This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Generative Multimodal Models are In-Context Learners

Quan Sun[1*] Yufeng Cui[1*] Xiaosong Zhang[1*] Fan Zhang[1*] Qiying Yu[2,1*] Yueze Wang[1]
Yongming Rao[1] Jingjing Liu[2] Tiejun Huang[1,3] Xinlong Wang[1†]

[1] Beijing Academy of Artificial Intelligence [2] Tsinghua University [3] Peking University

[*]equal contribution [†]project lead

code & models: https://github.com/baaivision/Emu

Abstract

The human ability to easily solve multimodal tasks in context (i.e., with only a few demonstrations or simple instructions), is what current multimodal systems have largely struggled to imitate. In this work, we demonstrate that the task-agnostic in-context learning capabilities of large multimodal models can be significantly enhanced by effective scaling-up. We introduce ***Emu2****, a generative multimodal model with 37 billion parameters, trained on large-scale multimodal sequences with a unified autoregressive objective.* ***Emu2*** *exhibits strong multimodal in-context learning abilities, even emerging to solve tasks that require on-the-fly reasoning, such as visual prompting and object-grounded generation. The model sets a new record on multiple multimodal understanding tasks in few-shot settings. When instruction-tuned to follow specific instructions,* ***Emu2*** *further achieves new state-of-the-art on challenging tasks such as question answering benchmarks for large multimodal models and open-ended subject-driven generation. These achievements demonstrate that* ***Emu2*** *can serve as a base model and general-purpose interface for a wide range of multimodal tasks. Code and models are publicly available to facilitate future research.*

1. Introduction

Multimodal tasks [25, 41] encompass anything involving understanding and generation in single or multiple modalities [5, 19, 58], which can be highly diverse and long-tail. Previous multimodal systems largely rely on designing task-specific architecture and collecting a sizable supervised training set, both of which are difficult to scale, particularly when this process needs to be repeated for each new task encountered. By contrast, humans can solve a new task in context, *i.e.*, with only a few demonstrations or simple instructions – a capability that current multimodal models have yet to learn.

[†]Correspondence to *wangxinlong@baai.ac.cn*

Recently, generative pretrained language models have demonstrated strong in-context learning abilities [11, 21, 73]. By training a 37-billion-parameter model **Emu2** and thoroughly evaluating it on diverse multimodal tasks, we demonstrate that a scaled-up multimodal generative pretrained model can harness similar in-context learning abilities and effectively generalize to unseen multimodal tasks. **Emu2** is trained with a unified autoregressive objective: predict-the-next-multimodal-element (either visual embeddings or textual tokens). In this unified generative pretraining process, large-scale multimodal sequences (*e.g.*, text, image-text pairs, and interleaved image-text-video) are used for model training.

We measure **Emu2**'s capabilities of learning from a few examples or instructions on standard multimodal datasets, as well as new tasks unseen in the training set. Specifically, **Emu2** is evaluated under two scenarios: (a) *few-shot setting*, where we allow as many examples as possible to fit the context window of the model; and (b) *instruction tuning*, where the model is tuned to follow specific instructions.

Emu2 achieves promising results in the few-shot setting on a wide range of vision-language tasks. For example, it demonstrates state-of-the-art few-shot performance on multiple visual question-answering datasets. We observe a performance improvement when the number of examples in context increases. Figure 1 illustrates **Emu2**'s strong multimodal reasoning capabilities for tasks in the wild, *e.g.*, recognition and counting in a specific format. **Emu2** also learns to follow visual prompting in context (*e.g.*, the circles laid on the images in Figure 1), even although it struggles at a smaller scale or at zero shot.

As **Emu2** is inherently equipped to handle interleaved text-image-video at both input and output, it serves as a powerful and versatile base model for diverse multimodal tasks, by following specific task instructions. For example, after instruct tuning with conversational data,

