MIO: A FOUNDATION MODEL ON MULTIMODAL TOKENS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we introduce MIO, a novel foundation model built on multimodal tokens, capable of understanding and generating speech, text, images, and videos in an end-to-end, autoregressive manner. While the emergence of large language models (LLMs) and multimodal large language models (MM-LLMs) propels advancements in artificial general intelligence through their versatile capabilities, they still lack true any-to-any understanding and generation. Recently, the release of GPT-4o has showcased the remarkable potential of any-to-any LLMs for complex real-world tasks, enabling omnidirectional input and output across images, speech, and text. However, it is closed-source and does not support the generation of multimodal interleaved sequences. To address this gap, we present MIO, which is trained on a mixture of discrete tokens across four modalities using causal multimodal modeling. MIO undergoes a four-stage training process: (1) alignment pre-training, (2) interleaved pre-training, (3) speech-enhanced pre-training, and (4) comprehensive supervised fine-tuning on diverse textual, visual, and speech tasks. Our experimental results indicate that MIO exhibits competitive, and in some cases superior, performance compared to previous dual-modal baselines, any-to-any model baselines, and even modality-specific baselines. Moreover, MIO demonstrates advanced capabilities inherent to its any-to-any feature, such as interleaved video-text generation, chain-of-visual-thought reasoning, visual guideline generation, instructional image editing, etc. Anonymous codes and supplemental materials are available at `https://anonymous.4open.science/r/anonymous_MIO-DDE5`.

1 INTRODUCTION

The advent of Large Language Models (LLMs) is commonly considered the dawn of artificial general intelligence (AGI) (OpenAI et al., 2023; Bubeck et al., 2023), given their generalist capabilities such as complex reasoning (Wei et al., 2022), role playing (Wang et al., 2023c), and creative writing (Wang et al., 2024a). However, original LLMs lack multimodal understanding capabilities. Consequently, numerous multimodal LLMs (MM-LLMs) have been proposed, allowing LLMs to understand images (Li et al., 2023b; Alayrac et al., 2022), audio (Borsos et al., 2023; Rubenstein et al., 2023; Tang et al., 2023; Das et al., 2024), and other modalities (Lyu et al., 2023; Zhang et al., 2023d; Moon et al., 2023). These MM-LLMs typically involve an external multimodal encoder, such as EVA-CLIP (Sun et al., 2023b) or CLAP (Elizalde et al., 2022), with an alignment module such as Q-Former (Li et al., 2023b) or MLP (Liu et al., 2023b) for multimodal understanding. These modules align non-textual-modality data features into the embedding space of the LLM backbone.

Another line of work involves building **any-to-any** and end-to-end MM-LLMs that can input and output non-textual modality data. Typically, there are four approaches: (1) Discrete-In-Discrete-Out (DIDO): Non-textual modality data is discretized using vector quantization techniques (van den Oord et al., 2017; Esser et al., 2020) and then fed into LLMs (Ge et al., 2023b; Zhan et al., 2024; Liu et al., 2024). (2) Continuous-In-Discrete-Out (CIDO): The LLM backbones intake densely encoded non-textual modality data features and generate their quantized representations (Diao et al., 2023; Team et al., 2023). (3) Continuous-In-Continuous-Out (CICO): The LLMs both understand and generate non-textual modality data in their densely encoded representations (Sun et al., 2023c;a; Dong et al., 2023; Zheng et al., 2023; Wu et al., 2023). (4) Autoregression + Diffusion (AR + Diff): The autoregressive and diffusion modeling are integrated in a unified LLM (Zhou et al., 2024; Xie

Table 1: The comparison between previous models and MIO (ours). **I/O Consistency** indicates whether the model ensures that the input and output representations for the same data remain consistent. **Uni. Bi. SFT** refers to whether the model undergoes a unified (Uni.) supervised fine-tuning (SFT) for both multimodal understanding and generation (Bi.=Bidirectional). **Multi-Task SFT** assesses whether the model undergoes a comprehensive SFT that includes diverse tasks, with at least visual question answering tasks. **MM. Inter. Output** evaluates whether the model supports the generation of multimodal interleaved (MM. Inter.) sequences. We refer readers to §1 for the definitions of the different modeling approaches.

Models	Emu1 (Sun et al., 2023c)	Emu2 (Sun et al., 2023a)	SEED-LLaMA (Ge et al., 2023b)	AnyGPT (Zhan et al., 2024)	CM3Leon (Yu et al., 2023), Chameleon (Team, 2024)	Gemini (Reid et al., 2024)	Transfusion (Zhou et al., 2024)	MIO (ours)
I/O Consistency	✗	✓	✓	✓	✓	✗	✗	✓
Uni. Bi. SFT	✗	✗	✓	✓	✓	✓	✗	✓
Multi-Task SFT	✓	✓	✓	✗	✓	✓	✗	✓
Speech I/O	✗/✗	✗/✗	✗/✗	✓/✓	✗/✗	✓/✗	✗	✓/✓
Video I/O	✓/✓	✓/✓	✓/✓	✗/✗	✗/✗	✓/✗	✗	✓/✓
Voice Output	✗	✗	✗	✗	✗	✗	✗	✓
MM. Inter. Output	✗	✗	✓	✗	✗	✗	✗	✓
Modeling	CICO	CICO	DIDO	DIDO	DIDO	CIDO	AR+Diff	DIDO

et al., 2024; Li et al., 2024b). Although these works have succeeded in building MM-LLMs unifying understanding and generation, they exhibit some drawbacks, as illustrated in Table 1. For example, Emu1 (Sun et al., 2023c) and Emu2 (Sun et al., 2023a) explore the autoregressive modeling of three modalities: text, images, and videos. SEED-LLaMA (Ge et al., 2023b) proposes a new image quantizer aligned with LLMs' embedding space and trains the MM-LLMs on images and videos. However, neither considers the speech modality, which is heterogeneous from visual modalities like videos and images. Although AnyGPT (Zhan et al., 2024) has explored settings involving four modalities, including text, image, speech, and music, it lacks video-related abilities, voice synthesis, and comprehensive multi-task supervised fine-tuning, leading to limited multimodal instruction-following and reasoning capabilities. Furthermore, AR + Diff approaches, such as Transfusion (Zhou et al., 2024), suffer from limited multimodal understanding capabilities because the multimodal inputs are noised for denoising modeling, and the image tokenizer used (*i.e.*, VAE (Kingma & Welling, 2013)) is suitable for image generation rather than image understanding.

Moreover, most of current MM-LLMs are typically dual-modal, combining text with another modality, such as images. Although previous works, such as Meta-Transformer (Zhang et al., 2023d) and Unified-IO 2 (Lu et al., 2023), have explored omni-multimodal understanding settings with more than two non-textual modalities, they still lag significantly behind their dual-modal counterparts, especially in terms of multimodal instruction-following capabilities. Moreover, these MM-LLMs are typically focused on understanding only, neglecting the important aspect of multimodal generation. Several works have enabled LLMs to call external tools to address this issue. For example, HuggingGPT (Shen et al., 2023) generates textual image descriptions for external diffusion models to synthesize images. GPT-4 (OpenAI et al., 2023) can utilize either an image generator like DALL-E 3 (Betker et al., 2024) or a text-to-speech (TTS) tool like Whisper (Radford et al., 2022) to support multimodal generation.[1] However, these methods are not end-to-end, relying on the text modality as an interface.

Recently, the release of GPT-4o has demonstrated the capabilities of any-to-any and end-to-end foundation models.[2] It is the first foundational model to accept multimodal tokens as inputs and generate multimodal tokens within a unified model while also demonstrating strong abilities in complex multimodal instruction-following, reasoning, planning, and other generalist capabilities. Furthermore, as the continuous scaling up of LLMs in the community depletes high-quality language tokens, GPT-4o verifies a new source of data for LLM training: multimodal tokens. This approach suggests that the next generation AGI could derive more knowledge from multimodal tokens when language tokens are exhausted. However, GPT-4o is closed source and focuses primarily on end-to-end support for speech I/O, image I/O, 3D generation, and video understanding. Its recent open-source "alternatives", such as VITA (Fu et al., 2024), still lack the ability to *generate* data of all supported modalities, particularly for the generation of multimodal interleaved sequences.

