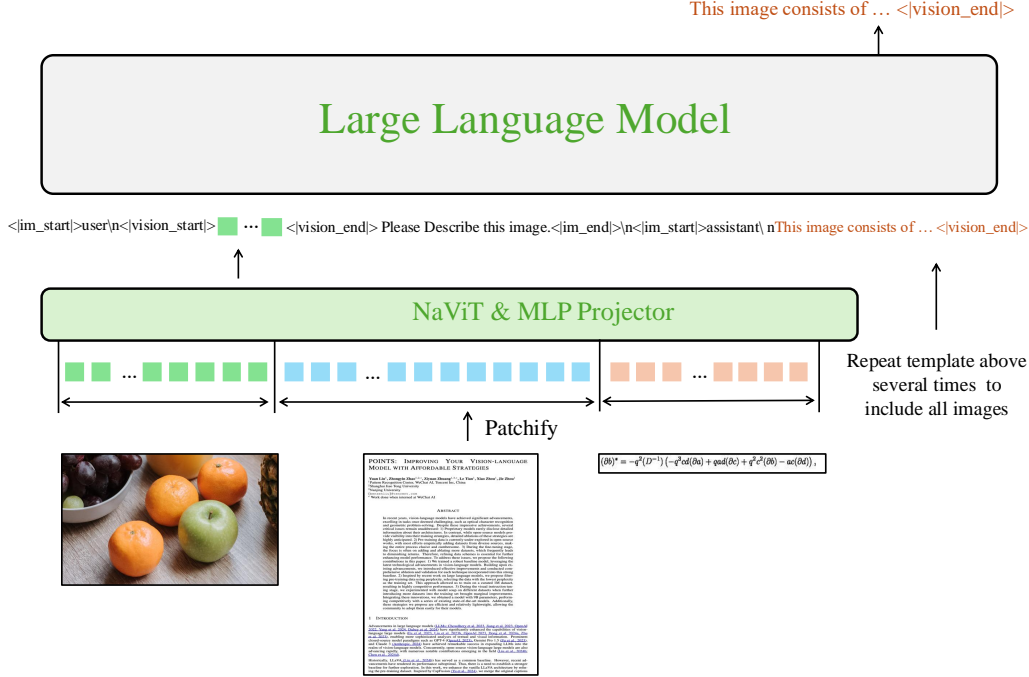


Figure 3: POINTS1.5 uses the conventional LLaVA-style architecture, consisting of a vision encoder, a MLP projector and a LLM.

Figure 3 illustrates the architecture of POINTS1.5. This model adheres to the traditional LLaVA-style architecture [Liu et al., 2023b], which comprises a vision encoder, an MLP projector, and a language model (LLM).

Vision Encoder As discussed in the previous section, training a vision-language model using the LLaVA-style architecture is akin to the continual post-training of the LLM, enabling it to process tokens from the image modality. Therefore, starting with a high-quality vision encoder is crucial for the LLM to accurately interpret images. To support images with arbitrary resolutions, POINTS1.0 follows recent works such as LLaVA-Next [Liu et al., 2024a] and InternVL [Chen et al., 2024c], which split a large image into several tiles that the vision encoder can process. However, this method has inherent drawbacks, as it disrupts the spatial relationships between patches within an image. Although strategies like adding line splitters [Dong et al., 2024b] and incorporating a global view alongside the split patches [Chen et al., 2024c] can mitigate this issue, the problem persists. Consequently, POINTS1.5 replaces the CLIP vision encoder used in POINTS1.0 with a NaViT-style vision encoder [Dehghani et al., 2024], following recent advancements [Wang et al., 2024b, Laurençon et al., 2024]. Unlike the CLIP vision encoder, the NaViT-style vision encoder can natively handle images with arbitrary resolutions without the need for splitting.

Batch Forwarding with NaViT With the introduction of NaViT, a new challenge arises in batch forwarding. Unlike the CLIP vision encoder, where images can be concatenated along the batch size dimension, NaViT processes images that vary in sequence length after being patchified. To address this, we adopt a strategy inspired by large language models (LLMs): we pack multiple image sequences into a single, long sequence. We then record the start and end indices of each image

sequence to ensure that self-attention is applied only within the boundaries of the current image sequence [Dao, 2024].

Projector In accordance with POINTS1.0, the projector consists of a two-layer MLP with a GELU activation function [Hendrycks and Gimpel, 2016] between the layers to introduce non-linearity.

Large Language Model In alignment with POINTS1.0, we have selected Qwen2.5-7B-Instruct. After the release of this paper, we plan to introduce POINTS1.5 with bigger language model.

3 Bilingual Support

In this section, we will discuss the curation of the Chinese dataset used in POINTS1.5. But before the discussion, we will refine the chat template used in the pre-training of POINTS1.0.

Chat Template As discussed in previous sections, training a LLaVA-style vision-language model involves the continual post-training of the LLM. Following POINTS1.0, the LLM of POINTS1.5 is also initialized from the instruction-tuned version of Qwen2.5-7B². However, during the pre-training stage of POINTS1.0, we use a continuation template to pack the data, similar to the one used during the pre-training process of the LLM, which deviates from the template used in the initialized LLM. In this version, we employ the conversation template used in Qwen2.5-7B-Instruct and observe improved performance compared to the continuation template. Since the pre-training data are the image-caption pairs, we add a prompt, similar to *Please describe this image.*, to each data sample. To diversify the prompts, we create a candidate prompt pool (Figure 5) and randomly sample one for each data sample. Additionally, to distinguish visual tokens from text tokens, we add image prefix and suffix tokens around the visual tokens. Figure 4 shows the difference between the chat template during pre-training between POINTS1.0 [Liu et al., 2024d] and POINTS1.5.

