
∞ GPT: Training Large Language Models For Any-To-Any Generation

Abdul Waheed
Carnegie Mellon University
abdulw@andrew.cmu.edu

Abhigyan Kishor
University of Pittsburgh
abk171@pitt.edu

Liangyu Wang
Carnegie Mellon University
liangyu2@andrew.cmu.edu

Abstract

Abstract

Recent advancements in large language models (LLMs) have demonstrated impressive capabilities in text-based understanding and generation. However, there remains a significant opportunity to extend these capabilities to support any-to-any generation, allowing models to process and produce outputs across text, images, and audio. To address this challenge, we propose a two-stage approach: first, we pre-train existing unimodal models, such as Phi-1.5, on a diverse array of multimodal data to build foundational cross-modal representations. This pre-training equips the model with a robust understanding of the interactions between different modalities. In the second stage, we fine-tune the model on an extensive collection of multimodal instructions, enabling it to perform complex tasks that require generating responses in all three modalities. To ensure seamless integration across modalities, we utilize quantization techniques that convert various data types into a unified representation space, facilitating the application of standard text tokenization methods. Our system will be rigorously evaluated on multiple multimodal benchmarks to assess its performance and generalization capabilities. Furthermore, we plan to investigate training strategies that do not depend on quantization, aiming to enhance the model’s fidelity and efficiency in multimodal interactions. This research aims to advance the development of versatile LLMs capable of dynamic any-to-any generation across diverse forms of data.

1 Introduction

2 Introduction

The principled scaling of language models, particularly with respect to both model size and data volume, has yielded powerful systems that demonstrate impressive generalization capabilities and understanding across a wide range of text-based tasks [1]. Recent advancements, notably models such as LLaMA [2, 3], Mistral [4], Qwen [5, 6], Yi [7], and Gemma [8], illustrate the significant progress achieved through such scaling laws. These systems have reached new benchmarks results in language understanding and generation, benefiting from vast data and substantial compute resources. Yet, despite their successes, these models are inherently limited by their reliance on text as the sole modality, both for input and output. This constraint limits their applicability in real-world, multimodal settings where users may need to process and generate across various forms of data, including images, audio, and video.

To bridge this gap, efforts have emerged to extend large language models (LLMs) to multimodal capabilities, enabling them to understand and generate insights from different data types. Models like MiniGPT-4 and others have pushed the boundaries of multimodal comprehension [9, 10, 11, 12, 13, 14]. These models, however, largely focus on multimodal understanding rather than generation, thereby limiting their generative capabilities in practical applications that demand versatile output across modalities.

In this work, we aim to address these limitations by training language models to handle multimodal data as both inputs and outputs—a concept we refer to as “any-to-any generation.” Our methodology involves tokenizing non-text data (e.g., images, audio) to transform them into discrete token representations, effectively creating a unified data format that can be processed by LLMs using standard text tokenizers [15, 16]. By applying text-based tokenization across these diverse data types, we establish a consistent framework for handling various modalities within a single model architecture. Our approach draws inspiration from prior work on multimodal processing in LLMs [17], but we aim to advance this by introducing a multi-stage process that enhances the model’s capability to generate outputs across all included modalities.

Specifically, our training approach is structured into two key stages:

1. **Stage 1 - Continuous Pretraining on Multimodal Data:** We initiate this stage by extending existing unimodal LLMs, using Phi-1.5 [18] as our baseline, to train on multimodal data. Through this pretraining phase, the model begins to develop an understanding of patterns and representations across different data types, laying the groundwork for multimodal generation.
2. **Stage 2 - Instruction-Focused Fine-Tuning with Any-to-Any Data:** In the second stage, we compile and curate a vast dataset of diverse instruction-based multimodal interactions. By fine-tuning on this rich instruction data, we enable the model to generate responses for inputs across any combination of images, text, and audio. Early observations from this stage indicate that the model effectively handles instructional prompts involving all three modalities, showcasing its flexibility in generating contextually appropriate responses.