[1] https://openai.com/index/chatgpt-can-now-see-hear-and-speak/
[2] https://openai.com/index/hello-gpt-4o/

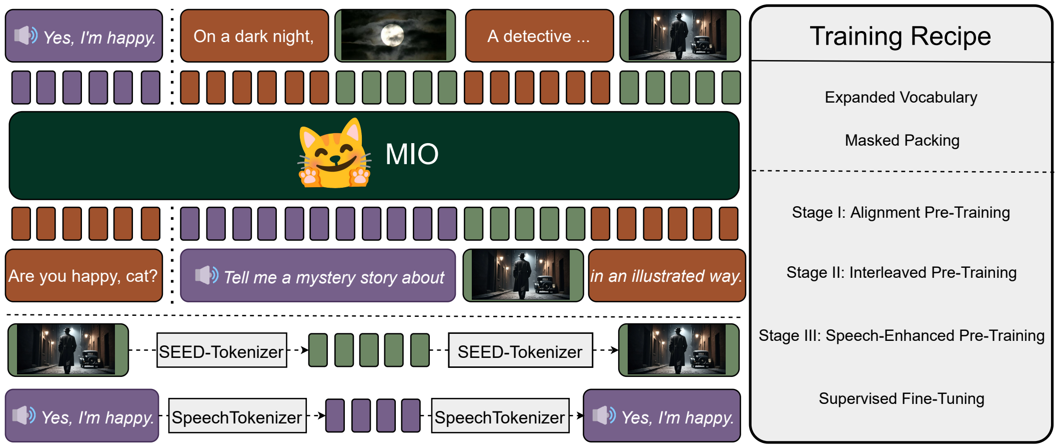

Figure 1: The framework of MIO and its training recipe.

To address the aforementioned issues, we introduce MIO (Multimodal Input and Output, or Multimodal Interleaved Output), the first open-source any-to-any foundation model that unifies multimodal understanding and generation across four modalities–text, image, speech (with voice), and video, while enabling the generation of multimodal interleaved sequences. Specifically, MIO is built on discrete multimodal tokens that capture both semantic representations through contrastive loss and low-level features via reconstruction loss (Ge et al., 2023a; Zhang et al., 2023b) from raw multimodal data. Due to the consistent data format shared with textual corpora, the model can treat non-textual modalities as "foreign languages", allowing it to be trained with the next-token-prediction. Note that since the representation of an image remains the same whether it is used as an input or an output, our model flexibly supports multimodal interleaved sequence generation, where an image functions simultaneously for both understanding and generation. Moreover, we employ three-stage pre-training with an additional SFT stage to effectively train the model for modality scaling.

Our experimental results show that MIO, trained on a mixture of four modalities, demonstrates competitive performance compared to its dual-modal counterparts and previous any-to-any multimodal language model baselines. Additionally, MIO is the first model to demonstrate interleaved video-text generation, chain-of-visual-thought reasoning, and other emergent abilities relying on any-to-any and multimodal interleaved output features (*c.f.*,§3.5).

2 METHOD

Firstly, we elaborate on our modeling approach, which supports multimodal token input and output, as well as causal language modeling (CausalLM), in §2.1. Secondly, we describe our three-stage pre-training procedures in §2.2. Thirdly, we provide details of our comprehensive supervised fine-tuning on diverse multimodal understanding and generation tasks in §2.3.

2.1 MODELING

As illustrated in Figure 1, the framework of MIO involves three parts: (1) multimodal tokenization, (2) causal multimodal modeling, and (3) multimodal de-tokenization.

Multimodal Tokenization. In our work, we use SEED-Tokenizer (Ge et al., 2023a) as our image tokenizer and SpeechTokenizer (Zhang et al., 2023b) as our speech tokenizer. SEED-Tokenizer encodes images using a ViT (Dosovitskiy et al., 2021) derived from BLIP-2 (Li et al., 2023b), and then converts the encoded features into fewer tokens with causal semantics via Q-Former (Li et al., 2023b). These features are subsequently quantized into discrete tokens that are well-aligned with the language model backbone's textual space. The codebook size for these discrete image tokens is 8192. SEED-Tokenizer transforms each image into a 224x224 resolution and quantizes it into 32 tokens. We use two special tokens, <IMAGE> and </IMAGE>, to indicate the start and end of the image tokens per image, respectively.

As for videos, we first apply specific frame-cutting methods to convert videos into image sequences. In our training data processing procedures, the number of frames for each video is dynamically determined by its duration, the length of its context, or its scene switching[3] to (1) avoid exceeding the LLM backbone's context window limit, and (2) capture complete but concise information of the video. Each frame is then tokenized in the same manner as an image.

In terms of speech, SpeechTokenizer (Zhang et al., 2023b) leverages an 8-layer RVQ (Lee et al., 2022) to tokenize speech into tokens with 8 codebooks, with each codebook derived from one layer. Since the first layer's quantization output is distilled from HuBERT (Hsu et al., 2021), which encodes more semantic information, SpeechTokenizer can separate content tokens and timbre tokens from a quantized speech. The first-layer quantization is treated as content quantization, while the remaining layers' quantization is treated as timbre quantization. SpeechTokenizer encodes speech into 50 tokens per second for each codebook, resulting in 400 tokens per second with all eight codebooks. To improve context efficiency, we drop the last four layers' codebooks and only use the content codebook and the first three timbre codebooks. Our vocabulary size for the speech modality is $1024 \times 4 = 4096$.

Since the open-source pretraining-level speech data is collected from individuals with diverse voices, the timbre tokens exhibit a relatively random and noisy pattern, while the content tokens are more fixed-pattern and better aligned with the corresponding transcriptions. Given these priors in speech tokens, it is important to choose the proper interleaving mode of speech tokens (Copet et al., 2023). We denote the four codebooks as $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$, where $\mathcal{A}$ is the codebook for content tokens and the remaining three are for timbre tokens. For simplicity, assuming that we have only two tokens for each codebook in a tokenized speech sequence (i.e., a_1a_2, b_1b_2, c_1c_2, and d_1d_2), there are two interleaving patterns for causal multimodal modeling: (1) sequential interleaving pattern: $a_1a_2b_1b_2c_1c_2d_1d_2$ and (2) alternating interleaving pattern: $a_1b_1c_1d_1a_2b_2c_2d_2$.

In our preliminary experiments, we observed that text-to-speech generation (TTS) training is difficult to converge when using the alternating interleaving pattern because the noisy and random timbre tokens ($b_1c_1d_1$) tend to mislead the continuations. Moreover, the speech-to-text understanding (ASR) performance improves much more slowly during training with the alternating interleaving pattern due to the sparsity of semantic information in the timbre tokens. As a result, we drop the timbre tokens for speech understanding and use the sequential interleaving pattern for speech generation. We use <SPCH> and </SPCH> as special tokens to indicate the start and end of the speech token sequence.

Causal Multimodal Modeling. As illustrated in Figure 1, the speech and images, including video frames, are tokenized by SpeechTokenizer (Zhang et al., 2023b) and SEED-Tokenizer (Ge et al., 2023a), respectively. We add the 4096 speech tokens and 8192 image tokens to the LLM's vocabulary. In addition, we introduce four new special tokens, namely <IMAGE>, </IMAGE>, <SPCH>, and </SPCH>, to the vocabulary. Consequently, the embedding layer of the LLM backbone and the language modeling head are extended by $4096 + 8192 + 4 = 12292$ to support the embedding and generation of these new tokens. The image tokens contain *causal* semantics due to the use of a *Causal* Q-Former (Ge et al., 2023a), and the speech tokens are intrinsically causal due to their temporal nature. Therefore, these multimodal tokens are as suitable for autoregressive training as textual tokens, allowing us to unify the training objectives for understanding and generation of multimodal tokens into next-token-prediction with cross-entropy loss. The training objective is thus:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t}; \theta) \tag{1}$$

where x_t represents the discrete multimodal tokens, and θ denotes the parameters of the LLM backbone. We use the pre-trained model, Yi-6B-Base (AI et al., 2024), for initialization.

Furthermore, to eliminate the computational inefficiency caused by <PAD> tokens, we use the masked packing strategy (Lu et al., 2023; Liu et al., 2024; Dehghani et al., 2023). Specifically, the samples are concatenated along the sequence length dimension until the context window is full. Then, we

[3] https://github.com/Breakthrough/PySceneDetect

construct the causal attention mask for the tokens of each sample and mask out all the tokens of the other samples.

Multimodal De-Tokenization. After the generation of multimodal tokens, it is essential to use modality-specific decoders to reconstruct the images or speech from the codes. Specifically, for image tokens, we directly utilize SEED-Tokenizer's decoder, which involves an MLP projection to convert the discrete codes into dense latents. These latents condition an off-the-shelf diffusion model (Rombach et al., 2022) to generate the images in the pixel space (Ge et al., 2023a). The vanilla SpeechTokenizer (Zhang et al., 2023b) involves generating timbre tokens through a *non-autoregressive* model outside the language model, and then feeding the concatenated content and timbre tokens into the SpeechTokenizer decoder to synthesize speech. In our work, to inject the timbre priors into the multimodal language model itself, the timbre tokens are also generated by the *autoregressive* language model.

2.2 Pre-Training

As shown in Table 2, we use a three-stage strategy for pre-training, with each stage targeting different objectives. The three stages are: (1) Alignment Pre-training: This stage focuses on learning a multimodal representation more aligned with the language space. (2) Interleaved Pre-training: This stage aims to obtain a multimodal representation with richer contextual semantics. (3) Speech-enhanced Pre-training: This stage specifically enhances the model's speech-related capabilities, while concurrently replaying data from other modalities. For more details on the pre-training data and its processing procedures, we refer the readers to Appendix A.