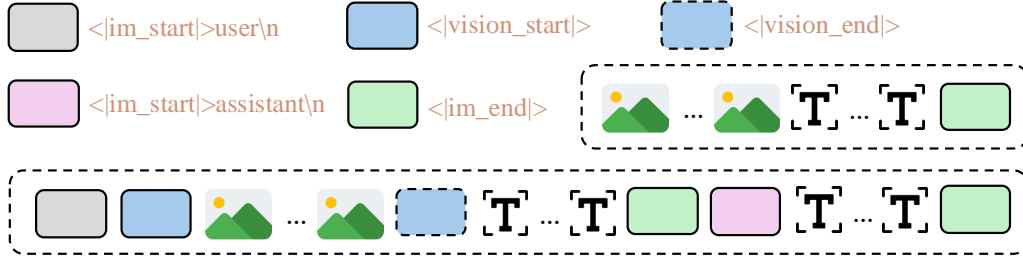


Figure 4: The chat template during pre-training in POINTS1.0 (above) and POINTS1.5 (below)

Chinese Pre-training Dataset Following POINTS1.0, we employ a two-step procedure to create the pre-training dataset: i) We use CapFusion [Yu et al., 2024] to merge the caption generated by a vision-language Model (VLM) with the image’s original caption, resulting in the final caption. ii) We filter the generated caption using perplexity. The CapFusion process is described by the following formula:

$$\text{Caption} = \mathcal{G}(c, \mathcal{F}(I)) \quad (1)$$

\mathcal{G} represents a large language model, \mathcal{F} denotes a vision-language model, c is the original caption, and I is the corresponding image. The Chinese captions were generated during the development of POINTS1.0. For this purpose, we utilize InternLM2 [Cai et al., 2024] as the large language model and InternLM-XComposer2 [Dong et al., 2024a] as the vision-language model. In the future, we plan to generate captions using more advanced models, such as POINTS1.5, which is also expected to further improve performance. Subsequently, we employ perplexity to filter these data:

²<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

$$\text{Perplexity}(s) = \exp\left(-\frac{1}{N} \sum_{i=1}^N \log P(w_i | w_1, w_2, \dots, w_{i-1})\right) \quad (2)$$

Let $\{w_1, \dots, w_N\}$ denote the sequence of text tokens for s . We arrange these tokens in ascending order and select the first 20% (approximately 1 million) for the pre-training phase. This subset of the Chinese dataset is then combined with the original 1 million English dataset to pre-train POINTS1.5.

<p>Could you provide a detailed description of this image? Can you describe this image thoroughly? Please give a detailed account of this image. Would you mind describing this image in detail? Could you explain this image in detail? Please offer a detailed description of this image. Can you give a comprehensive description of this image? Would you please describe this image in detail? Could you describe this image with detail? Please provide a thorough description of this image.</p>	<p>你能详细描述一下这张图片吗？ 你能彻底描述一下这张图片吗？ 请详细描述一下这张图片。 你介意详细描述一下这张图片吗？ 你能详细解释一下这张图片吗？ 你能给出这张图片的全面描述吗？ 请提供这张图片的详细描述。 请你详细描述一下这张图片好吗？ 你能详细描述一下这张图片吗？ 请提供这张图片的全面描述。</p>
--	--

Figure 5: Prompts used in the chat template during pre-training stage.

Chinese Visual Instruction Tuning Dataset We inherit all visual instruction tuning datasets from POINTS1.0, except for those generated in this section. To create Chinese visual instruction tuning datasets, we employ several strategies: i) Translate existing English datasets (both questions and answers) into Chinese. ii) Use images and questions from existing datasets, and generate corresponding answers using a powerful VLM, such as Qwen2-VL-72B. This strategy is only applied to caption datasets. iii) Collect images from the internet, manually design questions (Figure 6), generate answers using a powerful VLM, and verify the answers with human labelers. This strategy is primarily used for Chinese OCR datasets. The following table shows the datasets and the corresponding strategies used to construct the Chinese datasets.

Datasets	Strategy
VQAv2[Goyal et al., 2017], GQA[Hudson and Manning, 2019] OKVQA[Marino et al., 2019]	Translate English into Chinese
LVIS-Instruct4V[Wang et al., 2023], LAION-GPT4V Images collected from Internet	Question translation&VLM VLM&Human Check

Table 1: Datasets and corresponding strategies to generate Chinese datasets.

<p>Please extract all the text from the image Please extract text from the image Please extract all the text from the image Please extract all the text from the image Please extract the text from all the images Please get all the text from the image What is the text in this image What is the text content on this picture What is the text in the image Can you tell me what the text on this image is</p>	<p>请提取图片中的所有文字 请从图片中提取文字 请从图片中提取全部文字 请从图片中提取出所有文字 请提取所有图片中的文字 请从图片中获取所有文字 这张图片里的文字是什么 这幅图上的文字内容是什么 请问图片中的文字是什么 你能告诉我这张图片上的文字是什么吗</p>
---	---

Figure 6: Prompts to create the Chinese OCR datasets.

After the creation of Chinese datasets, we obtain the distribution across 9 categories and the English&Chinese distribution for the final visual instruction tuning datasets we used in POINTS1.5.

We observe a significant imbalance among the different categories. However, we have not yet identified effective methods to balance the data across these categories, and we leave this challenge for future work.

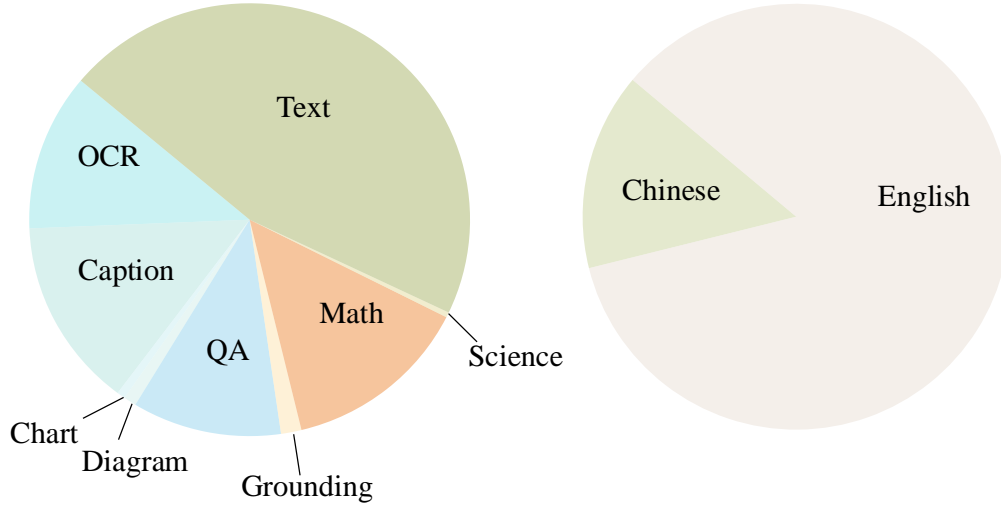


Figure 7: **Distribution of visual instruction tuning data in POINTS1.5.** The left figure shows the distribution across different categories, and the right figure shows the distribution between English and Chinese.