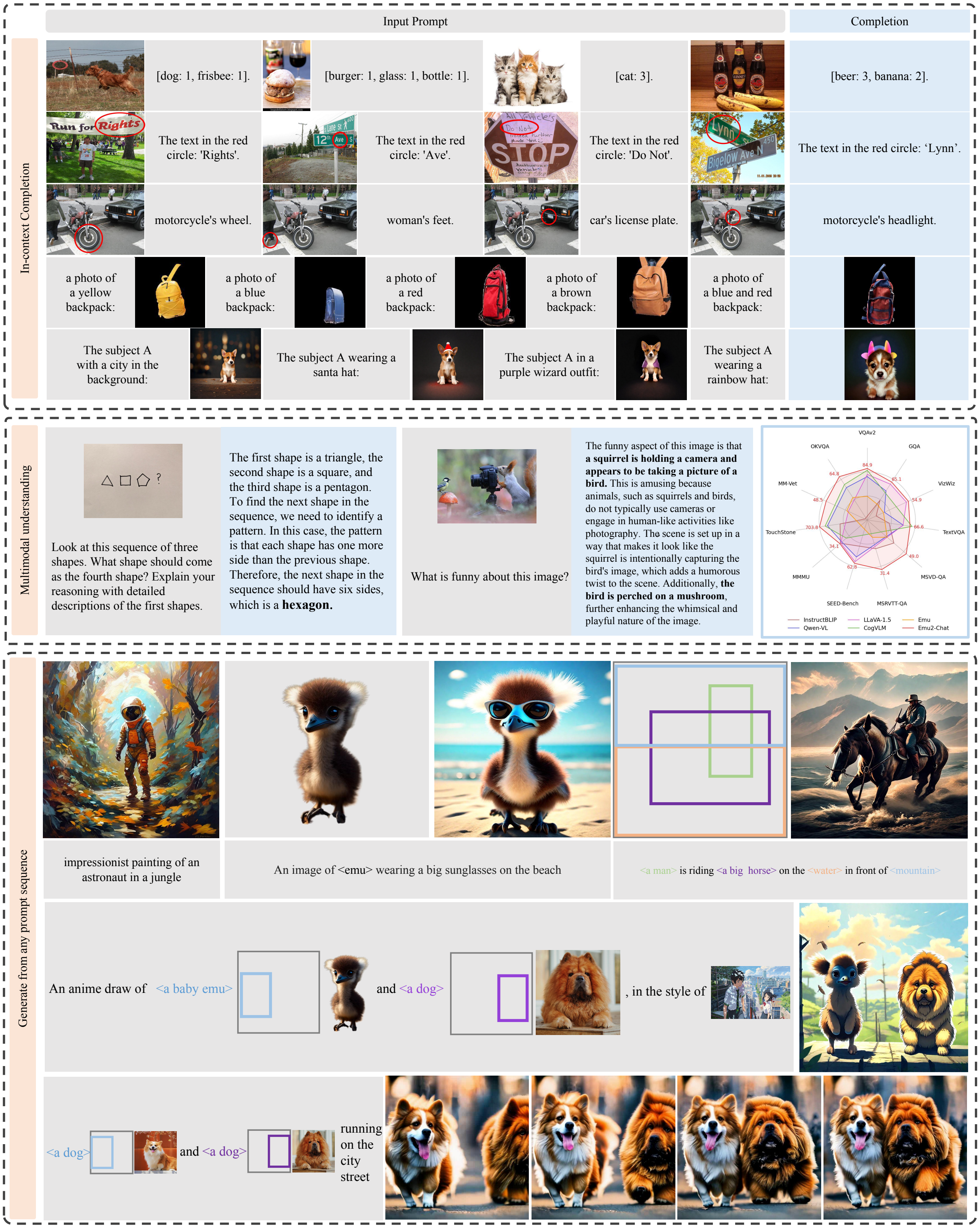

Figure 1. **Emu2** is a large generative multimodal model that serves as a foundation and a general-purpose interface for a broad range of multimodal tasks across understanding and generation, with remarkable in-context learning abilities.

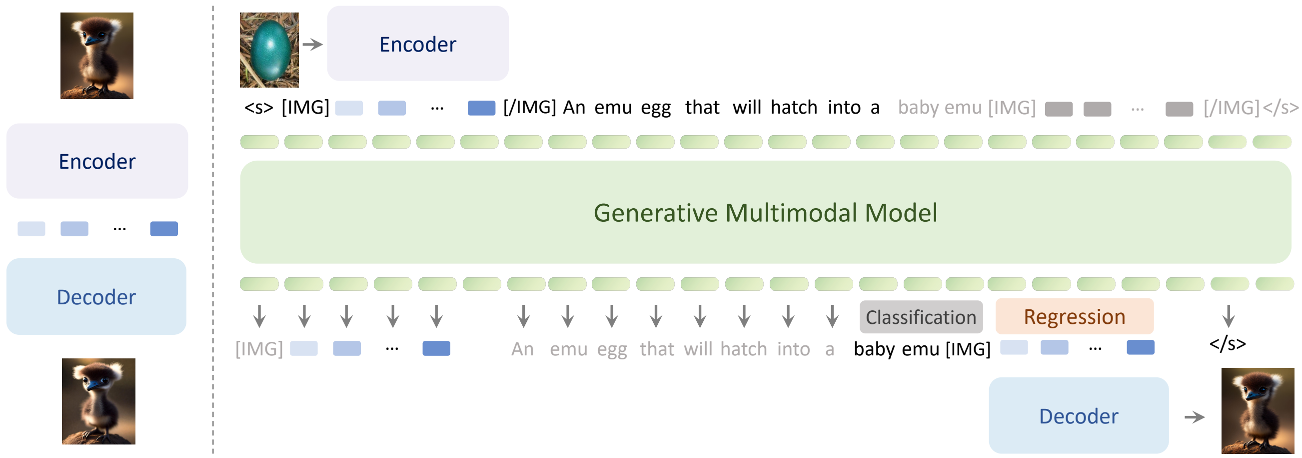

Figure 2. Overview of **Emu2** architecture. **Emu2** learns with a predict-the-next-element objective in multimodality. Each image in the multimodal sequence is tokenized into embeddings via a visual encoder, and then interleaved with text tokens for autoregressive modeling. The regressed visual embeddings will be decoded into an image or a video by a visual decoder.

