Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action

Jiasen Lu[1*] Christopher Clark [1*] Sangho Lee[1*] Zichen Zhang[1*]
Savya Khosla[2] Ryan Marten[2] Derek Hoiem[2] Aniruddha Kembhavi[13]
[1]Allen Institute for AI [2] University of Illinois Urbana-Champaign [3] University of Washington
{jiasenl, chrisc, sanghol, chralesz, anik}@allenai.org
unified-io-2.allenai.org

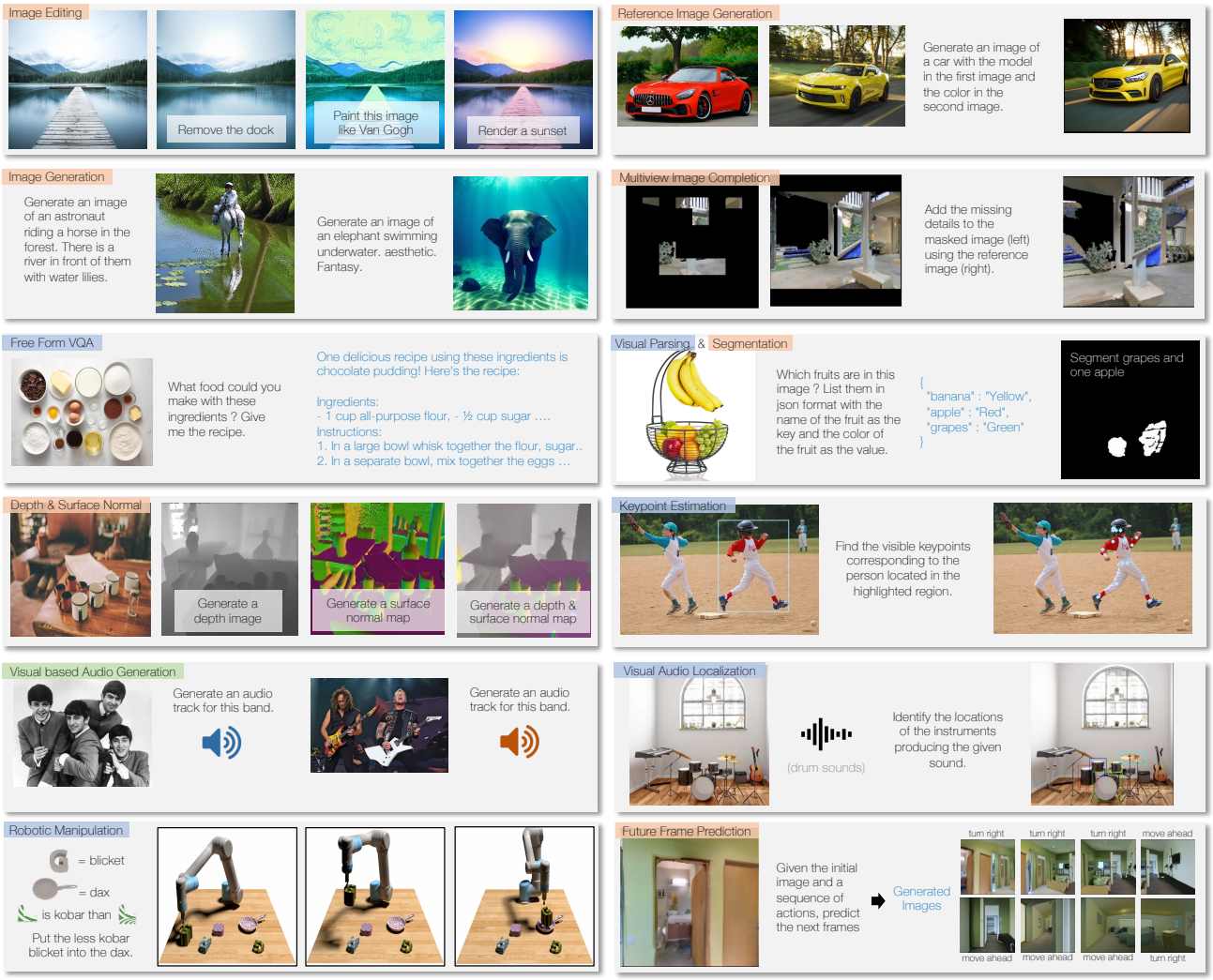

Figure 1. UNIFIED-IO 2 is an instruction-following model with a huge breadth of abilities and supported modalities. It can generate images (red box), including image editing, image generation, depth estimation, surface normal estimation, and future frame prediction *etc*. It can also generate texts (blue box), including long-form answers to queries, keypoint estimation, visual audio localization, predicting actions for robotic manipulation *etc*. It can generate audio (green box) from images or text. Click and for the corresponding audio samples.

* Leading Authors, equal contribution. A description of each author's contribution is available in Appendix A. Corresponding to Jiasen Lu.

Abstract

We present UNIFIED-IO 2*, the first autoregressive multimodal model that is capable of understanding and generating image, text, audio, and action. To unify different modalities, we tokenize inputs and outputs – images, text, audio, action, bounding boxes* etc.*, into a shared semantic space and then process them with a single encoder-decoder transformer model. Since training with such diverse modalities is challenging, we propose various architectural improvements to stabilize model training. We train our model from scratch on a large multimodal pre-training corpus from diverse sources with a multimodal mixture of denoisers objective. To learn an expansive set of skills, such as following multimodal instructions, we construct and finetune on an ensemble of 120 datasets with prompts and augmentations. With a single unified model,* UNIFIED-IO 2 *achieves state-of-the-art performance on the GRIT benchmark and strong results in more than 35 benchmarks, including image generation and understanding, natural language understanding, video and audio understanding, and robotic manipulation. We release all our models to the research community.*

1. Introduction

As AI researchers, we seek to build intelligent agents that can perceive their environment, communicate with others, act in the world, and reason about their interactions. The world is multimodal, so our agents must partake in rich interactions that are multimodal in nature via vision, language, sound, action *etc*. Psychologists have argued that the redundancy of our sensory systems serves as supervisory mechanisms to improve each other [48, 144, 167]. This provides a natural motivation to create models with similar learning capabilities, supporting many different modalities that can supervise each other during training.

Building models that can parse and produce many modalities is a complex undertaking. Training Large Language Models (LLMs) with billions of parameters, despite only supporting a single modality, is extremely challenging across many fronts – from sourcing and processing massive datasets, ensuring data quality and managing biases, designing effective model architectures, maintaining stable training processes, and instruction tuning to enhance the model's ability to follow and understand user instructions. These challenges are hugely amplified with the addition of each new modality.

In light of these difficulties, a line of recent works in building multimodal systems has leveraged pre-trained LLMs, with some augmenting with new modality encoders [5, 46, 119], some adding modality specific decoders [14, 96] and others leveraging the LLM's capabilities to build modular frameworks [64, 166, 173]. Another line of works on training multimodal models from scratch has focused on generating text output [81, 143] with a few recent works supporting the understanding and generation of two modalities – text and images [123, 125]. Building generative models with a wider coverage of modalities, particularly when training from scratch, remains an open challenge.

In this work, we present UNIFIED-IO 2, a large multimodal model (LMM) that can encode text, image, audio, video, and interleaved sequences and produce text, action, audio, image, and sparse or dense labels. It can output free-form multimodal responses and handle tasks unseen during training through instruction-following. UNIFIED-IO 2 contains 7 billion parameters and is pre-trained from scratch on an extensive variety of multimodal data – 1 billion image-text pairs, 1 trillion text tokens, 180 million video clips, 130 million interleaved image & text, 3 million 3D assets, and 1 million agent trajectories. We further instruction-tune the model with a massive multimodal corpus by combining more than 120 datasets covering 220 tasks across vision, language, audio, and action.