Our next steps involve a thorough evaluation of the model across various multimodal benchmarks to quantify its performance, robustness, and generalization capabilities across modalities. In addition to assessing our model’s performance, we plan to conduct comparative analyses with other leading any-to-any generative models, evaluating strengths and weaknesses across different benchmarks and datasets. Furthermore, as part of our ongoing research, we aim to develop models capable of processing image and speech modalities directly, circumventing the need for quantization. This would further streamline multimodal processing, offering a more seamless and efficient approach to any-to-any generation.

In the following sections, we provide an in-depth review of related work in Section 3 and outline our experimental design and progress in Section 4.

3 Related Work

Large Language Models (LLMs) have revolutionized natural language processing, demonstrating unprecedented capabilities across various tasks. While their success in unimodal scenarios is well-established, extending these capabilities to multimodal domains remains a frontier of active research. The development of multimodal LLMs has progressed through several significant stages, from early foundational works to recent unified frameworks.

The initial phase of multimodal LLM development during 2020-2021 was marked by several pioneering works that established core methodological foundations. CLIP [19] introduced a paradigm-shifting approach by leveraging contrastive learning to bridge the visual-linguistic divide, successfully creating unified representations across modalities. This was complemented by DALL-E [20], which demonstrated the first successful large-scale text-to-image generation system, validating the feasibility of cross-modal generative tasks. In parallel, architectural innovations emerged through works like UNITER [21], which introduced a unified framework for vision-language understanding through extensive pre-training. ViLT [22] further refined this approach by proposing a streamlined architecture that efficiently integrated visual and linguistic features while maintaining strong performance across diverse multimodal tasks.

The field subsequently progressed toward more sophisticated architectural explorations during 2022-2023. Flamingo [23] introduced a novel integration approach, combining pre-trained language models with visual encoders to achieve exceptional performance in visual question-answering. CoCa [24] advanced this trajectory by developing a unified framework for vision-language pre-training and image-text generation through contrastive captioning. This period also saw the expansion of multimodal capabilities into practical applications. PaLM-E [25] demonstrated the feasibility of extending multimodal processing to robotic control, while GPT-4V [26] showcased sophisticated visual reasoning capabilities within a large language model framework. BLIP-2 [27] introduced an innovative bootstrapping approach that effectively addressed the modality gap in vision-language pre-training.

Recent advances in multimodal generation have explored diverse architectural approaches for handling multiple modalities. Transfusion [28] represents a novel direction by integrating a predict-then-diffuse paradigm within a unified transformer architecture, enabling joint processing of both discrete tokens and continuous features. This approach demonstrates the potential of combining language modeling capabilities with diffusion-based generation in a single framework. Building upon the idea of unified multimodal processing, CoDi [29] introduced a composable diffusion framework that creates a shared latent space across modalities, allowing for flexible generation between different modality pairs through carefully designed diffusion paths and attention mechanisms.

For Any-to-Any Generation with LLMs, AnyGPT [17] represents a significant breakthrough by introducing a unified multimodal processing framework. Its innovative use of discrete representations enables seamless integration of diverse modalities—including speech, text, images, and music—within a single LLM architecture. The introduction of AnyInstruct-108k, comprising 108,000 multi-turn dialogue samples, marks a crucial contribution to large-scale multimodal instruction data. Complementing these architectural advances, MixEval [30] has introduced a novel evaluation paradigm that efficiently combines existing benchmarks. Its methodology achieves a remarkable 0.96 correlation with human assessments while significantly reducing evaluation costs to 6% of traditional methods.

Despite these advances, several key challenges persist in multimodal LLM development. These include ensuring effective alignment across different modalities, balancing architectural sophistication with computational efficiency, and managing the substantial computational resources needed for training and deployment. Recent works have proposed innovative solutions to these challenges. AnyGPT addresses them through sophisticated data preprocessing and discrete representations, while CoDi employs novel composable generation strategies to create a shared multimodal space.

4 Experiments

4.1 Datasets

We train our model in two stages. In stage one we continuously pre-train and existing text-only LLMs on multi-modal data with the pipeline briefly described in 1. We then fine-tune (instruction tuning) the multimodal pre-trained LLM on the vast amount of instruction data. We discuss the datasets we used subsequently.

