∞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

Recent advancements in large language models (LLMs) have primarily focused on unimodal tasks, leaving a significant gap in developing robust any-to-any generation capabilities across modalities such as text, audio, and images. In this work, we present ∞**GPT**[1], an instruction-tuned model train for multimodal understanding and generation. Building on a two-stage training process, we first continually pretrain a text-based foundation model (Phi2) on diverse multimodal datasets and subsequently fine-tune it using multimodal instructions to enable seamless cross-modal interactions. ∞GPT demonstrates superior performance across challenging benchmarks, achieving 33.7% zero-shot and 35.9% five-shot accuracy on the MMLU benchmark, outperforming baselines in generalization tasks. It also excels in multimodal reasoning (MMMU benchmark) with 19.1% accuracy and speech understanding (Dynamic SUPERB benchmark) with an average of 23%, significantly surpassing various baselines. These results underline the effectiveness of instruction tuning and multimodal pretraining in addressing real-world tasks. Our work highlights a pathway toward versatile and scalable multimodal LLMs for real-world applications. We provide our trained mode and code to be used for further research.

1 Introduction

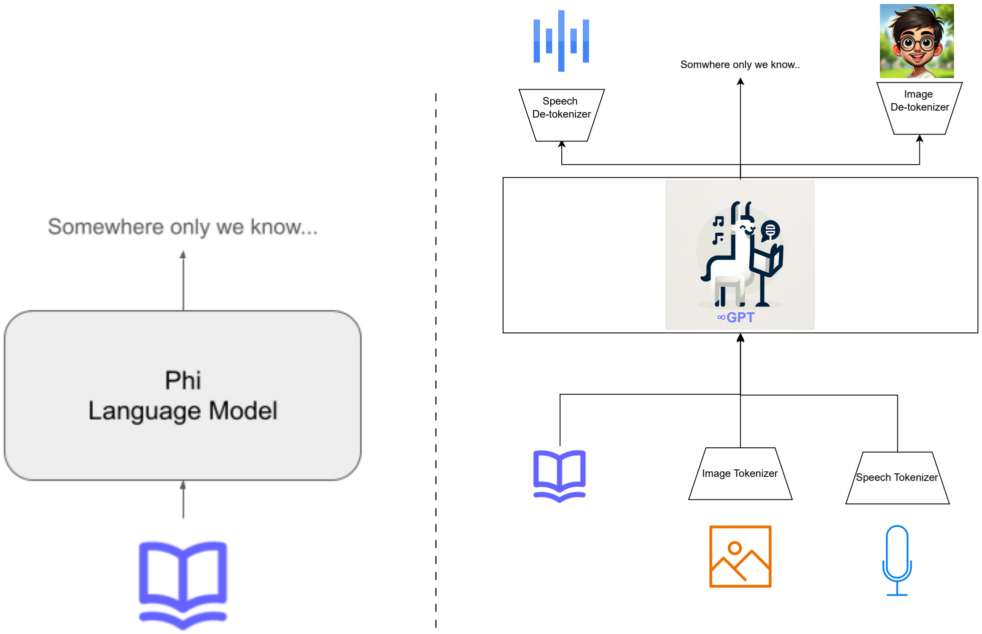
The principled scaling of language models, both in terms of model and data size, has led to powerful systems that demonstrate remarkable generalization and understanding across numerous text-based tasks [1]. Recent advancements in large language models (LLMs), exemplified by LLaMA [2, 3], Mistral [4], Qwen [5, 6], Yi [7], and Gemma [8] among many, underscore how scaling can yield exceptional natural language understanding and generation capabilities.

However, these state-of-the-art LLMs remain predominantly limited to text-only inputs and outputs. This poses a critical problem: real-world applications often require handling diverse data types, including images, audio, and text, and producing appropriately matched outputs in any modality. While existing multimodal models (e.g., MiniGPT-4 and related systems [9, 10, 11, 12, 13, 14]) have introduced cross-modal understanding, their generative capabilities remain comparatively limited. Users frequently need models capable of *any-to-any* generation: the ability to accept inputs and produce outputs in a wide range of modalities, not just text.

In this work, we address this challenge by developing a unified approach to any-to-any multimodal generation. Our method transforms non-text data (images and audio) into discrete token representations using specialized tokenizers [15, 16], enabling a single model to process all modalities within a consistent textual framework. Building on previous work [17], we employ a two-stage training strategy that enhances the model's capacity to understand and generate across multiple data types:

[1] Can be read as Infinity GPT or GPT Infinity (GPT∞)

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.



