An Open Source Path to Embodied Multimodal Intelligence

Sushant Karki
Science Academy
University of Maryland College Park
skarki@umd.edu

Chandramani
Science Academy
University of Maryland College Park
cm05@umd.edu

***Abstract*—The recent development in large embodied multimodal mode like RT2, GPT4, Palm-E has demonstrated extraordinary multi-modal abilities, such as multiple embodied tasks including sequential robotic manipulation planning, visual question answering, and captioning. These features are rarely observed in previous vision- language models. Large language models have been demonstrated to perform complex tasks. However, enabling general inference in the real world, e.g. for robotics problems, raises the challenge of grounding.**

In our project, we focus on crafting a Vision-Language-Action (VLA) model, similar to to RT-2, with a distinctive twist — the integration of open-source models and innovative parameter-efficient fine-tuning techniques. Our core methodology involves fine-tuning the Llava model using QLoRA, a method designed to enhance parameter efficiency. Our approach includes novel aspects such as direct action representation as tokens, and token association for efficient action representation.

***Index Terms*—Large Language Model, Robotic Action, Table Top, VQA, VLM**

I. Introduction

The recent introduction of Google's PaLM-E, a massive 562-billion embodied multimodal language model, and RT2 an improvement over RT2, marks a significant leap in the realms of linguistics and robotics. However, the closed-source nature of this model poses challenges in terms of accessibility and costs. In response to these challenges, we advocate for the development of an open-source, parameter-efficient implementation of PaLM-E and RT2. Such an alternative holds immense potential to democratize access to this groundbreaking technology, enabling a diverse community of researchers and developers to explore and expand its capabilities. Moreover, by addressing the cost-intensive nature of fine-tuning a model of this scale, our initiative aims to make multi-modal language model research more affordable and widely accessible. Through this endeavor, we aspire to foster collaboration, innovation, and progress in the dynamic field of multi-modal language models.

We have developed a new approach to enhance the capabilities of large language models, making them more useful in real-world scenarios like robotics. The challenge we addressed is how to connect the language understanding of these models with the real-world environment they operate in.

Our primary objective is to provide a solution that not only addresses the closed-source limitations of PaLM-E but also mitigates the significant expenses associated with fully fine-tuning a model of this scale. This endeavor is poised to accelerate progress in multimodal language models, empowering a wider community to contribute to and benefit from advancements in this dynamic field. In the project's initial phase, our primary emphasis lies in crafting an open-source implementation of PaLM-E [1] and RT2 [2], with a particular emphasis on augmenting parameter efficiency. This implementation is custom-designed for the Tabletop Manipulation task within the Language Table environment.

Our Contribution

1) We prioritize the integration of open-source models, promoting transparency and accessibility in our Vision-Language-Action (VLA) model. By tapping into existing open-source tools, we actively contribute to a collaborative and inclusive development and research environment.
2) Our distinctive approach involves fine-tuning the BaK-Llava model using QLoRA, a method tailored to enhance model efficiency with reduced computational costs. This innovative technique optimizes parameter performance, making our model resource-efficient..

II. Related Work

A. Visual Language Model

The field of vision-language pre-training has witnessed the development of diverse model architectures to enhance performance across a spectrum of vision and language tasks. Notable among these are the dual-encoder architecture, as demonstrated in works by Radford et al. (2021) [3] and Jia et al. (2021) [4], the fusion-encoder architecture introduced by Tan and Bansal (2019) [5] and further explored by Li et al. (2021) [6], the encoder-decoder architecture as employed by Cho et al. (2021) [7], and the more recent unified transformer architecture proposed by Li et al. (2022) [8] and Wang et al. (2022b) [9]. Over the years, various pre-training objectives have been put forth, converging towards established methodologies such as image-text contrastive learning, image-text matching, and (masked) language modeling .Additionally, BLIP-2 emerges as a novel vision-language pre-training method, employing a Querying Transformer (Q-Former) that undergoes two distinct stages of pre-training: vision-language representation learning with a frozen image encode, followed by vision-to-language generative learning with a frozen Large Language Model (LLM). This innovative approach aims to bridge the modality gap, offering a unique perspective in the evolving landscape of vision-language pre-training methodologies. [10]