Our pre-training and instruction tuning data, totaling over 600 terabytes, presents significant challenges for training due to its diversity and volume. To effectively facilitate self-supervised learning signals across multiple modalities, we develop a novel multimodal mixture of denoiser objective that combines denoising and generation across modalities. We also develop dynamic packing – an efficient implementation that provides a 4x increase in training throughput to deal with highly variable sequences. To overcome the stability and scalability issues in training, we propose to apply key architectural changes, including 2D rotary embeddings, QK normalization, and scaled cosine attention mechanisms on the perceiver resampler. For instruction tuning, we ensure every task has a clear prompt, either using existing ones or crafting new ones. We also include open-ended tasks and create synthetic tasks for less common modalities to enhance task and instruction variety.

We evaluate UNIFIED-IO 2 on over 35 datasets across the various modalities it supports. Our single model sets the new state of the art on the GRIT [66] benchmark, which includes diverse tasks such as keypoint estimation and surface normal estimation. On vision & language tasks, it matches or outperforms the performance of many recently proposed VLMs that leverage pre-trained LLMs. On image generation, it outperforms the closest competitor [174] that leverages the pre-trained stable diffusion model [154], especially in terms of faithfulness as per the metrics defined in [76]. It also shows effectiveness in video, natural language, audio, and embodied AI tasks, showcasing versatility despite its broad capability range. Moreover, UNIFIED-IO 2 can follow free-form instructions, including novel ones. Figure 1 offers a glimpse into how it handles various tasks. Further examples, along with the code and models, are accessible on our project website.

2. Related Work

Inspired by the success of language models as general-purpose text processing systems [20, 122, 177], there has been a recent wave of multimodal systems trying to achieve similar general-purpose capabilities with additional modalities. A common approach is to use a *vision-encoder* to build features for input images and then an *adapter* to map those features into embeddings that can be used as part of the input to an LLM. The network is then trained on paired image/language data to adapt the LLM to the visual features. These models can already perform some tasks zero-shot or with in-context examples [109, 132, 178], but generally a second stage of visual instruction tuning follows using instructions, visual inputs, and target text triples to increase zero-shot capabilities [25, 34, 118, 119, 205, 218, 225].

Building upon this design, many researchers have expanded the breadth of tasks these models can support. This includes creating models that can do OCR [12, 220], visual grounding [12, 26, 143, 189, 207, 212, 219], image-text-retrieval [97], additional languages [112], embodied AI tasks [17, 135, 140, 152] or leverage other expert systems [52]. Other efforts have added new input modalities. This includes video inputs [110, 126], audio [80] or both [216]. PandaGPT [170] and ImageBind-LLM [69] use the universal encoder ImageBind [56] to encode many kinds of input modalities, and ChatBridge [222] uses a similar universal encoder based on language. While these efforts are effective for understanding tasks, they do not allow complex multimodal generation and often exclude modalities long considered central to computer vision (*e.g.*, ImageBind cannot support sparse annotation of images).

Fewer works have considered multimodal generation. UNIFIED-IO [123], LaVIT [88], OFA [186], Emu [172] and CM3Leon [210] train models to generate tokens that a VQ-GAN [49, 179] can then decode into an image, while GILL [96], Kosmos-G [141] and SEED [53] generate features that a diffusion model can use, and JAM [4] fuses pre-trained language and image generation models. UNIFIED-IO 2 also uses a VQ-GAN, but supports text, image, and audio generation. See Appendix C for more discussion about related work.

3. Approach

In this section, we discuss the unified task representation (3.1), the model architecture and techniques to stabilize training (3.2), the multimodal training objective (3.3) and the efficiency optimizations (3.4) used in UNIFIED-IO 2.

3.1. Unified Task Representation

UNIFIED-IO 2 processes all modalities with a single, unified encoder-decoder transformer [181]. This is achieved by encoding various inputs and outputs – images, text, audio, action, boxes *etc*., into sequences of tokens in a shared representation space. Our encoding procedure follows the design of UNIFIED-IO [123], with several modifications to improve performance and new encoders and decoders for additional modalities. Figure 2 shows an overview of the model. A high-level overview of how modalities are encoded is given below, see Appendix D.1 for additional details.

Text, Sparse Structures, and Action. Text inputs and outputs are tokenized using the byte-pair encoding [161] from LLaMA [177], which we chose since it supports Unicode symbols and preserves whitespace. Sparse structures such as bounding boxes, keypoints, and camera poses are discretized and then encoded using 1000 special tokens added to the vocabulary [27, 123]. Points are encoded with a sequence of two such tokens (one for x and one for y), boxes are encoded with a sequence of four tokens (upper left and lower right corners), and 3D cuboids are represented with 12 tokens that encode the projected center, virtual depth, log-normalized box dimension, and continuous allocentric rotation [16]. For embodied tasks, discrete robot actions [17] are generated as text commands (*e.g.*, "move ahead" to command the robot to move forward in navigation). Special tokens are used to encode the robot's state, such as its position and rotation.

Images and Dense Structures. Images are encoded with a pre-trained Vision Transformer (ViT) [84]. We concatenate the patch features from the second and second-to-last layers of the ViT to capture both low and high-level visual information. These features are passed through a linear layer to get embeddings that can be used as part of the input sequence for the transformer. To generate images, we use VQ-GAN [49] to convert images into discrete tokens. These tokens are added to the vocabulary and then used as the target output sequence in order to generate an image. For better image quality, we use a dense pre-trained VQ-GAN model with 8×8 patch size that encodes a 256×256 image into 1024 tokens with a codebook size of 16512.

Following [123], we represent per-pixel labels, which include depth, surface normals, and binary segmentation masks, as RGB images that can be generated or encoded with our image generation and encoding abilities. For segmentation, UNIFIED-IO 2 is trained to predict a binary mask given a class and bounding box. An entire image can be segmented by first doing detection, and then querying the model for a segmentation mask for each detected bounding box and class.

Audio. UNIFIED-IO 2 encodes up to 4.08 seconds of audio into a spectrogram (See Appendix D.1 and Table 7). The spectrogram is then encoded with a pre-trained Audio Spectrogram Transformer (AST) [57], and the input embeddings are built by concatenating the second and second-to-last layer features from the AST and applying a linear layer