Emu2 achieves state-of-the-art results on visual question-answering tasks, and surpasses previous models of more complex designs. In addition, **Emu2** can be fine-tuned to function as a controllable visual generation model of high quality. It is capable of accepting a mixture of text, locations and images as conditions, and generating images that are grounded as specified.

Given the broad spectrum of capabilities displayed by **Emu2**, we conduct a thorough analysis of its potential societal implications and discuss in detail potential concerns over misuse. By identifying further tasks where **Emu2**'s in-context learning can further improve, we highlight the necessity for continuous enhancement of the model and the importance of deploying **Emu2** responsibly.

2. Approach

2.1. Model Architecture

Emu2 is a generative multimodal model that learns with a predict-the-next-element objective in multimodal context. As illustrated in 2, the architecture of **Emu2** consists of three components: Visual Encoder, Multimodal Modeling, and Visual Decoder. Each image in the input multimodal sequence is tokenized into continuous embeddings via the Visual Encoder and then interleaved with text tokens for autoregressive Multimodal Modeling. The regressed visual embeddings are then decoded into an image or a video by the Visual Decoder. Specifically, we leverage pretrained EVA-02-CLIP-E-plus [70], LLaMA-33B [73] and SDXL [58] to initialize the Visual Encoder, Multimodal Modeling, and Visual Decoder, respectively. Compared to Emu [71], **Emu2** embraces a simpler framework which connects the Visual Encoder and Multimodal Modeling through mean pooling each image to 8×8 image patches, followed by a linear projection, instead of using an additional C-Former [71].

2.2. Pretraining

2.2.1 Data

The pretraining data for **Emu2** comprises several publicly accessible datasets, including image-text pairs from LAION-2B [65] and CapsFusion-120M [87], video-text pairs from WebVid-10M [8], interleaved image-text data from Multimodal-C4 (MMC4) [95], interleaved video-text data from YT-Storyboard-1B [71], grounded image-text pairs from GRIT-20M introduced by Kosmos-2 [57] and CapsFusion-grounded-100M curated by CapsFusion-120M. Additionally, language-only data from Pile [26] is included to retain textual reasoning capability.

2.2.2 Training

Similar to Emu [71], **Emu2** learns with the predict-the-next-element objective within a multimodal sequence. Each image is encoded into $N = 64$ dimension-fixed visual embeddings and then interleaved with text tokens to construct a multimodal sequence. The interleaved sequence is then fed into a Transformer decoder for autoregressive modeling.

Emu2 is first pretrained on image-text and video-text pair data with only captioning loss on the text tokens. The input images are resized to 224×224. We adopt the AdamW optimizer [50] with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1 \times 10^{-6}$. The maximum learning rate is 1×10^{-4} for the linear projection layer, 3×10^{-5} for Multimodel Modeling, and 5×10^{-5} for Visual Encoder. We pretrain **Emu2** on 162 million image-text samples and 7 million video-text samples for 35,200 iterations. The global batch size is 6,144 for the image-text pairs and 768 for video-text pairs. The training process is then restarted at a higher 448-pixel resolution for an additional 4,000 iterations.

Then, we freeze the Visual Encoder and only optimize the linear projection layer and Multimodel Modeling with both text classification loss and image regression loss. Additional datasets including image-text interleaved data, video-text interleaved data, grounded image-text pair data, and language-only data are used in the training. All images are resized to 448×448, and the maximum learning rate is 1×10^{-5}. We use a global batch size of 12,800 for image-text pair data, 6,400 for video-text pair data, 3,200 for image-text and video-text interleaved data, and 800 for language-only data. The training process spans 20,350 iterations and consumes about 160 million samples of image-text data and 3.8B tokens of language-only data.

2.2.3 Visual Decoding

We train the Visual Decoder to directly decode visual embeddings generated by the Visual Encoder into image. We use SDXL-base[58] as the initialization of our Visual Decoder, which is fully trained to solve the new task of autoencoding. Specifically, we use N visual embeddings as the condition input to the Visual Decoder and adjust the dimension of the projection layers in cross-attention modules to match the dimension of visual embeddings.

Unlike Emu [71] where each optimization step of its Visual Decoder requires an autoregressive inference of the language model, **Emu2**'s visual decoding can be considered as training a detokenizer, which can be trained off-the-shelf without the language model. Once trained, the Visual Decoder together with the Visual Encoder works as an image autoencoder that can tokenize an image into embeddings and detokenize back. During **Emu2** inference, it generates N image embeddings and decodes to an image on the fly.