4.2 Pre-training Dataset.

For pre-training we follow the approach described [17]. We continuously pre-train Phi-1.5 after extending its tokenizer to support audio and image tokens for roughly same amount of data as AnyGPT [17].

4.3 Instruction Tuning Dataset.

We collect our instruction tuning data from wide range of sources. It consist multiple multimodal configurations that involves both understanding and generating multimodal data. We present summary of our data in Table 1. However, we intend to refine and extend this data further.

Dataset Type	Details
Text2Image (215K)	AnyInstruct ¹
Image2Text (273K)	Llava-instruction-mix ²
Text2Text (312K)	Alpaca [31] instruct-human-assistant-prompt ³ prosocial-dialog [32] Share-GPT-GPT4 ⁴ GPT4-LLM-Cleaned ⁵
Speech2Text(450K)	DynamicSuperb [33] - 10 tasks train splits only
Text2Speech (144K)	DynamicSuperb/Text2Speech_LibriTTS-TestOther [33] DynamicSuperb/Text2Speech_LibriTTS-TestClean [33] Single Speaker - 145K [33] LibriTTS-R ⁶
Total	1395K instruction-response pairs
Combined Dataset	∞ -Instruct-1.4M

Table 1: Any-to-Any Instruction Dataset Overview.

4.4 Baselines

For comparing our model to existing baselines, we utilize the AnyGPT[17] framework, a unified multimodal language model that leverages discretization techniques to enable cross-modal generation between text, audio, and image spaces. Built upon the LLaMA-2[3] backbone, AnyGPT provides robust natural language understanding and generation capabilities. To handle multimodal inputs effectively, the model incorporates specialized tokenizers: the SEED tokenizer[34] for image discretization, the SpeechTokenizer[15] for audio processing.

For benchmarking and evaluation, we employ the Multimodal Multilingual Machine Understanding (MMMU) dataset[35], which provides a diverse suite of multimodal data across various languages and formats. This dataset allows us to rigorously test AnyGPT’s performance in any-to-any modality generation tasks, such as text-to-image, audio-to-text, and image-to-audio generation. Using the MMMU dataset, we can effectively benchmark our model, highlighting its strengths and identifying areas for improvement in modality versatility and multilingual understanding.

Note. For mid-term update, we consider language model with extended pretraining (referred as Pretrained) as baseline and we compare against our instruction tuned model ∞ GPT on MMLU [36] in zero-shot and five-shot setting.

4.5 Evaluation

We initially aim to do the evaluation on the following benchmark, however, we intend to expand this over time to make our evaluation more comprehensive.

- **Text2Text:** We will use MMLU [36] and IFEval [37] for instruction-following evaluation.
- **Image2Text:** Initially, we will evaluate image understanding tasks using MMMU [35], with plans to expand to other multimodal benchmarks over time.
- **Text2Image:** We intend to conduct a human evaluation to assess the quality of generated images from text prompts.
- **Speech2Text:** For speech understanding, we plan to evaluate our models on over 10 tasks from DynamicSUPERB [33] that are not included in the instruction-tuning dataset.

- **Text2Speech:** We will evaluate our model’s ability to generate speech from text prompts on the test split of LibriTTS [38].

5 Results

5.1 Text Understanding

The Massive Multitask Language Understanding (MMLU) [36] benchmark is a comprehensive tool designed to evaluate language models across diverse subjects such as math, science, humanities, and social sciences through multiple-choice questions, emphasizing zero-shot performance. This benchmark offers key insights into models’ generalization abilities in real-world tasks, making it crucial for natural language processing research. In our evaluation on MMLU, we compared the zero-shot accuracy of pretrained and instruction-tuned models across varied subjects, finding that instruction tuning notably enhances performance across domains. These findings underscore the impact of tuning on improving language models’ multitask capabilities and point to areas for further model development.

The evaluation of zero-shot and five-shot performances on MMLU tasks across various domains shows that instruction-tuned models (∞ GPT) consistently outperform the pretrained baseline. In the zero-shot setting, ∞ GPT achieved the highest average accuracy of 33.7%, with particularly strong results in domains like Business, Geography, and Culture. These improvements are also evident in the five-shot setup, where our ∞ GPT reached an average accuracy of 35.9%, demonstrating robust gains across tasks such as World Religions, Health, and Computer Science. This trend indicates that instruction tuning enhances the model’s adaptability and performance in both few-shot and zero-shot settings, especially in knowledge-intensive tasks.