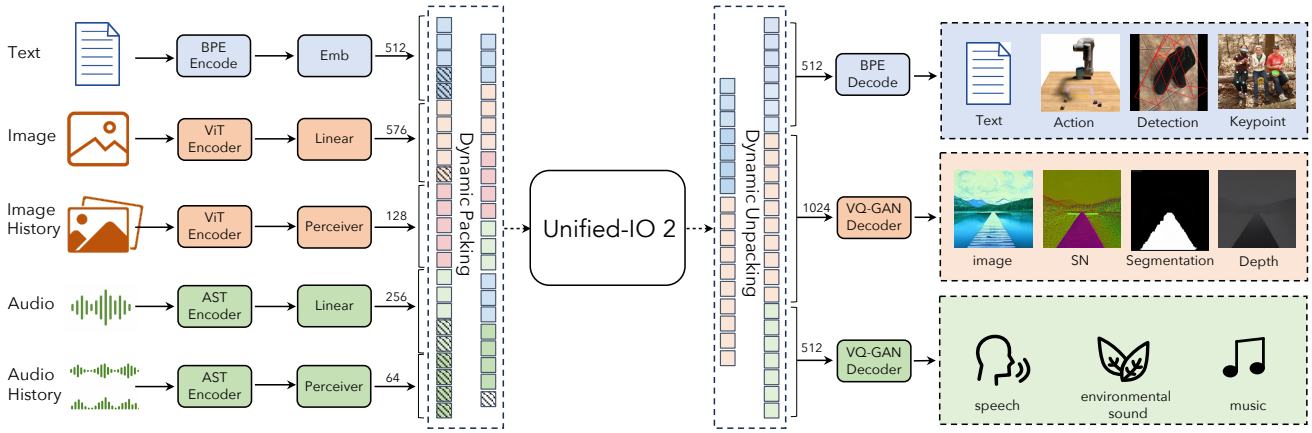

Figure 2. UNIFIED-IO 2 architecture. Input text, images, audio, or image/audio history are encoded into sequences of embeddings which are concatenated and used as input to an encoder-decoder transformer model. The transformer outputs discrete tokens that can be decoded into text, an image, or an audio clip.

just as with the image ViT. To generate audio, we use a ViT-VQGAN [208] to convert the audio into discrete tokens. Since there is no public codebase, we implement and train our own ViT-VQGAN with 8×8 patch size that encodes a 256×128 spectrogram into 512 tokens with a codebook size of 8196.

Image and Audio History. We allow up to four additional images and audio segments to be given as input, which we refer to as the image or audio history. These elements are also encoded using the ViT or AST, but we then use a perceiver resampler [5, 86], see Table 7 for hyperparameters, to further compress the features into a smaller number of tokens (32 for images and 16 for audio). This approach greatly reduces the sequence length and allows the model to inspect an image or audio segment in a high level of detail while using elements in the history for context. This history is used to encode previous video frames, previous audio segments, or reference images for tasks such as multi-view image reconstruction or image-conditioned image editing. Eight special tokens are added to the text vocabulary and used to reference the individual elements in these histories in the text input or output.

3.2. Architecture

UNIFIED-IO 2 uses a transformer encoder-decoder architecture. However, we observe that using a standard implementation following UNIFIED-IO leads to increasingly unstable training as we integrate additional modalities. To address this, we include various architectural changes that significantly stabilize multimodal training.

2D Rotary Embedding. We extend RoPE to two-dimensional positions: For any 2D indexes (i, j), we split each of the query and key embeddings of the transformer attention heads in half and apply separate rotary embeddings constructed by each of the two coordinates to the halves, see Appendix D.2.

QK Normalization. We observe extremely large values in the multi-head attention logits and find that applying LayerNorm [10] to the queries and keys following [38] mitigates the problem.

Scaled Cosine Attention. Even with QK normalization we observe extreme values in the logits of the perceive resampler used for the history inputs. Therefore, we apply more strict normalization in the perceiver by using scaled cosine attention [121], which significantly stabilizes training.

To avoid numerical instabilities, we also enable float32 attention logits. Jointly updating the pre-trained ViT and AST can also cause instabilities. Thus, we freeze the ViT and AST during pretraining and finetune them at the end of instruction tuning. More details and loss curves are in Appendix D.7.

Table 1 gives the details of our different models, see Appendix D.5 for additional hyperparameter details and Appendix D.7 for loss curves.

3.3. Training Objective

A strong multimodal model has to be exposed to solving diverse sets of problems during pre-training. UL2 [175] proposed the Mixture of Denoisers (MoD), a unified perspective to train LLMs, which combines the span corruption [147] and causal language modeling [19] objectives. Motivated by this, we propose a generalized and unified perspective for multimodal pre-training.

Multimodal Mixture of Denoisers. MoD uses three paradigms: `[R]` – standard span corruption, `[S]` – causal language modeling, and `[X]` – extreme span corruption. For text targets, we follow the UL2 paradigms. For image and audio targets, we define two analogous paradigms: `[R]` – masked denoising where we randomly mask $x\%$ of the input image or audio patch features and task the model

Model	model dims	mlp dims	encoder lyr	decoder lyr	heads	Params
UIO-2_{L}	1024	2816	24	24	16	1.1B
UIO-2_{XL}	2048	5120	24	24	16	3.2B
UIO-2_{XXL}	3072	8192	24	24	24	6.8B

Table 1. Size variant of UNIFIED-IO 2.

to re-construct it and `[S]` – where we ask the model to generate the target modality conditioned only on other input modalities. During training, we prefix the input text with a modality token (`[Text]`, `[Image]`, or `[Audio]`) and a paradigm token (`[R]`, `[S]`, or `[X]`) to indicate the task.

3.4. Efficient Implementation

Training on heavily multimodal data results in highly variable sequence lengths for the transformer's inputs and outputs, both because modalities are often missing for individual examples and because the number of tokens used to encode particular modalities can vary from just a few tokens (for a sentence) to 1024 tokens (for an output image). To handle this efficiently, we use packing, a process where the tokens of multiple examples are packed into a single sequence, and the attentions are masked to prevent the transformer from cross-attending between examples.

A complication in the multi-modal setting is that many modality-specific encoders cannot be run on packed inputs (*e.g.* the image ViT). As a solution, we apply the modality-specific encoders before packing, and then dynamically arrange the resulting features into packed sequences for the transformer. During training, we use a heuristic algorithm to re-arrange data being streamed to the model so that long examples are matched with short examples they can be packed with. This overall setup lets us train efficiently while using fixed-size tensors in the computation graph, as required by the Jax neural network library that our implementation uses [15]. Packing optimization was also explored in [100], but not in the streaming setup. Dynamic packing leads to an almost 4x increase in training throughput (Details in Appendix D.4).

4. Multimodal Data

One critical difference between UNIFIED-IO 2 and prior work is that we train the model with a diverse set of multimodal data from scratch. This requires curating high-quality, open-source multimodal data for both pre-training (4.1) and instruction tuning (4.2).

4.1. Pre-training Data

Our pre-training data comes from various sources and covers many modalities. We provide a high-level overview and details in Appendix E.

NLP [33%]. We use the publicly available datasets that were employed to train MPT-7B [176]. This dataset emphasizes English natural language text but also contains code and markdown. It includes text from the RedPajama dataset [32], C4 [68], Wikipedia, and stack overflow. We follow the proportion suggested by [176] and remove multilingual and scientific data.

Image & Text [40%]. Text and image paired data comes from LAION-400M [159], CC3M [163], CC12M [23], and RedCaps [42]. To help train the image-history modality, we also use the interleaved image/text data from OBELICS [104]. We use the last image as the image input and the remaining images as the image history. Special tokens are used to mark where those images occur in the text.

Video & Audio [25%]. Video provides strong self-supervisory signals with high correlations between audio and visual channels. We sample audio and video data from various public datasets including YT-Temporal-1B [215], ACAV100M [105], AudioSet [54], WebVid-10M [13], HD-VILA-10M [200] and Ego4D [60].

3D & Embodiment [1%]. For self-supervised 3D and embodiment pre-training, we use CroCo [194] for cross-view generation and denoising; Objaverse [40] for view synthesis; and random trajectories in ProcTHOR [39] and Habitat [157] for the next action and frame predictions.