Table 2: Pre-training stages and their details. We use "Inter" to denote "Interleaved" for short. We provide batch sizes for each data type per GPU in image-text pair data:language-only data:(image-text interleaved data + video data):speech-text pair data. See Appendix A and Appendix B for more details including pre-training data sources, data cleaning procedures, pre-training hyperparameters, etc.

Pre-training Stage **Objective**	**Stage I** Multimodal Alignment	**Stage II** Multimodal Interleaving	**Stage III** Speech Enhancement
Image-Text Pair	SBU, CC3M, LAION-COCO, JourneyDB	SBU, CC3M, LAION-COCO, JourneyDB	CC3M LAION-COCO
Language-Only	RefinedWeb	RefinedWeb	RefinedWeb
Image-Text Inter	-	OBELICS, MMC4-core-ff	MMC4-core-ff
Video-Text Pair	-	WebVid-10M	WebVid-10M
Video-Text Inter	-	HowTo-100M, YT-Temporal-180M	HowTo-100M, YT-Temporal-180M
Speech-Text Pair	Libriheavy	Libriheavy	Libriheavy
GPUs **Training Steps** **Batch Size**	128 A800-80GB 24,800 12:2:0:2	128 A800-80GB 12,800 2:2:6:6	8 A800-80GB 32,200 2:1:1:12

Stage I: Alignment Pre-Training. To fully leverage the superior capabilities of the pre-trained LLM backbone, it is essential to align the non-textual modality data representations with text. There are two types of pre-training data for image-text multimodal learning: (1) Image-text paired data: This data has well-aligned dependencies between images and text. (2) Image-text interleaved data: This data features more natural and contextual dependencies but is less aligned. Note that in our setting, video-text paired and interleaved data can be treated as image-text interleaved data, with videos being sequential images interleaved with text. Therefore, in this stage, we exclude the image-text interleaved data and video data to ensure the most aligned pattern between images and text.

Stage II: Interleaved Pre-Training. In this stage, we extend the data used for pre-training to include image-text interleaved data (including video-text data) as a novel image-text dependency pattern. The image-text interleaving pattern has a different nature compared to pairing patterns. Although Li et al. (2023b) and Sun et al. (2023c) argued that interleaved image-text data mainly serves for *multimodal in-context learning*, we argue that it is also essential for context-aware image generation where images are generated based on specific context, rather than a precise description of the image content. For example, in image-text interleaved data, the text might serve as the image's preceding or continuing context, rather than its description. This pattern significantly differs from the previous descriptive image generation demonstrated in image-text paired data, where images are generated based on precise and detailed text that clearly describe the content of the images (Team et al., 2023). Therefore, context-aware image generation is essential for tasks such as *chain-of-visual-thought reasoning* or *visual storytelling* (Team et al., 2023; Huang et al., 2016), where images are generated without textual descriptions. Due to the lack of benchmarks and evaluation metrics for context-aware image generation, we provide some demonstrations in §3.5 to showcase the potential of our model in visual storytelling, interleaved video-text generation, instructional image editing, chain-of-visual-thought reasoning, multimodal in-context learning, etc.

Moreover, in this stage, due to the extensive training on image-text paired data in Stage I, we can reduce its mixing ratio to the minimal essential scale for replay to avoid catastrophic forgetting. This allows us to increase the batch size for image-text interleaved data, video data, and speech data.

Stage III: Speech-Enhanced Pre-Training. The speech tokenizer that we use generates 200 tokens for each second of audio. Given that the duration of a speech sample can be 15 seconds, this results in around 3,000 tokens per sample. In comparison, the image tokenizer produces only 32 tokens per image. This creates a significant disparity in the number of tokens among different modalities. Consequently, our training data is dominated by speech tokens. If we mix all the different modalities according to their original proportions for training, the model would likely become overly focused on speech, at the expense of other modalities.

To address this issue, we implement a three-stage strategy that gradually increases the proportion of speech tokens. In Stage I, speech-text data accounts for 12.5% of the training tokens, which rises to 37.5% in Stage II, and finally reaches 75.0% in Stage III. This incremental increase in the proportion of speech tokens ensures that the model's performance in non-speech modalities is not compromised by the speech modality, while also allowing for the optimization of the model's speech capabilities.

Furthermore, we keep the data mixing ratio for other modalities of pre-training data at the minimal essential scales for replay, and we only use the high-quality subsets of them in this stage. This stage requires significantly fewer compute resources, due to the foundation laid in the previous stages.

We refer the reader to Appendix B for more details about the hyperparameters and prompt templates.

2.3 Supervised Fine-Tuning

As shown in Table 9, our model undergoes comprehensive and systematic supervised fine-tuning (SFT) with 16 different tasks and 34 diverse open-source datasets. The chat template used for SFT is the same as that used for Yi-6B-Chat (AI et al., 2024), and only the assistant responses are supervised. We refer the reader to Appendix C for more details about the hyperparameters and prompt templates.

3 Experiments

In this section, we present our quantitative evaluation results across various domains: image-related tasks (§3.1), speech-related tasks (§3.2), and video-related tasks (§3.3). Due to the lack of benchmarks for several advanced and emergent abilities of any-to-any multimodal LLMs, we also provide numerous qualitative demonstrations (§3.5) to demonstrate these capabilities. We refer the reader to Appendix D for more details, including the decoding hyperparameters and prompt templates.

3.1 Image-Related Tasks

Image Understanding. We compare our models with Emu (Sun et al., 2023c), SEED-LLaMA (Ge et al., 2023b), AnyGPT (Zhan et al., 2024), Flamingo (Alayrac et al., 2022), Kosmos-1 (Huang et al., 2023), MetaLM (Hao et al., 2022), IDEFICS (Laurençon et al., 2023), CM3Leon (Yu et al., 2023), InstructBLIP (Dai et al., 2023), Qwen-VL-Chat (Bai et al., 2023), and LLaVA 1.5 (Liu

Table 3: Experimental results for image understanding abilities. "Imagen" denotes whether the model is capable of generating images. "Speech" denotes whether the model supports speech modality. "I" denotes the instruction tuned version. The metrics used are CIDEr for COCO, MCQ accuracy for the SEED Bench, and VQA accuracy for the other tasks, following the standard procedures. In all cases, higher scores indicate better performance.

Models	Imagen	Speech	COCO	VQAv2	OKVQA	VizWiz	SEED Bench
Emu-Base (14B)	✓	✗	112.4	52.0	38.2	34.2	47.3
Emu-I (14B)	✗	✗	120.4	57.2	43.4	32.2	58.0
SEED-LLaMA-I (8B)	✓	✗	124.5	66.2	45.9	55.1	51.5
AnyGPT (8B)	✓	✓	107.5	-	-	-	-
Flamingo (9B)	✗	✗	79.4	51.8	44.7	28.8	42.7
Flamingo (80B)	✗	✗	84.3	56.3	31.6		-
Kosmos-1 (1.6B)	✗	✗	84.7	51.0	-	29.2	-
MetaLM (1.7B)	✗	✗	82.2	41.1	11.4	-	-
IDEFICS-I (80B)	✗	✗	117.2	37.4	36.9	26.2	53.2
CM3Leon (7B)	✓	✗	61.6	47.6	23.8	37.6	-
InstructBLIP (8.1B)	✗	✗	-	-	-	34.5	58.8
Qwen-VL-Chat (13B)	✗	✗	-	78.2	56.6	38.9	58.2
LLaVA 1.5 (7B)	✗	✗	-	78.5	-	50.0	58.6
MIO-Instruct (7B)	✓	✓	120.4	65.5	39.9	53.5	54.4

et al., 2023a). We evaluate our models in diverse tasks, including: (1) image captioning on MS-COCO (Lin et al., 2014) Karpathy test split with CIDEr score (Vedantam et al., 2014) as the metric, (2) three visual question-answering benchmarks, *i.e.*, VQAv2 (Goyal et al., 2016) (test-dev split), OK-VQA (Marino et al., 2019) (val split), and VizWiz (Gurari et al., 2018), with VQA accuracy as the metric, and (3) SEED-Bench (Li et al., 2023a), a comprehensive visual question-answering benchmark including 9 dimensions with MCQ accuracy as the metric. The scores for all baselines are copied from their reports. As shown in Table 3, our MIO-Instruct is ranked in the top group among all baselines, demonstrating its competitive image understanding performance. Although SEED-LLaMA achieved better scores compared to our model, we additionally support the speech modality. It is also noteworthy that MIO, with a size of approximately 7 billion parameters, outperforms several larger models such as Emu-14B and even IDEFICS-80B.