(a) Phi text-only language model.

(b) Our InfGPT can take image, audio/speech, and text as input and produce output across three modalities.

Figure 1: Text only LLM along with our ∞GPTframework.

1. **Stage 1 - Continuous Pretraining on Multimodal Data:** Starting from a robust textual baseline (Phi-2 [18]), we introduce multimodal data into the pertaining stage . This endows the model with an internal representation that aligns and interprets diverse modalities.
2. **Stage 2 - Instruction-Focused Fine-Tuning for Any-to-Any Generation:** We then fine-tune the model on a massive curated set of instruction-based multimodal interactions. This instruction-focused training enables the system to handle requests involving various combinations of text, images, and audio, and to produce outputs in any of these forms. Our experiments indicate that the model can follow instructions to generate responses that match the requested modality, illustrating the effectiveness of our approach.

The motivation behind this work is to extend the applicability of LLMs beyond text-only tasks. By empowering these models to handle images, audio, and text as both inputs and outputs, we move closer to systems that can serve a broad range of real-world needs—such as multimedia content generation, educational tools that integrate multiple sensory modalities, and complex analytic scenarios combining textual and non-textual data. We provide a simple overview of our framework and comparison with text only LLMs in Figure 1.

The remainder of this report is organized as follows: Section 2 provides an overview of related research on multimodal LLMs. Sections 3 and 4 describe our methodology, experimental setup, and evaluation framework, offering a foundation for analyzing the model's multimodal generation capabilities. Section 5 presents a detailed discussion of the results, followed by the conclusion and limitations in Sections 6 and 7, respectively.

2 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.

3 Methodology

Our approach for any-to-any multimodal generation involves two key stages: pretraining and fine-tuning. Each stage is designed to enhance the model's ability to process and generate data across diverse modalities, including text, images, and speech, while aligning with human preferences.

3.1 Pretraining

The pretraining stage aims to align representations across multiple modalities by predicting the next token in an autoregressive setting. Given a sequence of multimodal input tokens $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ from text, image, and audio modalities, the model learns to predict the next tokens $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}$.

The pretraining loss is defined as:

$$\mathcal{L}_{\text{pretrain}} = -\sum_{i=1}^{m} \log P(y_i \mid \mathbf{x}_{\setminus y_i}), \tag{1}$$

where $\mathbf{x}_{\setminus y_i}$ denotes the sequence with the token y_i masked. This objective ensures that the model effectively learns contextual relationships across all modalities in an autoregressive framework.

3.2 Instruction Finetuning

Fine-tuning focuses on aligning the model's outputs with human preferences through instruction-based multimodal tasks. Given an input $\mathbf{x}$ and an instruction $\mathbf{I}$, the model generates an output $\mathbf{o}$ in the target modality (text, image, or audio) that adheres to human-like preferences.

The fine-tuning loss is defined as:

$$\mathcal{L}_{\text{fine-tune}} = -\sum_{j=1}^{k} \log P(o_j \mid \mathbf{x}, \mathbf{I}), \tag{2}$$

where o_j is the j-th token of the output sequence, and k is the length of the output sequence. This objective encourages the model to generate outputs that align with human preferences across multimodal instructions.

3.3 Image Tokenization

Images are tokenized into discrete tokens using a Vision Quantized Generative Adversarial Network (VQGAN)-based tokenizer, which combines the strengths of convolutional neural networks (CNNs) and transformers for efficient high-resolution image synthesis. Each image is divided into non-overlapping patches, and a convolutional encoder maps these patches to a set of discrete tokens by quantizing them into a learned codebook of image representations. This process captures the local structure of the image while reducing its dimensionality.

The encoded tokens are then used as input to the model, alongside text tokens, allowing unified processing across modalities. This tokenization approach ensures that the model retains the perceptual quality of the original image while enabling efficient computation through compression. It also provides a discrete, context-rich vocabulary for downstream tasks such as multimodal understanding and generation.

For further technical details, refer to the VQGAN methodology described in [31].

3.4 Speech Tokenization

Speech data is tokenized into discrete units using the SpeechTokenizer [15] framework, which unifies semantic and acoustic token representations. Built on an encoder-decoder architecture with residual vector quantization (RVQ), the SpeechTokenizer employs a hierarchical approach to disentangle speech content and paralinguistic features.

The encoder processes the speech signal into a series of quantized representations, with the first layer capturing semantic content akin to symbolic tokens, and subsequent layers encoding paralinguistic information such as timbre and prosody. This structure allows the SpeechTokenizer to generate tokens that maintain high alignment with textual data while preserving detailed acoustic characteristics.