Augmentation [1%]. While there is a lot of unsupervised data on the web for images, text, video, and audio, options are much more limited for dense and sparse annotations. We propose to solve this through large-scale data augmentation. We consider two types of data augmentation: 1. Automatically generated segmentation data from SAM [94] to train the model to segment an object given a point or bounding box. 2. Synthetic patch-detection data which tasks the model to list the bounding boxes of synthetically added shapes in an image. We additionally train the model to output the total number of patches in the image to pre-train its counting abilities.

Training Sample Construction. During pre-training, most of our data contains various modalities without a supervised target. In these cases, we randomly pick one of the modalities present to be the target output. Then, we either remove that modality from the example or replace it with a corrupted version. Other modalities that might be present in the example are randomly kept or masked to force the model to make predictions using whatever information is left. An example is shown in E.2.

4.2. Instruction Tuning Data

Multimodal instruction tuning is the key process to equip the model with diverse skills and capabilities across various modalities and even adapt to new and unique instructions. We construct the multimodal instruction tuning dataset by combining a wide range of supervised datasets and tasks. We ensure every task has a clear prompt, either using exist-

ing ones or writing new ones. We also include open-ended tasks and create synthetic tasks for less common modalities to enhance task and instruction variety. Our mixture includes 220 tasks drawn from over 120 external datasets. We provide a high-level overview and examples here and leave details and a visualization of the distribution in Appendix F.

Overall, our instruction tuning mixture is composed of 60% prompting data, meaning supervised datasets combined with prompts. To avoid catastrophic forgetting, 30% of the data is carried over from pre-training. Additionally, 6% is task augmentation data we build by constructing novel tasks using existing data sources, which enhances existing tasks and increases task diversity. The remaining 4% consists of free-form text to enable chat-like responses.

5. Experiments

In this section, we evaluate our pre-trained and instruction-tuned models on a broad range of tasks that require parsing and producing all modalities: images, video, audio, text, and actions. **We do not perform task-specific finetuning in any experiments.** Details about experimental setups, additional result details, results on natural language tasks, results on 3D object detection, and additional studies for UNIFIED-IO 2's instruction capabilities are in Appendix G.

5.1. Pre-training Evaluation

We demonstrate the effectiveness of our pre-training by evaluating UNIFIED-IO 2 on commonsense natural language inference (HellaSwag [214]), text-to-image generation (TIFA [76]) and text-to-audio generation (AudioCaps [93]). We also assess spatial and temporal understanding on SEED-Bench [106], a benchmark for comprehensively evaluating perception and reasoning on image and video modalities. Table 2 shows that UNIFIED-IO 2 achieves comparable or even better performance on both generation and comprehension tasks compared to the task-specific specialist [154] or the universal multimodal model [9].

Results on HellaSwag suggest that UNIFIED-IO 2 has decent language modeling capabilities, but is behind dedicated language models. This may be due to the fact that the model sees far fewer tokens compared to language-only LLMs – approximately 250 billion tokens in total. Qualitative results of pre-training are in Appendix G.1.

5.2. GRIT Results

We evaluate on the General Robust Image Task (GRIT) Benchmark [66], which includes seven tasks: categorization, localization, VQA, referring expression, instance segmentation, keypoint, and surface normal estimation. Completing all 7 tasks requires understanding image, text, and sparse inputs and generating text, sparse, and dense outputs. Although this is a subset of the modalities UNIFIED-IO 2 supports, we evaluate on GRIT because it provides a standardized and comprehensive benchmark on this set of capabilities. See Appendix G.4 for additional inference details on GRIT.

Method	HellaSwag↑	TIFA↑	SEED-S↑	SEED-T↑	AudioCaps↓
LLaMA-7B [177]	**76.1**	-	-	-	-
OpenLLaMa-3Bv2 [55]	70.0	-	-	-	-
KOSMOS- [143] 2	49.4	-	-	-	-
SD v1.5 [154]	-	78.4	-	-	-
OpenFlamingo-7B [9]	-	-	34.5	33.1	-
UIO-2_{L}	38.3	70.2	37.2	32.2	3.08
UIO-2_{XL}	47.6	77.2	40.9	34.0	3.10
UIO-2_{XXL}	54.3	**78.7**	**40.7**	**35.0**	**3.02**

Table 2. Zero-shot performance on commonsense sentence completion (HellaSwag [214]), text-to-image generation (TIFA [76]), spatial and temporal comprehension (Seed-Bench [106]), and text-to-audio generation (AudioCaps [93]).

	Method	Cat.	Loc.	Vqa	Ref.	Seg.	KP	Norm.	All
Ablation	UIO-2_{L}	70.1	66.1	67.6	66.6	53.8	56.8	44.5	60.8
	UIO-2_{XL}	74.2	69.1	69.0	71.9	57.3	68.2	**46.7**	65.2
	UIO-2_{XXL}	**74.9**	**70.3**	**71.3**	**75.5**	**58.2**	**72.8**	45.2	**66.9**
Test	GPV-2 [89]	55.1	53.6	63.2	52.1	-	-	-	-
	UIO$_{XL}$ [123]	60.8	67.1	**74.5**	**78.9**	56.5	67.7	44.3	64.3
	UIO-2_{XXL}	**75.2**	**70.2**	71.1	75.5	**58.8**	**73.2**	**44.7**	**67.0**

Table 3. Results on the GRIT ablation and test sets [66].

Results are shown in Table 3. Overall, UNIFIED-IO 2 is state-of-the-art on GRIT, surpassing the previous best model, UNIFIED-IO, by 2.7 points. On individual tasks, we can observe gains in localization (3 points), categorization (14 points), segmentation (2 points), and keypoint (5 points). On VQA, our GRIT evaluations show UNIFIED-IO 2 is better on same-source (84.6 *vs*. 81.2) questions, suggesting the gap is due to reduced performance on the new-source questions that were constructed from Visual Genome; see Appendix G.4 for additional discussion. Despite being slightly behind UNIFIED-IO, UNIFIED-IO 2 still obtains strong referring expression scores that compare favorably to prior work on generalist multimodal models, see Table 5. Surpassing UNIFIED-IO while also supporting much higher quality image and text generation, along with many more tasks and modalities, illustrates the impressive multi-tasking capabilities of our model. UNIFIED-IO 2 even maintains better overall performance with the 3-billion parameter model (65.2 *vs*. 64.5), which is roughly equal in size to UNIFIED-IO. Ablation results show average performance, and all individual tasks improve with model size, showing that UNIFIED-IO 2 benefits from scale.

5.3. Generation Results

Table 4 shows results on tasks that require generating image, audio, and action outputs. We evaluate using

Method	Image		Audio			Action
	FID↓	TIFA↑	FAD↓	IS↑	KL↓	Succ.↑
minDALL-E [37]	-	79.4	-	-	-	-
SD-1.5 [154]	-	78.4	-	-	-	-
AudioLDM-L [117]	-	-	1.96	8.13	1.59	-
AudioGen [101]	-	-	3.13	-	2.09	-
DiffSound [203]	-	-	7.75	4.01	2.52	-
VIMA [87]	-	-	-	-	-	72.6
VIMA-IMG [87]	-	-	-	-	-	42.5
CoDi [174]	11.26	71.6	1.80	8.77	1.40	-
Emu [172]	11.66	65.5	-	-	-	-
UIO-2_{L}	16.68	74.3	2.82	5.37	1.93	50.2
UIO-2_{XL}	14.11	80.0	2.59	5.11	1.74	54.2
UIO-2_{XXL}	13.39	81.3	2.64	5.89	1.80	56.3

Table 4. Results on text-to-image generation (MS COCO [115] and TIFA [76]), text-to-audio generation (AudioCaps [93]) and action generation (VIMA-Bench [87]).