B. Leveraging Pre-trained LLMs in Vision-Language Tasks.

In recent years, there has been a notable surge in the adoption of autoregressive language models as decoders in vision-language tasks, as evidenced by studies like Chen et al. (2022) [11], Huang et al. (2023) [12], Yang et al. (2022) [13] [10]. This strategy leverages the power of cross-modal transfer, facilitating the sharing of knowledge between language and multimodal domains. Pioneering works such as VisualGPT [11] and Frozen [14] have illustrated the advantages of employing a pre-trained language model as a vision-language model decoder. Subsequent advancements, including Flamingo, [15] employed gated cross-attention to align a pre-trained vision encoder and language model, exhibiting remarkable in-context few-shot learning capabilities. The introduction of BLIP-2 [10] further optimized the alignment of visual features with the language model, utilizing a Flan-T5 [16] with a Q-Former. Notably, the recent development of PaLM-E [1], featuring 562 billion parameters, signifies a groundbreaking effort to integrate real-world continuous sensor modalities into a Large Language Model (LLM), establishing a direct link between real-world perceptions and human languages. Additionally, the release of GPT-4 [17] showcases enhanced visual understanding and reasoning capabilities, achieved through extensive pre-training on a vast collection of aligned image-text data.

III. Dataset

A. Language Table Dataset

The Language Table dataset, a significant contribution to natural language-instructable robots, is introduced with an open-sourced framework, including datasets, environments, benchmarks, and policies. Trained using behavioral cloning on a large dataset, the resulting policy demonstrates exceptional proficiency, achieving a 93.5% success rate on diverse language instructions for real-world visuo-linguo-motor skills. Notably, the policy exhibits adaptability, responding to real-time human guidance for precise rearrangement tasks. The dataset, comprising nearly 600,000 labeled trajectories, is a substantial advancement, incorporating real robot data and various simulation scenarios. This comprehensive resource aims to advance natural language interaction with robots by providing valuable insights and diverse components for research and development. [18]

B. LLaVA Visual Instruct

The LLaVA Visual Instruct 150K dataset serves as a valuable resource for fine-tuning and enhancing the visual instruction capabilities of language models, specifically tailored towards the multimodal capabilities of GPT-4. Constructed as an augmentation of the COCO dataset, this dataset comprises GPT-generated multimodal instruction-following data. The dataset encompasses a total of 158,000 unique language-image instruction-following samples, distributed across various contexts. [19]

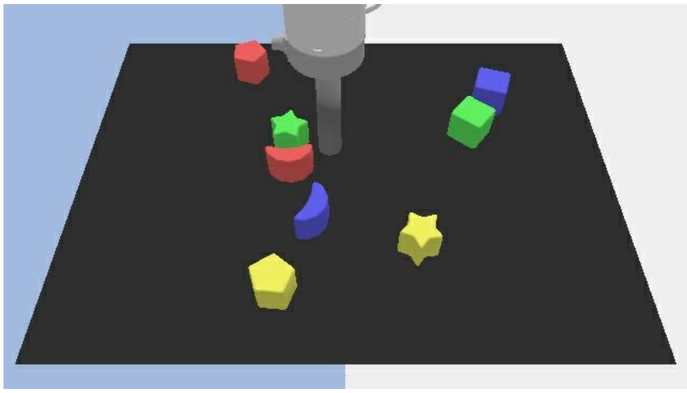


Fig. 1. Language Table Dataset

One of the example from the dataset involves an image of a man with a truck and associated conversations between a human and the GPT model. The interactions cover a range of inquiries related to the image content, including the activities of a man in the back of a pickup truck. The conversations include questions about the man's actions, details about the pickup truck, and speculative reasoning about the possible reasons for the depicted scenario.

