Context-aware chatbot using MLLMs for Cultural Heritage

Pavan Kartheek Rachabatuni
Università degli Studi di Firenze
Italy
pavankartheek.rachabathuni@unifi.it

Filippo Principi
Università degli Studi di Firenze
Italy
filippo.principi@unifi.it

Paolo Mazzanti
Università degli Studi di Firenze
Italy
paolo.mazzanti@unifi.it

Marco Bertini
Università degli Studi di Firenze
Italy
marco.bertini@unifi.it

ABSTRACT

Multi-modal Large Language Models (MLLMs) are currently an extremely active research topic for the multimedia and computer vision communities, and show a significant impact in visual analysis and text generation tasks. MLLM's are well-versed in integrated understanding, analysis of complex data from cross modalities (i.e. text-image) and text generation with chat abilities. Almost all MLLM's, focus on alignment of image features to textual features for downstream text generation tasks includes detailed image description, visual question answering, stories and poems generation, phrase grounding, etc.. However, when focusing on visual question answering, questions that are highly relevant to the context of an image may not be answered correctly with the existing MLLM's, contrary to questions that are related to visual aspects. Moreover, generating meta data (context) for an image using present day MLLM's is hard task due to hallucinating characteristic of underlying Large Language Models (LLM's), and adequate contextual information cannot be directly derived from an image based perspective.

Considering the cultural heritage domain, these issues hamper the introduction of multimedia chatbots as tools to support learning and understanding artworks, since contextual information is typically needed to better understand the content of the artworks themselves, and museum curators require that scientifically accurate information is provided to the users of such systems. In this paper we present a system that combines contextual description of the artworks to enhance the contextual visual question answering task.

CCS CONCEPTS

• **Computing methodologies** → Computer vision; Natural language processing; • **Applied computing** → Arts and humanities.

MMSys '24, April 15–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0412-3/24/04
https://doi.org/10.1145/3625468.3652193

KEYWORDS

Visual Question Answering, Chatbot, Cultural Heritage, Museums, Digital Learning

ACM Reference Format:
Pavan Kartheek Rachabatuni, Filippo Principi, Paolo Mazzanti, and Marco Bertini. 2024. Context-aware chatbot using MLLMs for Cultural Heritage. In *ACM Multimedia Systems Conference 2024 (MMSys '24), April 15–18, 2024, Bari, Italy.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3625468.3652193

1 INTRODUCTION AND PREVIOUS WORK

Artificial intelligence, notably in the form of personal assistants and chatbots, has become increasingly sophisticated and prevalent in specific task domains. At the core of this evolution are Multimodal Language-and-Image Models (MLLMs), which have significantly advanced the way images are perceived and understood along with the text. These models analyze pixels, colors, object recognition, and the spatial relationships between objects, demonstrating adeptness at interpreting and generating detailed information about the content and, to some extent, also to the context of images. They go beyond mere identification by describing scenes, identifying events, and even inferring the mood or atmosphere.

This advancement is particularly evident in the realms of Computer Vision and Natural Language Processing (NLP), where tasks such as Visual Question Answering [1] (VQA) and caption generation test the ability of technology to understand and articulate visual content. These tasks demand not just recognition but a nuanced interpretation of images, setting the stage for a deeper discussion of current capabilities. Such capabilities are central to the challenges posed by VQA and caption generation, as they require a comprehensive understanding of both the visual elements and the narrative they convey.

In spite of these advancements, a critical limitation has been observed: while MLLMs excel at interpreting the visual surface and immediate context of images, they often lack a deep understanding of historical and cultural significance of images; in addition, these models suffer from hallucination due to over generalization characteristic by underlying language model. The contextual (historical) content or meta information of the image cannot be decoded form the image itself. This limitation becomes starkly evident when these models encounter questions or tasks that require an understanding of historical or contextual knowledge extending beyond the visible.

In the cultural heritage domain, this issue is particularly pronounced. Despite notable progress in VQA, these models primarily

focus on visual content, often providing limited insights into the cultural assets they depict. Their proficiency in interpreting visual elements does not extend to the intricate historical and cultural contexts that are vital for a comprehensive understanding of cultural heritage.