For the decoding of video data, we train a diffusion-based decoder [67]. Similar to [46, 74], we adapt a 2D denoising U-Net to 3D style by inserting a 1D temporal convolution following each 2D spatial convolutional layer and extending the spatial attention to spatial-temporal attention. This video decoder is initialized via Stable Diffusion 2.1 [62] and fully trained to generate video clips conditioned on visual embeddings from **Emu2**.

Training Setup. We use the images in LAION-COCO [2] and LAION-Aesthetics [1] to train the Visual Decoder under the task of image autoencoding. The Visual Encoder and VAE in SDXL are frozen, and only the U-Net is updated during training. We adopt AdamW optimizer [50] with $\beta_1 = 0.9, \beta_2 = 0.999$ and the weight decay of 0.01. We use log learning rate warm-up and linear learning rate decay with a peak learning rate of 1×10^{-4} for 2,000 and 6,000 steps, respectively. We filter out images whose resolution is lower than 512×512. The input to the Visual Encoder is set to 448×448, while the output of the Visual Decoder is set to 1024×1024. We also employ the classifier-free guidance [30], which randomly discards image embeddings with the probability of 10%. The batch size is set to 2,048 in total.

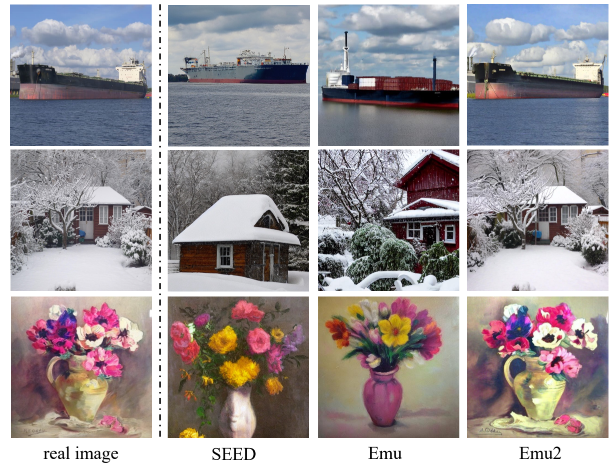

Figure 3. Comparison of autoencoding results among different methods [27, 71]. **Emu2**'s Visual Encoder and Visual Decoder in the architecture of CLIP-Diffusers form a strong autoencoder.

2.3. Instruction Tuning

Emu2 can be efficiently aligned to follow specific task instructions. We fine-tune the base model with conversational data to yield **Emu2-Chat**, which is capable of following multimodal questions and making responses in dialogue. Similarly, we derive a controllable visual generation model **Emu2-Gen**, which is capable of accepting a mix of text, locations, and images as conditions, and generating images that are grounded in the specified text or subject.

2.3.1 Instruction-Following Chat

Training Data. We adopt a uniform approach to train on both academic-task-oriented datasets and multimodal chat data to empower **Emu2-Chat** with the instruction-following ability while retaining rich visual knowledge. As academic-task-oriented datasets have brief annotations that limit the model's capacity to provide more comprehensive and helpful responses, we distinguish between these two data categories by employing different system messages and including instructions with output-format control information as used in [48]. A summary of data used is as follows: (a) Academic-task-oriented data: image captioning [18, 66], visual question answering [28, 32, 68], knowledgeable question answering [51, 53], multimodal classification [45], and referring expression comprehension [34, 52]. (b) Multimodal chat data: GPT-assisted visual instruction [49, 93], language instruction [4, 72], clock reading [83], and video chat [44].

Training Objective. In instruction tuning of **Emu2-Chat**, two special tokens, `[USER]` and `[ASSISTANT]`, are incorporated into the model to denote roles. These tokens help organize different data types in the following format: "`<Sys.Msg.> [USER]: <Instruction> [ASSISTANT]: <Answer>`". Here `<Sys.Msg.>` represents system message and varies between the two major task categories (academic-task-oriented and multimodal chat). The `<Instruction>` section comprises multimodal tokens, including images, videos, and text. Only tokens in the `<Answer>` section will be supervised by cross-entropy loss during training.

Training Setup. We use a global batch size of 768 and train for 8k steps. The learning rate linearly warms up to 1×10^{-5} in the first 100 steps, then decays to zero with a cosine schedule. The model is trained using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1\times10^{-6}$, and a gradient clipping of 5.0. The sequence length during training is limited to 2048, and any excess beyond that is truncated directly. We consistently employed an input image/video resolution of 448×448. For video data, we uniformly sample frames in time as input to the model. The number of sampled frames for each video is randomly chosen from 8, 12, and 16. To capture more intricate spatial details, following the visual encoder stage, we apply mean-pooling to each static image, dividing it into 16×16 tokens during instruction fine-tuning. This differs from the pre-training phase, where 8×8 tokens were utilized.