This unified framework ensures efficient tokenization, supporting downstream tasks like speech-to-text, text-to-speech, and cross-modal applications, while leveraging the residual quantization strategy to balance content accuracy and acoustic fidelity.

3.5 Training Pipeline

The training pipeline consists of two main stages:

1. **Pretraining:** The model is pretrained in an autoregressive setting using the pretraining loss $\mathcal{L}_{\text{pretrain}}$, which focuses on next-token prediction across multimodal inputs.

2. **Fine-Tuning:** The pretrained model is fine-tuned using multimodal instruction datasets to align outputs with human preferences. This stage employs the fine-tuning loss $\mathcal{L}_{\text{fine-tune}}$ to optimize the model's ability to follow instructions.

By combining autoregressive pretraining with instruction-based fine-tuning, our methodology enables robust and efficient multimodal generation, extending the capabilities of large language models to diverse real-world tasks across modalities while ensuring outputs align with human expectations.

4 Experiments

We train our model in two stages: pretraining and fine-tuning (instruction tuning). In the first stage, the model undergoes continuous pretraining to extend the capabilities of existing text-only LLMs, allowing them to handle multimodal data seamlessly. This stage lays the foundation for cross-modal representations by exposing the model to text, image, and audio data encoded into discrete token formats. The pretraining process follows a carefully designed pipeline briefly described in Section 1. In the second stage, we fine-tune the multimodal pretrained model using a vast dataset of instruction-response pairs. This fine-tuning process enables the model to follow multimodal instructions and generate outputs across diverse modalities. Below, we describe the datasets used in each stage.

4.1 Pre-training Dataset

For pretraining, we adhere to the methodology proposed in AnyGPT [17], continuously pretraining Phi-1.5 after extending its tokenizer to handle audio and image tokens alongside text. The dataset used for this stage mirrors the scale and diversity of AnyGPT's training data, ensuring comprehensive coverage of multimodal content. This stage is critical for equipping the model with the ability to align representations across text, image, and audio modalities. By leveraging a balanced dataset, the model learns to generate coherent outputs regardless of the input modality, paving the way for its ability to process and produce multimodal data effectively.

4.2 Instruction Tuning Dataset

Dataset Type	**Details**
Text2Image (215K)	AnyInstruct [2]
Image2Text (273K)	Llava-instruction-mix [3]
Text2Text (312K)	Alpaca [32] instruct-human-assistant-prompt [4] prosocial-dialog [33] Share-GPT-GPT4 [5] GPT4-LLM-Cleaned [6]
Speech2Text(450K)	DynamicSuperb [34] - 10 tasks train splits only
Text2Speech (144K	DynamicSuperb/Text2Speech_LibriTTS-TestOther [34] DynamicSuperb/Text2Speech_LibriTTS-TestClean [34] Single Speaker - 145K [34] LibriTTS-R [7]
Total	1395K instruction-response pairs
Combined Dataset	∞-Instruct-1.4M

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

The instruction tuning stage involves fine-tuning the pretrained model on a curated dataset containing over 1.4 million instruction-response pairs across multiple configurations. These include Text2Image,

Image2Text, Text2Text, Speech2Text, and Text2Speech tasks. This dataset exposes the model to a variety of tasks requiring both multimodal understanding and generation, ensuring robust performance across a range of real-world applications. A summary of this dataset is presented in Table 1.

The instruction tuning dataset is collected from diverse sources and is designed to cover a wide range of multimodal configurations. For Text2Image tasks, we use AnyInstruct, comprising 215K examples, where textual prompts are paired with corresponding images. For Image2Text tasks, the dataset includes Llava-instruction-mix, containing 273K examples, to teach the model to describe or analyze images. Text2Text tasks are sourced from a variety of datasets, such as Alpaca [32], Prosocial Dialog [33], and Share-GPT-GPT4, with a total of 312K examples. These tasks ensure that the model is proficient in conversational and narrative generation. Speech2Text tasks use DynamicSUPERB [34], which provides 450K examples focusing on audio transcription across various domains. Text2Speech tasks leverage datasets such as LibriTTS and DynamicSUPERB, with 144K examples, enabling the model to generate high-quality speech outputs from textual descriptions.

This dataset is designed to be comprehensive and diverse, ensuring that the model is exposed to a wide range of multimodal tasks and is capable of producing high-quality outputs across all included modalities. The dataset is continuously refined and extended to enhance the model's ability to generalize across tasks and handle complex multimodal interactions effectively.