Image Generation. We compare our models with Emu (Sun et al., 2023c), SEED-LLaMA (Ge et al., 2023b), GILL (Koh et al., 2023), and AnyGPT (Zhan et al., 2024) for image generation. We use two benchmarks, *i.e.*, MS-COCO (Lin et al., 2014) Karpathy test split and Flickr30K (Plummer et al., 2015). Following GILL (Koh et al., 2023) and SEED-LLaMA (Ge et al., 2023b), we use CLIP-I as the metric that evaluates the similarity between the generated images and the ground-truth images with the image encoder in CLIP (Radford et al., 2021). As shown in Table 4 and Table 12 the pre-trained model and instruction-tuned model of MIO both have competitive image generation capabilities. Note that beyond single image generation abilities, our model can also exhibit multi-image generation capabilities such as generating visual stories, image sequences, and even visual thoughts as illustrated in §3.5.

Table 4: Image generation evaluation by CLIP-I score. "I" denotes the instruction tuned version. Higher values are better.

Models	MS-COCO	Flickr30K
Emu-Base	66.46	64.82
SEED-LLaMA	69.07	65.54
SEED-LLaMA-I	70.68	66.55
GILL	67.45	65.16
AnyGPT	65.00	-
MIO-Base	64.15	62.71
MIO-Instruct	67.76	68.97

3.2 Speech-Related Tasks

We evaluate the speech understanding and generation abilities of MIO on ASR and TTS tasks. Wav2vec 2.0 (Baevski et al., 2020), Whisper Large V2 (Radford et al., 2023), and AnyGPT (Zhan et al., 2024) are the baselines for ASR tasks, while VALL-E (Wang et al., 2023a), USLM (Zhang et al., 2023b) , and AnyGPT (Zhan et al., 2024) are the baselines for TTS tasks. The test set used for ASR evaluation is LibriSpeech (Panayotov et al., 2015), while the test set used for TTS evaluation is

VCTK (Veaux et al., 2017) following AnyGPT (Zhan et al., 2024)'s practice. The Whisper medium model is used to transcribe the speech generated for the TTS task. The WER (word error rate) is computed by comparing the generated transcribed text with the ground-truth transcription after text normalization[4].

As shown in Table 3.2, our models exhibit speech performance comparable to the speech-specific baselines and outperform the AnyGPT baseline. It is important to note that although AnyGPT is capable of generating content tokens for speech, it lacks the ability to generate timbre tokens, which necessitates the use of an additional voice cloning model. In contrast, our models generate both content and timbre tokens, making the TTS tasks more challenging for our models compared to AnyGPT. Nonetheless, after instruction tuning, our model still achieves better TTS performance. More evaluations of the TTS and Speech-to-Speech generation performance are provided in Appendix E.3 and E.2.

Table 5: Speech ability evaluation. "WER" denotes word error rate. Lower values are better.

Models	ASR WER	**Models**	TTS WER
Wav2vec	2.7	VALL-E	7.9
Whisper	2.7	USLM	6.5
AnyGPT	8.5	AnyGPT	8.5
MIO-Base	6.3	MIO-Base	12.0
MIO-Instruct	10.3	MIO-Instruct	4.2

3.3 Video-Related Tasks

We compare MIO with Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023b), InstructBLIP (Dai et al., 2023), Emu (Sun et al., 2023c), and SEED-LLaMA (Ge et al., 2023b) for video understanding. The models are evaluated on the MSVDQA (Chen & Dolan, 2011a) and MSRVTT-QA (Xu et al., 2017). The results are presented in Table 6. Our model achieves the highest scores compared to all baselines. Due to the lack of video (frame sequence) generation benchmarks in our setting, we provide video generation examples in §3.5. These results demonstrate the superior performance of our models in both video understanding and video generation.

Table 6: Video understanding evaluation using top-1 accuracy for both benchmarks. "I" denotes the instruction-tuned version.

Models	MSVDQA	MSRVTT-QA
Flamingo (9B)	30.2	13.7
BLIP-2 (4.1B)	33.7	16.2
InstructBLIP (8.1B)	41.8	22.1
Emu-Instruct (14B)	32.4	14.0
SEED-LLaMA-I (8B)	40.9	30.8
MIO-Instruct	42.6	35.5

Table 7: Language-only evaluation. "I" denotes the instruction-tuned version.

Models	**MMLU**
LLAMA-1-7B-Base	33.0
LLAMA-2-7B-Chat	47.9
SEED-LLAMA-8B-I	36.1
AnyGPT-Base	26.4
AnyGPT-Chat	27.4
MIO-Instruct	45.7

Table 8: Results for trimodal comprehension (text, image, and speech).

Models	**OmniBench**
Gemini-1.5-Pro	42.67
Reka-Core-20240501	31.52
AnyGPT (8B)	17.77
video-SALMONN (13B)	34.11
Unified-IO 2 (6.8B)	34.24
MIO-Instruct (7B)	36.96

3.4 Language-only Tasks

We evaluate our models on MMLU (Hendrycks et al., 2021). The baselines are two LLaMA variants (Touvron et al., 2023a;b), the instruction-tuned SEED-LLaMA (Ge et al., 2023b), and AnyGPT (Zhan et al., 2024). For the MMLU benchmark, we conduct zero-shot evaluation experiments using the official evaluation code. The experimental results are shown in Table 7. We can observe that our models have superior language-only performance compared with all any-to-any MM-LLM baselines and even surpass LLaMA-1-7B-Base, an advanced pure language model.

[4] `https://github.com/openai/whisper/blob/main/whisper/normalizers/english.py`

3.5 Demonstrations

We illustrate the basic and advanced abilities of MIO in Figure 5 and 4. The basic abilities of MIO involve image understanding and generation, video understanding and generation, ASR, and TTS. The advanced abilities of MIO are based on its any-to-any and multimodal interleaved sequence generation features. These abilities involve visual storytelling (*i.e.*, interleaved video-text generation), chain of visual thought, speech-in-speech-out, instructional image editing, visual guideline generation, etc. We refer the readers to Appendix E.5 for more demonstrations including multimodal chain of thought and multimodal in-context learning.

3.6 Ablation Studies

Generality for Trimodal Understanding. We evaluate our model using the OmniBench (Li et al., 2024d), which incorporates text, image, and speech modalities as inputs, requiring the model to choose one of four options as the correct answer to determine accuracy. Although MIO acquires its multimodal understanding capabilities through dual-modal training, the evaluation results in Table 8 indicate that MIO also exhibits superior trimodal comprehension abilities.

Effect of Different Image Tokenizers. The image tokenizer has a significant impact on image modality alignment. In Figure 2, we compare the image generation performance under a controlled setting after training for solely 3K steps in stage 1, using various image tokenizers. The image tokenizers used for comparison include a VQGAN (Esser et al., 2020) with a vocabulary size of 1024 and a compression rate of 16 (VQGAN-1024), as well as the VQGAN-Gumbel with a vocabulary size of 8192 (VQGAN-8192)[5]. Our results indicate that the SEED-Tokenizer, which captures more semantic and higher-level image information, exhibits faster convergence. In contrast, both VQGAN tokenizers show slower convergence due to their lower-level image information.

4 Related Works

4.1 Multimodal LLMs

Figure 2: Comparing different image tokenizers for image generation within a controlled setting (limited to 3K training steps).

With the rapid success of Large Language Models (LLMs), current multimodal LLMs (MM-LLMs) are typically built on a pre-trained LLM backbone and are endowed with the ability to understand multiple modalities (Li et al., 2019; Lu et al., 2019; Kim et al., 2021; Zeng et al., 2022; Zhou et al., 2022; Wang et al., 2023b; 2024e). Generally, these MM-LLMs align the representations of images obtained from visual encoders with the text embedding space, thereby leveraging the powerful capabilities of the foundational models. For example, BLIP-2 (Li et al., 2023b) uses CLIP-ViT (Radford et al., 2021) to extract high-level features from images and then employs a Q-Former to compress the number of image tokens and further align image tokens with the LLM embeddings. In contrast, LLaVA (Liu et al., 2023b; Li et al., 2024a) utilizes a simple linear projection or MLP as the connector between the image encoder and the LLM backbone. These models demonstrate strong multimodal understanding abilities, achieving significant progress in tasks such as visual question answering, visual commonsense reasoning, chart understanding, etc.

Additionally, beyond images, other MM-LLMs have also focused on modalities such as speech and video. For instance, LLaSM (Shu et al., 2023) and InternVideo (Wang et al., 2022; 2024c) are MM-LLMs designed for speech and video understanding, respectively. These models adopt a similar architectural design to BLIP-2 or LLaVA but redesign modality-specific encoders.