2.3.2 Controllable Visual Generation

Training Data. We leverage a mix of high-quality datasets to unleash the potential of controllable generation in context. We use a grounded image-text pair dataset CapsFusion-grounded-100M and GRIT [57] for grounded text-to-image generation. To mitigate the impact of image backgrounds on the effectiveness of multi-entity subject-driven generation, we employ SAM [36] to preprocess the grounding data, yielding a subset of approximately 5 million samples with segmentation results. Additionally, we leverage InstructPix2Pix constructed by [10] for image editing tasks. For the text-to-image task, we use a filtered subset of CapsFusion [87], LAION-Aesthetics [1], SA-1B [36], and LAION-High-Resolution [3].

We also collect data from premium sources (*e.g.*, Unsplash [20]) and outputs from advanced text-to-image systems (*e.g.*, Midjourney-V5 [54] and DALL-E-3 [9]) for quality fine-tuning. This diverse dataset includes around 500k high-quality image-text pairs. For all the data above, during the training, only samples with image resolutions higher than 448×448 were retained to ensure generation quality. More details can be found in the supplementary.

Model	Shot	VQAv2	OKVQA	VizWiz	TextVQA	Hateful Memes
Kosmos-1 (1.6B)	0	51.0	-	29.2	-	-
	4	51.8	-	35.3	-	-
	8	51.4	-	39.0	-	-
Flamingo (9B)	0*	51.8	44.7	28.8	31.8	57.0
	4	56.3	49.3	34.9	33.6	62.7
	8	58.0	50.0	39.4	33.6	63.9
	16	59.4	50.8	43.0	33.5	64.5
Flamingo (80B)	0*	56.3	50.6	31.6	35.0	46.4
	4	63.1	57.4	39.6	36.5	68.6
	8	65.6	57.5	44.8	37.3	**70.0**
	16	66.8	**57.8**	48.4	37.6	**70.0**
IDEFICS (80B)	0*	60.0	45.2	36.0	30.9	60.6
	4	63.6	52.4	40.4	34.4	57.8
	8	64.8	55.1	46.1	35.7	58.2
	16	65.4	56.8	48.3	36.3	57.8
Emu (14B)	0*	52.9	42.8	34.4	-	-
	4	58.4	-	41.3	-	-
	8	59.0	-	43.9	-	-
	16	-	-	-	-	-
Emu2 (37B)	0	33.5	26.7	40.4	26.4	52.2
	4	67.0	53.2	54.6	48.2	62.4
	8	67.8	54.1	54.7	49.3	65.8
	16	**68.8**	57.1	**57.0**	**50.3**	66.0

Table 1. Zero-shot and few-shot evaluations of **Emu2**. 0^* denotes text two-shot and image zero-shot results following Flamingo [5]. The best results are in **bold** and the second best are underlined.

Training Objective. We use the same unified generative pretraining objective to adapt to diverse generation tasks in context. Specifically, a training sample for generation is formulated as: "`<s>A photo of <p>a man</p><coor>image embedding of object localization image</coor>[IMG]image embedding of man[/IMG]sitting next to <p>a dog</p><coor>image embedding of object localization image</coor>[IMG]image embedding of dog[/IMG][IMG]image embedding of the whole image[/IMG]</s>`". We represent the coordinates of each object directly in image form by drawing the bounding box of each object at its specified location on a black image. Our **Emu2-Gen** conducts unified multimodal modeling of the text, object image, and corresponding object localization image. The regression loss only applies to the visual embeddings of the last image. We freeze the Visual Encoder during fine-tuning. We randomly drop tokens of entities and object localization image to enhance model adaptability and robustness. Additionally, we apply data augmentation to each object image, incorporating random background variations and random crop, aiming to reduce the reliance on image backgrounds.

Training Setup. We use a global batch size of 4,096 and

Model	Visual Question Answer							LMM Benchmarks			
	VQAv2 [28]	OKVQA [53]	GQA [32]	VizWiz [29]	TextVQA [68]	MSVD [82]	MSRVTT [82]	SEED [39]	MM-Vet [88]	TS [7]	MMMU [89]
Flamingo-9B [5]	51.8	44.7	-	28.8	-	30.2	13.7	-	-	-	-
Flamingo-80B [5]	56.3	50.6	-	31.6	-	35.6	17.4	-	-	-	-
Kosmos-1 [31]	51.0	-	-	29.2	-	-	-	-	-	-	-
Kosmos-2 [57]	51.1	-	-	-	-	-	-	50.0	-	-	26.6
BLIP-2-13B [42]	-	-	41.0	19.6	42.5	20.3	10.3	46.4	22.4	-	-
InstructBLIP-13B [22]	-	-	49.5	33.4	50.7	41.2	24.8	-	25.6	552.4	-
IDEFICS-9B [38]	50.9	38.4	-	35.5	25.9	-	-	-	-	-	-
IDEFICS-80B [38]	60.0	45.2	-	36.0	30.9	-	-	-	-	-	-
Shikra-13B [15]	77.4*	47.2	-	-	-	-	-	-	-	-	-
Qwen-VL-13B-Chat [6]	78.2*	56.6*	57.5*	38.9	61.5*	-	-	58.2	-	645.2	-
LLaVA-1.5-13B [48]	80.0*	-	63.3*	53.6	61.3	-	-	61.6	35.4	-	33.6
CogVLM [77]	83.4*	58.9*	-	-	**68.1***	-	-	-	-	662.6	30.1
Emu-I [71]	62.0	49.2	46.0	38.3	-	37.0	21.2	-	36.3	-	-
Emu2-Chat	**84.9***	**64.8***	**65.1***	**54.9**	66.6*	**49.0**	**31.4**	**62.8**	**48.5**	**703.8**	**34.1**