IV. Method and Tools

Our project is dedicated to developing a Vision-Language-Action (VLA) model similar to RT-2, but distinctively using open-source models and focusing on parameter-efficient fine-tuning techniques. The centerpiece of our methodology is fine-tuning the recent BakLlava [20] [19] model with QLoRA, a method that enhances parameter efficiency.

1) **Action Encoding and Tokenization:**
   - **Direct Action Representation:** n alignment with the RT-2 model, our approach trains the model to output actions as tokens, which are integrated with language tokens for a cohesive action-language output.
   - **Discretization of Action Space:**Our model's action space, in contrast to RT-1's [21] 6-DoF action space, is confined to the X and Y coordinates (delta values) of robot arm movements. These dimensions are discretized into 256 uniform bins, each represented as an integer, for a simplified and manageable action space.
   - **Token Association:** Following a concept introduced in "Symbol Tuning Improves In-Context Learning in Language Models" [22] and implemented in RT-2 [2], we reserve 256 tokens from the Llava model's tokenizer to represent action tokens. This strategy involves mapping the discrete bin ordinals to specific tokens, allowing for an efficient

representation of robot actions within the model's output.

2) **Model Co-Fine-Tuning:**
   - **Data Integration:** The model processes inputs comprising robot camera images and textual task descriptions, formatted akin to standard VQA style. The outputs are strings of tokens representing robot actions.
   - **VQA Dataset Generation:** We also generated a custom VQA dataset using images from the language table dataset, created in a format similar to the llava-instruct-150k dataset, with the assistance of ChatGPT. This enriches the model's capability in image-grounded reasoning and aligns with the high-quality structure of the LLaVA-Instruct-80K dataset.
   - **Co-Fine-Tuning :** An essential part of our training involves co-fine-tuning with the Llava Instruct dataset alongside the language table robotics data, to help enhance the model's generalizability across diverse visual concepts and specific robot actions
   - **Training Batch Balance:** A balanced representation of language table dataset and the Llava Instruct dataset is maintained in each training batch, adjusting the sampling weight to ensure adequate exposure to both types of data.

V. Experiments

Our experimental framework was designed to iteratively improve the integration of Vision-Language Models (VLMs) in Visual Question Answering (VQA) tasks and robotic action generation. Each experiment built upon the learnings of its predecessors, refining our approach and methodology

- **Initial VLM with Frozen Components:** We initiated our experiments by training a VLM that incorporated a frozen Vision Transformer (ViT) [23] and a frozen Llama-2-7B-chat [24] as the vision and language models, respectively. A linear layer was introduced to project vision embeddings to dimensions compatible with the language model. This model was trained using 10,000 data points each from the COCO and CLEVR datasets. Despite the strategic selection of CLEVR for its object similarity to the language table dataset, this model exhibited poor performance in VQA tasks, particularly lacking in generating image-grounded text.
- **Adoption of MiniGPT4 Architecture:** Following the architectural insights from the MiniGPT4 study, we incorporated a QFormer and a Linear layer between the vision and language models. Despite training this model on the same datasets as the first experiment, the performance improvements were minimal, with similar issues in VQA tasks persisting. [25]
- **Expanded Dataset with LoRA Adapters:** To enhance the model's capabilities, we expanded the training dataset to 20,000 data points from each of the COCO [26], CLEVR, and [27] COCO-VQA datasets [28]. This expansion aimed to provide a well-rounded and diversified set of visual and textual data for training. The model maintained the linear layer strategy and incorporated LoRA adapters into the ViT, while keeping the language model frozen. This version showed a notable improvement in describing image contents but still faced challenges in reasoning with images from the language table dataset.
- **Integration of Language Table VQA Dataset:**
  To specifically address the reasoning limitations, we generated a VQA dataset for the language table images using ChatGPT and added it to our training set. This model exhibited improved performance in reasoning about languagetable images, marking a significant advancement from previous iterations.
- **Experiment with Robotic Actions Dataset** We further experimented by generating a dataset for robotic actions using the language table dataset. This involved discretizing actual X and Y effector delta values into action tokens using our action tokenizer. Despite integrating this new multimodal VQA dataset into our training, the model consistently produced identical action tokens for varying prompts, indicating a critical issue in diverse action token generation.
- **Fine-Tuning Llava VLM with QLoRA:**
  Recognizing the potential of the advanced Llava VLM, we decided to fine-tune it, focusing on a subset of the self-attention and multimodal projection layers using QLoRA. Initial memory challenges were overcome by employing additional GPUs. This final model outperformed all previous versions, retaining robust VQA capabilities while also being able to output robot action tokens. However, there is still room for improvement in the accuracy and diversity of the robot action token generation. Through these experiments, we have progressively refined our approach to developing a model that adeptly combines VQA reasoning with robotic action generation. While the final model presents significant advancements, it also highlights areas for future research, particularly in enhancing the quality and accuracy of robot action token generation.