[5] `https://github.com/CompVis/taming-transformers`

Recently, increasing attention has been paid to unifying multiple modalities within a single MM-LLM. For example, ImageBind (Girdhar et al., 2023) develops encoders suited for multiple modalities such as images, videos, audio, heat maps, among others, while OmniBind (Wang et al., 2024d) trains an omni-representation model by aligning encoders across four modalities: audio, language, images, and 3D objects. OmniBench (Li et al., 2024d) is proposed to evaluate the models' abilities for visual, acoustic, and textual understanding.

However, these models focus primarily on multimodal understanding and often overlook the important aspect of multimodal generation.

4.2 Any-to-Any MM-LLMs

To enable multimodal generation in MM-LLMs, a straightforward approach is to allow these models to call external multimodal generation tools, such as Stable Diffusion (Rombach et al., 2022) or text-to-speech (TTS) tools (Shen et al., 2023; Li et al., 2024c; OpenAI et al., 2023). However, as highlighted in the Gemini technical report (Team et al., 2023), relying on an intermediate natural language interface can limit the model's ability to express images. If a model cannot natively output images, it will not be able to generate images with prompts of interleaved sequences of image and text. This claim is in line with our distinction between descriptive image generation and context-aware image generation, as discussed in §2.2.

As a result, recent works focus on the unification of multimodal understanding and generation in a single model (*i.e.*, any-to-any MM-LLMs), enabling the generation of multimodal tokens without natural language as an interface. These models typically follow different approaches, depending on how images are represented in both input and output sides. For example, the Discrete-In-Discrete-Out (DIDO) approach has been explored in works such as SEED-LLaMA (Ge et al., 2023b), AnyGPT (Zhan et al., 2024), and Chameleon (Team, 2024). Continuous-In-Discrete-Out (CIDO) methods have been implemented in models like DaVinCi (Diao et al., 2023), Gemini (Team et al., 2023), and Unified-IO 2 (Lu et al., 2023). The Continuous-In-Continuous-Out (CICO) approach is used in models such as Emu (Sun et al., 2023c;a), and DreamLLM (Dong et al., 2023). Another approach, the integration of autoregression and diffusion (AR + Diff), can be seen in models like Transfusion (Zhou et al., 2024), Show-o (Xie et al., 2024), and Li et al. (2024b)'s.

However, these models face specific limitations. DreamLLM (CICI, Dong et al. (2023)) and CIDO models suffer from inconsistencies between input and output forms for multimodal data, making it difficult for them to natively support the generation of interleaved multimodal sequences where an image functions in a coupled way as both input and output. Emu2 (CICO, Sun et al. (2023a)) struggles with the challenges of the mean square error (MSE) loss used for training continuous output representations, as well as with the uni-modal assumption of the Gaussian distribution in the MSE loss. Transfusion (AR + Diff, Zhou et al. (2024)) applies noise to images from the input side to support multimodal generation with diffusion modeling, and relies on VAE (Kingma & Welling, 2013) features rather than CLIP (Radford et al., 2021) features for denoising, which largely trade off the multimodal understanding abilities.

To mitigate these issues, we adopt the DIDO approach. A comprehensive comparison of our models with other any-to-any MM-LLMs is presented in Table 1.

5 Conclusion

In conclusion, MIO represents an advancement in the realm of multimodal foundation models. By employing a rigorous four-stage training process, MIO successfully integrates and aligns discrete tokens across text, image, video, and speech modalities. This comprehensive approach enables MIO to understand and generate multimodal content in an end-to-end, autoregressive manner, addressing the limitations of current multimodal large language models. Our experimental results showcase its competitive performance across a variety of benchmarks compared to the dual-modality baselines and other any-to-any multimodal large language models. With the any-to-any and multimodal interleaved output features, MIO exhibits novel emergent abilities such as interleaved video-text generation, chain-of-visual-thought reasoning, etc.

REFERENCES

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. *arXiv preprint arXiv: 2403.04652*, 2024.

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.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4211–4215, 2020.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv preprint*, abs/2308.12966, 2023.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions, 2024. URL `https://cdn.openai.com/papers/dall-e-3.pdf`.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv: 2311.15127*, 2023.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv: 2303.12712*, 2023.

David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 190–200, Portland, Oregon, USA, June 2011a. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1020`.

David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, Portland, OR, June 2011b.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv: 2106.06909*, 2021.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre D'efossez. Simple and controllable music generation. *Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2306.05284. URL `https://arxiv.org/abs/2306.05284v3`.

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

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv: 2307.08691*, 2023.

Tri Dao, Daniel Y. Fu, Stefano Ermon, A. Rudra, and Christopher R'e. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Neural Information Processing Systems*, 2022.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, David Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, Xilai Li, Karel Mundnich, Monica Sunkara, Sundararajan Srinivasan, Kyu J Han, and Katrin Kirchhoff. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv: 2405.08295*, 2024.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, J. Heek, Matthias Minderer, Mathilde Caron, A. Steiner, J. Puigcerver, Robert Geirhos, Ibrahim M. Alabdulmohsin, Avital Oliver, Piotr Padlewski, A. Gritsenko, Mario Luvci'c, and N. Houlsby. Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. *Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2307.06304.

Shizhe Diao, Wangchunshu Zhou, Xinsong Zhang, and Jiawei Wang. Write and paint: Generative vision-language models are unified modal learners. In *The Eleventh International Conference on Learning Representations*, 2023.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv: 2309.11499*, 2023.

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.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision. *arXiv preprint arXiv: 2206.04769*, 2022.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12868–12878, 2020. URL `https://api.semanticscholar.org/CorpusID:229297973`.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*, 2024.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv: 2312.11370*, 2023.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023a.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023b.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. *arXiv preprint arXiv: 2305.05665*, 2023.

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. *International Journal of Computer Vision*, 2016. doi: 10.1007/s11263-018-1116-0.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. *arXiv preprint arXiv: 1802.08218*, 2018.

Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *ArXiv preprint*, abs/2206.06336, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv: 2106.07447*, 2021.

Shaohan Huang, Li Dong, Wenhui Wang, Y. Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, O. Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. *Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2302.14045.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1233–1239, 2016.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv: 2405.01483*, 2024.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. Libriheavy: a 50,000 hours asr corpus with punctuation casing and context, 2023.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5583–5594, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv: 1312.6114*, 2013.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *NeurIPS*, 2023.

LAION. Laion coco: 600m synthetic captions from laion-2b-en. `https://laion.ai/blog/laion-coco/`, 2022.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization, 2022.

Bo Li*, Peiyuan Zhang*, Kaichen Zhang*, Fanyi Pu*, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024. URL `https://github.com/EvolvingLMMs-Lab/lmms-eval`.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023a.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024a.

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*, abs/2301.12597, 2023b.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023c.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *ArXiv*, 2019.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv: 2406.11838*, 2024b.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024c.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv: 2409.15272*, 2024d.

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*, pp. 740–755. Springer, 2014.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint*, 2024.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485, 2023b.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2017.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. URL `https://arxiv.org/abs/1908.02265`.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action, 2023.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3195–3204, 2019.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, Kavya Srinet, Babak Damavandi, and Anuj Kumar. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv: 2309.16058*, 2023.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted

Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*, 2023.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card. *arXiv preprint arXiv: 2410.21276*, 2024.

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

Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Journeydb: A benchmark for generative image understanding, 2023.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv: 2306.01116*, 2023.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

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. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv: 2212.04356*, 2022.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022. doi: 10.1109/cvpr52688.2022.01042. URL `http://dx.doi.org/10.1109/cvpr52688.2022.01042`.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv: 2306.12925*, 2023.

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

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *arXiv preprint arXiv: 2303.17580*, 2023.

Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. Llasm: Large language and speech model. *arXiv preprint arXiv: 2308.15930*, 2023.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023a.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv: 2303.15389*, 2023b.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality, 2023c.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv: 2310.13289*, 2023.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv e-prints*, pp. arXiv–2405, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-

David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham

Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv: 2312.11805*, 2023.

Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288, 2023b.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *ArXiv*, abs/1711.00937, 2017. URL https://api.semanticscholar.org/CorpusID:20282961.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. 2017.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *arXiv preprint arXiv: 1411.5726*, 2014.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023a.

Tiannan Wang, Wangchunshu Zhou, Yan Zeng, and Xinsong Zhang. EfficientVLM: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13899–13913, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.873. URL https://aclanthology.org/2023.findings-acl.873.

Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, Yibin Liu, Jialong Wu, Shengwei Ding, Long Li, Zhiwei Huang, Xinle Deng, Teng Yu, Gangan Ma, Han Xiao, Zixin Chen, Danjun Xiang, Yunxia Wang, Yuanyuan Zhu, Yi Xiao, Jing Wang, Yiru Wang, Siran Ding, Jiayang Huang, Jiayi Xu, Yilihamu Tayier, Zhenyu Hu, Yuan Gao, Chengfeng Zheng, Yueshu Ye, Yihang Li, Lei Wan, Xinyue Jiang, Yujie Wang, Siyu Cheng, Zhule Song, Xiangru Tang, Xiaohua Xu, Ningyu Zhang, Huajun Chen, Yuchen Eleanor Jiang, and Wangchunshu Zhou. Weaver: Foundation models for creative writing. *arXiv preprint arXiv: 2401.17268*, 2024a.