Table 2. Results on visual question answering and LMM benchmarks. * indicates that samples from this task's training set have been trained. SEED and TS respectively represent SEED-Bench [39] and TouchStone [7]. For MM-Vet, we present the average result of five scoring runs.

train for 3k steps. The learning rate linearly warms up to 5×10^{-5} in the first 100 steps, then decays to zero with a cosine schedule. We further fine-tune for 900 steps using the 500k high-quality pairs with a batch size of 2048.

3. Evaluation

3.1. Pretrained Base Model

We evaluate zero-shot and few-shot abilities of **Emu2** on OKVQA [53], VQAv2 [28], VizWiz [29], TextVQA [68], and HatefulMemes [35] tasks. Details of the datasets and prompts can be found in supplementary materials. The results are presented in Table 1. **Emu2** demonstrates remarkable in-context ability, showcasing improved performance with more in-context samples seen. Specifically, on VQAv2, VizWiz and TextVQA datasets, **Emu2** outperforms Flamingo-80B and IDEFICS-80B under all few-shot settings with a much smaller model scale (37B).

Figure 1 demonstrates **Emu2**'s few-shot capabilities in the wild. For example, the model learns to classify and count simultaneously in a specific format via a few examples (row 1). Additionally, **Emu2** is capable of following visual prompts in context, *e.g.*, the red circles laid on the images (row 2 and 3).

3.2. Instruction-Following Chat

Our **Emu2-Chat** is evaluated on academic-task-oriented benchmarks including image question-answering datasets (VQAv2 [28], OKVQA [53], GQA [32], VizWiz [29], TextVQA [68]) and video question-answering datasets (MSVD [82] and MSRVTT [82]). The evaluation also encompassed recent benchmarks for large multimodal models, including SEED-Bench [39], MM-Vet [88], TouchStone [7] and MMMU [89]. When evaluated on SEED-Bench, we followed the setup of LLaVa-1.5 [48] by presenting options to the model for completing multiple-choice tasks.

As shown in Table 2, **Emu2-Chat** consistently outperforms other models in image question-answering tasks, encompassing well-established benchmarks like VQAv2 and GQA. Notably, it shows a noticeable improvement in the OKVQA task, which requires the utilization of external knowledge, showcasing the advantage of our model for mastering real-world knowledge. For video question-answering, **Emu2-Chat** demonstrated advantages even though it did not use video question-answering data for training. It achieved an accuracy of 49.0 and 31.4 on the MSVD-QA and MSRVTT-QA tasks, respectively, surpassing InstructBLIP and the larger Flamingo-80B. More importantly, our model has also achieved better results on LMM benchmarks. LMM benchmarks such as MM-Vet provide a more comprehensive evaluation of model abilities, including solving complicated tasks. **Emu2-Chat** achieves a score of 48.5 in MM-Vet and 703.8 in TouchStone, confirming its superior capability in understanding and solving multimodal problems.

In addition, we demonstrated the visual grounding capability of our model using the refer expression comprehension benchmarks. In Table 3, **Emu2-Chat** achieved the best results among generalist models on RefCOCO [34], RefCOCO+ [52] and RefCOCOg [52]. Its most notable advantage was observed in RefCOCO+, which focused solely on purely appearance-based descriptions without allowing the use of position references. This highlights our model's powerful perceptual abilities in capturing intricate details.

Figure 4. Visualization of **Emu2-Gen**'s controllable generation capability. The model is capable of accepting a mix of text, locations and images as input, and generating images in context. The presented examples include text- and subject-grounded generation, stylization, multi-entity composition, subject-driven editing, and text-to-image generation.

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
OFA-L [75]	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58
Shikra-7B [15]	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19
Shikra-13B [15]	87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16
Qwen-VL-7B [6]	89.36	92.26	85.34	83.12	88.25	77.21	85.58	85.48
Emu2-Chat	**90.40**	**93.88**	**85.97**	**87.05**	**91.43**	**80.47**	**87.64**	**88.11**
CogVLM [77]	92.51	93.95	88.73	87.52	91.81	81.43	89.46	90.09

Table 3. Results on referring expression comprehension. We grayed out CogVLM because its generalist grounding-enhanced model was specialist trained on high-quality grounding data.

