

SUPPLEMENTARY MATERIAL - MMSEARCH: BENCHMARKING THE POTENTIAL OF LARGE MODELS AS MULTI-MODAL SEARCH ENGINES

Anonymous authors

Paper under double-blind review

OVERVIEW

- Section **A**: Related work.
- Section **B**: Additional experiments and analysis.
- Section **C**: Additional experimental details.
- Section **D**: More dataset details.
- Section **E**: Qualitative examples.

A RELATED WORK

Large Multimodal Models. Recently, multimodal models (Radford et al., 2021; Li et al., 2022; OpenAI, 2023; Rombach et al., 2022; Jiang et al., 2024b) has gained unparalleled attention. Building on the success of Large Language Models (LLMs) (Touvron et al., 2023a;b) and large-scale vision models (Radford et al., 2021), Large Multimodal Models (LMMs) are gaining prominence across diverse domains. These models extend LLMs to handle tasks involving various modalities, including mainstream 2D image processing (Liu et al., 2023a; Zhu et al., 2023; Lin et al., 2023; Gao et al., 2023), as well as 3D point clouds (Xu et al., 2023; Guo et al., 2023; 2024), and videos (Li et al., 2023; Chen et al., 2023a; Zhang et al., 2023; Fu et al., 2024). Among these LMMs, OpenAI’s GPT-4o (OpenAI, 2024b) and Anthropic’s Claude 3.5 Sonnet (Anthropic, 2024) demonstrate outstanding visual reasoning and comprehension capability, setting new standards in multi-modal performance. However, their closed-source nature limits broader adoption and development. In contrast, another research trajectory focuses on open-source LMMs for the community. Pioneering works like LLaVA (Liu et al., 2023a; 2024; Li et al., 2024b;a), LLaMA-Adapter (Zhang et al., 2024b; Gao et al., 2023), and MiniGPT-4 (Zhu et al., 2023; Chen et al., 2023b) incorporate a frozen CLIP (Radford et al., 2021) model for image encoding and integrate visual information into LLM for multi-modal instruction tuning. Later, works such as mPLUG-Owl (Ye et al., 2023a;b; 2024), SPHINX (Gao et al., 2024; Lin et al., 2023), and InternLM-XComposer (Dong et al., 2024) further advanced the field by incorporating diverse visual instruction tuning data and generalizing to more scenarios. More recent developments in the field have taken diverse directions. For example, several studies (Zong et al., 2024; Tong et al., 2024) explore multiple vision encoders design. Meanwhile, other works (Liu et al., 2024; Chen et al., 2024c; Qwen Team, 2024) incorporate high-resolution image input. Multi-image instruction data (Li et al., 2024b; Jiang et al., 2024a) is also integrated to enable perception across multiple images. While various benchmarks, both in the general (Fu et al., 2023; Liu et al., 2023b; Yu et al., 2023) and expert (Zhang et al., 2024c; Lu et al., 2023; 2022) domain, has been proposed, the potential of LMM to function as a multimodal search engine remains largely unexplored. To this end, we introduce the MMSEARCH benchmark, which evaluates LMMs’ zero-shot abilities of multimodal search, offering valuable insights for future research.

Large models with Retrieval Augmented Generation (RAG). RAG (Retrieval-Augmented Generation) is an effective strategy for enhancing model knowledge by retrieving relevant information from external sources (Fan et al., 2024). RAG has been leveraged in various scenarios including knowledge-intensive question answering (Borgeaud et al., 2022; Guu et al., 2020), machine translation (He et al., 2021), and hallucination elimination (Béchar & Ayala, 2024). Current works has

focused on improving specific aspects of RAG. RG-RAG (Chan et al., 2024) proposes to refine the query for retrieval by decomposition and disambiguation. Self-RAG (Asai et al., 2023) incorporates the self-reflection of LLM to enhance the generation quality. The AI search engine could be viewed as a form of RAG with the Internet serving as the external knowledge source. Recently, MindSearch (Chen et al., 2024b) proposes an AI search engine framework to simulate the human minds in web information seeking. Meanwhile, multiple benchmarks of RAG (Yang et al., 2024; Chen et al., 2024a) have been introduced to comprehensively evaluate a RAG system. However, both the current AI search engine and RAG benchmark are limited to the text-only setting, leaving the multimodal search engine and evaluation largely unexplored. To bridge this gap, we introduce MMSEARCH-ENGINE and MMSEARCH, a multimodal AI search engine pipeline and dataset designed to evaluate various multimodal scenarios.

B ADDITIONAL EXPERIMENTS AND ANALYSIS

B.1 SCALING TEST-TIME COMPUTE VS SCALING MODEL SIZE

Recent works such as OpenAI o1 (OpenAI, 2024a) and Li et al. (2024c) have highlighted the critical role of scaling test-time computation in enhancing model performance. Our end-to-end task, which requires multiple Internet interactions, presents an opportunity to investigate the potential of scaling test-time computation compared to scaling model size. To explore this, we conduct experiments using LLaVA-OneVision-7B (Li et al., 2024a), focusing on scaling test-time computation, and compare against LLaVA-OneVision-72B scaling in model size, which aims to provide insights into the relative benefits of increased inference computation versus increased model parameters.

For scaling up the test-time computation, we adopt a multi-modal search strategy similar to best-of-N solution, where ‘N’ denotes 25 in our settings. Specifically, for LLaVA-OneVision-7B, we first prompt the model to generate a requery 5 times, from which we selected the one with the highest requery score S_{req} . This requery is then used to retrieve brief results from 8 websites from a search engine. The model is again prompted 5 times to select the most informative website. After removing duplicates from the selected websites, we extract the full website content from the remaining ones and prompt the model to answer 5 times, obtaining 25 end-to-end outputs in total. We compute the F1 score for each answer against the ground truth and take the maximum as the model’s end-to-end score for the query. Table 1 shows that LLaVA-OneVision-7B (TTC) achieves the score of 55.2% in the end-to-end task, significantly enhancing the original score of 29.6%, which surpasses LLaVA-OneVision-72B’s 44.9% and GPT-4V’s 52.1%. This result reveals the substantial potential of scaling test-time computation, validating the effectiveness of this technique as introduced by OpenAI o1. Our findings provide valuable insights for future research in this domain, suggesting that increased inference computation may offer comparable or superior performance improvements to increased model size not only in math and code tasks, but also in multimodal search tasks.

Table 1: **Scaling Test-Time Compute vs Scaling Model Size.** ‘TTC’ and S_{e2e} denote Test-Time Computation and the score of end-to-end task.

Model	Inference Cost	S_{e2e}
LLaVA-OV-7B	1	29.6
LLaVA-OV-7B (TTC)	~25	55.2
LLaVA-OV-72B	~6	44.9

B.2 DEFINITION OF ERRORS IN THE REQUERY AND SUMMARIZATION TASKS

Five types of requery error:

- *Lacking Specificity*, where the model fails to include all the specific information in the requery and therefore leads to sub-optimal search results. For example, the query is asking the release date of Vision Pro in China. However, the model omits the condition of China and directly asks about the release date of Vision Pro.
- *Inefficient Query*, where the model does not consider the real scenario and the requery is inefficient for the search engine to find the answer. For example, the query is asking whether the Van Gogh’s Sunflowers and Antoni Clavé’s Grand Collage are both oil paintings. Clearly, it is a commonsense that Van Gogh’s Sunflowers is an oil painting and Antoni Clavé’s Grand Collage is much less well-known. An efficient query should be asking about

the images of Antoni Clavé’s Grand Collage and further determine if it is also an oil painting by directly looking at it. However, the model directly asks the original query to the search engine. There is very little chance that an exact same question has ever been raised so probably this requery will bring very little helpful information.