TIFA [76], which measures faithfulness to the prompt using VQA models and has been shown to correlate well with human judgments, and FID [73] on MS COCO [115]. On TIFA, we find that UNIFIED-IO 2 scores close to minDALL-E [37], and about 10 points ahead of other generalist models such as CoDi [174] and Emu [172]. We attribute this strong image generation ability to extensive pre-training and the use of a fine-grained VQ-GAN. We include examples of our generation results from the TIFA benchmark in the Appendix G.6. UNIFIED-IO 2's FID scores are slightly higher than the compared models, although we note that qualitatively the generated images are still very smooth and detailed.

For text-to-audio generation, we evaluate on the AudioCaps [93] test set. AudioCaps consists of 10-second audio clips, while our model can generate 4.08-second audio at a time, so we cannot do a direct evaluation on this benchmark. Instead, we generate an audio segment based on the text description and previous audio segments as additional input; see Appendix G.7 for more details. While this is not a directly comparable setup to related work, it still gives a reasonable quantitative measure of our audio generation abilities. UNIFIED-IO 2 scores higher then specialist models except the recent latent diffusion model [117], which shows it's competitive audio generation ability.

For action, we evaluate using VIMA-Bench [87], a robot manipulation benchmark containing 17 tasks with text-image interleaved prompts. Since VIMA's action space is action primitives, UNIFIED-IO 2 directly predicts all actions at once given the initial observation and multimodal prompt. We report the average success rate for 4-level evaluation protocol [87] and compare with the original casual VIMA policy with object-centric inputs, as well as VIMA-IMG, a Gato [152]-like policy with image inputs like ours.

5.4. Vision Language Results

We evaluate vision language performance and compare it against other vision/language generalist models, *i.e.*, models that are also designed to perform many tasks and can follow instructions. Results on a collection of 12 vision/language benchmarks are shown in Table 5. SoTA results from specialist models are shown for reference.

UNIFIED-IO 2 achieves strong results on VQA, only passed by the much larger 13B LLaVa model [118] on VQA v2 [59], and ahead of all other generalist models on ScienceQA [124] and TallyQA [1]. OK-VQA [130] is the exception. We hypothesize that because it requires external knowledge, extensive language pre-training is important for this task, and therefore our reduced performance is since UNIFIED-IO 2 was not pre-trained as extensively on text as the dedicated language models used by Qwen-VL [12] and mPLUG-Owl2 [206].

On referring expression, UNIFIED-IO 2 is ahead of Shikra [26] and Ferret [207] and matches the scores achieved by Qwen-VL. On captioning, UNIFIED-IO 2 also achieves a strong CIDEr score [182] of 130.3, ahead of Shikra and InstructBLIP [34] but behind Qwen-VL and mPLUG-Owl2.

Finally, we evaluate using three recently proposed evaluation-only benchmarks. MMB (MMBench [120]) tests multiple facets of vision language understanding with multiple choice questions, while SEED-Bench additionally tests video understanding. We show a detailed breakdown of our score in the Appendix G.5. Regarding the overall score, UNIFIED-IO 2 has the strongest score of any 7B model on the SEED-Bench leaderboard[1], and scores the highest on MMB by 3.8 points. Notably, it excels LLaVa-1.5 13B model in both benchmarks. UNIFIED-IO 2 also reaches 87.7 on the POPE object hallucination benchmark [113], showing that it is not very prone to object hallucination.

Overall, UNIFIED-IO 2 can match or surpass other vision & language generalist models on these benchmarks despite encompassing many more modalities and supporting high-quality image and audio generation. This shows that its wide breadth of capabilities does not come at the expense of vision/language performance.

5.5. Video, Audio and other Results

UNIFIED-IO 2 shows reasonable performance on audio and video classification and captioning, as well as video question answering, as shown in Table 6. Notably, UNIFIED-IO 2 outperforms BLIP-2 [109] and InstructBLIP [34] on Seed-Bench Temporal [106] by 8.5 points. UNIFIED-IO 2 also achieves better performance on Kinetics-Sounds [7] than MBT [137], which is trained

[1] as of 11/17/23

Method	VQA^{v2}	OKVQA	SQA	SQA^{I}	Tally-QA	RefCOCO	RefCOCO+	RefCOCO-g	COCO-Cap.	POPE	SEED	MMB
InstructBLIP (8.2B)	-	-	-	79.5	68.2†	-	-	-	102.2	-	53.4	36
Shikra (7.2B)	77.4	47.2	-	-	-	87.0	81.6	82.3	117.5	84.7	-	58.8
Ferret (7.2B)	-	-	-	-	-	87.5	80.8	83.9	-	85.8	-	-
Qwen-VL (9.6B)	78.8	58.6	-	67.1*	-	89.4	83.1	85.6	131.9	-		38.2
mPLUG-Owl2 (8.2B)	79.4	57.7	-	68.7*	-	-	-	-	137.3	86.2	57.8	64.5
LLaVa-1.5 (7.2B)	78.5	-	-	66.8*	-	-	-	-	-	85.9	58.6	64.3
LLaVa-1.5 (13B)	80.0	-	-	71.6*	72.4†	-	-	-	-	85.9	61.6	67.7
Single Task SoTA	86.0 [29]	66.8 [77]	90.9 [119]	90.7 [34]	82.4 [77]	92.64 [202]	88.77 [187]	89.22 [187]	149.1 [29]	-	-	-
UIO-2$_{L}$ (1.1B)	75.3	50.2	81.6	78.6	69.1	84.1	71.7	79.0$^{\diamond}$	128.2	77.8	51.1	62.1
UIO-2$_{XL}$ (3.2B)	78.1	53.7	88.8	87.4	72.2	88.2	79.8	84.0$^{\diamond}$	130.3	87.2	60.2	68.1
UIO-2$_{XXL}$ (6.8B)	79.4	55.5	88.7	86.2	75.9	90.7	83.1	86.6$^{\diamond}$	125.4	87.7	61.8	71.5

Table 5. Vision-language results on nine tasks [1, 28, 59, 91, 124, 129, 130, 136, 209] and three evaluation-only benchmarks [106, 113, 120]. Results marked with * are zero-shot and † are evaluated with the open-source releases, and $^{\diamond}$ indicates that our RefCOCO-g results are on the Google split rather than the UMD split.

	Video							Audio		
Method	Kinetics-400 [90]	VATEXCaption [190]	MSR-VTT [199]	MSRVTT-QA [198]	MSVD-QA [198]	STAR [196]	SEED-T [106]	VGG-Sound [24]	AudioCaps [93]	Kinetics-Sounds [7]
MBT [137]	-	-	-	-	-	-	-	52.3	-	85.0
CoDi [174]	-	-	74.4	-	-	-	-	-	78.9	-
ImageBind [69]*	50.0	-	-	-	-	-	-	27.8	-	-
BLIP-2 [109]*	-	-	-	9.2	18.3	-	36.7	-	-	-
InstructBLIP [34]*	-	-	-	22.1	41.8	-	38.3	-	-	-
Emu [172]**	-	-	-	24.1	39.8	-	-	-	-	-
Flamingo-9B [5]**	-	57.4	-	29.4	47.2	41.2	-	-		
Flamingo-80B [5]	-	84.2	-	47.4	-	-	-	-	-	-
UIO-2$_{L}$	68.5	37.1	44.0	39.6	48.2	51.0	37.5	37.8	45.7	86.1
UIO-2$_{XL}$	71.4	41.6	47.1	39.3	50.4	52.0	45.6	44.2	45.7	88.0
UIO-2$_{XXL}$	73.8	45.6	48.8	41.5	52.1	52.2	46.8	47.7	48.9	89.3

Table 6. Results on action classification, video captioning, VQA, visual comprehension, audio classification, and audio captioning. *: zero-shot, **: few-shot in-context learning.

solely on that dataset.