Wenbin Wang, Yang Song, and Sanjay Jha. Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech. *arXiv preprint arXiv: 2406.14875*, 2024b.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024c.

Zehan Wang, Ziang Zhang, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Hengshuang Zhao, and Zhou Zhao. Omnibind: Large-scale omni multimodal representation via binding spaces. *arXiv preprint arXiv: 2407.11895*, 2024d. URL `https://arxiv.org/abs/2407.11895v1`.

Zekun Wang, Jingchang Chen, Wangchunshu Zhou, Haichao Zhu, Jiafeng Liang, Liping Shan, Ming Liu, Dongliang Xu, Qing Yang, and Bing Qin. SmartTrim: Adaptive tokens and attention pruning for efficient vision-language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14937–14953, Torino, Italia, May 2024e. ELRA and ICCL. URL `https://aclanthology.org/2024.lrec-main.1300`.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv: 2310.00746*, 2023c.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004. doi: 10.1109/TIP.2003.819861. URL `https://ieeexplore.ieee.org/document/1284395`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv: 2309.05519*, 2023. URL `https://arxiv.org/abs/2309.05519v2`.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv: 2408.12528*, 2024. URL `https://arxiv.org/abs/2408.12528v1`.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pp. 1645–1653, 2017.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 5288–5296, 2016.

Zhiyang Xu, Trevor Ashby, Chao Feng, Rulin Shao, Ying Shen, Di Jin, Qifan Wang, and Lifu Huang. Vision-flan:scaling visual instruction tuning, Sep 2023. URL https://vision-flan.github.io/.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. Scaling autoregressive multi-modal models: Pretraining and instruction tuning, 2023.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal neural script knowledge models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=CRFSrgYtV7m.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. X ^{2}-vlm: All-in-one pre-trained model for vision-language tasks. *arXiv preprint arXiv:2211.12402*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *IEEE International Conference on Computer Vision*, 2023. doi: 10.1109/ICCV51070.2023.01100.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. Anygpt: Unified multimodal llm with discrete sequence modeling. *ArXiv*, abs/2402.12226, 2024. URL https://api.semanticscholar.org/CorpusID:267750101.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities, 2023a.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech language models, 2023b.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023c.

Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv: 2307.10802*, 2023d.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *ArXiv preprint*, abs/2309.07915, 2023.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens, 2023.

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. *arXiv preprint arXiv: 2408.11039*, 2024. URL https://arxiv.org/abs/2408.11039v1.

Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. VLUE: A multi-task multi-dimension benchmark for evaluating vision-language pre-training. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27395–27411. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhou22n.html.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.

A PRE-TRAINING DATA

Pre-training Data Sources. The pre-training data sources involve six types:

1. Image-text paired data: SBU (Ordonez et al., 2011), CC3M (Sharma et al., 2018), LAION-COCO (LAION, 2022), and JourneyDB (Pan et al., 2023), where JourneyDB only serves for image generation.
2. Language-only data: RefinedWeb (Penedo et al., 2023).
3. Image-text interleaved data: OBELICS (Laurençon et al., 2023), MMC4-core-ff (Zhu et al., 2023).
4. Video-text paired data: WebVid-10M (Bain et al., 2021).
5. Video-text interleaved data: HowTo-100M (Miech et al., 2019), Youtube-Temporal-180M (Zellers et al., 2021).
6. Speech-text paired data: Libriheavy (Kang et al., 2023).

Pre-training Data Processing. We have different data processing procedures for different data types illustrated in §A following Emu (Sun et al., 2023c) and Qwen-VL (Bai et al., 2023):

1. Image-text paired data: we remove pairs with more than 2:1 aspect ratio or smaller than 224×224 resolution of the image. We remove pairs with more than 0.27 CLIP scores. We remove non-English pairs. We randomly place the image or text at the forefront for generating captions based on images and vice versa.
2. Language-only data: we use the same data processing pipeline as used in Yi (AI et al., 2024).
3. Image-text interleaved data: we filter the data using a CLIP score threshold of 0.25, and follow the same procedure as illustrated in Emu (Sun et al., 2023c).
4. Video-text paired data: we randomly place the frames or text at the forefront for generating captions based on frames and vice versa. 60% of the pairs are text-to-video, while 40% of the pairs are video-to-text. We sample 4 to 8 frames of each video for training according to the text lengths.
5. Video-text interleaved data: We first use PySceneDetect to extract key frames from the video based on scene changes, following the practice of Stable Video Diffusion (Blattmann et al., 2023). Then, for each video clip between two key frames, we extract a central frame for textual caption generation with BLIP-2 (Li et al., 2023b). Additionally, the video clips between key frames are processed using ASR (automatic speech recognition) tools to extract subtitles. The ASR text and captions are then integrated and refined using Yi-34B-Chat (AI et al., 2024), resulting in a single text segment. These text segments, along with the key frames and central frames, form the video-text interleaved data.
6. Speech-text paired data: we remove speechs with more than 15 seconds.

B PRE-TRAINING DETAILS

Hyperparameters. We enable Flash Attention (Dao et al., 2022; Dao, 2023) during pre-training. Gradient clipping is set to 1.0 for all stages. The maximum sequence length for training is 2800 tokens. We use a cosine learning rate scheduler with a peak learning rate of 3e-5 and a warmup ratio of 0.03. The optimizer used is AdamW (Loshchilov & Hutter, 2017).

Prompt Templates. The prompt template is only necessary for paired datasets. For image-text paired data, we use the prompt templates of "{image} The caption of this image is: {caption}" and "Please generate an image of "{caption}": {image}". For video-text paired data: we use the prompt templates of "Please describe the following video: {image} {description}" and "Please generate a video for "{description}": {video}". For speech-text paired data: we use the prompt templates of "{speech} Transcribe this speech: {transcription}" and "Please generate a speech of "{transcription}": {speech}" during Stage I and Stage II. While for Stage III, we change the ASR prompt template into '{speech} The transcription of this speech is: {transcription}".

C SUPERVISED FINE-TUNING DETAILS

Table 9: Supervised Fine-Tuning Data. "ICL" denotes In-Context Learning, and "CoT" denotes Chain of Thought.

Task	Dataset
Language Only	OpenHermes (Teknium, 2023)
Multimodal ICL	MMICL (Zhao et al., 2023)
Multimodal CoT	ScienceQA (Lu et al., 2022)
Chart Understanding	Geo170K (Gao et al., 2023)
Instructional Image Generation	InstructPix2Pix (Brooks et al., 2023), MagicBrush (Zhang et al., 2024)
ASR	LibriSpeech (Panayotov et al., 2015), GigaSpeech (Chen et al., 2021), Common Voice (Ardila et al., 2020)
Video Dialogue	VideoChat2-IT (Li et al., 2023c)
Image QA	Vision-Flan (Xu et al., 2023), VizWiz (Gurari et al., 2018), LAION-GPT4V[6], LLaVAR (Zhang et al., 2023c), OCR-VQA (Mishra et al., 2019), VQA (Goyal et al., 2016), TextVQA (Singh et al., 2019), OK-VQA (Marino et al., 2019), Mantis-Instruct (Jiang et al., 2024)
Speech Generation	SpeechInstruct (Zhang et al., 2023a)
Speech Understanding	SpeechInstruct (Zhang et al., 2023a)
Image Captioning	Flickr30K (Plummer et al., 2015), MS-COCO (Lin et al., 2014)
Descriptive Image Generation	Flickr30K (Plummer et al., 2015), MS-COCO (Lin et al., 2014)
TTS	GigaSpeech (Chen et al., 2021), Common Voice (Ardila et al., 2020)
Video Generation	MSR-VTT (Xu et al., 2016), MSVD (Chen & Dolan, 2011b)
Video Understanding	MSR-VTT (Xu et al., 2016), MSVD (Chen & Dolan, 2011b), MSVD-QA (Chen & Dolan, 2011a), MSRVTT-QA (Xu et al., 2017)
Visual Storytelling	VIST (Huang et al., 2016)

Supervised Fine-Tuning Data. As shown in Table 9, we use 16 tasks with 34 datasets for a comprehensive supervised fine-tuning.

Prompt Templates. The chat template is the same as used in Yi (AI et al., 2024). The system prompt is unified as: "You are MIO, an AI assistant capable of understanding and generating images, text, videos, and speech, selecting the appropriate modality according to the context." except for speech generation and TTS whose system prompts are "You are MIO, an AI assistant capable of understanding images, text, videos, and speech, and generating speech. Please respond to the user with speech only, starting with <spch> and ending with </spch>." to avoid randomness of the output modality.