These tasks typically involve translating visual content into relevant language representations. However, the cultural heritage domain demands a deeper approach. It necessitates a thorough understanding of cultural assets within their historical contexts. This understanding is crucial for downstream applications such as context-aware visual question answering and contextual caption generation, where a mere surface-level interpretation falls short of providing meaningful insights.

In essence, VQA and caption generation must be enhanced with contextual knowledge, particularly historical context, to meet the nuanced requirements of the cultural heritage domain and, considering the educational role [6] of museums, must provide scientifically accurate information regarding the artworks. This holistic approach ensures a more comprehensive understanding of cultural assets. Thus, the development of intelligent systems for knowledge-dependent tasks in the field of cultural heritage is both significant and necessary in today's context.

Models like LLaVA [10] and MiniGPT-4 [13] utilize frozen pre-trained visual encoders (specifically, CLIP VIT-L/14 and EVA-CLIP VIT-G/14 [7]) to extract visual features, which are then mapped into the input embedding space of a frozen language model. The alignment of visual and language features is accomplished through MLP/linear layers, enabling these models to perform exceptionally well in downstream vision-language tasks. However, it's essential to clarify that the use of shallow alignment methods leads to hallucination.

In contrast, CogVLM [12] stands out by incorporating a trainable visual expert module. While it also freezes visual (EVA2-CLIP-E) and language (vicuna-7b-1.5) encoders, it introduces a visual expert into each language model layer. This innovative architecture facilitates deep alignment between vision and language features, enhancing accuracy in aligning visual features with a static understanding of language. Consequently, the model excels in focusing on visual content rather than being overly concerned with image context during text generation.

OtterHD-8B [8] builds on the Fuyu-8B[3] model and distinguishes itself as a decoder-only model, eliminating the need for a fixed-resolution vision encoder. This unique design enables OtterHD-8B to process images of various resolutions, even up to 1024x1024, while directly incorporating pixel-level information into the language decoder. Empirical evaluations reveal state-of-the-art performance, particularly when instruction-tuned for processing higher-resolution images. However, it primarily prioritizes intricate image details for text generation, sometimes at the expense of considering the broader image context.

In this work we present a context-aware chatbot system based on a MLLM that has been trained specifically for the cultural heritage domain. The chatbot is based on an extension of LLaVa with GLIP, and uses Retrieval Augmented Generation (RAG) to reduce hallucinations in the answer, so to produce scientifically accurate answers using curated artwork information, providing broader historical and cultural context. The system is open source and, in addition, a novel dataset for training new multimedia chatbots is also provided[1].

2 THE SYSTEM

2.1 Design Considerations

The use of AI technologies is emerging as a topical issue in the current scenario regarding the latest trends on digital innovation, inclusion, and participation in museums and Cultural Heritage. Thematic reports [2], dedicated Networks[2] and recent European projects[3] focus on the opportunities and challenges of using AI in different application areas: archiving and cataloguing, museum management of visitor information and audience engagement activities. Among all these use cases, museums primarily employ AI technology to re-imagine and re-interpret collections, engage diverse audiences, personalize visitor experiences, and enhance user interactions [5]. A changing museum vision fosters this novelty, a new mindset focuses on people and visitors, their interactions with museum collections, and the tools used to enhance the learning experience also using AI-Gen and chatbots. Digital learning, playful approach, user-generated content, and participatory storytelling are relevant topics in the museum debate and all related to the use of AI-based tools to increase visitor engagement.

The use of AI, specifically in Natural Language Understanding (NLU) and Natural Language Processing (NLP), alongside Chatbot-powered educational tools in museums, extends audience reach beyond traditional methods. These AI tools enhance the visitor experience by creating a welcoming and non-intimidating space for questions, fostering trust and interaction, particularly among young people, families, and children. Chatbots motivate visitors to explore exhibits more deeply, encouraging active participation, especially for non-expert visitors. The automated and personalized dialogue improves visitor engagement, contributing to a more comprehensive understanding of audience preferences [11]. As part of the European H2020 ReInHerit project, an open-source toolkit with AI-based applications has been developed, including a multimedia chatbot[4]. This web-based application answers questions about artwork visual content or context, aiming to overcome limitations of existing Visual Question Answering (VQA) approaches. The design is influenced by the popularity of chat-based interaction, exemplified by ChatGPT. Additionally, a chatbot engine using a GPT-based neural network has been added to the backend, following the success of such architectures and training methods.