- *Excluding Image Search Results*, where the model totally ignores the information in the screenshot of the image search results and therefore lacks important specific information in the requery. For example, the query is ‘When did this football player obtain the gold medal?’ and provides an image of the player. The model is supposed to find out the player’s name by viewing the image search result and raise a requery like ‘[PLAYER NAME] obtained the gold medal time’. However, the model fails to incorporate the player’s name in the requery and definitely the retrieved websites will not include any helpful information.
- *No Change*, where the model just uses the question as the query input to the search engine.
- *Irrelevant*, where the model either matches wrong information from the image search result or mistakenly understands the query and outputs an irrelevant requery.

Five types of summarization error:

- *Text Reasoning Error*, where the model fails to extract the answer from the website textual information.
- *Image-text Aggregation Error*, where obtaining the answer needs combining the information from both images and texts. The model fails to do so.
- *Image reasoning Error*, where the model fails to extract the answer from the image, and the answer can only be obtained from the image.
- *Hallucination* (Huang et al., 2023), where the model provides an unfaithful answer that cannot be grounded in the given content.
- *Informal*, the output format does not follow the prompt specifications, the same error type in the end-to-end task.

C ADDITIONAL EXPERIMENTAL DETAILS

More Implementation Details All our experiments of open-source models are conducted without any fine-tuning on search data or tasks. As for the prompts, the requery prompt contains 3 examples to better guide LMMs to output a valid requery. While prompts for other tasks are all in a zero-shot setting. We prompt the LMM to output as few words as possible for a better match with the ground truth. We employ the metric introduced in Section 2 in the main paper. Besides, we recruit eight qualified college students and ask them to solve the problems in MMSEARCH independently, following the same pipeline of MMSEARCH-ENGINE. This score serves as a baseline for human performance. We conduct all experiments on NVIDIA A100 GPUs.

The input image dimensions for the webpage’s top section screenshot were set to 1024×1024 pixels. For the full-page screenshot, we set the initial webpage width to 512 pixels, although the actual width of a small portion of webpages may vary due to its layout settings. Furthermore, considering that a full-page screenshot can be extremely lengthy, directly inputting it as a single image into an LLM would result in excessive downsizing, making the content too vague for accurate identification. To address this, we segmented the full-page screenshot into multiple images, starting from the top, with each segment measuring 512 pixels in height. Because of the context length limitations of LMMs, the maximum number of full-page screenshot segments is therefore restricted to 10.

Full-page Screenshot Slimming. For the full-page screenshot, we compute the Sobel gradients (Kanopoulos et al., 1988) to detect the edges and generate a gradient magnitude image. We iteratively remove the areas with gradients below a threshold, which represent the blank areas. This approach, shown in Fig. 1, effectively reduces image size while maintaining the document content.

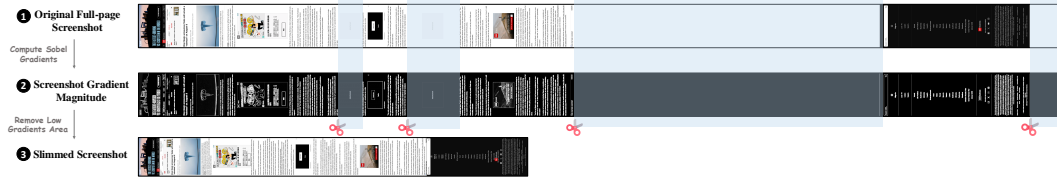


Figure 1: **Illustration of the Screenshot Slim Process.** We leverage Sobel gradients (Kanopoulos et al., 1988) to identify blank areas and remove them. After slimming, the screenshot size is largely reduced without any information loss.

Model Sources. For different LMMs, we select their latest models with size around 7B for evaluation to fully reveal their multimodal search proficiency. Table 2 presents the release time and model sources of LMMs used in MMSEARCH.

Table 2: **The Release Time and Model Source of LMMs Used in MMSEARCH.**

Model	Release Time	Source
GPT-4V (OpenAI, 2023)	2023-09	https://platform.openai.com/
GPT-4o (OpenAI, 2024b)	2024-05	https://platform.openai.com/
Claude 3.5 Sonnet (Anthropic, 2024)	2024-06	https://www.anthropic.com/news/claude-3-5-sonnet
InternLM-XC2.5 (Zhang et al., 2024a)	2024-07	https://github.com/InternLM/InternLM-XComposer
Mantis (Jiang et al., 2024a)	2024-05	https://tiger-ai-lab.github.io/Mantis/
LLaVA-NeXT-Interleave (Li et al., 2024b)	2024-06	https://github.com/LLaVA-VL/LLaVA-NeXT
InternVL2 (Chen et al., 2024c)	2024-07	https://github.com/OpenGVLab/InternVL
mPlug-Owl3 (Ye et al., 2024)	2024-08	https://github.com/X-PLUG/mPLUG-Owl
Idefics3 (Laurençon et al., 2024)	2024-08	https://huggingface.co/HuggingFaceM4/Idefics3-8B-Llama3
LLaVA-OneVision (Li et al., 2024a)	2024-08	https://llava-vl.github.io/blog/2024-08-05-llava-onevision/
Qwen2-VL (Qwen Team, 2024)	2024-08	https://github.com/QwenLM/Qwen2-VL

Input Prompts of LMM for Response Generation. We showcase the input prompts of LMM for the three tasks respectively in Table 3-5. We adopt two types of prompts for queries with an image and without images. For query with an image, we specifically require the LMM to leverage the image search result to solve the task.

Table 3: **Input Prompt of LMMs for Requery.** We adopt two different prompts for the query with image input and without image input. We leverage a 3-shot prompt to guide the LMM to generate a reasonable requery.

Question	Prompt
Query without image	<p>You are a helpful assistant. I am giving you a question, which cannot be solved without external knowledge. Assume you have access to a text-only search engine (e.g., google). Please raise a query to the search engine to search for what is useful for you to answer the question correctly. Your query needs to consider the attribute of the query to search engine. Here are 3 examples:</p> <p>Question: Did Zheng Xiuwen wear a knee pad in the women’s singles tennis final in 2024 Paris Olympics? Query to the search engine: Images of Zheng Xiuwen in the women’s singles tennis final in 2024 Paris Olympics</p> <p>Question: When will Apple release iPhone16? Query to the search engine: iPhone 16 release date</p> <p>Question: Who will sing a French song at the Olympic Games closing ceremony? Query to the search engine: Singers at the Olympic Games closing ceremony, French song.</p> <p>Question: <i>{question}</i>.</p> <p>Query to the search engine (do not involve any explanation):</p>
Query with image	<p>You are a helpful assistant. I am giving you a question including an image, which cannot be solved without external knowledge. Assume you have access to a search engine (e.g., google). Please raise a query to the search engine to search for what is useful for you to answer the question correctly. You need to consider the characteristics of asking questions to search engines when formulating your questions. You are also provided with the search result of the image in the question. You should leverage the image search result to raise the text query. Here are 3 examples:</p> <p>Question: Did Zheng Xiuwen wear a knee pad in the women’s singles tennis final in 2024 Paris Olympics? Query to the search engine: Images of Zheng Xiuwen in the women’s singles tennis final in 2024 Paris Olympics</p> <p>Question: When will Apple release iPhone16? Query to the search engine: iPhone 16 release date</p> <p>Question: Who will sing a French song at the Olympic Games closing ceremony? Query to the search engine: Singers at the Olympic Games closing ceremony, French song</p> <p>Question: <i>{query_image}{question}</i>. The image search result is: <i>{image_search_result}</i></p> <p>Query to the search engine (do not involve any explanation):</p>

Table 4: **Input Prompt of LMMs for Rerank.** We adopt two different prompts for the query with image input and without image input.