In COCO object detection, excluding the 'stuff' categories, our model reached an average precision (AP) of 47.2, with AP50 at 57.7 and AP75 at 50.0. However, it has difficulties with images containing many objects. Previous research, like Pix2Seq [27], suggests that autoregressive models face similar challenges, which can be improved with extensive data augmentation. Our model's data augmentation on object detection is comparatively more limited.

Our model shows weak performance in depth estimation, with an RMSE of 0.623 on NYUv2 depth dataset [138]. However, fine-tuning specifically for this task improved the RMSE to 0.423. In our experiment, we simply normalize the depth map with the max depth value in each dataset. Due to the incompatibility of dense ground-truth depth across different datasets [150], our model failed to capture the exact scale in the current prompt, which could potentially be solved by using better normalization and metric evaluation.

Appendix G shows qualitative visualizations of other tasks, such as single object tracking, future state prediction of robotic manipulation, and image-based 3D view synthesis, *etc*.

6. Conclusion

We introduced UNIFIED-IO 2, the first autoregressive multimodal model that is capable of understanding and generating image, text, audio, and action. This model was trained from scratch on a wide range of multimodal data and further refined with instruction tuning on a massive multimodal corpus. We developed various architectural changes to stabilize the multimodal training and proposed a multimodal mixture of denoiser objective to effectively utilize the multimodal signals. Our model achieves promising results across a wide range of tasks. We show that going from LLMs to LMMs enables new capabilities and opportunities. In the future, we would like to extend UNIFIED-IO 2 from the encoder-decoder model to a decoder-only model. Additionally, we plan to expand the model's size, enhance the data quality, and refine the overall model design.

Acknowledgement We thank Klemen Kotar for helping gather Embodied AI pre-training data, Jonathan Frankle from MosaicML for suggesting the mixture of NLP pre-training data, Jack Hessel for interleaved image & text dataset and Micheal Schmitz for helping support the compute infrastructure. We also thank Tanmay Gupta for helpful discussions, as well as Hamish Ivison, and Ananya Harsh Jha for their insightful discussions about model design. We additionally thank Oscar Michel, Yushi Hu and Yanbei Chen for their help editing the paper, and Matt Deitke for help setting up the webpage. Savya Khosla and Derek Hoiem were supported in part by ONR award N00014-23-1-2383. This research was made possible with cloud TPUs from Google's TPU Research Cloud (TRC).

References

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering Complex Counting Questions. In *AAAI*, 2019. 7, 8, 28

[2] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. MusicLM: Generating Music From Text. *arXiv preprint arXiv:2301.11325*, 2023. 25, 28

[3] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A Large Scale Dataset of Object-Centric Videos in the Wild with Pose Annotations. In *CVPR*, 2021. 32, 33, 37

[4] Emanuele Aiello, Lili Yu, Yixin Nie, Armen Aghajanyan, and Barlas Oguz. Jointly Training Large Autoregressive Multimodal Models. *arXiv preprint arXiv:2309.15564*, 2023. 3

[5] 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. In *NeurIPS*, 2022. 2, 4, 8, 20, 36, 37

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 25

[7] Relja Arandjelovic and Andrew Zisserman. Look, Listen and Learn. In *ICCV*, 2017. 7, 8, 25, 30

[8] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program Synthesis with Large Language Models. *arXiv preprint arXiv:2108.07732*, 2021. 28

[9] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*, 2023. 6, 20

[10] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. In *NeurIPS Deep Learning Symposium*, 2016. 4

[11] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and Exploiting the Aleatoric Uncertainty in Surface Normal Estimation. In *ICCV*, 2021. 32

[12] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*, 2023. 3, 7, 33

[13] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *ICCV*, 2021. 5, 24

[14] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. AudioLM: A Language Modeling Approach to Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. 2

[15] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. Jax: composable transformations of python+numpy programs, 2018. 5, 21

[16] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3D: A Large Benchmark and Model for 3D Object Detection in the Wild. In *CVPR*, 2023. 3, 19, 25, 29, 32, 33

[17] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *CoRL*, 2023. 3

[18] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*, 2023. 25, 28

[19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. In *NeurIPS*, 2020. 4, 36, 37

[20] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023. 3

[21] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A Naturalistic Open Source Movie for Optical Flow Evaluation. In *ECCV*, 2012. 26, 29

[22] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 32

[23] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *CVPR*, 2021. 5, 24

[24] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A Large-Scale Audio-Visual Dataset. In *ICASSP*, 2020. 8, 25, 30

[25] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: Large Language Model As a Unified Interface for Vision-Language Multi-task Learning. *arXiv preprint arXiv:2310.09478*, 2023. 3

[26] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*, 2023. 3, 7, 33

[27] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A Language Modeling Framework for Object Detection. In *ICLR*, 2022. 3, 8

[28] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*, 2015. 8

[29] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. PaLI-X: On Scaling up a Multilingual Vision and Language Model. *arXiv preprint arXiv:2305.18565*, 2023. 8

[30] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *NAACL-HLT*, 2019. 33

[31] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018. 33

[32] Together Computer. RedPajama: an Open Dataset for Training Large Language Models. https://github.com/togethercomputer/RedPajama-Data, 2023. 5, 24

[33] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm, 2023. 25, 28

[34] 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. In *NeurIPS*, 2023. 3, 7, 8, 33

[35] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130:33–55, 2022. 25, 29

[36] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *CVPR*, 2017. 28

[37] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khac, Luke Melas, and Ritobrata Ghosh. DALL·E Mini, 2021. 7, 34

[38] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling Vision Transformers to 22 Billion Parameters. In *ICML*, 2023. 4

[39] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *NeurIPS*, 2022. 5, 24, 25

[40] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A Universe of Annotated 3D Objects. In *CVPR*, 2023. 5, 24, 25, 28, 37

[41] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 19, 25

[42] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks Track*, 2021. 5, 24

[43] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. CogView: Mastering Text-to-Image Generation via Transformers. In *NeurIPS*, 2021. 34

[44] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning Optical Flow with Convolutional Networks. In *ICCV*, 2015. 26, 29

[45] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 19

[46] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An Embodied Multimodal Language Model. In *ICML*, 2023. 2, 31

[47] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An Audio Captioning Dataset. In *ICASSP*, 2020. 25, 28

[48] Gerald M Edelman. Neural Darwinism: Selection and reentrant signaling in higher brain function. *Neuron*, 10(2):115–125, 1993. 2

[49] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*, 2021. 3, 19

[50] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, et al. LaSOT: A High-quality Large-scale Single Object Tracking Benchmark. *International Journal of Computer Vision*, 129:439–461, 2021. 28, 29, 36, 37

[51] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 2021. 32

[52] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He,

Xiangyu Yue, et al. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv preprint arXiv:2304.15010*, 2023. 3

[53] 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*, 2023. 3

[54] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *ICASSP*, 2017. 5, 20, 24, 25, 30