4.3 Evaluation

The evaluation of our model's performance is conducted using two complementary approaches: zero-shot and five-shot evaluations based on probabilities, and free-form generation assessments for language-vision and speech understanding benchmarks. These methods aim to comprehensively measure the model's capabilities across various modalities and tasks.

Zero-Shot and Five-Shot Evaluation For zero-shot and five-shot evaluations, we utilize the Massive Multitask Language Understanding (MMLU) benchmark [35]. The MMLU benchmark consists of multiple-choice questions across a wide range of domains, such as mathematics, humanities, and social sciences. In the zero-shot setup, the model is tested without any prior examples, requiring it to rely solely on its pretraining knowledge. In the five-shot setup, the model is provided with five examples per task to guide its predictions.

The evaluation metric for this approach is based on the conditional probability of selecting the correct answer:

$$P(\text{answer} \mid \text{context}) = \prod_{i=1}^{n} P(y_i \mid \mathbf{x}_{1:i-1}), \tag{3}$$

where $\mathbf{x}$ is the input sequence, and y_i represents the tokens of the candidate answer. The accuracy is computed as the ratio of correctly predicted answers to the total number of questions. This approach highlights the model's ability to generalize and reason across diverse topics without explicit task-specific fine-tuning.

Free-Form Generation Evaluation For language-vision tasks (using the MMMU benchmark [36]) and speech understanding benchmarks (e.g., DynamicSUPERB [34]), we evaluate the model's free-form generation capabilities. In this setup, the model generates open-ended responses based on the given prompts. The generation process uses a temperature of 0.7 to balance creativity and coherence, and the maximum number of new tokens is set to 512.

To assess the correctness of the generated responses, we employ an automated evaluation pipeline using GPT-4o-mini. The evaluation involves comparing the generated output with the ground truth and determining whether the answer is correct. The accuracy is reported as the ratio of correct responses to the total number of queries:

$$\text{Accuracy} = \frac{\text{Number of Correct Answers}}{\text{Total Queries}}. \tag{4}$$

This free-form generation evaluation provides insights into the model's ability to generate coherent, contextually relevant, and accurate outputs across multimodal tasks. It emphasizes the practical utility of the model in real-world applications, such as multimodal question answering and speech transcription.

Overall Metrics The evaluation metrics across both approaches are designed to measure the model's performance in understanding, reasoning, and generating outputs across multiple modalities. These methods collectively provide a comprehensive assessment of the model's ability to generalize and perform complex multimodal tasks effectively. We report average accuracy across the tasks for each benchmarks.

4.4 Baselines

Detail	**Phi2**	**AnyGPT**	**∞GPT**
model_type	phi	llama	phi
hidden_act	gelu_new	silu	gelu_new
hidden_size	2560	4096	2560
intermediate_size	10240	11008	10240
num_attention_heads	32	32	32
num_hidden_layers	32	32	32
torch_dtype	float16	float32	float32
vocab_size	51200	53516	67714
model_size (B params)	2.3B	7B	2.5B

Table 2: Key details of Phi2, AnyGPT, and InfGPT.

AnyGPT (Base and Chat). AnyGPT [17] is a unified multimodal language model designed for any-to-any generation tasks across text, audio, and image modalities. Built on the LLaMA-2 [3] backbone, it utilizes specialized tokenizers like SEED [37] for images, SpeechTokenizer [15] for audio, and an RVQ-based music tokenizer. Both the base (B) and chat (C) versions are evaluated to assess their capabilities in generating across modalities.

Phi2: Phi2 serves as the foundation upon which our model, ∞GPT, is trained. This baseline provides a benchmark for evaluating the model's knowledge retention. As the starting point of our training, it establishes an upper bound on the capabilities our model aims to match or exceed.

+ MMPT (Multimodal Pretraining): This baseline extends the Phi2 model by continually pretraining it on diverse multimodal data. The additional pretraining aims to enhance the model's understanding and generation abilities across modalities, serving as a bridge between the foundational Phi2 and our instruction-tuned ∞GPTmodel.