4 Visual Instruction Tuning Set Filtering

Before filtering the visual instruction tuning datasets, we manually check each of these datasets used in POINTS1.0, and identify two significant issues: i) Some questions could be answered without referring to the image (Figure 9). and ii) A large number of grammatical errors in some datasets (Figure 8).

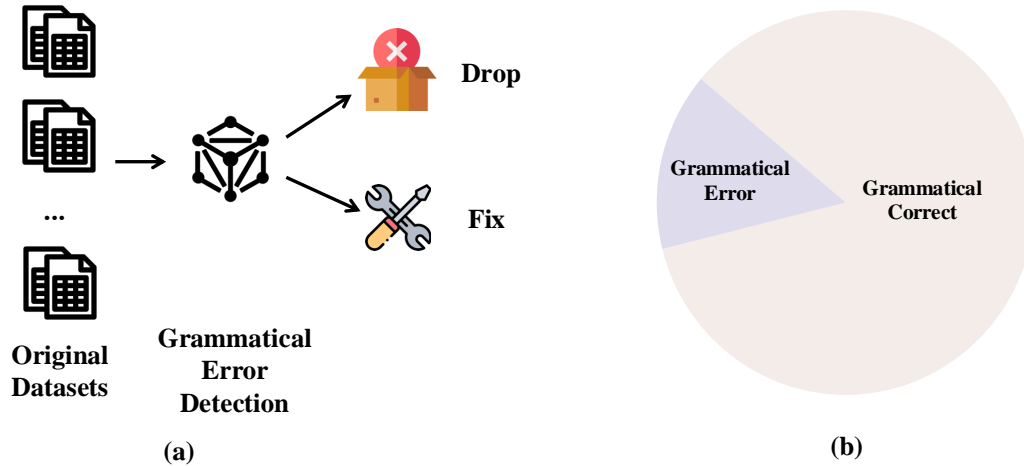


Figure 8: **Procedure to filter out samples containing grammatical errors (a) and distribution between grammatically correct samples and samples containing grammatical errors (b).**

Questions can be answered without images. It is common sense that the data used to train a vision-language model should enable the model to solve problems based on images. If questions can be answered without images, they degenerate into pure-text data [Liu et al., 2023c]. To filter out such data, we use a powerful open-source LLM, such as Qwen2.5-72B-Instruct, to answer the questions without the images. If the LLM provides the correct answer, the corresponding data sample is discarded. This filtering strategy is applied only to datasets containing fixed and definite answers, such as AI2D [Kembhavi et al., 2016]. We then train the model with the filtered dataset but observe slightly degraded performance. This phenomenon is consistent with previous works [Dai et al., 2024, Zhang et al., 2024, Yao et al., 2024], which suggest that pure-text data is helpful in maintaining the capability of the pre-trained LLM.

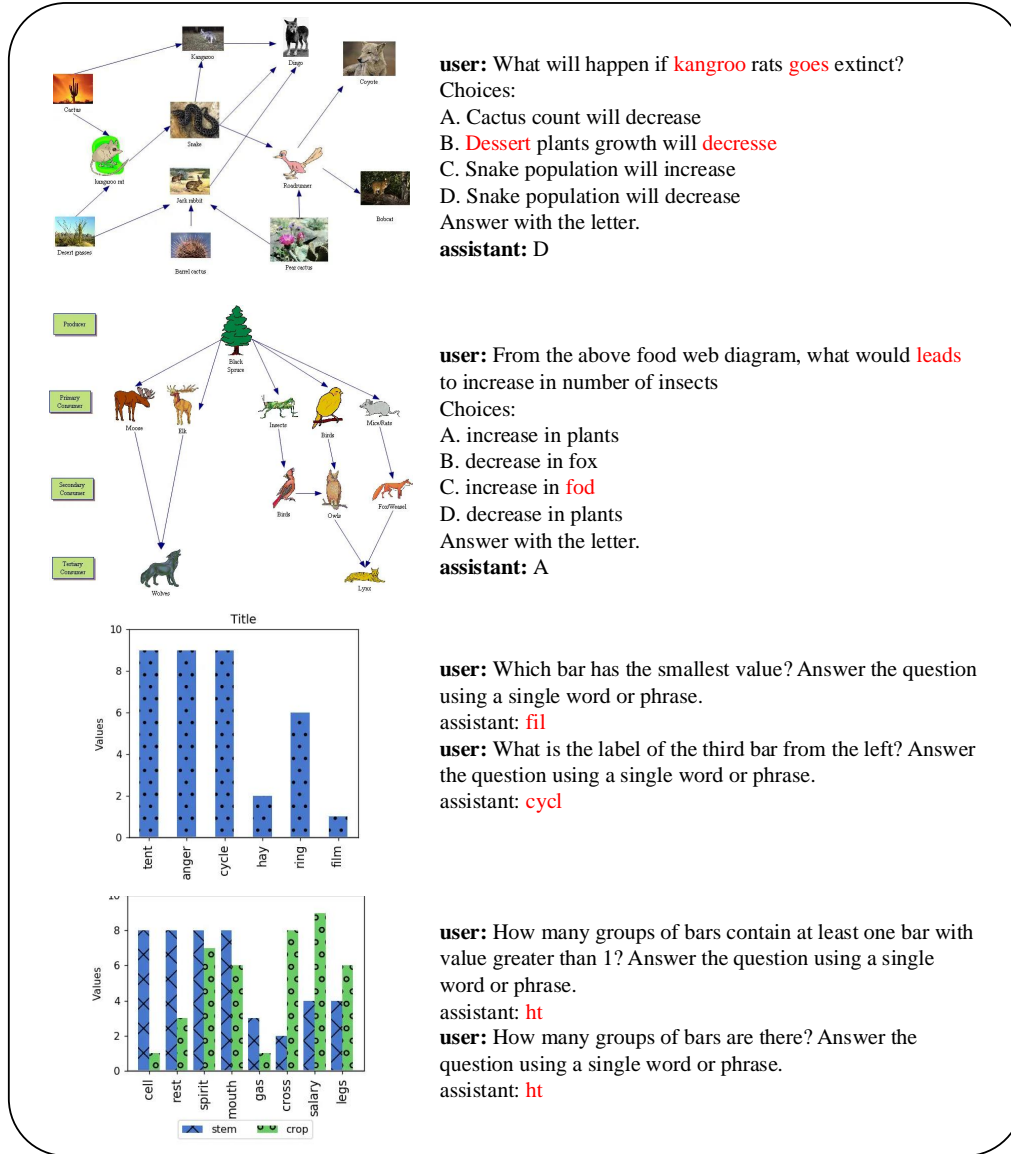


Figure 10: Data samples containing grammatical errors (marked with red) in visual instruction tuning set.