[55] Xinyang Geng and Hao Liu. OpenLLaMA: An Open Reproduction of LLaMA. https://github.com/openlm-research/open_llama, 2023. 6

[56] 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. In *CVPR*, 2023. 3

[57] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Interspeech*, 2021. 3, 20

[58] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 29

[59] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*, 2017. 7, 8

[60] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *CVPR*, 2022. 5, 24

[61] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. In *CVPR*, 2018. 28, 29

[62] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *CVPR*, 2019. 28

[63] Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay Policy Learning: Solving Long-Horizon Tasks via Imitation and Reinforcement Learning. In *CoRL*, 2019. 28, 30

[64] Tanmay Gupta and Aniruddha Kembhavi. Visual Programming: Compositional visual reasoning without training. In *CVPR*, 2023. 2

[65] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards General Purpose Vision Systems: An End-to-End Task-Agnostic Vision-Language Architecture. In *CVPR*, 2022. 32

[66] Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. GRIT: General Robust Image Task Benchmark. *arXiv preprint arXiv:2204.13653*, 2022. 2, 6, 20

[67] 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. In *CVPR*, 2018. 28

[68] Ivan Habernal, Omnia Zayed, and Iryna Gurevych. C4Corpus: Multilingual Web-size Corpus with Free License. In *LREC*, 2016. 5, 24

[69] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. ImageBind-LLM: Multi-modality Instruction Tuning. *arXiv preprint arXiv:2309.03905*, 2023. 3, 8

[70] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 32

[71] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, 2022. 22

[72] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *ICLR*, 2021. 33

[73] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*, 2017. 7, 34

[74] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021. 28, 34

[75] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In *ICLR*, 2020. 28

[76] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. In *ICCV*, 2023. 2, 6, 7, 34, 35

[77] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models. *arXiv preprint arXiv:2312.03052*, 2023. 8

[78] Jingwei Huang, Yichao Zhou, Thomas Funkhouser, and Leonidas J Guibas. FrameNet: Learning Local Canonical Frames of 3D Surfaces from a Single RGB Image. In *ICCV*, 2019. 26, 29

[79] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2019. 28, 29

[80] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. *arXiv preprint arXiv:2304.12995*, 2023. 3

[81] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language Is Not

All You Need: Aligning Perception with Language Models. In *NeurIPS*, 2023. 2, 18

[82] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual Storytelling. In *NAACL-HLT*, 2016. 28

[83] Drew A Hudson and Christopher D Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *CVPR*, 2019. 28

[84] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. 3

[85] Keith Ito and Linda Johnson. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. 20, 25, 28

[86] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In *ICLR*, 2022. 4

[87] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: General Robot Manipulation with Multimodal Prompts. In *ICML*, 2023. 7, 19, 28, 30, 36, 37

[88] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, et al. Unified Language-Vision Pretraining in LLM with Dynamic Discrete Visual Tokenization. *arXiv preprint arXiv:2309.04669*, 2023. 3

[89] Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly Supervised Concept Expansion for General Purpose Vision Models. In *ECCV*, 2022. 6, 32

[90] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*, 2017. 8, 34

[91] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*, 2014. 8, 25, 29

[92] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Interspeech*, 2019. 34

[93] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in The Wild. In *NAACL-HLT*, 2019. 6, 7, 8, 25, 28, 34

[94] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *ICCV*, 2023. 5, 24, 25, 32

[95] Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. The Stack: 3 TB of permissively licensed source code. *Transactions on Machine Learning Research*, 2023. 24

[96] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating Images with Multimodal Language Models. In *NeurIPS*, 2023. 2, 3

[97] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding Language Models to Images for Multimodal Inputs and Outputs. In *ICML*, 2023. 3

[98] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *NeurIPS*, 2020. 20

[99] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. OpenAssistant Conversations – Democratizing Large Language Model Alignment. In *NeurIPS Datasets and Benchmarks Track*, 2023. 28

[100] Mario Michael Krell, Matej Kosec, Sergio P Perez, and Andrew Fitzgibbon. Efficient Sequence Packing without Cross-contamination: Accelerating Large Language Models without Impacting Performance. *arXiv preprint arXiv:2107.02027*, 2021. 5

[101] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually Guided Audio Generation. In *ICLR*, 2023. 7

[102] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 2017. 25, 28

[103] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 20, 25, 28

[104] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *NeurIPS Datasets and Benchmarks Track*, 2023. 5, 22, 24

[105] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. ACAV100M: Automatic Curation of Large-Scale Datasets for Audio-Visual Video Representation Learning. In *ICCV*, 2021. 5, 20, 24

[106] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 6, 7, 8, 33

[107] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu.

MIMIC-IT: Multi-Modal In-Context Instruction Tuning. *arXiv preprint arXiv:2306.05425*, 2023. 25, 28

[108] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A Multi-Modal Model with In-Context Instruction Tuning. *arXiv preprint arXiv:2305.03726*, 2023. 28, 33, 34

[109] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, 2023. 3, 7, 8

[110] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-Centric Video Understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3, 33, 34

[111] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. UniFormerV2: Unlocking the Potential of Image ViTs for Video Understanding. In *ICCV*, 2023. 25, 29, 34

[112] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M^3IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning. *arXiv preprint arXiv:2306.04387*, 2023. 3, 25, 28

[113] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *EMNLP*, 2023. 7, 8, 33

[114] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Russ Salakhutdinov. High-Modality Multimodal Transformer: Quantifying Modality & Interaction Heterogeneity for High-Modality Representation Learning. *TMLR*, 2023. 18

[115] 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 *ECCV*, 2014. 7, 25, 28, 34

[116] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual Spatial Reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 28

[117] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *ICML*, 2023. 7, 34

[118] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. *arXiv preprint arXiv:2310.03744*, 2023. 3, 7, 33

[119] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *NeurIPS*, 2023. 2, 3, 8, 25, 28

[120] Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281*, 2023. 7, 8

[121] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin Transformer V2: Scaling Up Capacity and Resolution. In *CVPR*, 2022. 4

[122] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. In *ICML*, 2023. 3, 25, 28

[123] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks. In *ICLR*, 2023. 2, 3, 6, 19, 25

[124] 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 *NeurIPS*, 2022. 7, 8, 28

[125] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models. In *NeurIPS*, 2023. 2

[126] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Ming-Hui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video Assistant with Large Language model Enhanced abilitY. *arXiv preprint arXiv:2306.07207*, 2023. 3

[127] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive Language: Talking to Robots in Real Time. *IEEE Robotics and Automation Letters*, 2023. 28, 30

[128] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv preprint arXiv:2306.05424*, 2023. 34

[129] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR*, 2016. 8

[130] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *CVPR*, 2019. 7, 8, 28

[131] Irene Martin Morato and Annamaria Mesaros. Diversity and Bias in Audio Captioning Datasets. In *DCASE*, 2021. 25, 28

[132] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly Mapping from Image to Text Space. In *ICLR*, 2023. 3

[133] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual Question Answering by Reading Text in Images. In *ICDAR*, 2019. 28

[134] Utkarsh Mishra, Shangjie Xue, Yongxin Chen, and Danfei Xu. Generative Skill Chaining: Long-Horizon Skill Planning with Diffusion Models. In *CoRL*, 2023. 31

[135] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. EmbodiedGPT: Vision-Language Pre-Training via Embodied Chain of Thought. In *NeurIPS*, 2023. 3

[136] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling Context Between Objects for Referring Expression Understanding. In *ECCV*, 2016. 8, 29

