M-MiniGPT4: Multilingual VLLM Alignment via Translated Data

Seung Hun Han [⋆†*] MBZUAI
Youssef Mohamed [⋆] KAUST
Mohamed Elhoseiny KAUST
{firstname.lastname}@kaust.edu.sa
[⋆] Equal Contribution [†] Work done while being an intern at KAUST

Abstract

This paper presents a Multilingual Vision Large Language Model, named M-MiniGPT4. Our model exhibits strong vision-language understanding (VLU) capabilities across 11 languages. We utilize a mixture of native multilingual and translated data to push the multilingual VLU performance of the MiniGPT4 architecture. In addition, we propose a multilingual alignment training stage that uses parallel text corpora to further enhance the multilingual capabilities of our model. M-MiniGPT4 achieves 36% accuracy on the multilingual MMMU benchmark, outperforming state-of-the-art models in the same weight class, including foundation models released after the majority of this work was completed. We open-source our models, code, and translated datasets to facilitate future research in low-resource and multilingual settings.

1 Introduction

With the rise of powerful general-purpose Large Language Models (LLMs), multimodal extensions have begun to garner significant attention. In particular, Vision Large Language Models (VLLMs) combine the reasoning capabilities of LLMs with visual perception using a vision encoder. Notable early examples of open-source VLLMs include (Zhu et al., 2023a), (Liu et al., 2023), and (Dai et al., 2023a).

However, as early LLM development focused primarily on the English language, the derived VLLMs also tended to possess limited multilingual capabilities. Consequently, the benefiting audience has been restricted to English speakers, leaving approximately 75% of the global population[1] unable to benefit from advancements in these powerful open-source VLLMs.

Recently, open-source LLMs trained on multilingual data, such as Llama 3.1 (Dubey et al., 2024), Qwen (Team, 2024), and Command R (AI, 2024), have emerged. Similarly, Qwen-VL (Bai et al., 2023), based on Qwen 2.5 (Yang et al., 2024), has demonstrated improved multilingual capabilities. However, many of these models were not developed with multilinguality as their core objective, leading to limitations in language coverage. In this work, we explore the development and benchmarking of massively multilingual VLLMs using both synthetic and human-translated data. Furthermore, we demonstrate that the plug-and-play framework of MiniGPT4 is well-suited for multilingual learning, allowing it to scale with independent advancements in text-only LLMs and vision encoders.

VLLMs are typically trained in three stages. The first stage is designed to align the vision and language modalities and involves large-scale paired image-language data. Models following stage 1 tend to show weak reasoning performance; thus, a second stage involving high-quality instruction data is required to produce performant models. While data for both stages are readily available for English, high-quality multilingual multimodal data is scarce, and parallel multilingual multimodal datasets are virtually nonexistent. To mitigate this issue, we utilize state-of-the-art translation models to translate popular vision-language datasets. We show that translated data improves the model's multilingual performance without any notable degradation in English performance.

However, translated data does not account for the cultural and linguistic nuances that manifest only in natively collected datasets. As a result, relying solely on translation can result in sub-optimal multilingual VLLMs. To address this, we leverage parallel text corpora used for training machine translation models, as well as multilingual non-parallel text-only data, to improve the multilingual alignment of our models.

[*]Corrsponding author: eddieseunghunhan@gmail.com
[1]cochrane.org/news/cochrane-evidence-different-languages

To summarize, our contributions are:

- We translate multiple vision-language datasets to create new multilingual resources.
- We demonstrate the use of parallel text corpora to improve the multilingual performance of VLLMs.
- We train a state-of-the-art (SOTA) multilingual VLLM based on the MiniGPT4 architecture.
- We translate the MMMU benchmark to assess the multilingual reasoning performance of VLLMs.
- We open-source all translated datasets and models to support the community.

2 Related Work

Large Language Models and Multilinguality. LLMs have emerged as a transformative force in artificial intelligence, with success attributed to advances in GPU capabilities and large-scale training data. The field witnessed a paradigm shift with GPT-3 (Brown et al., 2020), demonstrating remarkable zero-shot capabilities. This sparked the development of numerous models, including open-source alternatives such as Bloom and OPT (Scao et al., 2022; Zhang et al., 2022), and proprietary models like Chinchilla (Hoffmann et al., 2022), PaLM (Chowdhery et al., 2022), and Megatron-Turing NLG (Smith et al., 2022).

While LLaMA (Touvron et al., 2023) introduced an approach with fewer parameters but more extensive training data, and the field continues to evolve with Llama 2/3, GPT-4, and Mistral (Jiang et al., 2023), these models primarily focus on English. To address this, Multilingual Large Language Models (MLLMs) have emerged, excelling in cross-lingual transfer tasks. XLM-R (Conneau et al., 2019) pioneered cross-lingual capabilities, followed by models like Bloom and mT5 (Xue et al., 2020) that intentionally incorporate substantial non-English data. Their instruction-tuned variants, Bloomz and mT0 (Muennighoff et al., 2022), have further advanced multilingual capabilities. Our research leverages these robust cross-lingual transfer capabilities to extend Vision and Language models into multilingual applications.