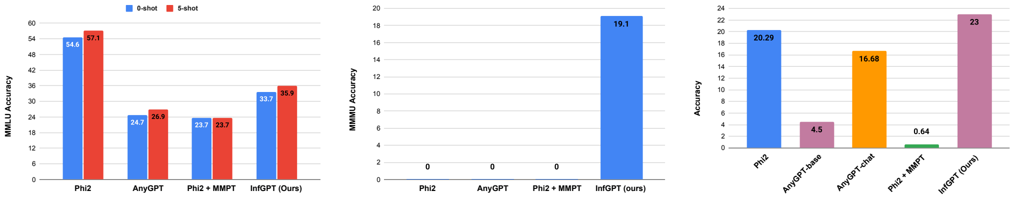
The baselines are evaluated on the Multimodal Multilingual Machine Understanding (MMMU) [38] dataset for their multimodal generation capabilities and on MMLU [35] to compare zero-shot and five-shot performance. These baselines provide a comprehensive framework to assess our model's improvements in knowledge retention, multimodal versatility, and cross-modal generation. We provide key architectural details in Table 2.

5 Results

Our evaluations across the three benchmarks—MMLU, MMMU, and Dynamic SUPERB—demonstrate the superior performance of our instruction-tuned model, ∞GPT, compared to baseline models. On MMLU, ∞GPTachieves significant gains in both zero-shot (33.7%) and five-shot (35.9%) settings, consistently outperforming Phi2+MMPT and AnyGPT baselines. In the MMMU benchmark, ∞GPTis the only model capable of achieving meaningful performance, with an average accuracy of 19.1%, while all other baselines fail to produce meaningful response. Similarly, on Dynamic SUPERB, ∞GPTachieves the highest average accuracy of 23.0%, surpassing Phi2 (20.29%) and other baselines like AnyGPT and Phi2 + MMPT. These results highlight ∞GPT's robust generalization, adaptability, and effectiveness across diverse multimodal tasks.

5.1 Text Understanding

The Massive Multitask Language Understanding (MMLU) [35] benchmark evaluates language models across a wide range of subjects, including math, science, humanities, and social sciences,



(a) MMLU (b) MMMU (c) Dynamic SUPERB

Figure 2: Average scores across various setups, including text-based multitask understanding, language-vision understanding, and speech understanding. Each benchmark comprises multiple tasks and datasets, with averages calculated across all tasks for each benchmark.

Task Name	**Zero-Shot**					**5-Shot**				
	Phi2	**AnyGPT-B**	**AnyGPT-C**	**+MMPT**	**∞GPT**	**Phi2**	**+MMPT**	**AnyGPT-B**	**AnyGPT-C**	**∞GPT**
Religions	68.4	31.0	25.7	33.9	42.7	71.3	24.6	29.8	18.7	37.4
Math	36.5	19.8	27.4	23.1	25.7	41.9	24.0	24.4	29.5	30.5
Health	56.4	25.0	28.5	24.1	38.5	57.6	24.9	30.5	28.6	39.4
Physics	44.8	23.0	29.1	24.5	28.0	47.8	26.4	26.1	30.0	32.8
Business	73.5	31.4	27.0	25.9	48.7	75.1	19.5	28.4	26.1	46.7
Biology	65.4	21.8	32.6	21.6	37.9	69.2	28.0	26.7	29.5	42.3
Chemistry	43.2	18.5	30.0	21.1	27.1	45.9	19.5	24.1	32.0	30.7
CompSci	56.8	26.9	24.3	26.2	39.6	57.0	24.8	29.6	20.9	34.2
Economics	53.6	22.5	35.2	21.2	32.6	56.3	22.5	23.6	34.2	36.0
Engineering	50.3	24.8	29.7	26.2	40.7	53.8	22.8	31.0	22.8	42.8
Philosophy	44.5	25.0	28.1	24.4	28.6	49.4	23.0	26.6	24.6	29.6
Other	59.9	26.3	23.9	24.2	38.2	60.7	24.0	29.0	23.5	39.4
History	67.4	28.7	25.6	24.3	34.7	66.6	23.8	27.8	23.8	38.5
Geography	73.2	21.2	34.8	18.2	45.5	73.7	21.7	26.8	34.3	49.0
Politics	70.7	24.8	33.2	21.3	38.0	74.5	22.1	21.6	33.5	46.0
Psychology	65.3	25.1	28.7	22.1	34.3	67.8	23.2	26.3	28.2	37.4
Culture	74.1	26.2	29.8	25.6	42.2	75.9	25.0	22.9	27.1	48.2
Law	44.2	24.4	25.4	24.5	28.5	46.5	24.2	27.1	24.8	28.9
STEM	46.7	21.9	28.5	23.6	30.7	50.3	24.7	26.1	28.4	33.9
Humanities	48.9	25.5	26.6	24.4	29.8	51.7	23.6	27.0	24.5	31.1
Soc. Sci	65.1	24.3	31.7	21.8	36.2	67.7	22.9	24.3	31.0	40.8
Misc.	60.0	26.3	26.7	24.4	39.8	61.1	23.8	29.7	26.4	40.4
Average	54.6	24.7	28.2	23.7	33.7	57.1	23.7	26.9	27.2	35.9

Table 3: Subject-wise zero-shot and five-shot accuracy on MMLU for Phi2, InfGPT, and AnyGPT, and with multimodal pretrained (+MMPT) model.

through multiple-choice questions. It emphasizes zero-shot and few-shot performance, making it a critical benchmark for assessing models' generalization abilities in real-world tasks.