Question	Prompt
Query without image	<p>You are a helpful assistant. I am giving you a question and 8 website information related to the question (including the screenshot, snippet and title). You should now read the screenshots, snippets and titles. Select 1 website that is the most helpful for you to answer the question. Once you select it, the detailed content of them will be provided to help you correctly answer the question. The question is $\{question\}$. The website informations is $\{website_information\}$. You should directly output 1 website’s index that can help you most, and enclose the website in angle brackets. The output format should be: <Website Index >. An example of the output is: <Website 1 >. Your answer:</p>
Query with image	<p>You are a helpful assistant. I am giving you a question including an image. You are provided with the search result of the image in the question. And you are provided with 8 website information related to the question (including the screenshot, snippet, and title). You should now read the screenshots, snippets and titles of these websites. Select 1 website that is the most helpful for you to answer the question. Once you select it, the detailed content of them will be provided to help you correctly answer the question. The question is $\{query_image\}\{question\}$. The image search result is $\{image_search_result\}$. The website information is $\{website_information\}$. You should directly output 1 website’s index that can help you most, and enclose the website in angle brackets. The output format should be: <Website Index >. An example of the output is: <Website 1 >. Your answer:</p>

Table 5: **Input Prompt of LMMs for Summarization.** We adopt two different prompts for the query with image input and without image input.

Question	Prompt
Query without image	<p>You are a helpful assistant. I am giving you a question and 1 website information related to the question. Please follow these guidelines when formulating your answer: 1. If the question contains a false premise or assumption, answer “invalid question”. 2. When answering questions about dates, use the yyyy-mm-dd format. 3. Answer the question with as few words as you can.</p> <p>You should now read the information of the website and answer the question. The website information is <i>{website_information}</i>. The question is <i>{question}</i>. Please directly output the answer without any explanation:</p>
Query with image	<p>You are a helpful assistant. I am giving you a question including an image. You are provided with the search result of the image in the question. And you are provided with 1 website information related to the question. Please follow these guidelines when formulating your answer: 1. If the question contains a false premise or assumption, answer “invalid question”. 2. When answering questions about dates, use the yyyy-mm-dd format. 3. Answer the question with as few words as you can.</p> <p>You should now read the information of the website and answer the question. The website information is <i>{website_information}</i>. The image search result is <i>{image_search_result}</i>. The question is <i>{query_image}{question}</i>. Please directly output the answer without any explanation:</p>

D MORE DATA DETAILS

D.1 DATA EXAMPLE OF 4 EVALUATION TASKS




Query Information	Image	Image Search Result	Question: In which day of 2024 was the lunar rover project in the picture announced to be canceled?
Task1 Requery			
Query Information			
LMM Requery: VIPER was announced to be cancelled		Requery Annotation: NASA Ends VIPER Project	
Task2 Rerank			
Query Information			
Brief Result			
 <p><Website 1>: Title: NASA Ends VIPER Project, Continues Moon Exploration Snippet: NASA Ends VIPER Project, Continues Moon Exploration. Tiernan P. Doyle. Jul 17, 2024. RELEASE 24-094. ... (Volatiles Investigating Polar Exploration Rover) project. NASA stated cost increases, delays to the launch date, and the risks of future cost growth as the reasons to stand down on the mission. The rover was originally planned to launch in ...</p> <p><Website 2>: Title: NASA cancels VIPER lunar rover - SpaceNews Snippet: The VIPER lunar rover. Credit: NASA. BUSAN, South Korea 12/01/2014 NASA has canceled a robotic lunar rover mission that would have searched for ice at the south pole of the moon, citing development ...</p>	 <p><Website 3>: Title: NASA Ends VIPER Project, Continues Moon Exploration - Yahoo Finance Snippet: NASA Ends VIPER Project, Continues Moon Exploration. PR Newswire. Wed, Jul 17, 2024, 4:13 PM 3 min read. Link Copied. 0. WASHINGTON, July 17, 2024 /PRNewswire/ -- Following a comprehensive ...</p> <p><Website 4>: Title: NASA axes robotic lunar rover project VIPER due to rising costs - MSN Snippet: NASA has ended its VIPER project, which was hoping to launch the agency's first robotic lunar rover to the moon, due to the increasing costs of the program. "Decisions, of course, like this are ...</p>		
LMM Rerank: <Website 2>		Rerank Annotation: Valid: [< Website 1>, <Website 2>, <Website 3>] Unsure: [<Website 4>, <Website 5>] Invalid: [<Website 6>, <Website 7>, <Website 8>]	
Task3 Summarization			
Query Information			
Full Website Content			
 <p>Content: would have searched for ice at the south pole of the moon, citing development delays and cost overruns. NASA announced July 17 that it would end development of the Volatiles Investigating Polar Exploration Rover (VIPER) mission. The rover, to be sent to the south polar region of moon on a commercial lander called Griffin from Astrobotic Technology, would have explored terrain including permanently shadowed regions to better understand the extent and form of water ice there. At a briefing to announce the cancellation, agency officials said costs of VIPER had grown. Posted in Civil NASA cancels VIPER lunar rover by Jeff Foust July 17, 2024 July 17, 2024 Click to share on X (Opens in new window) Click to share on Facebook (Opens in new window) Click to share on LinkedIn (Opens in new window) Click to share on Reddit (Opens in new window) Click to email a link to a friend (Opens in new window) Click to share on Clipboard (Opens in new window) BUSAN, South Korea — NASA has canceled a robotic lunar rover mission that 's lander that would deliver the rover to the moon under a CLPS task order worth \$ 322 million. NASA said Griffin was now expected to be ready for the mission no earlier than September 2025. With VIPER canceled, NASA will retain the task order for Griffin. The mission will instead become a technology demonstrator, carrying a mass simulator in place of the rover to test Griffin 's ability to land large payloads. Kearns said NASA considered flying science payloads instead, but since the lander was designed for carrying a rover by more than 30 %, triggering a termination review by the agency. NASA had confirmed VIPER in 2021 at a cost of \$ 433.5 million. Joel Kearns, deputy associate administrator for exploration in NASA 's Science Mission Directorate, said the latest estimate was \$ 609.6 ...</p>			
LMM Answer: July 17		Answer Annotation: 07-17	
Task4 End-to-end			
Query Information			
LMM Answer: Sep 12		Answer Annotation: 07-17	

Figure 2: Example Input, LMM Output, and Ground Truth for Four Evaluation Tasks. The color coding of each module corresponds to Fig. 4 in the main paper. Task1 Requery (green), Task2 Rerank (purple), Task3 Summarization (blue), and Task4 End-to-end (yellow) are shown. Image best viewed in color.