Based on project recommendations, designing digital learning tools for user engagement in museums must carefully consider the ethical implications of AI usage. Developing these applications requires training on large datasets, highlighting the importance of ethically sourced and unbiased training data. Recommendations for museum chatbots like ChatGPT prioritize ethical conduct, transparency, and regulatory compliance, addressing concerns about privacy, training data, and result accuracy. Obtaining user consent

[1] https://github.com/miccunifi
[2] https://themuseumsai.network
[3] https://reinherit-hub.eu/tools/apps
[4] https://reinherit-hub.eu/tools/apps/

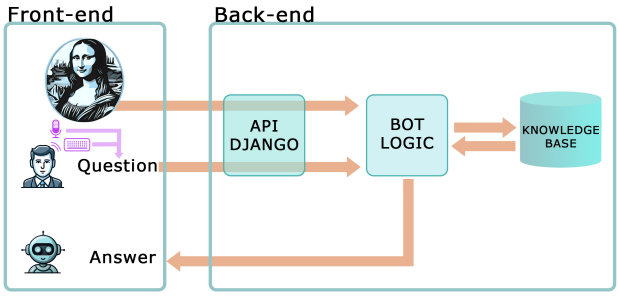

Figure 1: Overview of the chatbot system architecture.

and ensuring transparent, secure data storage is crucial before collecting personal data. To prevent errors and ensure quality content, user questions should be guided by specific instructions, compelling the chatbot to draw from curated knowledge provided by museum experts. This approach minimizes errors by relying on high-quality content from experts through the use of well-crafted prompts directing responses to the collected and validated dataset.

To improve chatbot systems, it's crucial to continuously refine their ability to handle uncertainties, enhance natural language understanding, and adapt to user preferences. This ensures a seamless, personalized experience. Optimizing responses, managing non-response scenarios, and refining dialogue flow enables a more human-like communication style in museums. Collecting unanswered questions guides curators in enhancing descriptive content based on user interests and curiosities. Diversifying datasets by integrating reliable quality content about collections in addition to Wikipedia or open data ensures the capability to correctly answer a wide range of visitor questions regarding artworks and their broader contextual references.

2.2 System Architecture

The application's frontend is constructed using Javascript, featuring a responsive design adaptable to both desktop web browsers and mobile applications, thereby delivering an innovative smart guide. To facilitate user interaction, particularly in a mobile context, speech recognition is integrated, eliminating the need for users to input extensive queries via the device's compact keyboard. Examples for the interface are shown in Fig. 4.

The back-end is developed in Python, using the Django web framework to provide the REST API to the front-end. Additionally, a chat bot engine has been seamlessly incorporated into the back-end, employing a neural model like GPT. This system enables the generation of more elaborate responses and a nuanced understanding of contextual information related to cultural assets, departing from the conventional Visual Question Answering (VQA) approach. The result is a more natural and interactive user experience.

2.3 The Dataset

In the domain of cultural heritage, the development of contextual visual question answering (VQA) models faces a significant challenge due to the lack of suitable datasets. A notable dataset, named VISCOUNTH [4], addresses this gap by providing approximately 500K images of Italian cultural assets, including paintings, statues, finds, prints, and churches, along with 6.5M question-answer pairs. This dataset emphasizes the reasoning of artwork images through associated knowledge to answer complex questions about the artworks. The questions are divided into three categories – visual, contextual, and mixed – and are spread across 43 question types. However, these questions appear to be based on predefined templates applicable to various cultural assets, and the answers often repeat words from the questions, limiting the model's ability to learn the context of the cultural assets. A more effective approach would be to generalize the answers with contextually associated knowledge, enabling the model to provide contextually accurate responses to complex questions about cultural heritage.