VLLMs: Recent advances in vision-language integration have focused on adapting LLMs to process visual information. Early approaches like Visual-GPT (Chen et al., 2022) and Flamingo (Alayrac et al., 2022) combined pre-trained LLMs with visual features. BLIP-2 (Li et al., 2023) introduced the Q-former to bridge visual and language representations. Building upon this, MiniGPT-4 (Zhu et al., 2023b) enhanced performance by incorporating the Vicuna model. LLaVA (Liu et al., 2023) aligned a frozen image encoder with LLaMA through instruction tuning, while InstructBLIP (Dai et al., 2023b) leveraged 26 diverse datasets. While effective for English, cross-lingual capabilities in these models remain largely unexplored.

Recently, models such as PALO (Rasheed et al., 2025) have tackled the multilingual aspect of VLLMs, supporting visual reasoning for 10 languages via translated instruction datasets. Although PALO showed promising visual understanding, it performed poorly on visual reasoning benchmarks. We observed that PALO models excelled at lengthy descriptions but failed at direct question answering. In this paper, we show that this issue stems from the limited size of the PALO dataset. Accordingly, we provide a more diverse translated dataset, resulting in significantly better performance in both understanding and reasoning tasks.

3 Datasets

VLLMs require high-quality multimodal data. We used No Language Left Behind (NLLB 1.3B) (Costa-jussà et al., 2022) to translate popular vision-language datasets into 10 languages: Chinese, Hindi, Spanish, French, Arabic, Bengali, Russian, Urdu, Japanese, and Korean. Additionally, we translated the MMMU visual reasoning benchmark (Yue et al., 2024) using the same model. We utilize both V&L datasets and text-only datasets.

3.1 V&L Datasets

- **Conceptual Captions** (Sharma et al., 2018), **SBU** (Ordonez et al., 2011), and **LAION** (Schuhmann et al., 2021) are weakly-labelled datasets consisting of English image-caption pairs. They are used for Stage 1 pretraining to align vision and text modalities. Collectively, they consist of roughly 5 million instances.

- **LLaVA-Instruct** consists of 6.8 million English image-text pairs used to train the LLaVA 1.5 model (Liu et al., 2023). The text consists of conversations and responses generated by GPT-4. We translated this dataset using

NLLB to all 10 target languages. We refer to the translated version as **LAVAM**.

- **PALO** consists of 2 million image-text pairs used to train the PALO model (Rasheed et al., 2025). The dataset providers translated the LLaVA-665K dataset into 9 different languages.

- **Cambrian Image (CI)** is a high-quality multimodal dataset released for the Cambrian-1 family of models (Tong et al., 2024), consisting of approximately 58 million English image-text pairs. We translated this dataset to all 10 target languages and refer to the translated version as **CI M**.

- **WIT** is a natively multilingual dataset derived from Wikipedia, consisting of 130K image-article pairs (Srinivasan et al., 2021). We filtered over 95% of the original datapoints due to quality issues (corrupted text, stub articles). We used the NLLB-CLIP model (Visheratin, 2023) to measure caption-image cosine similarity ($S_{I,C}$) and selected only pairs where $S_{I,C} \geq 0.0$.

3.2 Text Corpora

- **Cambrian Text (CT)** is a high-quality text dataset released for the Cambrian-1 models (Tong et al., 2024), consisting of 22 million datapoints. We translated this to the target languages, naming the version **CT M**.

- **Flores** is a natively multilingual dataset derived from Wikipedia, consisting of 110K translation pairs between 11 languages (Team, 2022).

- **XStoryCloze** is a human-translated paragraph completion dataset derived from StoryCloze (Lin et al., 2021), consisting of 20K datapoints across 10 languages.

We combine Flores and XStoryCloze into **MText** in our experiments.

3.3 Multilingual MMMU Benchmark

We utilized NLLB to translate the MMMU benchmark, which is designed to evaluate the reasoning capabilities of VLLMs. We validated the quality of our translation via evaluation before and after back-translation. specifically, we evaluated our model on the official English MMMU, translated MMMU to the target languages, and then back to English. Finally, we evaluated our model on the back-translated version. If the translation caused significant information loss, performance should drop; however, as shown in Table 1, performance remains consistent.

Model	E-MMMU	BT-MMMU
Pretrained	34.61	34.45
Finetuned	34.14	33.02

Table 1: **Performance validation via back-translation.** We report the average accuracy after back-translation from all target languages. **E-MMMU**: English MMMU; **BT-MMMU**: Back-Translated MMMU.

4 Experiments

4.1 Model Setup

We base our model on the MiniGPT4 (Zhu et al., 2023b) architecture. To support multilinguality, we replace the Vicuna LLM with Llama 3 (Dubey et al., 2024), which demonstrates superior performance on multilingual tasks. Our training pipeline consists of three stages, each designed to enhance specific aspects of model performance.