layer results in better performance (we fix the vision encoder in this stage, as we find unfreezing it degrades the performance (Figure 11). Our training configuration is summarized in Table 3. Notably, POINTS1.5 follows the path of POINTS1.0 [Liu et al., 2024d] to make computational resources more affordable, totaling less than 5 billion tokens, which is significantly fewer than most previous works [Chen et al., 2024c, Lu et al., 2024a, Wang et al., 2024b].

Model Soup over Best Performing Model Following POINTS1.0, we use model soup [Wortsman et al., 2022] to boost the performance of a single model. Model soup is conducted over models that perform best on our evaluation benchmark and mainly consist of models trained with different visual instruction tuning datasets and different visual instruction tuning epochs. The OpenCompass score of the best-performing single model is 66.5, and the final model obtained using model soup achieves a score of 67.4.

Model	Pre-training			Instruction Tuning		
	Vision	Projector	LLM	Vision	Projector	LLM
LLaVA-Next[Liu et al., 2024a]		✓			✓	✓
OneVision[Li et al., 2024]		✓		✓	✓	✓
POINTS[Liu et al., 2024d]	✓	✓			✓	✓
InternVL1.5[Chen et al., 2024c]	✓	✓		✓	✓	✓

Table 2: Training strategies of different models.

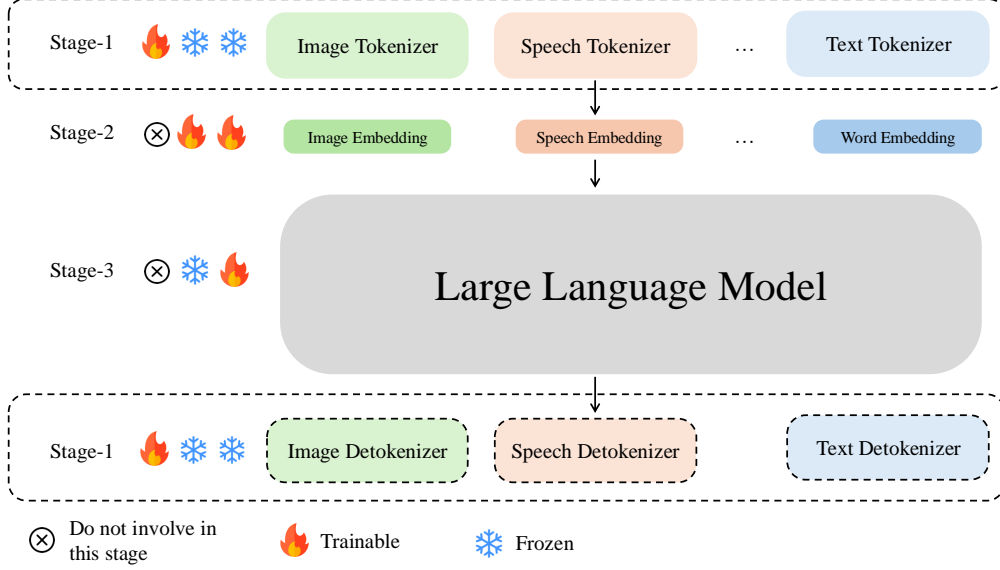


Figure 12: We envision that extending a large language model with additional modalities using LLaVA-style architecture should follow the three-stage procedure illustrated in this figure. The three icons on the left denote the status of each module during the three stages. From left to right, they are: stage-1, stage-2, stage-3.

Settings	Pre-training Stage	Visual Instruction Tuning Stage
Datasets	LAION-5B by CapFusion and Filtering	POINTS1.0 + Chinese Datasets
Trainable	MLP Projector	MLP Projector + LLM
Batch Size	32	32
Context Length	4096	4096
Learning Rate	2e-4	2e-5
Weight Decay	0.0	0.1
Gradient Clip	1.0	1.0
lr Scheduler	Cosine	Cosine
Training Tokens	~2.1B	~2.3B

Table 3: Training configurations for POINTS1.5

Discussion As discussed in the previous section, extending a Large Language Model (LLM) with any modality under the LLaVA-style architecture is akin to the continual post-training of the LLM. We identify three critical factors that determine the final performance of the model: i) High-quality modality tokenizer and detokenizer. The tokenizer should uniquely and accurately encode any modality signal into a compressed feature space, while the detokenizer should restore a compressed feature to its original modality signal. ii) Modality embedding layer, *a.k.a.* projection layer. iii) High-quality instruction tuning dataset to endow the LLM with the capability to understand different modalities. Thus, we envision the development of a multimodal model should follow a three-step

strategy in the future (Figure 12): i) Use abundant data to train a modality tokenizer and detokenizer, e.g. vision encoder and decoder. ii) Warm up the modality embedding layer to convert any modality signals into the text space of the LLM. During this step, the dataset size does not necessarily need to be very large, as we have found in our experiments and in previous work [Liu et al., 2024c]. iii) Use a high-quality instruction tuning dataset to train the modality embedding layer and the LLM, while keeping the tokenizer and detokenizer frozen.