Hyperparameters. Similar to pre-training (*c.f.*, Appendix B), we enable Flash Attention (Dao et al., 2022; Dao, 2023) during supervised fine-tuning. Gradient clipping is set to 1.0. The maximum sequence length for training is 2800 tokens. We use a cosine learning rate scheduler with a peak learning rate of 3e-5 and a warmup ratio of 0.03. The optimizer used is AdamW (Loshchilov & Hutter, 2017).

D EVALUATION DETAILS.

Hyperparameters. The decoding strategies and hyperparameters are quite important for a superior performance. As shown in Table 10, we use different sets of parameters for different output modalities.

Table 10: Decoding Hyperparameters.

Output Modality	Text	Image	Speech	Video
Beam size	5	1	1	1
Do Sampling	False	True	True	True
Top-P	-	0.7	0.7	0.7
Repetition Penalty	1.0	1.0	1.15	1.15
Temperature	1.0	1.0	1.0	1.0
Guidance Scale	1.0	1.0	1.0	1.0

Table 11: Prompt templates used for evaluating instruction-tuned models.

Task	Prompt Template
Image Captioning	Provide a one-sentence caption for the provided image. {image}
Image QA	(We use the prompt templates in LMMs-Eval (Li* et al., 2024)).
Image Generation	Please generate an image according to the given description. {description}
ASR	Please transcribe this speech.{speech_token}
TTS	Please generate a speech according to the given transcription. Start with <spch>. {transcription}
Text-only	The following are multiple choice questions (with answers) about {subject} {question}
Video QA	The goal is to use the visual information available in the image to provide an accurate answer to the question. This requires careful observation, attention to detail, and sometimes a bit of creative thinking.{video} Question: {question} Answer:

Prompt Templates. The prompt templates used for evaluating pre-training checkpoints are the same as used during pre-training. For SFT checkpoint evaluation, we list the prompt templates in Table 11.

E MORE EXPERIMENTS

E.1 IMAGE GENERATION EVALUATION

We compute two additional automatic metrics for evaluating image generation, i.e., SSIM (Wang et al., 2004) and Aesthetic Predictor v2.5[7] for the evaluation of structural integrity and aesthetics, respectively. SSIM (Structural Similarity Index Measure) evaluates the perceptual similarity between the generated images and the ground-truth images, focusing on luminance, contrast, and structure, with scores ranging from -1 (dissimilar) to 1 (identical). Aesthetic Predictor V2.5 is a SigLIP (Zhai et al., 2023)-based predictor that evaluates the aesthetics of an image on a scale from 1 to 10 (10 is the best). In addition, we randomly select 100 image descriptions from MS-COCO test set, and used each model to generate images accordingly for human preference evaluation. We ask 3 annotators to rank 3 images generated by the 3 models: "given the image description, which image is preferred?" The average ranking of MIO's, AnyGPT's, and Emu's generated images are 1.2 (MIO), 2.9 (AnyGPT), 1.9 (Emu). MIO aligns the best with the human preference. The percentage agreement between the three annotators (calculated as the number of cases with identical rankings by all annotators divided by 100) is 82.3%, indicating a high consistency in the human evaluation.

Dataset	MS-COCO		Flickr30K		MS-COCO Subset
Metric	SSIM (↑)	Aesthetic (↑)	SSIM (↑)	Aesthetic (↑)	Human Avg. Ranking (↓)
Emu	0.1749	3.733	0.1451	3.893	1.9
AnyGPT	0.1960	3.954	0.1585	4.251	2.9
MIO	0.2307	4.019	0.1727	4.326	1.2

Table 12: Image generation evaluation by SSIM, Aesthetic Predictor V2.5, and human preference.

Model	Supported Workflow	Content Score (1-5 points) (↑)
MIO	s2s	1.4
LLaMA-Omni (Fang et al., 2024)	s2t→t2s	2.4
AnyGPT	s2t→t2s	1.8

Table 13: Speech-to-Speech performance. "s2s" means "speech-to-speech", while "s2t" and "t2s" denote "speech-to-text" and "text-to-speech", respectively.

E.2 Speech-to-Speech Evaluation

Since there is a lack of speech to speech evaluation benchmarks, we randomly sample some conversations from the moss-002-sft dataset[8] and convert them into speech-to-speech format. Following the evaluation procedures outlined in LLaMA-Omni (Fang et al., 2024), we use the content score metric obtained from GPT-4o (OpenAI et al., 2024) to assess whether the model's response effectively addresses the user's instructions. The results are shown in Table 13.

Though the content score of MIO is slightly lower than LLaMA-Omni and AnyGPT, both LLaMA-Omni and AnyGPT first generate text replies and then convert these into voice. However, our model, MIO, is capable of directly generating speech responses to speech queries.

E.3 TTS Evaluation

Model	GLOBE		LibriSpeech test-clean	
	WER (↓)	Speech Similarity (↑)	WER (↓)	Speech Similarity (↑)
MIO	9.8	67.8	10.3	75.1
AnyGPT	27.9	67.3	28.1	71.3

Table 14: More automatic evaluations for the TTS performance.

We select two additional benchmarks, LibriSpeech test-clean (Panayotov et al., 2015) and GLOBE (Wang et al., 2024b), to evaluate the performance of TTS between our model and AnyGPT. For fair comparison, we don't specify the input voice prompt during evaluation of MIO and AnyGPT. WER (Word Error Rate) and speaker similarity are employed as the automatic metrics. The results are shown in Table 14. The results show that MIO performs significantly better than AnyGPT on both WER and speaker similarity across both benchmarks.

Additionally, we conduct a human evaluation to assess the speech quality of the outputs from MIO and AnyGPT. In this evaluation, participants are provided with the target speech, the speech generated by AnyGPT, and the speech generated by our model. They are tasked with determining which one sounded more natural and closer to the target speech. Evaluators could choose one of the two generated speeches or indicate that they find them equally natural.

Table 15: Human evaluation for the TTS performance.

MIO Win	54%
Tie	25%
MIO Lose	21%

[7] https://github.com/discus0434/aesthetic-predictor-v2-5?tab=readme-ov-file

[8] https://huggingface.co/datasets/fnlp/moss-002-sft-data

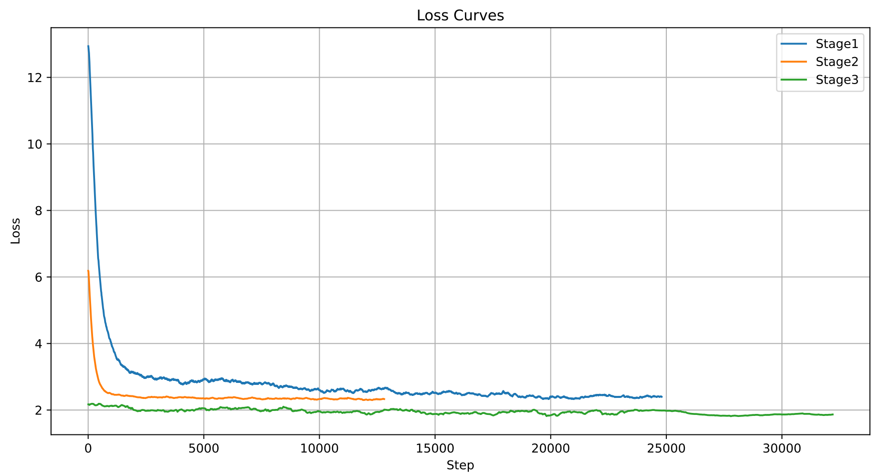

Figure 3: Loss curves of pretraing stages.

Each evaluation is rated by three independent human evaluators, and we report the average scores. The results are shown in Table 15. MIO significantly outperforms AnyGPT in the human evaluation, consistent with the results from the automatic evaluation.

E.4 Loss Curves

We plot the loss curves for each stage in Figure 3. We can observe that when introducing a new data type (i.e., image-text interleaved data) in stage 2, the training loss suddenly increases. However, in the third pretraining stage, i.e., the speech-enhancement stage, the training loss transitions more smoothly. Despite the fluctuations in loss between stages, which do have some impact on downstream performance during the fluctuation periods, we find that with continued training, the model's loss quickly recovers to its previous convergence level and continues optimizing effectively.

E.5 More Demonstrations.