Stage 1 aims to align the visual and language modalities. We use large-scale image-caption datasets (Conceptual Captions, SBU, and LAION). Our experiments indicate that incorporating additional datasets at this stage does not yield performance improvements; thus, this stage remains consistent with the original MiniGPT4 implementation.

Stage 2 enhances multilingual understanding by training with multilingual multimodal data. We leverage our translated datasets (ccSBU, LAION, LAVAM, PALO). We further experiment with the Cambrian Image (CI) dataset and its translated version (CI M).

Stage 3 focuses on boosting multilingual capabilities. We conduct ablation studies using CI and Cambrian Text (CT) datasets in both original and translated versions (CI M, CT M), as well as the parallel corpora used for translation (MText).

4.2 Results

Table 3 compares our model with state-of-the-art vision–language models on the MMMU and MMMU Multi benchmarks. On MMMU Multi, our model substantially outperforms other fine-tuning approaches built on the same base model, improv-

Stage 2	Stage 3	MMMU	MMMU Multi
(ccSBU, LAION, LAVAM, PALO)	-	31.02	29.83
+ CI	-	34.14	31.21
+ CI M	-	36.69	**33.57**
+ CI	CI M + CT M + MText	37.07	32.90
+ CI M	CI M + CT M + MText	**37.27**	33.45
+ CI	CI + CT	35.19	32.93
+ CI	CI + CT + MText	35.65	32.60

Table 2: Ablation Studies on Training Data Combinations. **CI**: Cambrian Image; **CT**: Cambrian Text; **M**: Translated/Multilingual version.

Model	MMMU	MMMU Multi
PALO	28.36	13.12
Qwen-VL 2.5	**52.89**	25.46
Our Model	37.27	**33.45**

Table 3: Comparison to SOTA Vision-Language Models.

ing from 13.12% (PALO) to 33.45%, despite both methods using the Llama-3 backbone. Our approach also exceeds the performance of the latest open-source foundational model, Qwen-VL 2.5, which achieves 25.46% on this benchmark, highlighting the effectiveness of the proposed method for multi-modal, multi-step reasoning.

On the standard MMMU benchmark, Qwen-VL 2.5 attains higher accuracy (52.89%) than our model. We attribute this gap primarily to differences in instruction tuning scale and data diversity, as Qwen-VL 2.5 benefits from more extensive instruction tuning than was applied in our setting.

Table 2 presents our ablation studies. Several key observations emerge:

- Adding the Cambrian Image dataset (CI) in Stage 2 improves performance on both benchmarks.
- Using the translated version (CI M) in Stage 2 yields further improvements (36.69% on MMMU and 33.57% on MMMU Multi).
- The optimal configuration combines CI M in Stage 2 with CI M + CT M + MText in Stage 3.
- Including multilingual text data (MText) in Stage 3 generally improves performance when combined with translated datasets.

These results demonstrate the effectiveness of our three-stage training approach and the importance of incorporating multilingual multimodal data to enhance cross-lingual vision-language understanding.

5 Conclusion

In this paper, we presented M-MiniGPT4, a multilingual vision-language model that demonstrates strong performance across 11 languages. Our approach leverages a three-stage training process that effectively combines native multilingual data with translated datasets to optimize cross-lingual vision-language understanding. We demonstrated that using translated vision-language data significantly improves multilingual performance and that incorporating parallel text corpora further enhances the model's capabilities.

Our experiments show that M-MiniGPT4 achieves state-of-the-art multilingual performance on the MMMU Multi benchmark (33.45%), substantially outperforming existing models like Qwen-VL 2.5 and PALO in multilingual visual reasoning tasks. By open-sourcing our models and translated datasets, we facilitate further research in multilingual multimodal AI, making these technologies more accessible to non-English speakers worldwide.

6 Limitations

Despite promising results, M-MiniGPT4 faces several limitations:

- **Translation Nuance:** Reliance on machine translation may not fully capture cultural nuances and linguistic subtleties present in natively collected multilingual data.
- **Language Coverage:** While our model was finetuned on 11 languages, this covers only a fraction of the world's languages.

- **Resource Disparity:** Translation quality varies, with high-resource languages (e.g., Spanish, French) benefiting from better translations compared to lower-resource languages (e.g., Bengali, Urdu).
- **Evaluation:** Our metrics may not comprehensively assess all aspects of cross-cultural understanding in visual reasoning.
- **Inherited Bias:** Reliance on pretrained LLMs inherits the biases and limitations inherent in the base models.

Future work should focus on expanding language coverage, incorporating more natively collected multilingual data, and developing nuanced evaluation frameworks for cross-cultural understanding.

References

Cohere AI. 2024. Cohere command r models. Accessed: 2024-12-24.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023a. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023b. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. 2025. Palo: A large multilingual multimodal language model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2025)*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449. ACM.

NLLB Team. 2022. No language left behind: Scaling human-centered machine translation.

Qwen Team. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Preprint*, arXiv:2406.16860.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alexander Visheratin. 2023. Nllb-clip – train performant multilingual image retrieval model on a budget. *Preprint*, arXiv:2309.01859.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. 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.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.