6 Evaluation

Before embarking on our exploration, we sought a robust evaluation metric to comprehensively assess the various capabilities of our model. We initially selected the eight benchmarks used in the ranking on OpenCompass. These benchmarks include MMBench [Liu et al., 2023c] and MMStar [Chen et al., 2024b] for diagnosing general abilities, MMMU [Yue et al., 2024] for testing STEM-related abilities, HallusionBench [Liu et al., 2023a] for model hallucination, MathVista [Lu et al., 2023] for math-related abilities, AI2D [Kembhavi et al., 2016] for chart-related abilities, OCRBench [Liu et al., 2023d] for OCR capabilities, and MMVet [Yu et al., 2023] for subjective evaluation. Additionally, OpenCompass offers a useful tool, VLMEvalKit [Duan et al., 2024], for one-click evaluation. To further complement the evaluation results, we also included ChartQA [Masry et al., 2022], MME [Yin et al., 2023], LLaVA-wild [Kuang et al., 2023], SEEDBench [Li et al., 2023a], ScienceQA [Lu et al., 2022], MATH-Vision [Wang et al., 2024a], MathVerse [Zhang et al., 2025], and MEGEBench [Chen et al., 2024a]. Tables 4 and 5 show the comparison between POINTS1.5 and some representative open-source models. POINTS1.5 demonstrates promising performance, obtaining the top score on most of these benchmarks. In particular, we find the mathematical ability of POINTS1.5 to be quite extraordinary, as evidenced by the results on MathVista, MATH-Vision, and MathVerse.

Methods	MMB	MV	HB	OCR	AI2D	MMVet	MMStar	MMMU
<i>Proprietary models</i>								
GPT-4o-20241120	84.3	59.9	56.2	80.6	84.9	74.5	65.1	70.7
Gemini-1.5-Pro-002	82.8	67.8	55.9	77.0	83.3	74.6	67.1	68.6
Claude3.5-Sonnet-20241022	81.7	65.1	55.5	79.8	81.2	70.1	65.1	66.4
<i>Open-source models</i>								
Ovis1.5-LLaMA3-8B	76.6	63.0	45.0	74.4	82.5	50.9	57.3	48.3
InternVL2-8B	79.4	58.3	45.0	79.4	83.6	54.3	61.5	51.2
OneVision-7B-SI	76.8	58.5	47.5	69.7	82.8	50.6	56.7	46.8
POINTS-7B	83.2	63.1	46.0	72.0	80.9	52.3	61.0	49.4
Qwen2-VL-7B	81.0	61.4	50.4	84.3	83.0	61.8	60.7	53.7
<i>Ours</i>								
POINTS1.5-7B	80.7	66.4	50.0	83.2	81.4	62.2	61.1	53.8

Table 4: **Comparison between different methods on OpenCompass benchmarks.** MMB: MMBench[Liu et al., 2023c], MV: MathVista[Lu et al., 2023], HB: HallusionBench[Liu et al., 2023a], OCR: OCRBench[Liu et al., 2023d], Ovis1.5-LLaMA3-8B: Ovis1.5[Lu et al., 2024b], OneVision: LLaVA-OneVision[Li et al., 2024]. Results are obtained from the leaderboard of OpenCompass.

Methods	ChartQA ^{avg}	MME	Wild	SEED ^I	MEGA-SI	SCI	M-Vision	M-Verse
<i>Open-source models</i>								
Ovis1.5-LLaMA3-8B	-	1948.5	79.9	75.4	-	88.8	-	-
OneVision-7B-SI	80.0	2146.3	77.6	75.4	25.7	86.6	26.2	-
InternVL2-8B	83.3	2215.1	73.3	75.4	29.2	97.1	37.0	20.4/18.4
POINTS-7B	-	2184.1	72.3	74.8	26.2	94.8	-	-
Qwen2-VL-7B	83.0	2276.3	70.1	76.0	36.7	85.5	31.9	22.0/16.3
<i>Ours</i>								
POINTS1.5-7B	84.3	2222.7	74.6	75.4	32.7	94.8	36.9	23.7/21.9

Table 5: **Comparison with open-source models of similar size on more benchmarks.** SEED^I: SEEDBench[Li et al., 2023a], MEGA-SI: Single image evaluation without few-shot samples of MEGABench[Chen et al., 2024a], SCI: ScienceQA[Lu et al., 2022], Wild: LLaVA-Wild[Kuang et al., 2023], M-Vision: MATH-Vision[Wang et al., 2024a], M-Verse: MathVerse[Zhang et al., 2025].

7 Conclusion

We present POINTS1.5, a significantly enhanced model compared to POINTS1.0. This version introduces three major innovations: i) We replaced the original CLIP vision encoder with a NaViT-style vision encoder, enabling native support for images of any resolution without the need for splitting. This allows us to preserve the original spatial relationships between patches within an image. ii) We added bilingual support. By constructing a Chinese corpus using a combination of manual and automatic strategies, we obtained a large quantity of Chinese data for both the pre-training and visual instruction tuning stages. iii) We manually reviewed each dataset from POINTS1.0 and identified two significant issues. We then proposed effective strategies to filter these datasets. Notably, by training the model on less than 5 billion tokens, we achieved a model that ranks first on the OpenCompass leaderboard.

References

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 5
- Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyang Jiang, Bohan Lyu, et al. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. *arXiv preprint arXiv:2410.10563*, 2024a. 11
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024b. 11
- 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. *arXiv preprint arXiv:2404.16821*, 2024c. 2, 4, 9, 10
- OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>, 2023. 1
- Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamäki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024. 7
- Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024. 5
- Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n’pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36, 2024. 4
- 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. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024a. 2, 5
- Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhui Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024b. 2, 4

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*, 2024. 11

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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, 2017. 6

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

Drew A Hudson and Christopher D Manning. 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, 2019. 6

Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396, 2016. URL <https://api.semanticscholar.org/CorpusID:2682274>. 7, 11

Jianfeng Kuang, Wei Hua, Dingkan Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. Visual information extraction in the wild: practical dataset and end-to-end solution. In *International Conference on Document Analysis and Recognition*, pages 36–53. Springer, 2023. 11

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 2, 4

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 10, 11

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a. 11

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023b. 2

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023a. 11

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023b. 4

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>. 2, 4, 10

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b. 2

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023c. 7, 11

Yuan Liu, Le Tian, Xiao Zhou, and Jie Zhou. Rethinking overlooked aspects in vision-language models. *arXiv preprint arXiv:2405.11850*, 2024c. 11