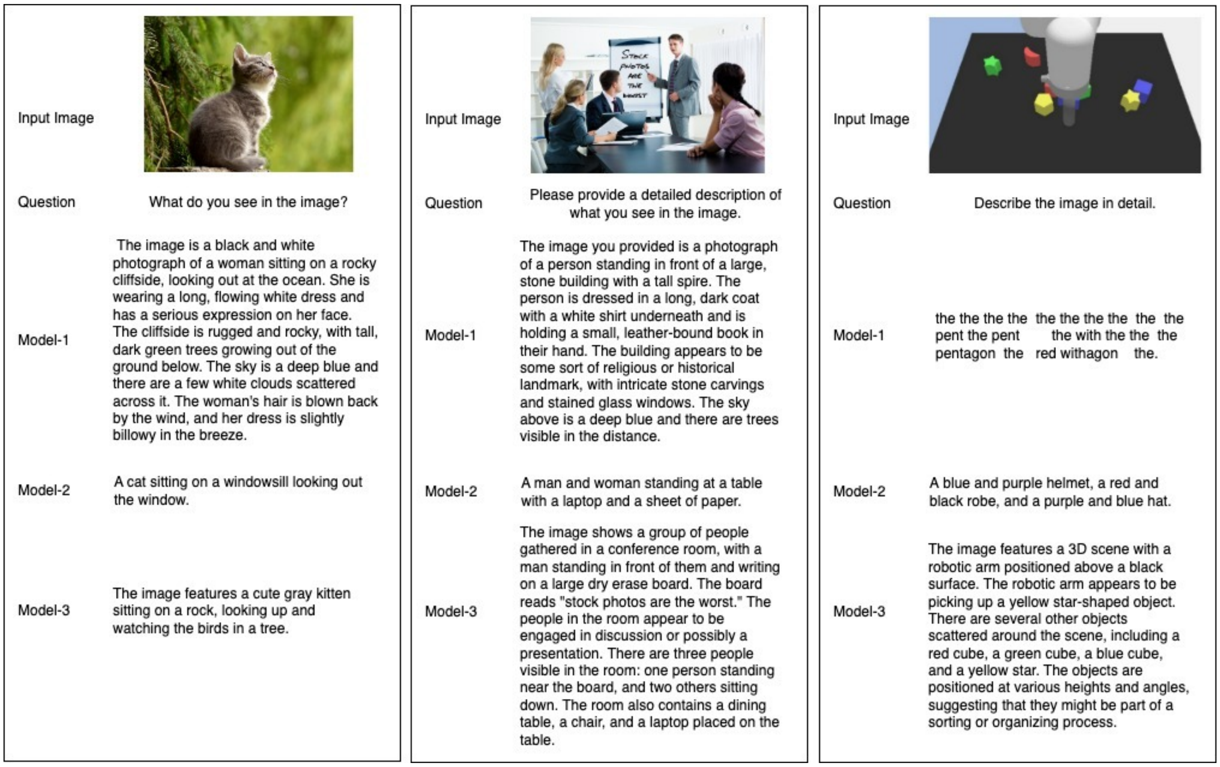
VI. Results

A. Original Objectives and Shortcomings

The primary goal of our project was to create a Vision-Language-Action (VLA) model capable of generating robotic action tokens and conducting simulations within the language table environment. This objective was set forth as a critical component of our robotics class assignment. Despite a concerted effort and innovative approaches, our project did not progress beyond the vision-language alignment phase to the action token generation phase, which was the intended milestone for the class.