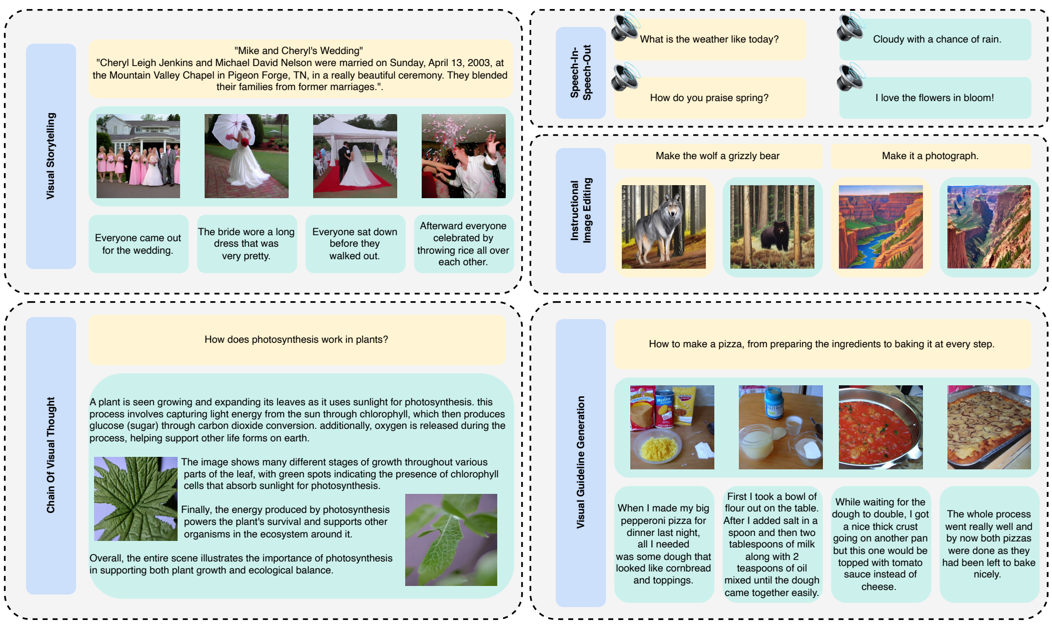

Figure 4: Demonstrations of MIO's advanced abilities. Yellow : inputs; Green : outputs.

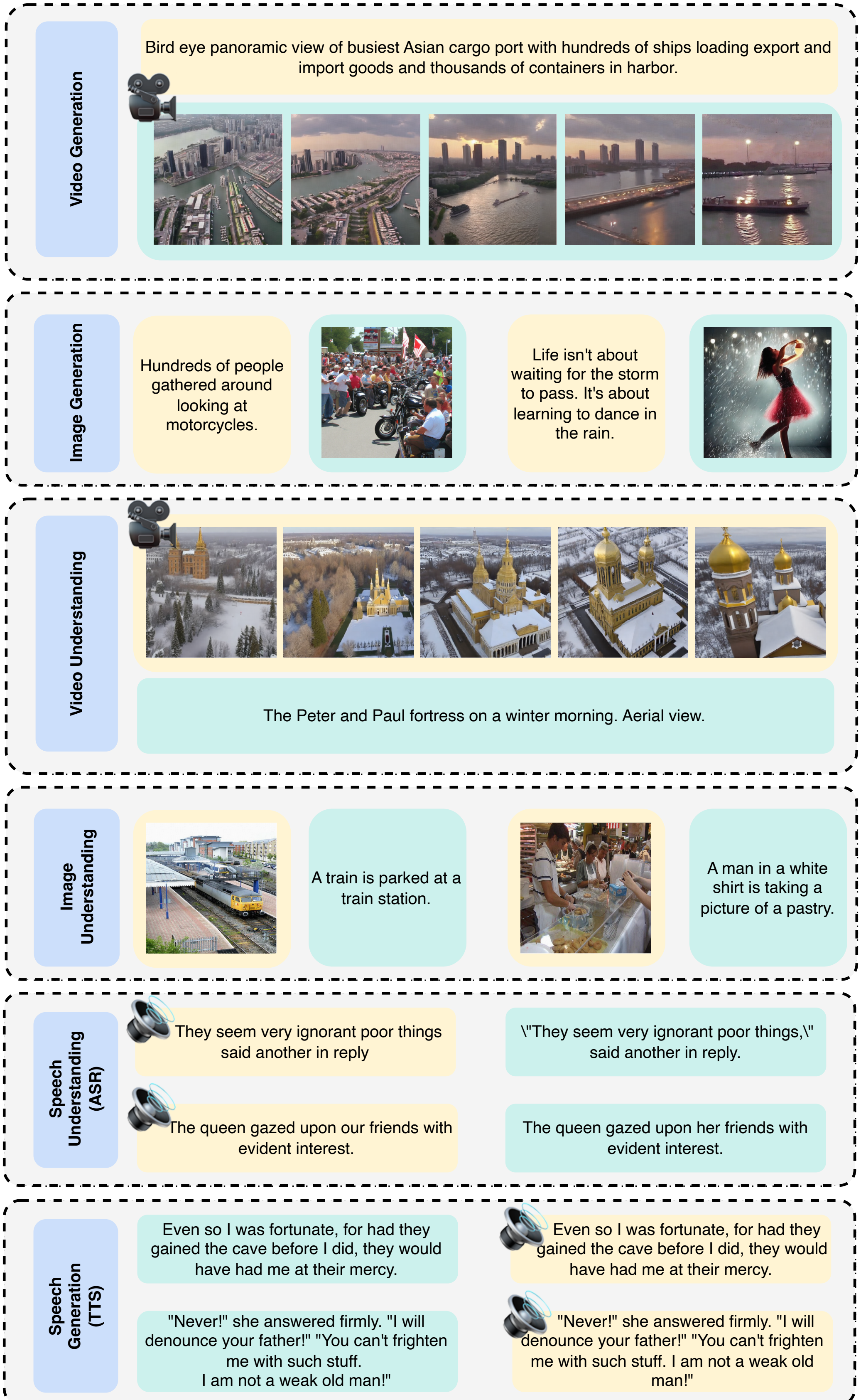

Figure 5: Demonstrations of MIO's basic abilities. Yellow : inputs; Green : outputs.

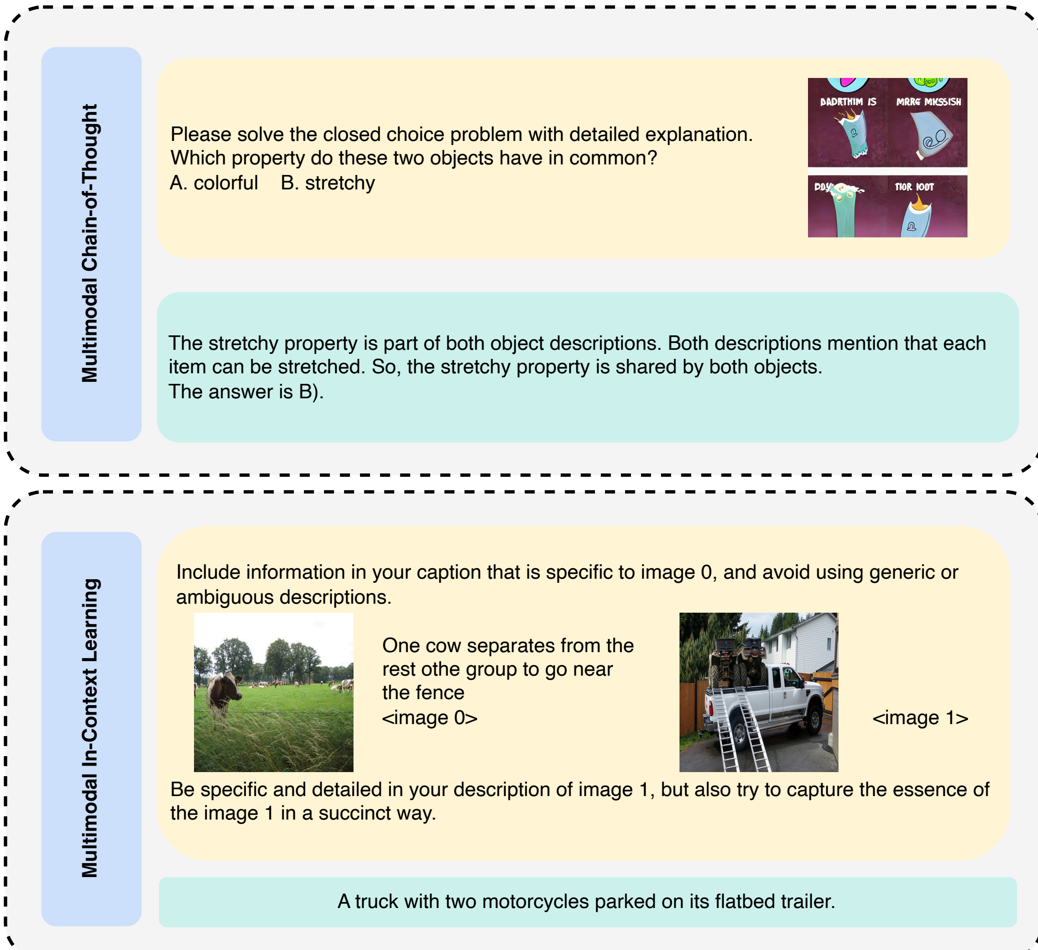

Figure 6: Multimodal Chain-of-Thought and Multimodal In-Context Learning Demos. Yellow : inputs; Green : outputs.