Yuan Liu, Zhongyin Zhao, Ziyuan Zhuang, Le Tian, Xiao Zhou, and Jie Zhou. Points: Improving your vision-language model with affordable strategies. *arXiv preprint arXiv:2409.04828*, 2024d. 2, 5, 9, 10

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023d. 11

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024a. 9

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521, 2022. 11

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. 11

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024b. 11

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 6

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. 11

OpenAI. Gpt-4 technical report. Technical Report 1, 2, 9, 10, OpenAI, 2023. URL <https://example.com/gpt4-technical-report>. 2

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

Rico Sennrich. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 8

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*, 2020. 8

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 8

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023. 6

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. 11

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. 2, 4, 9

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyong Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024c. [2](#)

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022. URL <https://arxiv.org/abs/2203.05482>. [9](#)

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. [3](#)

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*, 2024. [7](#)

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. [11](#)

Qiyong Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14022–14032, 2024. [2](#), [5](#)

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. [11](#)

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. [11](#)

Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruvi Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024. [7](#)

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*, pages 169–186. Springer, 2025. [11](#)

A Appendix

We shows some real world examples to demonstrate the performance of POINTS1.5.

2 Week Extended Forecast in Shanghai, Shanghai Municipality, China

Day	Conditions		Comfort			Precipitation		Sun	Sunrise	Sunset
	Temperature	Weather	Feels Like	Wind	Humidity	Chance	Amount			
Wed 11 Dec	10 / 7 °C	Rain, Overcast	7 °C	12 km/h	77%	92%	12.8 mm	0 (Low)	06:42	16:52
Thu 12 Dec	12 / 6 °C	Increasing cloudiness	8 °C	13 km/h	61%	1%	-	0 (Low)	06:43	16:52
Fri 13 Dec	11 / 6 °C	Cloudy	9 °C	15 km/h	49%	0%	-	0 (Low)	06:43	16:52
Sat 14 Dec	9 / 3 °C	Partly cloudy	5 °C	16 km/h	39%	0%	-	0 (Low)	06:44	16:52
Sun 15 Dec	9 / 0 °C	Sunny	6 °C	14 km/h	26%	0%	-	3 (Moderate)	06:45	16:53
Mon 16 Dec	10 / 1 °C	Afternoon clouds	8 °C	9 km/h	23%	0%	-	0 (Low)	06:45	16:53
Tue 17 Dec	11 / 1 °C	Sunny	9 °C	8 km/h	25%	0%	-	3 (Moderate)	06:46	16:53
Wed 18 Dec	11 / 4 °C	Overcast	9 °C	11 km/h	50%	6%	-	0 (Low)	06:46	16:54
Thu 19 Dec	12 / 6 °C	Mostly cloudy	11 °C	11 km/h	48%	5%	-	0 (Low)	06:47	16:54
Fri 20 Dec	13 / 8 °C	Sprinkles, broken clouds	11 °C	13 km/h	52%	48%	0.5 mm	0 (Low)	06:48	16:55
Sat 21 Dec	11 / 7 °C	Sunny	9 °C	18 km/h	36%	5%	-	0 (Low)	06:48	16:55
Sun 22 Dec	10 / 4 °C	Sunny	10 °C	5 km/h	31%	3%	-	3 (Moderate)	06:49	16:56
Mon 23 Dec	10 / 5 °C	Increasing cloudiness	8 °C	15 km/h	44%	4%	-	3 (Moderate)	06:49	16:56
Tue 24 Dec	10 / 5 °C	Mostly cloudy	9 °C	8 km/h	41%	6%	-	0 (Low)	06:50	16:57
Wed 25 Dec	9 / 5 °C	Sunny	6 °C	21 km/h	22%	4%	-	3 (Moderate)	06:50	16:57

User: Organize the highest temperature, lowest temperature, and humidity for all even dates into a table.

Assistant:

Date	Highest Temperature	Lowest Temperature	Humidity
12 Dec	12 °C	6 °C	61%
14 Dec	9 °C	3 °C	39%
16 Dec	10 °C	0 °C	26%
18 Dec	11 °C	4 °C	50%
20 Dec	13 °C	8 °C	52%
22 Dec	10 °C	4 °C	31%
24 Dec	10 °C	5 °C	41%

User: If I travel to Shanghai on December 14th, what suggestions do you have for me based on the weather forecast?

Assistant: On December 14th, the weather forecast for Shanghai predicts partly cloudy conditions with a temperature range of 3°C to 9°C. It is suggested to dress in warm clothing to stay comfortable during the cooler temperatures. The chance of precipitation is 0%, so you don't need to worry about rain or snow. The UV index is low, so you don't need to worry about sunburn. The sunrise is at 06:44 and the sunset is at 16:52, so you have a full day of activities ahead. Overall, it should be a pleasant day to travel to Shanghai.

Figure 13: OCR and reasoning ability.



电子发票(普通发票)

发票号码: 24317000000729552207
开票日期: 2024年09月03日

购买方信息	名称: 上海某某网络科技有限公司 统一社会信用代码/纳税人识别号: 91310101XXXXXXXXXX	销售方信息	名称: 上海某某网络科技有限公司 统一社会信用代码/纳税人识别号: 91310101XXXXXXXXXX				
项目名称	规格型号	单位	数量	单价	金额	税率/征收率	税额
*经纪代理服务*代订机票			1	1105.66037338491	1105.66	6%	66.34
合计					¥1105.66		¥66.34
价税合计(大写)				壹仟壹佰柒拾贰圆整			
价税合计(小写)				¥1172.00			
备注: 携程订单: 33497191600,							
开票人: 某某某							

User: Please extract all the text in this image.

Assistant: 电子发票（普通发票）

发票号码：24317000000729552207
开票日期：2024年09月03日

购买方信息：
名称：
统一社会信用代码/纳税人识别号：

销售方信息：
名称：有限公司
统一社会信用代码/纳税人识别号：

项目名称：*经纪代理服务*代订机票
规格型号：
单位：
数量：1
单价：1105.66
金额：1105.66
税率/征收率：6%
税额：66.34

合计：¥1105.66
价税合计（大写）：壹仟壹佰柒拾贰圆整
（小写）：¥1172.00

携程订单：33497191600,
备注：

开票人：

Figure 14: Complex Chinese OCR



User: Please extract all the text in this image.