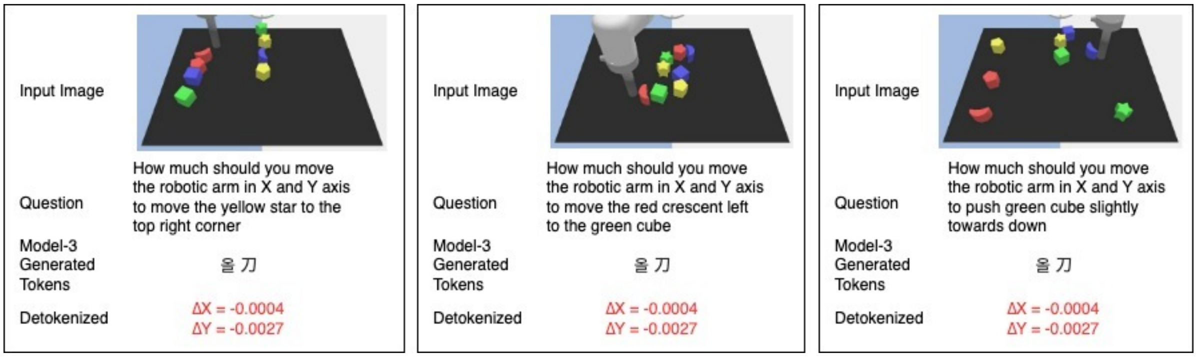
B. Evaluation of Vision-Language Model Configurations and Their Alignment Outcomes

1) **Model 1 (Frozen Vision Transformer + Trainable Linear Projection Layer + Frozen Llama2):**



Input Image		Input Image		Input Image	
Question	What do you see in the image?	Question	Please provide a detailed description of what you see in the image.	Question	Describe the image in detail.
Model-1	The image is a black and white photograph of a woman sitting on a rocky cliffside, looking out at the ocean. She is wearing a long, flowing white dress and has a serious expression on her face. The cliffside is rugged and rocky, with tall, dark green trees growing out of the ground below. The sky is a deep blue and there are a few white clouds scattered across it. The woman's hair is blown back by the wind, and her dress is slightly billowy in the breeze.	Model-1	The image you provided is a photograph of a person standing in front of a large, stone building with a tall spire. The person is dressed in a long, dark coat with a white shirt underneath and is holding a small, leather-bound book in their hand. The building appears to be some sort of religious or historical landmark, with intricate stone carvings and stained glass windows. The sky above is a deep blue and there are trees visible in the distance.	Model-1	the the the the the the the the the the pent the pent the with the the the pentagon the red withagon the.
Model-2	A cat sitting on a windowsill looking out the window.	Model-2	A man and woman standing at a table with a laptop and a sheet of paper.	Model-2	A blue and purple helmet, a red and black robe, and a purple and blue hat.
Model-3	The image features a cute gray kitten sitting on a rock, looking up and watching the birds in a tree.	Model-3	The image shows a group of people gathered in a conference room, with a man standing in front of them and writing on a large dry erase board. The board reads "stock photos are the worst." The people in the room appear to be engaged in discussion or possibly a presentation. There are three people visible in the room: one person standing near the board, and two others sitting down. The room also contains a dining table, a chair, and a laptop placed on the table.	Model-3	The image features a 3D scene with a robotic arm positioned above a black surface. The robotic arm appears to be picking up a yellow star-shaped object. There are several other objects scattered around the scene, including a red cube, a green cube, a blue cube, and a yellow star. The objects are positioned at various heights and angles, suggesting that they might be part of a sorting or organizing process.