D.2 SUBFIELD DEFINITION

News area encompasses a vast spectrum of information, ranging from everyday events to engaging entertainment content and specialized fields such as scientific discoveries and financial analysis. This comprehensive coverage serves as a rigorous assessment of the model’s ability to process information in diverse domains. We divide this expansive area into eight distinct subfields:

- **Traditional Sports:** Data concerning traditional athletic competitions, team performances, player statistics, and sporting events. This includes scores, league standings, player transfers, and analysis of various professional sports across different leagues and countries.
- **e-Sports:** Information about competitive video gaming, including tournament results, player rankings, and league information. This covers various game titles, team formations, streaming viewership statistics, and tournament information.
- **Technology:** Information about technological innovations, gadgets, software developments, and tech industry news. This includes product launches, software updates, cybersecurity issues, and artificial intelligence advancements.
- **Paper:** Content related to academic papers, research publications, and scholarly articles in various artificial intelligence fields. The queries include method explanation, figure understanding, and experiment settings.
- **Entertainment:** Data about movies, music, television, celebrities, and other forms of popular entertainment. It also includes data concerning video games.
- **Finance:** Information on financial markets, economic indicators, business news, and monetary policies. This covers stock prices, company earnings reports, company financial statements, and regulatory news regarding finance.
- **General News:** Broad coverage of various news topics not specific to any particular subfield. This includes a mix of local and global events, human interest stories, lifestyle articles, climate news, and general interest content that doesn’t fit neatly into other specialized news subfields.
- **False Premise:** Data related to misinformation or incorrect assumptions in the query. This subfield focuses on fact-checking capabilities. All the answers to the queries of this subfield are ‘invalid question’.

Knowledge area represents broad subfields of information and data related to general knowledge across various disciplines. This area concentrates on rare knowledge that most LMMs fail to answer. We categorize this area into five subfields:

- **Architecture:** Information about building design, architectural styles, building information, and construction projects. This includes city landmarks, the comparison of architectural styles, and multi-view architecture matchings.
- **Arts:** Data concerning visual arts, drawings, sculptures, badges, and other forms of creative expression. This covers artwork details, artist profiles, artwork history, and artwork style comparisons.
- **Fashion:** Content related to clothing trends, fashion brands, and designer collections. This includes retail price, clothing style, release date, and brand information.
- **Astronomy:** Information about celestial objects, space exploration, astronomical phenomena, and related research. This covers observational data from telescopes and image results from space missions. The questions focus on the background information of these celestial objects presented in the query image.
- **Anime:** Data about Japanese animation, including series storylines and character information. This encompasses character background, character appearance, voice actor information, and chapter information.
- **Auto:** Content related to automobiles, including vehicle specifications, industry trends, and automotive technology. This covers new car models, performance test results, coefficients of cars, and release date.

E QUALITATIVE EXAMPLES

Question: In the LPL 2024 Summer season, how many teams were in Group Ascend?

Answer: 9

Area: News

Subfield: ⚽ Sports

Round1 Requery

GPT-4o: LPL 2024 Summer season Group Ascend teams count

Qwen2-VL: LPL 2024 Summer season Group Ascend teams

LLaVA-OneVision: LPL 2024 Summer season Group Ascend teams

Round2 Rerank

GPT-4o Brief Results:



<Website 1>

Title: LPL 2024 Summer - Leaguepedia | League of Legends Esports Wiki - Fandom
Snippet: The 2024 League of Legends Pro League (LPL) is the second split of the twelfth year of China's professional League of Legends league. The summer season sees major changes of LPL structure, including introduction of "Fearless Draft" rules and return of group stages. Seventeen teams play against each other first in four double round robins, and then in two single round robins.



<Website 5>

Title: Group Stage / LPL 2024 Summer - schedule, results | u2014 Escorenews
Snippet: LPL 2024 Summer Match results, calendar, VODs, stream, team rosters, schedule ... VODs, stream, team rosters, schedule. LoL News Bets and predictions Matches Events Teams Players. LPL 2024 Summer LoL - u2022 2024-06-01 - 0000-00-00 u2022 \$578600. Stats. Event Placements Qualifier Group Stage Playoff Regional Finals. Group Stage Playoff - Phase 3 ...



<Website 2>

Title: LPL Summer 2024 - Liqupedia League of Legends Wiki
Snippet: The LPL Summer 2024 split is the second split of the 2024 LPL season. The league maintains all 17 teams from the Spring Split, and will be held across China, in cities such as Shanghai, Suzhou, Shenzhen, Xi'an and Beijing. Bilibili Gaming is the defending title champion. This split, the LPL will experiment with a different format, featuring a ...



<Website 6>

Title: 2024 LPL season - Wikipedia
Snippet: The 2024 LPL season is the 12th and ongoing season of the League of Legends Pro League ... The bottom two ascend group teams and the top four nirvana group teams will have to contest an additional match. [6] As per usual, the champion for Summer 2024 will qualify for the 2024 World Championship as China's number one seed. Spring. Regular Season



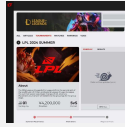
<Website 3>

Title: LPL 2024 Summer Placements - Leaguepedia | League of Legends Esports Wiki
Snippet: The 2024 League of Legends Pro League (LPL) is the second split of the twelfth year of China's professional League of Legends league. The summer season sees major changes of LPL structure, including introduction of "Fearless Draft" rules and return of group stages. Seventeen teams play against each other first in four double round robins, and then in two single round robins.



<Website 7>

Title: JD Gaming vs. EDward Gaming / LPL 2024 Summer Placements - Reddit
Snippet: JD Gaming vs. EDward Gaming / LPL 2024 Summer Placements - Week 4 - Group A / Post-Match Discussion LPL 2024 SUMMER ... With this win by JDG, FPX advance to Group Ascend alongside JDG for the LPL 2024 Summer Season. JDG | Leaguepedia | Liqupedia | Website | Twitter EDG | Leaguepedia | Liqupedia ... Team WE vs. Bilibili Gaming / LPL 2024 ...



<Website 4>

Title: LPL 2024 Summer LoL Coverage | GosuGamers
Snippet: The 2024 League of Legends Pro League (LPL) is the second split of the twelfth year of China's professional League of Legends league. The summer season sees major changes of LPL structure. Placements tournament will determine the groupings for the group stage. Jun 2024. 01.



<Website 8>

Title: League of Legends LPL 2024 Summer Split - Sportskeeda
Snippet: The LPL 2024 Summer Split format is vastly different, and a big change for a major region. The first stage of the Summer Split will be the Placements stage. Teams are divided into four groups with ...