3.3. Controllable Visual Generation

Qualitative Results. Figure 3 presents a visualization of **Emu2**'s autoencoding results. With **Emu2**'s Visual Encoder and Visual Decoder, we can tokenize an image into visual embeddings and detokenize them back. Compared with SEED [27] and Emu [71], **Emu2** shows significantly superior results. We also evaluate our image autoencoding results on MS-COCO [47] and achieve a strong 0.907 CLIP-I [59] score. More results are in the supplementary.

As depicted in Figure 4, **Emu2-Gen** is capable of accepting a mixture of text, locations and images as input, and generating images in context. The model skillfully engages in various controllable visual generation tasks in a zero-shot setting, capitalizing on the in-context learning capabilities in multimodality. Examples in Figure 4 show generated images of three dogs conditioned on different subjects, locations and scenarios. The presented visual samples demonstrate the model's proficiency in tasks such as re-contextualization, stylization, modification, region-controllable generation, and multi-entity composition.

Zero-shot Text-to-image Generation. We evaluate the zero-shot text-to-image generation capability on 30k randomly sampled data from the MS-COCO [47] validation set. We employ CLIP-ViT-B [60], following the approach in DALL-E 3[9], to calculate the CLIP-T score to assess prompt-following ability. Additionally, we utilize CLIP-ViT-L, as in GILL[37], to compute the CLIP-I score for measuring image similarity. A higher score means the generated image is more similar to the prompt or the real image. Table 4 shows that **Emu2-Gen** achieves the state-of-the-art performance in terms of both CLIP-I and CLIP-T scores compared to various unimodal generation models and multimodal models. More text-to-image generation cases can be found in supplementary.

Zero-shot Subject-driven Generation. Following Kosmos-G [56], we also evaluate our model's subject-driven image editing ability on DreamBench [63]. We generate four images for each prompt, resulting in a total of 3,000 images for a comprehensive evaluation. We employ

Models	CLIP-I ↑	CLIP-T ↑
unimodal generation models		
MUSE [13]	-	**0.320**
Imagen [64]	-	0.270
DALL-E 2 † [61]	-	0.314
DALL-E 3 † [9]	-	**0.320**
SDv1.5 [62]	0.667	0.302
SDXL [58]	0.674	0.310
multimodal generation models		
GILL [37]	0.684	-
SEED [27]	0.682	-
Emu [71]	0.656	0.286
Emu2-Gen	**0.686**	0.297

Table 4. Quantitative comparison of zero-shot text-to-image generation on MS-COCO [47] validation set. 30k samples are randomly sampled. †CLIP-T score is calculated on 4,096 samples. We also evaluate our image autoencoding results on MS-COCO which achieves a strong 0.907 CLIP-I score.

Methods	DINO ↑	CLIP-I ↑	CLIP-T ↑
Real Images (Oracle)	0.774	0.885	-
Fine-Tuning			
Textual Inversion [24]	0.569	0.780	0.255
DreamBooth [63]	0.668	0.803	**0.305**
BLIP-Diffusion [42]	0.670	0.805	0.302
Test Time Tuning Free			
Re-Imagen* [16]	0.600	0.740	0.270
SuTI [17]	0.741	0.819	0.304
BLIP-Diffusion* [42]	0.594	0.779	0.300
Kosmos-G* (single image input)	0.694	0.847	0.287
Emu2-Gen * (single image input)	**0.766**	**0.850**	0.287

Table 5. Quantitative comparison of zero-shot single-entity subject-driven generation on DreamBench. * denotes zero-shot methods.

DINO [12] and CLIP-I [59] to evaluate subject fidelity, and CLIP-T [59] to evaluate text fidelity, aligning with the methodology established by DreamBooth. Notably, **Emu2-Gen** excels in subject fidelity, as evidenced by its superior performance on DINO and CLIP-I metrics compared to methods like BLIP-Diffusion and Kosmos-G. **Emu2-Gen** impressively reconstructs subjects with just one image input in zero-shot setting, demonstrating superior subject fidelity through powerful visual decoding. Further illustrative cases are provided in the supplementary, showcasing **Emu2-Gen**'s proficiency in multi-entity generation.

4. Related Work

Large Multimodal Models. Recent years have witnessed the rapid growth of large multimodal models [5, 19, 31, 71]. CLIP [59] pioneered the learning of LMMs with a contrastive learning objective on massive image-text pair data. Flamingo [5] and Kosmos [31, 57] exhibit promising zero-shot and few-shot multi-modal understanding performance by training on large-scale image-text interleaved data. With the remarkable progress in open-sourced LLMs, [42, 43, 49, 94] show promising results by connecting vision encoders and LLMs with a small intermediate model. A school of successive efforts [69, 76, 84, 90, 91] further improves visual instruction tuning with better overall training pipelines [6, 40], grounding annotations [14, 15, 85, 92], and extra tasks [6]. There are early studies on training more unified large multimodal models [23, 27, 71, 86] that are capable of performing visual understanding and generation simultaneously. In this paper, we further explore the distinct solution proposed in Emu [71]: learning large multimodal models with generative objectives on both texts and images.