Fig. 2. Comparison of Model 1, Model 2 and Model 3 performances on different VQA Prompts and Images



Input Image		Input Image		Input Image	
Question	How much should you move the robotic arm in X and Y axis to move the yellow star to the top right corner	Question	How much should you move the robotic arm in X and Y axis to move the red crescent left to the green cube	Question	How much should you move the robotic arm in X and Y axis to push green cube slightly towards down
Model-3 Generated Tokens	올 刀	Model-3 Generated Tokens	올 刀	Model-3 Generated Tokens	올 刀
Detokenized	ΔX = -0.0004 ΔY = -0.0027	Detokenized	ΔX = -0.0004 ΔY = -0.0027	Detokenized	ΔX = -0.0004 ΔY = -0.0027

Fig. 3. Result of Model 3 on different Table top Scenarios. It is giving us similar action tokens for different robotic manipulation task. For tokens please refer Appendix B

Stage One - Vision-Language Alignment: The preliminary stage aimed to leverage the linear layer to instill image-grounded reasoning capabilities within the language model. However, this stage did not meet our expectations, as the model struggled to interpret images accurately. The inability to understand images prevented us from proceeding to the second stage of training for robotic action generation.

Stage One Results: The model's outputs during this stage showed a consistent inability to form accurate image descriptions, indicating a fundamental gap in vision-language alignment.

2) **Model 2 (Vision Transformer with QLoRA + Trainable Linear Projection Layer + Frozen Llama2):**

Stage One - Vision-Language Alignment: With an augmented dataset and architectural changes, including the incorporation of QLoRA in the Vision Transformer, this model demonstrated partial success in understanding images. The model began to show signs of grasping visual elements within the images, though it was not consistent across all dataset types. Specifically, it failed to comprehend the language table dataset, leading to the decision not to advance to stage two.

Stage One Results: The model's understanding of images varied, with some correct interpretations intermixed with inaccuracies, reflecting an incomplete but improved vision-language alignment compared to Model 1. For other results please refer AppendixA

3) **Model 3 (BakLlava Model Fine-Tuned on Various Datasets):**
Single-Stage Training for Robotic Actions: Given the setbacks with Models 1 and 2, we pivoted to leveraging the recently released open-source VLMs—Llava and BakLlava. Our strategy shifted to fine-tuning the BakLlava model on the robotic actions dataset with QLoRA, bypassing the initial alignment stage. We experimented with several dataset combinations to optimize performance.
Results: Encouragingly, the model retained its VQA capabilities post-fine-tuning, which was a positive indicator of its robustness. However, the model did not succeed in generating contextually varied robotic action tokens. Irrespective of the instruction provided, the model defaulted to outputting the same reserved tokens, highlighting a critical area for improvement. [Fig 5]

C. Interpretation of Results

The complexity of achieving accurate vision-language alignment was more pronounced than initially anticipated. The translation of visual and linguistic data into robotic action tokens proved difficult, indicating that further methodological refinements are needed. The consistent issue of repetitive action token generation suggests that our models may require a different approach to learn the variability and specificity required for robotic tasks.

VII. Future Work

Our research has achieved significant progress in integrating vision with language processing, enabling effective image interpretation. However, we face a persistent challenge with our model's tendency to overfit to robotic action data and its repetitive generation of action tokens. To enhance our model's capabilities, especially for practical robotic applications, we propose the following focused developments:

- **Integration of Advanced Image Segmentation Models:**
Objective: Implement models like Meta's Segment Anything Model (SAM) to improve object understanding in images. (SAM) [29] into our framework.
Challenges: Efficiently merging the segmentation output with the language model and managing the added computational complexity will be key. The segmentation model's accuracy will directly influence the overall system performance.
- **Continuous Value Generation for Robot Actions:**
Objective: Adopting the approach from the xVal paper [30] to enable the language model to generate continuous values for robotic actions, moving away from the current discretization method.
Challenges: Ensuring stable training and accurate continuous value prediction is crucial. This approach may require significant alterations to the traditional language model architecture to accommodate continuous outputs.
- **Ordinal Regression Classification Loss Implementation:**
Objective: To integrate an ordinal regression/classification loss alongside the traditional cross-entropy loss. This aims to exploit the ordinal nature of the action space, ensuring that even incorrect predictions are closer to the actual value.
Challenges: Balancing this new loss function with existing learning mechanisms and aligning token ordinality with action bins.