5.2 Multimodal Understanding

The Massive Multi-discipline Multimodal Understanding [39]

5.3 Speech Understanding

6 Division of Work

Abdul Waheed – Proposed the initial project idea and led the work on pretraining, followed by multimodal instruction data collection. Conducted experiments on instruction tuning and initial evaluations presented in this update. Also took the lead on drafting this report.

Abhigyan Kishor – Conducted essential research on baselines, contributed in-depth explanations for the report, and is currently experimenting with baseline evaluations on audio and image understanding and generation benchmarks. Actively participated in discussions about various project aspects.

Liangyu Wang – Conducted a thorough literature review and compiled findings for the paper, contributing critical insights for experiments. Engaged in project discussions and is exploring evaluations of baselines and ∞ GPT on novel benchmarks.

7 Conclusion

Task Name	Zero-Shot				5-Shot			
	LM-Base	AnyGPT	PT	∞ GPT	LM-Base	AnyGPT	PT	∞ GPT
Religions			33.9	42.7			24.6	37.4
Math			23.1	25.7			24.0	30.5
Health			24.1	38.5			24.9	39.4
Physics			24.5	28.0			26.4	32.8
Business			25.9	48.7			19.5	46.7
Biology			21.6	37.9			28.0	42.3
Chemistry			21.1	27.1			19.5	30.7
CompSci			26.2	39.6			24.8	34.2
Economics			21.2	32.6			22.5	36.0
Engineering			26.2	40.7			22.8	42.8
Philosophy			24.4	28.6			23.0	29.6
Other			24.2	38.2			24.0	39.4
History			24.3	34.7			23.8	38.5
Geography			18.2	45.5			21.7	49.0
Politics			21.3	38.0			22.1	46.0
Psychology			22.1	34.3			23.2	37.4
Culture			25.6	42.2			25.0	48.2
Law			24.5	28.5			24.2	28.9
STEM			23.6	30.7			24.7	33.9
Humanities			24.4	29.8			23.6	31.1
SocialSci			21.8	36.2			22.9	40.8
Misc.			24.4	39.8			23.8	40.4
Average			23.7	33.7			23.7	35.9

Table 2: Subject-wise zero-shot and five-shot accuracy on MMLU: Pretrained vs ∞ GPT models, including placeholders for LM-Base and AnyGPT.

References

- [1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.
- [2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
- [3] 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.
- [4] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. Mistral 7b, 2023.

- [5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023.
- [6] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.
- [7] 01. AI. :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2024.
- [8] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botet, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024.
- [9] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024.
- [10] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.
- [11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
- [12] Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. Parrot: Multilingual visual instruction tuning, 2024.
- [13] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023.

- [14] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.
- [15] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Spechtokenizer: Unified speech tokenizer for speech large language models, 2024.
- [16] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation, 2024.
- [17] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.
- [18] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report, 2023.
- [19] 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, 2021.
- [20] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.
- [21] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020.
- [22] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021.
- [23] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.
- [24] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022.
- [25] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023.
- [26] OpenAI. Gpt-4v(ision) system card, 2023.
- [27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.
- [28] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model, 2024.
- [29] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion, 2023.
- [30] Jinjie Ni, Yifan Song, Deepanway Ghosal, Bo Li, David Junhao Zhang, Xiang Yue, Fuzhao Xue, Zian Zheng, Kaichen Zhang, Mahir Shah, Kabir Jain, Yang You, and Michael Shieh. Mixeval-x: Any-to-any evaluations from real-world data mixtures, 2024.
- [31] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

- [32] Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. Prosocialdialog: A prosocial backbone for conversational agents. In *EMNLP*, 2022.
- [33] Chien yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung yi Lee. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech, 2024.
- [34] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model, 2023.
- [35] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024.
- [36] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.
- [37] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023.
- [38] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech, 2019.
- [39] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.