GPT-4o Rerank: <Website 6>

Figure 3: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593

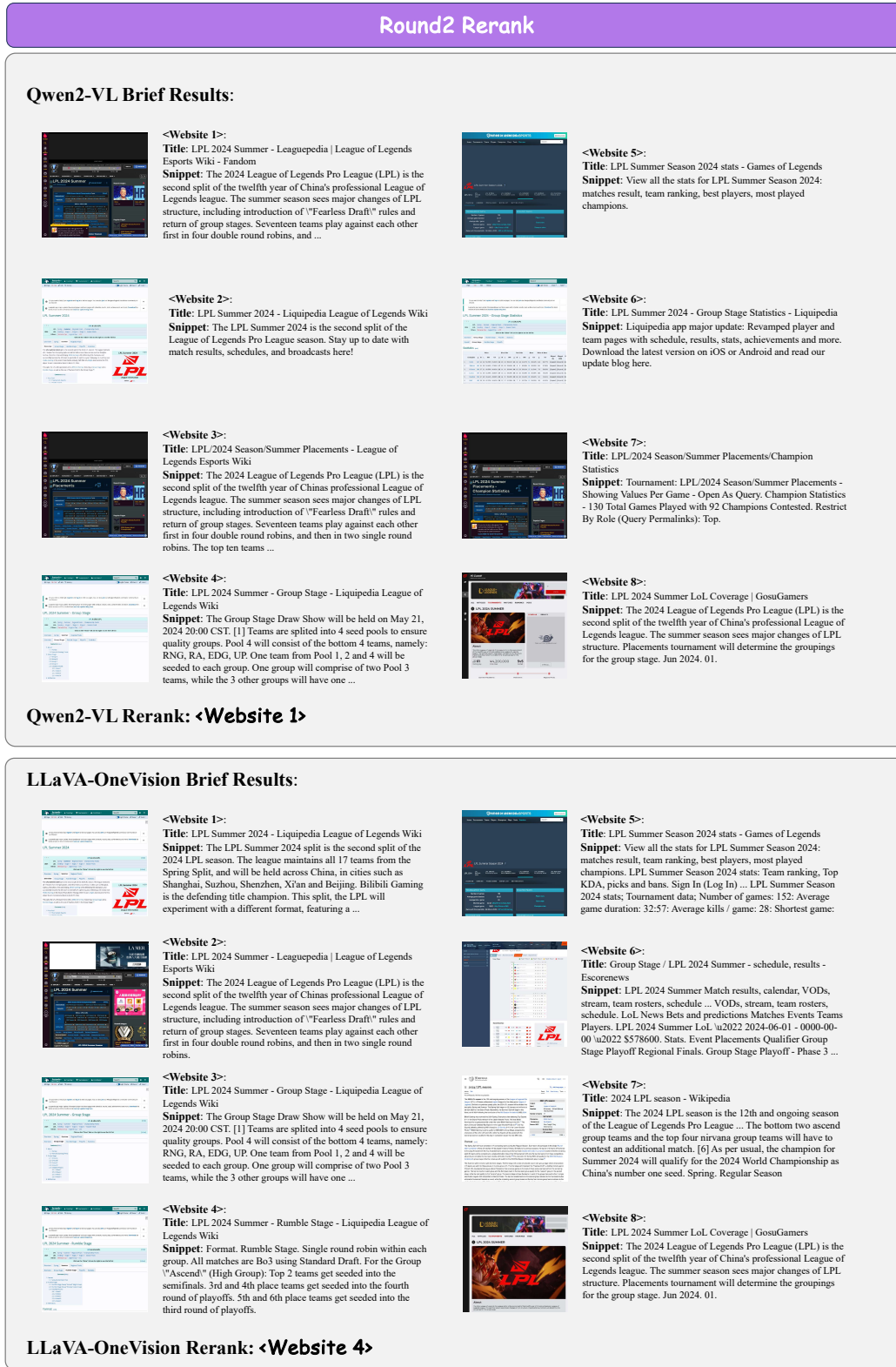


Figure 4: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647



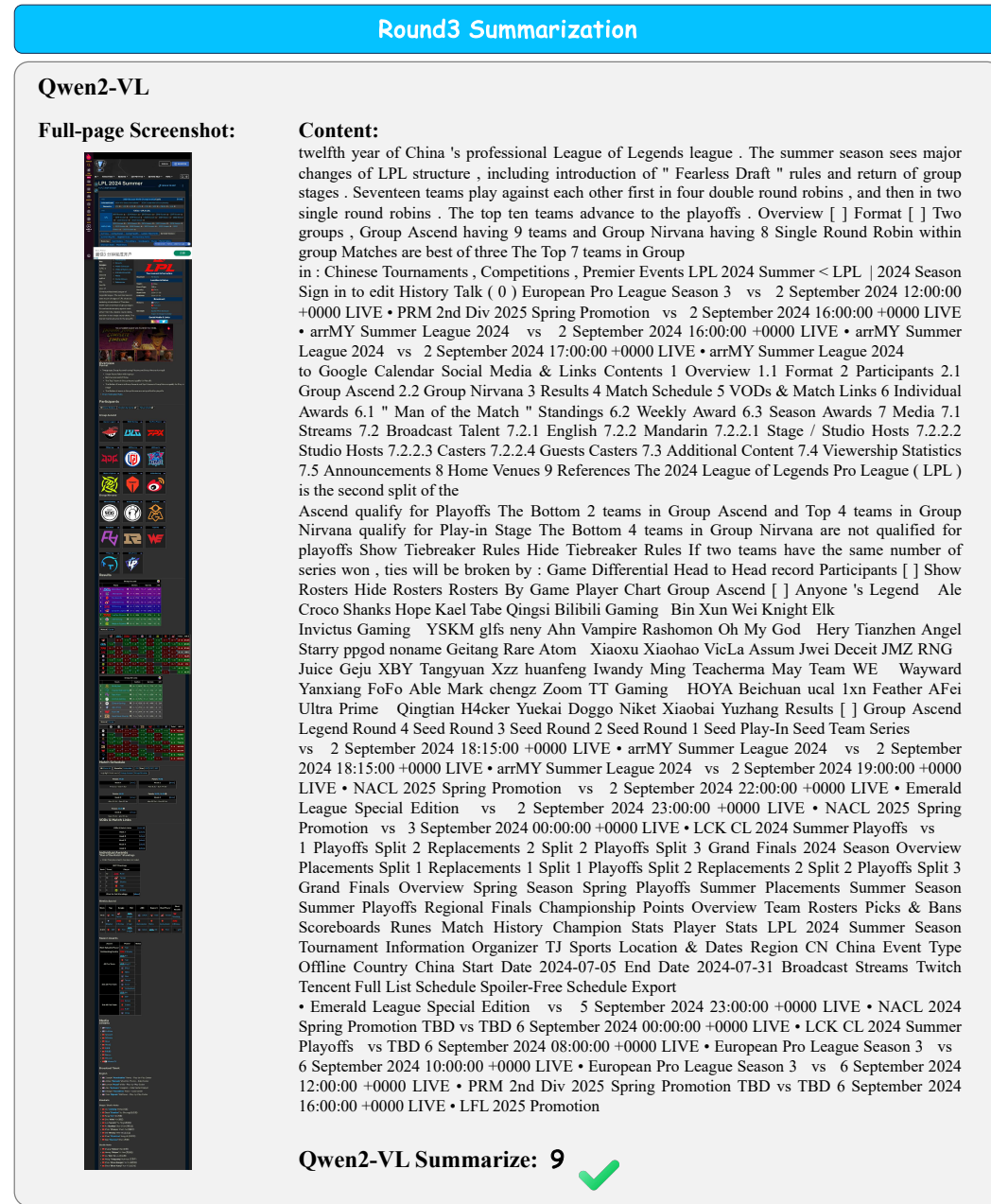


Figure 6: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.