VIII. Conclusion

In conclusion, our investigation into the development of Vision-Language-Action models has been a journey marked by both insights and obstacles. Model 1 was unable to establish a foundational understanding of images, a critical first step for subsequent action token generation. Model 2 offered incremental improvements in image comprehension, yet it did not achieve a comprehensive grasp of the language table dataset needed for the progression to robotic action training. Model 3, despite successfully retaining its Visual Question Answering capabilities, faced difficulties in producing a diverse set of robotic action tokens, a key requirement for our robotics class project.

These outcomes underscore the intricate challenges involved in synthesizing visual data interpretation with language processing to produce actionable outputs for robotics. The complexities encountered have shed light on the limitations of current methodologies and have opened avenues for future exploration, as detailed in our Future Work section.

Moving forward, the insights gained from this research lay a foundation for further advancements in the field. We anticipate that the next steps, guided by the reflections and proposed directions in our Future Work, will pave the way for breakthroughs in the creation of more sophisticated and capable VLA models. The lessons learned here will serve as valuable reference points for future endeavors in the dynamic and evolving landscape of the Multimodal Models.

IX. Contribution

- **Chandramani :** Led initial Vision-Language Model (VLM) training with ViT and Llama-2-7B-chat. Generated VQA dataset for language table images and expanded the MiniGPT4 architecture. Despite progress in image description, faced challenges in reasoning with language table images.
- **Sushant :** Addressed reasoning limitations, initially with ViT and Lora. Adopted MiniGPT4 architecture, showing improved reasoning performance. Spearheaded robotic actions dataset generation and fine-tuned BaKLlava VLM with QLoRA, enhancing VQA and robotic action capabilities.

References

[1] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023.

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

[4] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[5] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[6] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[7] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

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

[9] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

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

[11] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 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, 2022.

[12] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

[13] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022.

[14] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

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

[16] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[17] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[18] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time, 2022.

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

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

[21] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale, 2023.

[22] Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V. Le. Symbol tuning improves in-context learning in language models, 2023.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[24] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[25] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

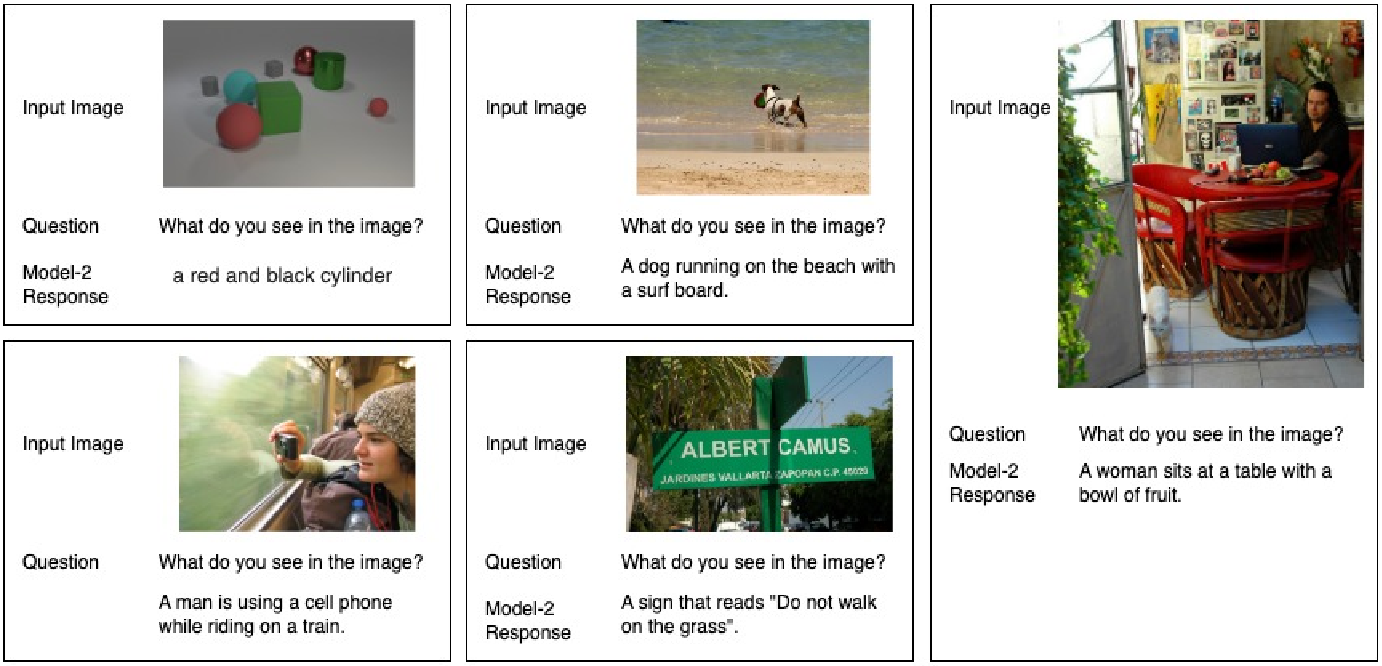
[27] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016.