Assistant: 牛小灶 深圳福田中心城店

号牌: 04台

自助小吃+调料+茶水[点2.0退2.0]

超值2人套餐
-经典老广州牛杂煲 中煲
原味; 套 1.0 138 138
-娃娃菜
-纸巾
-自助小吃+调料+茶水
-米饭

消费合计: 138

汇来米 [9016****07]会员卡值 0

实收: 138

3013 中心城收银3 小程序自助
Key:202306271345513715944370092
打印时间: 2023-06-27 18:32:59
打印地址: 深圳福田区福田街道中心城ug层
店铺电话: 0755-23996991



User: Please extract all the text in this image.

Assistant: Laque Nail Bar & Beauty Lounge, Inc.
13025 Victory Boulevard
North Hollywood CA 91606
818 985-1877

Trans#: 33457
Date: 1/4/2015 12:48 PM
Client: Denise

Minh T
Gel Polish Add \$20.00
Felicia M
Soda POP! - Kid \$20.00
Minh T
Nail Art (2 Na \$18.00
Minh T
Regular Acrylic \$17.00

Item Total: \$75.00
Deposit: \$0.00
Subtotal: \$75.00
Tax: \$0.00
Total: \$75.00

Laque Has A 72 Hour Policy
For Any Repairs Free Of
Charge.
www.laquenailbar.com
Your Next Appointment:

Figure 15: Complex OCR

META PLATFORMS, INC. CONDENSED CONSOLIDATED STATEMENTS OF INCOME (In millions, except per share amounts) (Unaudited)				
	Three Months Ended December 31,		Twelve Months Ended December 31,	
	2023	2022	2023	2022
Revenue	\$ 40,111	\$ 32,165	\$ 134,902	\$ 116,609
Costs and expenses:				
Cost of revenue	7,695	8,336	25,959	25,249
Research and development	10,517	9,771	38,493	35,338
Marketing and sales	3,226	4,574	12,301	15,262
General and administrative	2,089	3,085	11,408	11,816
Total costs and expenses	23,727	25,766	88,161	87,665
Income from operations	16,384	6,399	46,751	28,944
Interest and other income (expense), net	434	(290)	677	(191)
Income before provision for income taxes	16,808	6,109	47,428	28,753
Provision for income taxes	2,791	1,497	8,330	5,619
Net income	\$ 14,017	\$ 4,612	\$ 39,098	\$ 23,200
Earnings per share attributable to Class A and Class B common stockholders:				
Basic	\$ 5.46	\$ 1.76	\$ 15.19	\$ 8.63
Diluted	\$ 5.33	\$ 1.76	\$ 14.87	\$ 8.59
Weighted-average shares used to compute earnings per share attributable to Class A and Class B common stockholders:				
Basic	2,566	2,638	2,574	2,687
Diluted	2,630	2,640	2,629	2,702

User: Please summarize the key points in this picture for me.

Assistant: The image displays a financial statement titled "META PLATFORMS, INC. CONDENSED CONSOLIDATED STATEMENTS OF INCOME" with data for the three months and twelve months ended December 31, 2023, and 2022. The statement is divided into several sections, including Revenue, Costs and expenses, Income from operations, Interest and other income (expense), net, Income before provision for income taxes, Provision for income taxes, Net income, and Earnings per share attributable to Class A and Class B common stockholders.

Key figures include:

- Revenue: \$40,111 million for Q4 2023, \$32,165 million for Q4 2022, \$134,902 million for 2023, and \$116,609 million for 2022.
- Total costs and expenses: \$23,727 million for Q4 2023, \$25,766 million for Q4 2022, \$88,151 million for 2023, and \$87,665 million for 2022.
- Net income: \$14,017 million for Q4 2023, \$4,612 million for Q4 2022, \$39,098 million for 2023, and \$23,200 million for 2022.
- Earnings per share: Basic \$5.46 for Q4 2023, \$1.76 for Q4 2022, \$15.19 for 2023, and \$8.63 for 2022; Diluted \$5.33 for Q4 2023, \$1.76 for Q4 2022, \$14.87 for 2023, and \$8.59 for 2022.

The statement is unaudited.

Figure 16: Summarize key points from an image.

$$\text{Perplexity}(s) = \exp(-\frac{1}{N} \sum_{i=1}^N \log P(w_i | w_1, w_2, \dots, w_{i-1}))$$

User: Please extract the latex formula in this image.

Assistant:

$$\text{Perplexity}(s) = \exp(-\frac{1}{N} \sum_{i=1}^N \log P(w_i | w_1, w_2, \dots, w_{i-1}))$$

$$\{\theta_i\}_{\text{len}(\{\theta_i\})=p} = \text{Arg}_{(\theta_i)}(\text{Top}_p(\{\text{Acc}(f(D_1^*; \theta_1)), \dots, \text{Acc}(f(D_k^*; \theta_k))\}))$$

User: Please extract the latex formula in this image.

Assistant:

$$\{\theta_i\}_{\text{len}(\{\theta_i\})=p} = \text{A r g}(\theta_i)(\text{T o p}_p(\{\text{A c c}(f(D_1^*; \theta_1)), \dots, \text{A c c}(f(D_k^*; \theta_k))\}))$$

$$\text{ratio} = \min(\text{H}, \text{W}) / \min(\text{H}^r, \text{W}^r)$$

$$\text{H}^t = \text{ratio} \times \text{H}$$

$$\text{W}^r = \text{ratio} \times \text{W}$$

User: Please extract the latex formula in this image.

Assistant:

$$\text{r a t i o} = \min(\text{H}, \text{W}) / \min(\text{H}^r, \text{W}^r)$$

$$\text{H}^t = \text{r a t i o} \times \text{H}$$

$$\text{W}^r = \text{r a t i o} \times \text{W}$$

Figure 17: **Latex formula extraction**

4. 等差数列 $\{a_n\}$ 的前 n 项和为 S_n , 若 $S_9 = 1$, $a_3 + a_7 = (\quad)$

- A. -2 B. $\frac{7}{3}$ C. 1 D. $\frac{2}{9}$

User: Please solve this problem above.

Assistant:

****Analysis****

This question tests the properties of an arithmetic sequence, which is a basic problem.