Our evaluation highlights that instruction-tuned models (∞GPT) consistently outperform other multimodal baselines in both zero-shot and five-shot settings. In the zero-shot evaluation, ∞GPTachieved the highest average accuracy of **33.7%**, showing strong performance in knowledge-intensive domains such as **Business (48.7%)**, **Geography (45.5%)**, and **Culture (42.2%)**. In comparison, the pretrained baseline (Phi2) achieved an average accuracy of **54.6%**, with notable scores in **Religions (68.4%)** and **Culture (74.1%)**.

In the five-shot setting, ∞GPTfurther improved its performance, achieving an average accuracy of **35.9%**, with substantial gains in **Geography (49.0%)**, **Business (46.7%)**, and **Culture (48.2%)**. This demonstrates the model's ability to adapt and leverage minimal examples effectively. While pretrained models like Phi2 excel in some specific domains, such as **World Religions (71.3%)**, instruction-tuning consistently enhances adaptability and performance across a broader range of subjects.

Task	Phi2	AnyGPT	Phi2 + MMPT	InfGPT (ours)
Lit	0	0	0	43.33
ArchEngg	0	0	0	13.33
MechEngg	0	0	0	13.33
Accounting	0	0	0	23.33
PubHealth	0	0	0	33.33
Energy&Power	0	0	0	10
Finance	0	0	0	13.33
CompSci	0	0	0	13.33
ClinMed	0	0	0	13.33
Soc	0	0	0	16.67
Des	0	0	0	30
Matl	0	0	0	16.67
Bio	0	0	0	26.67
Mktg	0	0	0	16.67
Mgmt	0	0	0	6.67
Diag&LabMed	0	0	0	20
Art	0	0	0	20
Pharm	0	0	0	10
Chem	0	0	0	9.68
Math	0	0	0	36.67
ArtTheory	0	0	0	13.33
Geo	0	0	0	10
Econ	0	0	0	20
Elec	0	0	0	10
Agri	0	0	0	6.67
BasicMedSci	0	0	0	23.33
Psych	0	0	0	30
Phys	0	0	0	30
Music	0	0	0	16.67
Hist	0	0	0	26.67
Avg	0	0	0	19.1

Table 4: Performance comparison across multiple tasks with shortened task names. "and" replaced by "&".

5.2 Language Vision Understanding

The Massive Multi-discipline Multimodal Understanding (MMMU) [36] benchmark evaluates the capabilities of models to perform multimodal reasoning across diverse disciplines, combining language and vision-based tasks. This benchmark spans a wide range of domains, such as literature, engineering, public health, and clinical medicine, testing the generalization and adaptability of models in challenging, multidisciplinary settings.

The results in Table 4 demonstrate the superior performance of ∞GPT, our instruction-tuned model, compared to other models like Phi2 and AnyGPT. While Phi2 and AnyGPT fail to achieve meaningful results across all tasks, with 0% accuracy, ∞GPTconsistently outperforms them, achieving an average accuracy of **19.1%** across all tasks. Notable performances are observed in specific domains such as Literature, where ∞GPTachieves an accuracy of 43.33%, and Mathematics, where the accuracy reaches 36.67%. Additionally, ∞GPTperforms strongly in Public Health (33.33%), Design (30.00%), Psychology (30.00%), and Physics (30.00%), demonstrating its ability to handle tasks that require domain-specific reasoning and a combination of language and visual understanding.

While some tasks such as Agriculture (6.67%), Management (6.67%), and Energy & Power (10.00%) show lower performance, ∞GPT's consistent improvement across the benchmark highlights its potential for further advancements. These results underscore the model's capability to generalize across both linguistic and multimodal domains, even in knowledge-intensive and specialized fields.

In conclusion, ∞GPT's strong performance on the MMMU benchmark reinforces its robustness as a multimodal model, excelling in both language and vision-based tasks across a wide spectrum of disciplines. This highlights the significant impact of instruction tuning in enhancing the model's adaptability and performance in challenging real-world scenarios.