CHVQA Dataset Statistics

Number of Images	2890
Number of QA pairs	54986
Average question length (words)	16.43
Average answer length (words)	33.17
Average number of questions per image	19

Table 1: CHVA Dataset statistics in brief.

To address these limitations, we have introduced a new compact dataset CHVQA (Cultural Heritage Visual Question Answering) to facilitate the study of Visual Question Answering in the domain of cultural heritage. This dataset comprises approximately 55,000 question-answer pairs and features 2,890 cultural asset images (dataset statistics are shown in Tab. 1). Leveraging the capabilities of GPT-4, we have been able to generate question-answer pairs that are specifically tailored to historical content. These pairs are generated using by specific prompt (see Fig. 2) that is common to all the images, ensuring consistency across the dataset. Additionally, each image is accompanied by historical context information.

We sourced both the images and content from Wikipedia using the Wikipedia-API. This approach allowed us to curate a rich and diverse set of cultural heritage assets and their corresponding information.

2.4 The Chatbot System.

In this proposed system, shown in Fig. 3, is a 2 step model. Aa a first step, the workflow begins with processing the image through a Grounded Language Image Pre-training (GLIP) [9] and Visual encoder.

The pre-trained GLIP model identifies and categorizes various elements within an image, such as people and objects, by generating object regions and labels. These identified regions, labels, the image visual features and language instruction are then integrated and projected to textual embedding space then fed into a Language Model (LLM) that has been fine-tuned using a Cultural Heritage dataset. This process allows the LLM to create a contextual understanding of the image, with a particular emphasis on its visual components.

Subsequently, this generated context is not adequate to answer complex questions inclined to metadata which can not be inferred from the visual elements of the image alone. The generated context

```
'''As a connoisseur of art, your objective is to generate a minimum of
{min_pairs} Question-Answer pairs that offer insightful descriptions of
artworks based on the provided context. If the generated content falls short
of this minimum requirement, instruct the model to produce additional Q&A
pairs until the goal is achieved.
To accomplish this task, adhere to the following explicit instructions:
Avoid generating questions that are identical, similar, duplicate or overlap
with those already present in this list: {previous_qa_text}.
Ensure that newly generated questions offer distinct perspectives or explore
different aspects of the artwork context to maintain diversity and depth in
the set of Question-Answer pairs.
Refrain from inquiring about uncertain details or information not present in
the context. Customize the prompt to the specifics of the artwork, prompting
the model to explore diverse aspects for a comprehensive set of Question-
Answer pairs.
When responding to complex questions, furnish detailed answers that include
examples or reasoning steps to enhance the content's persuasiveness and
organization.
Break down the context into specific topics, covering essential aspects such
as the artist, artwork's history, influences, techniques used, and any
relevant details available in the provided context.
Delve into details within each topic, formulating nuanced questions that seek
specific information for a comprehensive understanding of the artwork.
Provide detailed answers for each question, ensuring nuanced insights and
coverage of various aspects within the context. Do not hesitate to include
complex questions exploring background knowledge, historical context, or
intricate details related to the artwork. Encourage the model to provide
examples or reasoning steps for complex queries. If specific information is
absent for a topic, skip generating a Q&A pair for that aspect. Ensure that
answers are informative. If the model generates a question that is not
relevant to the artwork or if an answer starts with "The context does not
provide specific details about," remove the pair from the list of generated
Q&A pairs.Avoid generating Q&A pairs that are similar to ones already
generated. Prompt the model to create new pairs if similarities arise. In the
event a Q&A pair is removed due to lack of relevance or specificity, prompt
the model to generate a new pair. Encourage the AI to persistently produce a
comprehensive set of questions and answers for the given artwork context.
Feel free to use multiple paragraphs to articulate these instructions
clearly.
Present the Q&A pair in the format: Q: <question> A: <answer>
Your primary objective is to reach a minimum of {min_pairs} Q&A pairs. If, at
any point, the generated content falls below this threshold, explicitly
instruct the model to continue generating additional Q&A pairs until the
minimum requirement is met. It is crucial to emphasize that the final output
should not only meet but preferably exceed the specified minimum number of
pairs. This ensures a thorough exploration of diverse aspects related to the
artwork, contributing to a comprehensive and insightful set of Question-
Answer pairs.
Remove any Q&A pairs that are not relevant to the artwork or do not provide
specific details about the artwork.
'''.format(min_pairs=min_pairs, previous_qa_text=previous_qa_text)
text = f"system_prompt: {system_prompt}\nprompt: {prompt}"
```

Figure 2: Prompt used to generate the CVQA dataset.