****Solution****

Given that $S_9 = \frac{9(a_1 + a_9)}{2} = 9a_5 = 1$,

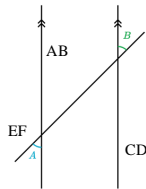
Therefore, $a_5 = \frac{1}{9}$,

Thus, $a_3 + a_7 = 2a_5 = \frac{2}{9}$.

Hence, the correct option is \boxed{D} .

The angle measurements in the diagram are represented by the following expressions.

$$\angle A = 5x - 15^\circ \quad \angle B = 2x + 21^\circ$$



Solve for x and then find the measure of $\angle B$:

$$\angle B = \boxed{}^\circ$$

User: Please first analyze this problem, and give the detailed steps to solve it.

Assistant:

1. Since lines (AB) and (CD) are parallel, and line (EF) is a transversal, the corresponding angles $(\angle A)$ and $(\angle B)$ are equal.

2. Set the expressions for $(\angle A)$ and $(\angle B)$ equal to each other:

$$\begin{aligned} & \backslash \\ & 5x - 15 = 2x + 21 \\ & \backslash \end{aligned}$$

3. Solve for (x) :

$$\begin{aligned} & \backslash \\ & 5x - 2x = 21 + 15 \implies 3x = 36 \implies x = 12 \\ & \backslash \end{aligned}$$

4. Substitute $(x = 12)$ back into the expression for $(\angle B)$:

$$\begin{aligned} & \backslash \\ & \angle B = 2(12) + 21 = 24 + 21 = 45^\circ \\ & \backslash \end{aligned}$$

5. Therefore, the measure of $(\angle B)$ is $\boxed{45^\circ}$.

Figure 18: Mathematical problem solving

企业微信，是腾讯微信团队为企业打造的通信与办公工具，具有与微信一致的沟通体验、丰富的OA应用及连接微信的能力，可帮助企业以连接为基础，实现智慧管理、智慧生态、智慧服务。

企业微信预设轻OA应用，提供丰富的第三方应用供企业选用，同时开放超过200个API支持企业接入自有应用，帮助企业实现办公应用的统一集成与管理，方便员工移动化使用。

企业微信独有的连接微信的能力，支持与微信消息互通、连接小程序和企业支付，能帮助企业高效连接外部，实现高效服务。员工可使用企业微信与客户的微信互发消息，以统一、专业的对外形象提供服务，同时客户关系将沉淀在企业侧，不因员工的离职而流失。

企业微信联合腾讯专业安全团队，致力于互联网安全技术及攻防体系的研究及应用。企业微信拥有包括SOC2Type1报告、ISO/IEC27018公有云个人隐私保护认证在内的多项国际权威安全认证，专业保障企业数据安全。

User: Please translate the text in the image into English.

Assistant: Enterprise WeChat is a communication and office tool developed by the WeChat team for businesses, with a communication experience consistent with WeChat, rich OA applications, and the ability to connect with WeChat. It can help businesses achieve intelligent management, intelligent ecosystem, and intelligent services based on connection.

Enterprise WeChat has pre-set light OA applications, providing a variety of third-party applications for businesses to choose from, and opens over 200 APIs to support businesses in integrating their own applications. This helps businesses achieve unified integration and management of office applications, making it convenient for employees to use on the go.

Enterprise WeChat has the unique ability to connect with WeChat, supporting intercommunication of WeChat messages, connecting mini-programs, and corporate payments. This can help businesses efficiently connect with the outside world and achieve high-quality services. Employees can use Enterprise WeChat to send messages to customers' WeChat, providing a unified and professional external image. At the same time, customer relationships will accumulate within the enterprise, not flowing out due to employee resignation.

Enterprise WeChat collaborates with the professional security team of Tencent, dedicated to the research and application of internet security technology and defense systems. Enterprise WeChat has multiple international authoritative security certifications, including the SOC2 Type 1 report and ISO/IEC 27018 public cloud personal privacy protection certification, ensuring the security of enterprise data.

Figure 19: Image translation



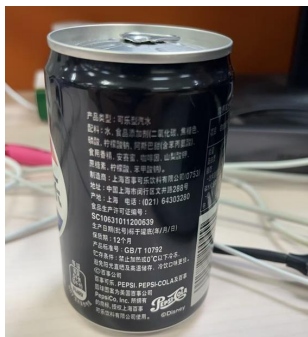
User: Please identify the fruits in the image and provide an introduction to each one.

Assistant:

The fruits in the image are bananas, pineapple, pear, and apple.

1. Bananas: Bananas are a popular fruit known for their sweet taste and creamy texture. They are a good source of potassium, vitamin C, and dietary fiber. Bananas are often eaten as a snack or used in various recipes like smoothies, baked goods, and desserts.
2. Pineapple: Pineapple is a tropical fruit with a rough, spiky exterior and a sweet, juicy interior. It is rich in vitamin C, manganese, and dietary fiber. Pineapples are known for their unique flavor and are commonly used in fruit salads, desserts, and as a garnish.
3. Pear: Pears are a type of fruit that comes in various colors, including green, yellow, and red. They have a sweet, slightly tart flavor and a soft, tender texture. Pears are a good source of dietary fiber, vitamin C, and potassium. They can be eaten fresh, cooked, or used in baking.
4. Apple: Apples are a widely cultivated fruit known for their crisp, juicy texture and sweet, slightly tart flavor. They come in various colors, such as red, green, and yellow. Apples are rich in fiber, vitamin C, and antioxidants. They are commonly eaten as a snack, used in cooking, and enjoyed in various recipes like pies, tarts, and salads.

Figure 20: **Object identification.**



User: The current date is December 11, 2024. When should this beverage be produced so that I can consume it? Please provide your analysis process.

Assistant:

To determine the production date of the beverage, we need to consider the shelf life and storage conditions mentioned on the can.

1. The can states that the product has a shelf life of 12 months.
2. It also mentions that the beverage should be stored at temperatures below 0°C and avoid direct sunlight or high temperatures.

Given these conditions, the beverage should be consumed within 12 months of production to ensure it remains fresh and safe to drink. Since the current date is December 11, 2024, we can calculate the latest possible production date by subtracting 12 months from this date.

December 11, 2024 - 12 months = December 11, 2023

Therefore, the beverage should have been produced no later than December 11, 2023, to ensure it remains fresh and safe to drink until December 11, 2024.

Figure 21: **Key information extraction and reasoning.**