[28] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[30] Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, Geraud Krawezik, Francois Lanusse, Michael McCabe, Ruben Ohana, Liam Parker, Bruno Régaldo-Saint Blancard, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. xval: A continuous number encoding for large language models, 2023.

APPENDIX A
OTHER RESULTS FOR MODEL 2



Input Image

Question What do you see in the image?

Model-2 Response a red and black cylinder

Input Image

Question What do you see in the image?

Model-2 Response A dog running on the beach with a surf board.

Input Image

Question What do you see in the image?

Model-2 Response A woman sits at a table with a bowl of fruit.

Input Image

Question What do you see in the image?

A man is using a cell phone while riding on a train.

Input Image

Question What do you see in the image?

Model-2 Response A sign that reads "Do not walk on the grass".

Fig. 4. Results of Model 2 on different set of Images

APPENDIX B
TOKENS

The Tokens we reserved for our action tokens are :

职 陽 亞 따 施 泰 載 벨 笑 華 迎 됨 豆 嘉 ĕ 庄
級 Ψ 気 责 អ 乱 休 約 ม ∑ 察 온 ర 乘 람 इ
ந โอ 亲 委 赤 됨 勝 怎 감 宋 調 짜 难 못 티 備
塞 险 旅 虚 笔 馆 唐 律 稍 散 ટ ヴ 副 尽 挂 県
⚠ 洋 鬼 암 孩 °C 扣 铁 闻 戳 む 秀 細 御 拖 좌
绍 참 향 끝 민 贵 纪 秋 ಕ 網 铺 恋 兼 羽 창 啟
弟 넌 慢 효 許 硬 잘 템 術 溪 暴 混 夢 랑 या 還
探 祖 织 軍 務 艺 應 擇 渡 業 령 決 刀 從 變 올
灣 평 衣 製 隊 ₱ 纳 賴 农 桥 阻 ជ 秘 박 伤 稿
拦 냉 宿 錄 镜 채 ə 党 급 洲 說 ĭ 尝 담 फ 哥
圣 萨 丁 虎 권 善 岩 커 抛 석 Έ 宣 拳 팅 枚 洛
証 陵 佐 館 돌 稱 聊 車 루 ძ 庫 지 統 련 ക 旗
励 紀 忠 杨 丹 却 舞 轉 ထ 丽 借 음 편 蒙 衡 叶
ς 谢 Ą ঔ à 既 济 ⊁ 準 답 ల 残 虑 急 招 막
產 垂 親 猫 刪 胡 晚 군 승 曾 論 ಹ 戰 鱼 實 특
崎 甘 該 링 頁 큰 毁 聖 麻 敏 運 될 쓰 ਟ 젝 復

Fig. 5. Tokens Used for action Tokenizer