In-Context Learning. Recent advancements in large language models [11, 21] underscore their capacity for in-context learning [11]. This phenomenon, particularly evident as LLMs scale up in size and data, has been exploited for complex challenges such as mathematical reasoning [81], signaling new emergent ability in model behavior [80]. Flamingo [5] integrates visual inputs to LLMs, enabling the in-context learning of visual-linguistic tasks such as image captioning and OCR through language-based interfacing. Painter [78] and SegGPT [79] conduct an early study of visual in-context learning. Inspired by the emerging abilities of large language models, in this work we study the problem of multimodal in-context learning by scaling up generative multimodal models and demonstrating strong results in broad understanding and generation tasks.

5. Conclusion

We present a 37 billion-parameter generative multimodal model **Emu2** that shows strong performance and versatility on many multimodal tasks in the in-context settings. **Emu2** serves as a base model and a general-purpose interface for a variety of multimodal tasks. We demonstrate state-of-the-art results on a broad range of benchmarks of multimodal understanding and generation. Specifically, our model largely surpasses prior work on the lately proposed LMM benchmarks that require more advanced capability compared to classic academic benchmarks. **Emu2** also shows remarkable capability of controllable visual generation in multimodal context, *e.g.*, subject-/text-grounded generation. Additionally, we review the limitations and broader social impact of **Emu2**. Despite discussed weaknesses, these results suggest that generative multimodal model at scale may be an important step towards the development of adaptable, general multimodal systems.

References

[1] Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/, . 4, 5, 2, 3

[2] Laion coco: 600m synthetic captions from laion2b-en. https://laion.ai/blog/laion-coco/, . 4, 2

[3] Laion-high-resolution. https://huggingface.co/datasets/laion/laion-high-resolution, . 5, 3

[4] Sharegpt. https://sharegpt.com/. 5, 2

[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. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 5, 6, 8, 3

[6] 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 arXiv:2308.12966*, 2023. 6, 7, 8

[7] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023. 6

[8] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 3, 1

[9] 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. 2023. 5, 7, 8, 3

[10] 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*, pages 18392–18402, 2023. 5, 3

[11] 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. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 8

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 8

[13] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 8

[14] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023. 8

[15] 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. 6, 7, 8

[16] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 8

[17] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 8

[18] 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. 4, 2

[19] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 1, 8

[20] Luke Chesser and Timothy Carbone. Unsplash. https://github.com/unsplash/datasets. 5, 3

[21] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1, 8

[22] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 6

[23] 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. 8

[24] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 8

[25] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022. 1

[26] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. 3, 2

[27] 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. 4, 7, 8

[28] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 4, 6, 2

[29] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 6

[30] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[31] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 6, 8

[32] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 4, 6, 2

[33] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 3

[34] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 5, 6, 2

[35] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020. 6

[36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 5, 3

[37] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023. 7, 8

[38] 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 *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 6

[39] 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

[40] 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. 8

[41] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 1(2):2, 2023. 1

[42] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 6, 8

[43] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 8

[44] 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*, 2023. 5, 2

[45] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M3it: A large-scale dataset towards multimodal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023. 4, 2

[46] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 4

[47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7, 8

[48] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 4, 6, 3

[49] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 5, 8, 2

[50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3, 4

[51] Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 4

[52] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 5, 6, 2

[53] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 4, 6, 2

[54] Midjourney. Midjourney. https://www.midjourney.com. 5, 3

[55] Leonardo Nicoletti and Dina Bass. Humans are biased: Generative ai is even worse. *Bloomberg Technology+ Equality. Accessed June*, 23:2023, 2023. 1

[56] 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. 7, 4

[57] 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. 3, 5, 6, 8, 1, 2

[58] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 1, 3, 4, 8, 2

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7, 8

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[61] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 8

[62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4, 8

[63] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 7, 8, 4

[64] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 8

[65] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 3, 1

[66] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 4, 2

[67] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 4

[68] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 4, 6, 2

[69] 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. 8

[70] 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*, 2023. 3

[71] 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, 4, 6, 7, 8, 1

[72] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023. 5, 2

[73] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 3

[74] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 4

[75] 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 *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 7

[76] 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. *arXiv preprint arXiv:2305.11175*, 2023. 8

[77] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan

Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 6, 7

[78] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 8

[79] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 8

[80] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 8

[81] 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. 8

[82] 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 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 6

[83] Charig Yang, Weidi Xie, and Andrew Zisserman. It's about time: Analog clock reading in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2508–2517, 2022. 5, 2

[84] 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. 8

[85] 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. 8

[86] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multimodal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023. 8

[87] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. *arXiv preprint arXiv:2310.20550*, 2023. 3, 5, 1

[88] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 6

[89] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 6

[90] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Transfer visual prompt generator across llms. *arXiv preprint arXiv:2305.01278*, 2023. 8

[91] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 8

[92] 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. 8

[93] 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. 5, 2

[94] 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. 8

[95] 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. 3, 1