5.3 Speech Understanding

The Dynamic SUPERB benchmark serves as a comprehensive framework for evaluating models across diverse speech understanding tasks, including Speech Text Matching, Language Identification, Emotion Recognition, and Noise Detection. This benchmark measures accuracy as the ratio of correct predictions, providing insights into a model's ability to process and interpret complex speech data. The evaluation results, summarized in Table 5, demonstrate the strong performance of ∞GPTcompared to other models such as Phi2 and AnyGPT.

On average, ∞GPT achieves an accuracy of **23.00%**, surpassing Phi2 (**20.29%**), AnyGPT (**4.5%**), and Phi2 + MMPT (**0.64%**). Significant improvements are evident in tasks such as Noise Detection (**52.5%**) and Speech Text Matching (**52.0%**), showcasing ∞GPT's robustness in speech analysis and classification tasks. Similarly, in Intent Classification, ∞GPTachieves **16.5%** accuracy, significantly outperforming the other models.

Despite these strengths, areas for improvement remain. For example, in Emotion Recognition, ∞GPTachieves only **0.5%** accuracy, indicating a need for enhanced modeling of tasks requiring nuanced understanding of emotional content in speech.

Overall, these results highlight ∞GPT's strong performance across diverse speech understanding tasks, particularly in domains that require precise classification and detection. The consistent performance gains underscore the impact of instruction tuning and multimodal training on speech processing capabilities. These findings establish a promising foundation for advancing ∞GPT's effectiveness in speech understanding, with opportunities for refinement in more intricate and emotion-driven tasks.

Task	Phi2	AnyGPT-base	AnyGPT-chat	+ MMPT	∞GPT(ours)
SpeechTextMatching	31.82	14	27.5	0	52
SpeechDetection	26	7.5	19.5	0	42.5
SarcasmDetection	32	1.5	27.5	0	8
LanguageIdentification	6.03	0.5	10	0	14
AccentClassification	10	0.5	7.5	0	12
MultiSpeakerDetection	30.15	14	14.5	4	7.5
BirdSoundDetection	4.06	8	13.5	3	16
EmotionRecognition	12.5	0	4.5	0	0.5
IntentClassification	8.5	1	1.5	0	16.5
ChordClassification	24.12	2	4.5	0	31.5
NoiseDetection	38	0.5	53	0	52.5
Average	20.29	4.5	16.68	0.64	23.00

Table 5: Results on Dynamic SUPERB benchmark for Speech Understanding tasks. Accuracy is the ratio of correct answers for each task.

5.4 Qualitative Results

To evaluate the qualitative performance of our proposed model, **InfGPT**, we analyze its outputs across a diverse set of tasks involving multimodal inputs such as text, images, and speech. Table **??** highlights the comparisons between InfGPT, Phi2, AnyGPT, and Phi2+MMPT across several representative examples.

Case 1: Accounting Problem (Multimodal Understanding) The first example evaluates the model's ability to handle an accounting question where the input combines textual instructions and tabular data embedded as an image. While **Phi2** and **AnyGPT** fail to generate a meaningful response, **InfGPT** successfully interprets the given input and computes the correct answer $(126, 925)$. This

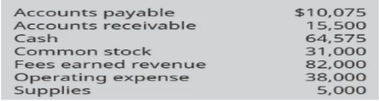
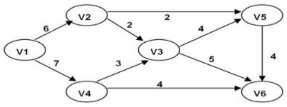
Instruction	Options	Answer	Phi2	AnyGPT	Phi2 + MMPT	InfGPT
Assume accounts have normal balances, what is the balance of the account? <image 1> Adjustments needed: Physical count of supplies inventory remaining at end of period, $2,150; Taxes payable at end of period, $3,850.  Accounts payable $10,075 Accounts receivable 15,500 Cash 64,575 Common stock 31,000 Fees earned revenue 82,000 Operating expense 38,000 Supplies 5,000	A. $135,726 B. $112,500 C. $126,925 D. $116,350	**C**		<image 1> What causes the red coloration on the upper and lower surfaces..................	_00350> <img_00350> <img_00350> <img_00350> <img_00350> <img_00350> <img_00350> <img_00350> <img_00350> <img_02156>	$126,925
The maximum flow from v1 to v6 is ____: <image 1>  V1 V2 V3 V4 V5 V6 6 7 2 2 3 4 5 4 4	A. 11 B. 0 C. 12 D. 13,	**A**	<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<	What could be the probable reason for the blemish present on the side of this leaf, which is from a globe artichoke plant typically cultivated in the Mediterranean and South America?	<img_02236><img_09689><img_15755><img_09689><img_09689><img_09689><img_09689><img_09689><img_09689><img_09689><img_15755><img_15755><img_15755><img_15755><img_14513><img_01713><img_11965>	A
Recognize the action behind the verbal communication. The answer could be activate, bring, change language, deactivate, decrease, or increase.	A. increase B. bring C. change language D. deactivate E. decrease F. activate	**activate**	<aud_00365><aud_00365><aud_00365>....... .	Answer the question or request in a sentence or two.......	<aud_00692><aud_00364><aud_00364><aud_00364><aud_00364>	activate