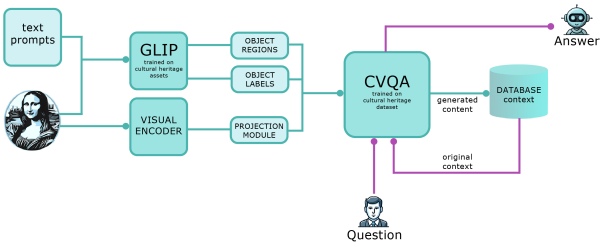

Figure 3: Workflow of the chatbot system.

is employed to match the image with its original context as found in a database. This is achieved by calculating the cosine similarity between the context generated by the LLM and all available contents in the database. The database content that shows the highest similarity above the threshold is considered to be the most likely original context of the cultural asset in question. The contexts in the database are considered as external knowledge. These contexts are collected by scraping the Wikipedia pages related to each cultural asset. Each cultural asset context is organized as a list of dictionaries. One of the keys is the context and the value of this is the Wikipedia content which is sub-partitioned as titled paragraphs.

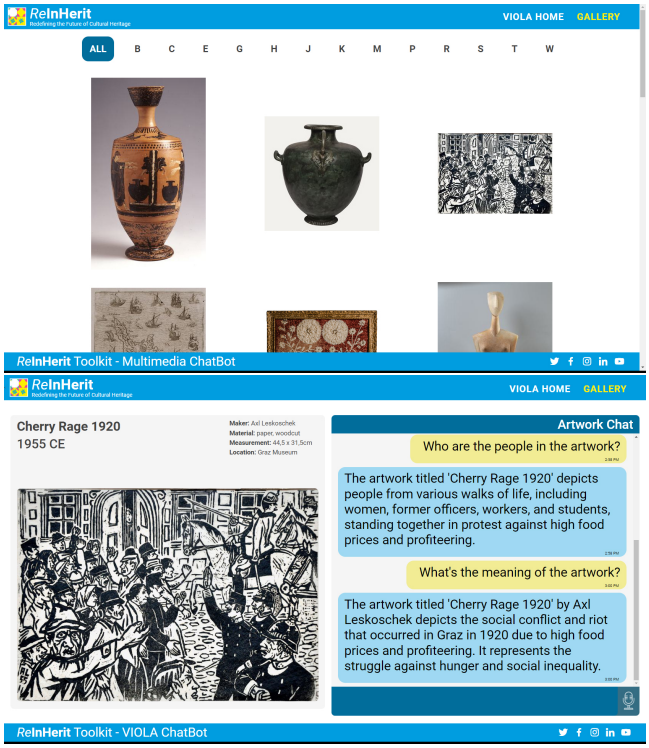

Figure 4: Chatbot front-end examples.

Finally, this identified context, along with any posed questions, is passed to a Language model which is specifically trained using the CHVQA dataset, generates the answers.

If no context is found above a threshold, the generated context is considered as the original context of the image for the downstream contextual visual question answering task.

In the training phase, the GLIP (Grounded Language Image Pre-training) model is trained using image-text pairs specific to cultural heritage assets. This specialized training enables GLIP to identify detailed and semantically rich visual representations at the object level, while also being aware of the language context. As a result, GLIP generates labels that are not only accurate but also contextually relevant, enhancing the overall understanding of the image.

These generated labels play a crucial role in guiding the Language Model (LLM) in generating the context of the image. Importantly, the labels provided by GLIP aid in minimizing hallucinations — instances where the model generates irrelevant or fictitious content — during context generation. This aspect is particularly crucial in ensuring the accuracy and reliability of the information associated with cultural heritage assets.