Figure 7: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

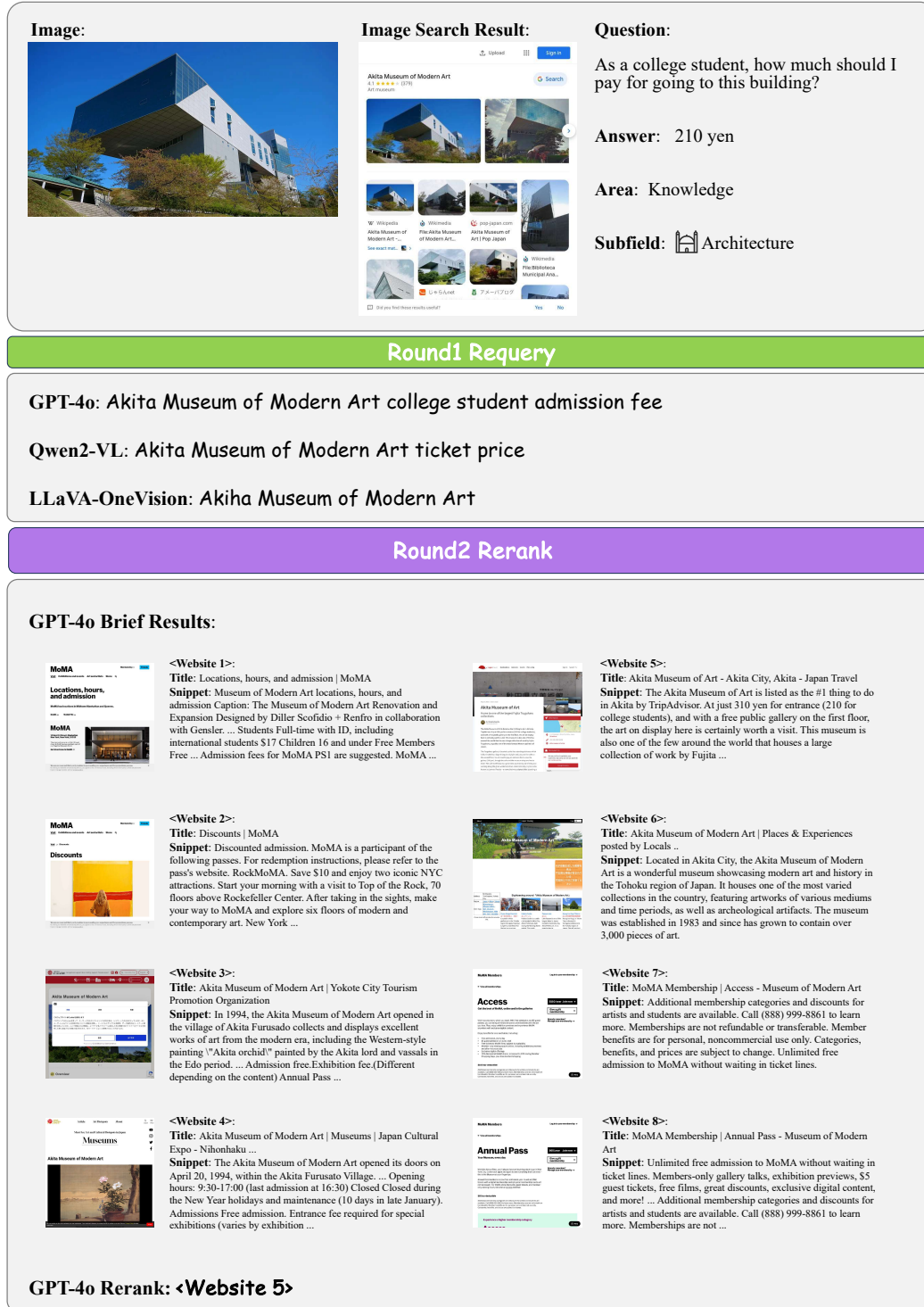


Figure 8: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

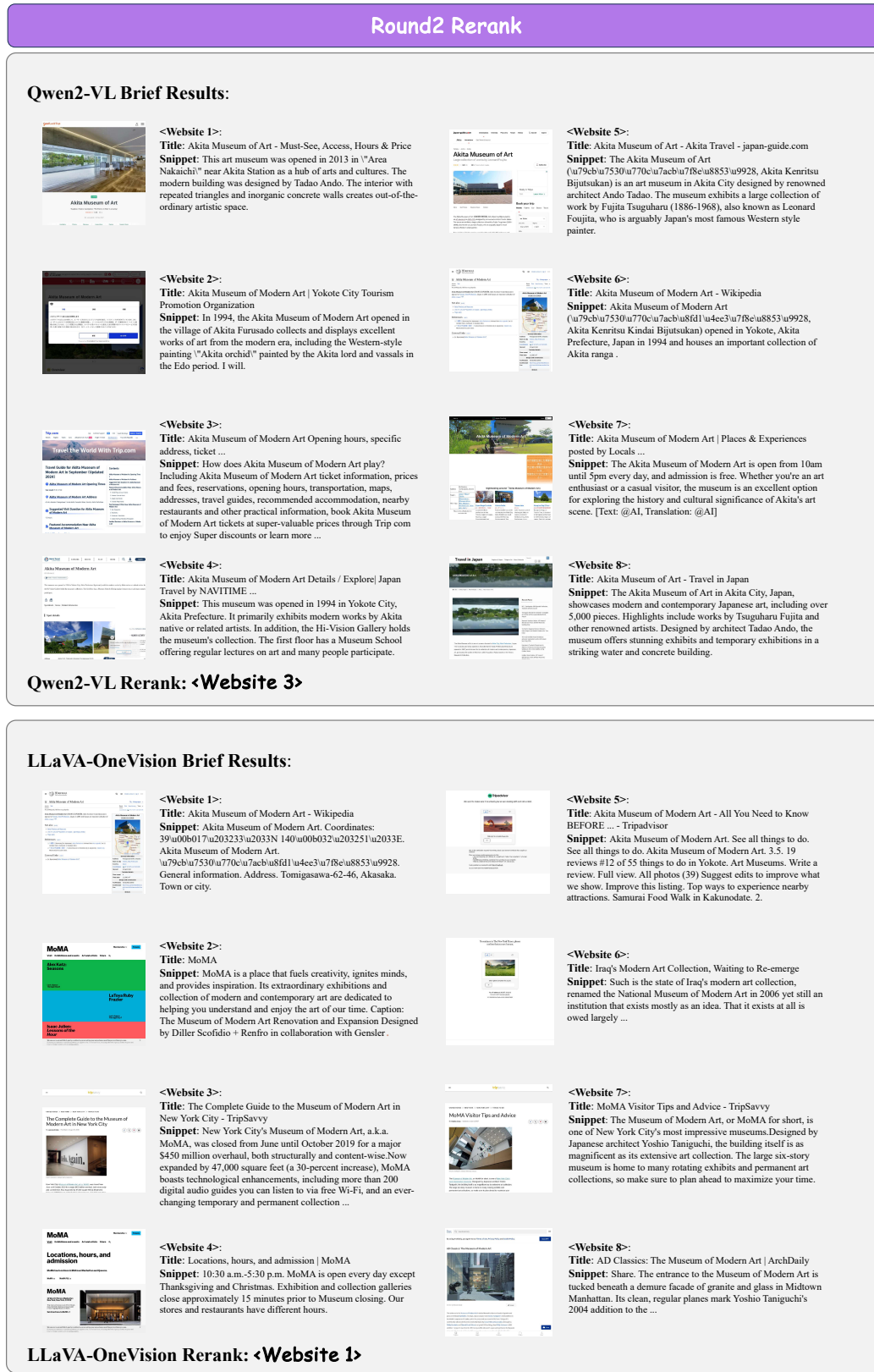
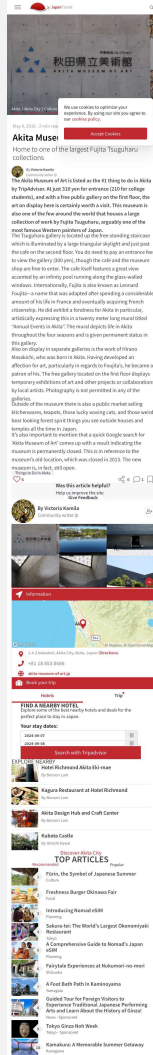


Figure 9: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

Round3 Summarization

GPT-4o

Full-page Screenshot:



Content:

Akita Akita City Culture Akita Akita City Culture May 9, 2016 - 2 min read Akita Museum of Art Home to one of the largest Fujita Tsuguharu collections By Victoria Kamila Community writer The Akita Museum of Art is listed as the # 1 thing to do in Akita by TripAdvisor . At just 310 yen for entrance (210 for college students) , and with a free public gallery on the first floor , the art on display here is certainly worth a visit . This museum is also one of the few around the world that Art ' comes up with a result indicating the museum is permanently closed . This is in reference to the museum ' s old location , which was closed in 2013 . The new museum is , in fact , still open . Things to Do in Akita 6 6 Share on Facebook Share on X (Twitter) Copy link to share 1 Was this article helpful ? Help us improve the site Give Feedback By Victoria Kamila Community writer Following Follow +6 Information 1-4-2 Nakadori , Akita City , Akita , Japan (Directions) +81 18 853 8686 akita-museum-of-art.jp Book your trip Hotels Trip Find a nearby hotel Explore some of the best nearby hotels and deals for the perfect place to stay in Japan . Your stay dates : Search with Tripadvisor Start your trip now When do you want to travel ? My dates are flexible Get started Explore nearby Hotel Richmond Akita Eki-mae By Bonson Lam Kagura Restaurant at Hotel Richmond By Bonson Lam Akita Design Hub and Craft Center By Bonson Lam Kubota Castle By Hitoshi Kawai Discover Akita City Top Articles Recommended Popular 1 Freshness Burger Okinawa Fair Food 2 Guided Tour Foujita ' s , he became a patron of his . The free gallery located on the first floor displays temporary exhibitions of art and other projects or collaborations by local artists . Photography is not permitted in any of the galleries . Outside of the museum there is also a public market selling kitchenwares , teapots , those lucky waving cats , and those weird bear looking forest spirit things you see outside houses and temples all the time in Japan . It ' s also important to mention that a quick Google search for ' Akita Museum of known as Leonard Foujita—a name that was adapted after spending a considerable amount of his life in France and eventually acquiring French citizenship . He did exhibit a fondness for Akita in particular , artistically expressing this in a twenty meter long mural titled “ Annual Events in Akita ” . The mural depicts life in Akita throughout the four seasons and is given permanent status in this gallery . Also on display in separate galleries is the work of Hirano Masakichi , who was born in Akita . Having developed an affection for art , particularly in regards to claim about Google saying it is closed . Wikipedia , Google maps , and the official website all say it is open (in English and in Japanese) . Perhaps there a few outdated websites that list it as closed . I used to go to the old one when I was in Akita 2011-2012 . The new one with poolside cafe looks nice . Reply Show all 0 replies 1 comment in total houses a large collection of work by Fujita Tsuguharu , arguably one of the most famous Western painters of Japan . The Tsuguharu gallery is located up the free-standing staircase which is illuminated by a large triangular skylight and just past the cafe on the second floor . You do need to pay an entrance fee to view the gallery (300 yen) , though the cafe and the museum shop are free to enter . The cafe itself features a great view accented by an infinity pool running along the glass-walled windows . Internationally , Fujita is also for Foreign Visitors to Experience Traditional Japanese Performing Arts and Learn About the History of Ginza ! News - Sponsored 3 Introducing Nomad eSIM Planning 4 Sakura-tei : The World ' s Largest Okonomiyaki Restaurant Tokyo 5 Kamakura : A Memorable Summer Getaway Kanagawa 6 A Foot Bath Path In Kaminoyama Yamagata 7 Fairytale Experiences at Nukumori-no-mori Shizuoka 8 A Comprehensive Guide to Nomad ' s Japan eSIM Planning 9 Tokyo Ginza Noh Week Tokyo - Sponsored 10 Fürin , the Symbol of Japanese Summer Culture 1 A Guide to Japanese Visas Planning 2 Guide to Bringing Medicines Into Japan Planning

GPT-4o Summarize: 210 yen ✓

Figure 10: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

Round3 Summarization

Qwen2-VL

Full-page Screenshot:



Content:

Reviews US \$ 66.00 View Top Restaurant Picks Near Akita Museum of Modern Art 1 . Bar Pasaporte Address : 204-1 Aza Takuboshita Fuke Otsutsumi Distance : 445m Bar Pasaporte No reviews yet Other Cuisine View 2 . Kuidoraku Price : \$ 8.00 Address : 7-2 Ekimaecho, Yokote, Akita 013-0036 Distance : 2.28km Kuidoraku No reviews yet Bars/Bistros US \$ 8.00 View 3 . Korakuen Yokoteten Price : \$ 5.00 Address : It is 28-1, Sanmaibashi in Maego, Yokote-shi, Akita character Distance : 2.03km Korakuen Yokoteten No reviews yet Other Chinese Cuisine US \$ 5.00 View 4 . Ganso Kamiya Yakisoba Restaurant Address : Nakano-117-67 Oyashinmachi, Yokote, Akita 013-0051, Japan Distance : 1.17km Ganso Kamiya Yakisoba Restaurant No reviews yet Fast Food View Verified Reviews of Akita Museum of Modern Art 第二号爱人 : 遇上秋田杆灯季, 还是挺热闹的一个节日在美术馆里面的话, 也有这种节日的气氛, 氛围都还是相当不错的, 而且的话里面虽然没有特别多的名画名品, 但是因为 是免费进入, 所以值得参观。Jedy Tan : 属于小众景点了, 秋田县本来就不大, 美术馆还在一个小 市里。但参观过能看出日本人近现代对西洋艺术的崇尚 xiaomoufa : 哈哈, 这是一个非常美丽的美术馆, 特别有意思的一个 景点。E30 * * * 67 : 蛮大的美术馆, 里面画有些我们国家古代的味道, 蛮不错 的 Also Popular With Visitors to Akita Museum of Modern Art 1 . Sendai Umino-Mori Aquarium Price : \$ 14.30 Discount : \$ 2.04 Recommended sightseeing time : : 4-5 hours Address : Japan , 〒983-0013 and free . everything related to the ninjas were very fun . Edo Wonderland Nikko Edomura 4.5 / 5 43 Reviews No.3 of Best Things to Do in Nikko Theme Parks From US \$ 37.43 View Contents Akita Museum of Modern Art Opening Times Akita Museum of Modern Art Address Suggested Visit Duration for Akita Museum of Modern Art Featured Accommodation Near Akita Museum of Modern Art 1 . Hotel Plaza Annex Yokote 2 . Yokote Central Hotel 3 . Quad Inn Yokote 4 . Yokote Plaza Hotel Top Restaurant Picks Near Akita Museum of Modern Art 1 . Bar Pasaporte Hotels Flights Trains Cars Car Rentals Airport Transfers App Customer Support USD Search Bookings Sign in / Register Travel with Trip.com Travel Guide for Akita Museum of Modern Art in September (Updated 2024) Akita Museum of Modern Art Opening Times Year round : 9:30-17:00 Akita Museum of Modern Art Address 62-46, Akasaka Tomigazawa | Inside Akita Furusato Mura, Yokote, Akita Prefecture Suggested Visit Duration for Akita Museum of Modern Art 1-2 hours Featured Accommodation Near Akita Museum of Modern Art 1 . Hotel Plaza Annex Yokote Address 2 . Kuidoraku 3 . Korakuen Yokoteten 4 . Ganso Kamiya Yakisoba Restaurant Verified Reviews of Akita Museum of Modern Art Also Popular With Visitors to Akita Museum of Modern Art 1 . Sendai Umino-Mori Aquarium 2 . Tsugaru-han Neputa mura Village 3 . Suntopia World 4 . Edo Wonderland Nikko Edomura Contents Akita Museum of Modern Art Opening Times Akita Museum of Modern Art Address Suggested Visit Duration for Akita Museum of Modern Art Featured Accommodation Near Akita Museum of Modern Art 1 . Hotel Plaza Annex Yokote 2 . Yokote Central Hotel 3 . Quad Inn Yokote 4 . Yokote Plaza Hotel Top Restaurant Picks Near Akita Museum of Modern Art 1 . Bar Pasaporte 2 . Kuidoraku 3 . Korakuen Yokoteten 4 . Ganso Kamiya Yakisoba Restaurant Verified Reviews of Akita Museum of Modern Art Also Popular With Visitors to Akita Museum of Modern Art 1 . Sendai Umino-Mori Aquarium 2 . Tsugaru-han Neputa mura Village 3 . Suntopia World 4 . Edo Wonderland Nikko Edomura Popular Travelogues Bangkok Travelogue | Manila Travelogue | Tokyo Travelogue | Taipei Travelogue | Hong Kong Travelogue | Seoul Travelogue | Kuala Lumpur Travelogue | Los Angeles Travelogue | Shanghai Travelogue \$ 14.30 View 2 . Tsugaru-han Neputa mura Village Price : \$ 3.38 Address : Japan , 〒036-8332 Aomori, Hirosaki, Kamenokomachi, 6 1 Distance : 0.63 mi Tsugaru-han Neputa mura Village No reviews yet From US \$ 3.38 View 3 . Suntopia World Price : \$ 8.85 Recommended sightseeing time : : 0.5-1 day Address : 1-1 Kubo, Agano, Niigata 959-2212, Japan Distance : 5.06 mi What travelers say : M515shunyi1618 : Very suitable for family outings of playground, Ferris wheel, pirate ship are fun . Suntopia World 5 / 5 9 Reviews Amusement Ekimaecho Price : \$ 53.00 Distance : 2.23km Hotel Plaza Annex Yokote 4.3 / 5 37 Reviews -5 % US \$ 53.00 View 2 . Yokote Central Hotel Address : Heiwacho 9-10 Price : \$ 50.00 Distance : 2.91km Yokote Central Hotel 3.7 / 5 18 Reviews -5 % US \$ 50.00 View 3 . Quad Inn Yokote Address : Sekibata-52-1 Yasuda Price : \$ 58.00 Distance : 1.9km Quad Inn Yokote 4.1 / 5 14 Reviews US \$ 58.00 View 4 . Yokote Plaza Hotel Address : 7-1 Ekimaecho Price : \$ 66.00 Distance : 2.27km Yokote Plaza Hotel 3.9 / 5 10