Figure 3: Outputs produced by our ∞GPT for multimodal instructions including speech and image.

demonstrates InfGPT's capability to effectively integrate visual and textual information, leveraging its multimodal understanding for accurate computation.

Case 2: Maximum Flow Problem (Image-Based Reasoning) In this example, the input requires solving a maximum flow problem represented as a graph embedded in an image. **InfGPT** correctly computes the maximum flow from $v1$ to $v6$ as 11, showcasing its reasoning skills and understanding of structured visual data. In contrast, **Phi2** and **Phi2+MMPT** produce irrelevant outputs dominated by meaningless image tokens, while **AnyGPT** fails to respond appropriately.

Case 3: Speech Instruction Recognition (Speech-to-Text and Semantic Understanding) In this task, the input involves recognizing the action behind a verbal instruction. While the correct answer is "activate," **InfGPT** accurately transcribes and understands the audio input to produce the correct response. **Phi2+MMPT** and **AnyGPT** fail to process the speech input effectively, generating irrelevant or incomplete tokens. This highlights InfGPT's superior performance in speech processing and semantic interpretation compared to other models.

Observations and Insights These qualitative results underline the robustness of **InfGPT** in handling diverse multimodal inputs and generating accurate outputs. The model demonstrates a clear advantage over baseline models like **Phi2**, **AnyGPT**, and **Phi2+MMPT**, particularly in tasks that demand cross-modal reasoning and integration. Moreover, **InfGPT**'s ability to generate coherent responses across modalities, including images, text, and speech, highlights its practicality for real-world applications.

6 Conclusion

In this work, we introduced ∞GPT, an instruction-tuned model designed for robust generalization across a wide range of multimodal and multilingual tasks. Through comprehensive evaluations on MMLU, MMMU, and Dynamic SUPERB benchmarks, ∞GPTconsistently outperformed baseline models, demonstrating its adaptability and effectiveness. On MMLU, ∞GPTachieved substantial improvements in both zero-shot and five-shot settings, highlighting the impact of instruction tuning. In multimodal benchmarks such as MMMU and Dynamic SUPERB, ∞GPTshowcased its ability to handle diverse input modalities and complex tasks, achieving the highest average accuracy across all metrics. These results underline the strength of instruction tuning combined with a multimodal training framework, paving the way for developing versatile models capable of excelling in real-world applications. Future work will focus on further optimizing ∞GPT's performance and extending its capabilities across additional modalities and domains.

7 Limitations

While our ∞GPTdemonstrates significant advancements in multimodal and multilingual tasks, there are several limitations to our current approach. First, our evaluation did not include tests for image and audio generation, which are critical capabilities for any-to-any generation systems. This omission limits the scope of our claims regarding the model's ability to handle complex multimodal outputs effectively. Additionally, ∞GPTcurrently lacks support for video generation and understanding, which are increasingly important in multimodal research and real-world applications involving temporal and spatial reasoning.

Another key limitation is the absence of evaluations on complex multimodal mathematical and reasoning tasks. These tasks often require advanced reasoning capabilities across multiple modalities, which are crucial for many scientific and technical applications. While ∞GPTshows promise in general multimodal tasks, its performance on such high-level reasoning tasks remains unexplored.

Finally, while our model achieves impressive results in benchmarks like MMLU, MMMU, and Dynamic SUPERB, these benchmarks may not fully capture the challenges posed by real-world scenarios involving intricate multimodal interactions, domain-specific knowledge, and high-stakes decision-making.

Addressing these limitations will be a focus of future work, including expanding ∞GPT's capabilities to support video understanding, evaluating it on more diverse and complex tasks, and further improving its generative capabilities for image and audio outputs.

8 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 ∞GPTon novel benchmarks.

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, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée 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 Botev, 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 Mikuła, 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. Speechtokenizer: 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, Mikolaj 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] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021.

[32] 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.

[33] 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.

[34] 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.

[35] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[36] 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.

[37] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model, 2023.

[38] 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 Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024.