[137] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention Bottlenecks for Multimodal Fusion. In *NeurIPS*, 2021. 7, 8

[138] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*, 2012. 8, 26, 29

[139] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. 32

[140] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. In *CoRL Workshop TGR*, 2023. 3

[141] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-G: Generating Images in Context with Multimodal Large Language Models. *arXiv preprint arXiv:2310.02992*, 2023. 3

[142] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction Tuning with GPT-4. *arXiv preprint arXiv:2304.03277*, 2023. 25, 28

[143] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv preprint arXiv:2306.14824*, 2023. 2, 3, 6, 18

[144] Jean Piaget, Margaret Cook, et al. *The Origins of Intelligence in Children*. International Universities Press New York, 1952. 2

[145] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild. In *NeurIPS*, 2023. 28

[146] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 32, 34

[147] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 21(140): 1–67, 2020. 4, 19

[148] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *ICML*, 2021. 21

[149] Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-Web: Learning Embodied Object-Search Strategies from Human Demonstrations at Scale. In *CVPR*, 2022. 25

[150] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2020. 8

[151] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video. In *CVPR*, 2017. 28, 29

[152] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A Generalist Agent. *Transactions on Machine Learning Research*, 2022. 3, 7, 36, 37

[153] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *ICCV*, 2021. 32

[154] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*, 2022. 2, 6, 7, 34

[155] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *NeurIPS*, 2016. 34

[156] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *ICLR*, 2022. 28

[157] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A Platform for Embodied AI Research. In *ICCV*, 2019. 5, 24, 25, 30

[158] Christoph Schuhmann. LAION-AESTHETICS. https://laion.ai/blog/laion-aesthetics/, 2022. 24

[159] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *NeurIPS Data-Centric AI Workshop*, 2021. 5, 24

[160] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. In *ECCV*, 2022. 28

[161] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *ACL*, 2016. 3

[162] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. RoboVQA: Multimodal Long-Horizon Reasoning for Robotics. *arXiv preprint arXiv:2311.00899*, 2023. 31

[163] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*, 2018. 5, 24

[164] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. In *ICML*, 2018. 22

[165] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *CVPR*, 2019. 28

[166] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. ProgPrompt: Generating Situated Robot Task Plans using Large Language Models. *ICRA*, 2023. 2

[167] Linda Smith and Michael Gasser. The Development of Embodied Cognition: Six Lessons from Babies. *Artificial life*, 11(1-2):13–29, 2005. 2

[168] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402*, 2012. 25, 29

[169] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 2023. 20

[170] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. PandaGPT: One Model To Instruction-Follow Them All. *arXiv preprint arXiv:2305.16355*, 2023. 3

[171] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A Corpus for Reasoning About Natural Language Grounded in Photographs. In *ACL*, 2019. 28

[172] 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. *arXiv preprint arXiv:2307.05222*, 2023. 3, 7, 8, 34

[173] Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual Inference via Python Execution for Reasoning. In *ICCV*, 2023. 2

[174] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-Any Generation via Composable Diffusion. In *NeurIPS*, 2023. 2, 7, 8, 18, 34

[175] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. UL2: Unifying Language Learning Paradigms. In *ICLR*, 2023. 4

[176] MosaicML NLP Team. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs, 2023. Accessed: 2023-05-05. 5, 22

[177] 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 arXiv:2302.13971*, 2023. 3, 6, 20

[178] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, Felix Hill, and Zacharias Janssen. Multimodal Few-Shot Learning with Frozen Language Models. In *NeurIPS*, 2021. 3

[179] Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In *NeurIPS*, 2017. 3

[180] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist Species Classification and Detection Dataset. In *CVPR*, 2018. 28

[181] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NeurIPS*, 2017. 3

[182] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*, 2015. 7, 36

[183] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images. *arXiv preprint arXiv:1601.07140*, 2016. 25, 29

[184] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, et al. BridgeData V2: A Dataset for Robot Learning at Scale. In *CoRL*, 2023. 28, 30

[185] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks for Action Recognition in Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (11):2740–2755, 2019. 29

[186] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *ICML*, 2022. 3, 18, 32

[187] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. ONE-PEACE: Exploring One General Representation Model Toward Unlimited Modalities. *arXiv preprint arXiv:2305.11172*, 2023. 8

[188] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. In *CVPR*, 2023. 18

[189] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks. In *NeurIPS*, 2023. 3

[190] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *ICCV*, 2019. 8, 25, 29

[191] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *EMNLP*, 2022. 28

[192] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *ICLR*, 2022. 18

[193] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned Language Models are Zero-Shot Learners. In *ICLR*, 2022. 28

[194] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion. In *NeurIPS*, 2022. 5, 24, 25, 28

[195] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 28

[196] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *NeurIPS Datasets and Benchmarks Track*, 2021. 8, 25, 29

[197] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. SUN Database: Large-scale Scene Recognition from Abbey to Zoo. In *CVPR*, 2010. 28

[198] 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 *ACM MM*, 2017. 8, 25, 29

[199] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *CVPR*, 2016. 8, 25, 29

[200] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions. In *CVPR*, 2022. 5, 24

[201] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *NAACL-HLT*, 2021. 24

[202] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Universal Instance Perception as Object Discovery and Retrieval. In *CVPR*, 2023. 8

[203] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete Diffusion Model for Text-to-sound Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023. 7

[204] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A Large-scale Dataset for Generalized Multi-view Stereo Networks. In *CVPR*, 2020. 26, 29

[205] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 3

[206] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. *arXiv preprint arXiv:2311.04257*, 2023. 7, 33

[207] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and Ground Anything Anywhere at Any Granularity. *arXiv preprint arXiv:2310.07704*, 2023. 3, 7, 33

[208] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized Image Modeling with Improved VQGAN. In *ICLR*, 2022. 4, 20

[209] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling Context in Referring Expressions. In *ECCV*, 2016. 8, 29

[210] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning. *arXiv preprint arXiv:2309.02591*, 2023. 3, 18

[211] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling Task Transfer Learning. In *CVPR*, 2018. 29

[212] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual Object Detection with Multimodal Large Language Models. *arXiv preprint arXiv:2305.18279*, 2023. 3

[213] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From Recognition to Cognition: Visual Commonsense Reasoning. In *CVPR*, 2019. 28

[214] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence? In *ACL*, 2019. 6, 33

[215] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound. In *CVPR*, 2022. 5, 20, 24, 28

[216] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3

[217] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing. In *NeurIPS Datasets and Benchmarks Track*, 2023. 25, 28

[218] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. *arXiv preprint arXiv:2303.16199*, 2023. 3

[219] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. *arXiv preprint arXiv:2307.03601*, 2023. 3

[220] 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*, 2023. 3, 25

[221] Minyi Zhao, Bingjia Li, Jie Wang, Wanqing Li, Wenjing Zhou, Lan Zhang, Shijie Xuyang, Zhihang Yu, Xinkun Yu, Guangze Li, et al. Towards Video Text Visual Question Answering: Benchmark and Baseline. In *NeurIPS Datasets and Benchmarks Track*, 2022. 29

[222] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst. *arXiv preprint arXiv:2305.16103*, 2023. 3

[223] Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. MoVQ: Modulating Quantized Vectors for High-Fidelity Image Generation. In *NeurIPS*, 2022. 20

[224] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 28

[225] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*, 2023. 3

[226] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis. In *CVPR*, 2019. 34