Qwen2-VL Summarize: US \$ 66.00



Figure 11: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

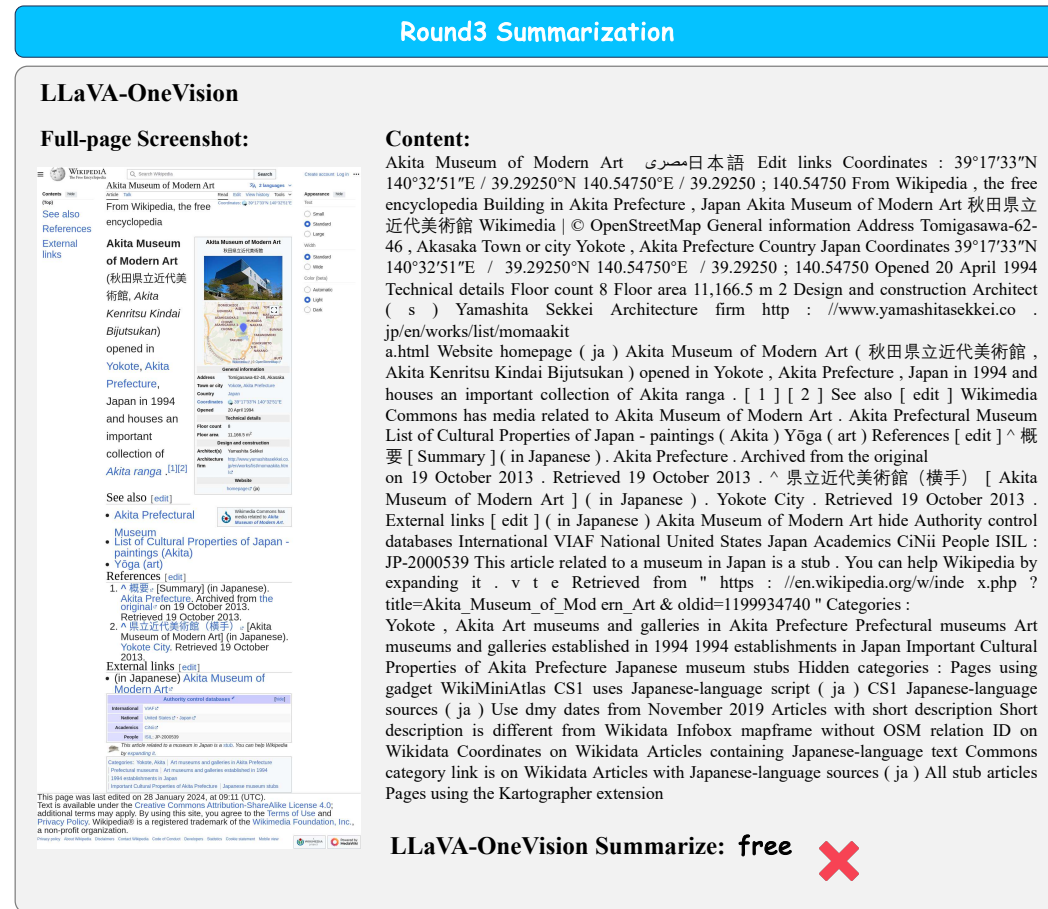


Figure 12: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

REFERENCES

- Anthropic. Claude-3.5. <https://www.anthropic.com/news/claude-3-5-sonnet>, 2024.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.
- Patrice Béchard and Orlando Marquez Ayala. Reducing hallucination in structured outputs via retrieval-augmented generation. *arXiv preprint arXiv:2404.08189*, 2024.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.
- Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*, 2024.
- Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023a.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024a.
- Jun Chen, Deyao Zhu¹ Xiaoqian Shen¹ Xiang Li, Zechun Liu² Pengchuan Zhang, Raghuraman Krishnamoorthi² Vikas Chandra² Yunyang Xiong, and Mohamed Elhoseiny. Minigt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023b.
- Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. Mindsearch: Mimicking human minds elicits deep ai searcher. *arXiv preprint arXiv:2407.20183*, 2024b.
- 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.
- Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.
- Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6491–6501, 2024.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- Chaoyou Fu, Yuhang Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

- Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *ICML 2024*, 2024.
- Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.
- Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Chengzhuo Tong, Peng Gao, Chunyuan Li, and Pheng-Ann Heng. Sam2point: Segment any 3d as videos in zero-shot and promptable manners. *arXiv preprint arXiv:2408.16768*, 2024.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.
- Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3170–3180, 2021.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.
- Dongfu Jiang, Xuan He, Huaye Zeng, Con Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024a.
- Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *arXiv preprint arXiv:2404.03653*, 2024b.
- Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988.
- Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions., 2024.
- 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*, 2024a.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.
- KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.
- Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*, 2024c.
- Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *ECCV 2024*, 2023.

- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a.
- 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 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- 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*, 2023b.
- 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.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv*, abs/2310.02255, 2023.
- OpenAI. GPT-4V(ision) system card, 2023. URL <https://openai.com/research/gpt-4v-system-card>.
- OpenAI. Openai o1. [Online], 2024a. <https://openai.com/index/learning-to-reason-with-llms/>.
- OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024b.
- Qwen Team. Qwen2-vl. 2024.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL <https://api.semanticscholar.org/CorpusID:231591445>.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.
- Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023.
- Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, et al. Crag—comprehensive rag benchmark. *arXiv preprint arXiv:2406.04744*, 2024.
- Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023a.
- Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023b.
- 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*, abs/2308.02490, 2023.
- Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.
- Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024a.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024b. URL <https://openreview.net/forum?id=d4UiXAHN2W>.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *ECCV 2024*, 2024c.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024.