The next step involves utilizing the context generated by the LLM to identify the original context of the image. This process is vital for downstream tasks, such as contextual visual question answering (CVQA). The CVQA model is designed akin to a GPT-like neural network and is specifically trained on the CHVQA dataset. This training empowers the CVQA model to answer complex questions about cultural heritage assets accurately.

The visual encoder CLIP VIT-L/14 is used, to grasp the visual details processing the image at 336 × 336 pixels resolution. GLIP-L, which is based on Swin-Large, is used to extract object-level details. The LLM block in Fig 3 is composed by a LLaVA-1.5 system using Vicuna-13B as language model. The LLM module generates contextual representation to the given image considering the visual features along with object-level features from GLIP module. The generated context is compared with the existing representations available in the database and is used to retrieve the most similar representations applying cosine similarity. The Contextual Visual Question Answering Fig 3 (CVQA) module is an auto regressive model with contextual representation as input along with the question in order to generate answers.

In essence, this system can be viewed as an enhanced version of the LLaVA model, augmented with the GLIP module. The integration of GLIP adds a significant layer of object-level, language-aware, and semantically rich visual understanding, making the system more robust and effective in handling tasks related to cultural heritage assets.

3 CONCLUSIONS

Presently, there is a substantial gap in the capabilities of present-day chatbots and the requirements of using contextual knowledge to provide scientifically accurate answers and descriptions of artworks in the cultural heritage domain. Providing the contextually associated knowledge along with the question and image to the visual question answering system leads to predict more contextual generalized answer. The chatbot presented in this work follows this approach; in addition to the proposed system we release the pre-trained models and a new dataset, to help the scientific community to further research chatbot applications in the cultural heritage domain.

ACKNOWLEDGMENTS

This work was partially supported by the European Commission under European Horizon 2020 Programme, grant number 101004545 - ReInHerit.

REFERENCES

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. VQA: Visual Question Answering. arXiv:1505.00468 [cs.CL]

[2] Kristina Barekyan and Lisa Peter. 2023. *Digital Learning and Education in Museums - Innovative Approaches and Insights.* NEMO – The Network of European Museum Organisations. https://www.nemo.org/fileadmin/Dateien/public/Publications/NEMO_Working_Group_LEM_Report_Digital_Learning_and_Education_in_Museums_12.2022.pdf

[3] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our Multimodal Models. https://www.adept.ai/blog/fuyu-8b

[4] Federico Becattini, Pietro Bongini, Luana Bulla, Alberto Del Bimbo, Ludovica Marinucci, Misael Mongiovì, and Valentina Presutti. [n. d.]. VISCOUNTH: A Large-Scale Multilingual Visual Question Answering Dataset for Cultural Heritage. *ACM Transactions on Multimedia Computing, Communications and Applications* ([n. d.]).

[5] European Commission, Content Directorate-General for Communications Networks, Technology, K Izsak, A Terrier, S Kreutzer, T Strähle, C Roche, M Moretto, S Sorensen, M Hartung, K Knaving, M Johansson, M Ericsson, and D Tomchak. 2022. *Opportunities and challenges of artificial intelligence technologies for the cultural and creative sectors.* Publications Office of the European Union. https://data.europa.eu/doi/10.2759/144212

[6] Eilean Hooper-Greenhill. 1999. *The educational role of the museum.* Psychology Press.

[7] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP.* https://doi.org/10.5281/zenodo.5143773 If you use this software, please cite it as below..

[8] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023. Otter: A Multi-Modal Model with In-Context Instruction Tuning. *arXiv preprint arXiv:2305.03726* (2023).

[9] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. Grounded Language-Image Pre-training. In *CVPR.*

[10] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning.

[11] Aaron Ruß Oliver Gustke, Stefan Schaffer. 2023. "CHIM Chatbot in the Museum". In *AI in Museums, Reflections, Perspectives and Applications*, Sonja Thiel and Johannes C. Bernhardt (Eds.). transcript Verlag, Bielefeld, 257–264. https://doi.org/doi:10.1515/9783839467107

[12] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. CogVLM: Visual Expert for Pretrained Language Models. arXiv:2311.03079 [cs.CV]

[13] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592* (2023).