

xGen-MM (BLIP-3): A Family of Open Large Multimodal Models

Le Xue^{1°} Manli Shu^{1°} Anas Awadalla^{1,2*} Jun Wang^{1*} An Yan^{1*}
 Senthil Purushwalkam^{1*} Honglu Zhou^{1*} Viraj Prabhu^{1*} Yutong Dai^{1*}
 Michael S Ryoo^{1*} Shrikant Kendre^{1*} Jieyu Zhang^{1,2*}
 Can Qin¹ Shu Zhang¹ Chia-Chih Chen¹ Ning Yu¹
 Juntao Tan¹ Tulika Manoj Awalgaonkar¹ Shelby Heinecke^{1†} Huan Wang^{1†}
 Jejin Choi^{2†} Ludwig Schmidt^{2†} Zeyuan Chen^{1†}
 Silvio Savarese^{1†} Juan Carlos Niebles^{1†} Caiming Xiong^{1†} Ran Xu^{1†}

¹Salesforce AI Research ²University of Washington

{lxue, ssavarese, jniebles, cxiong, ran.xu}@salesforce.com

[°]First authors; ^{*}Core authors; [†]Senior authors

[Project Page](#)

Abstract. This report introduces xGen-MM (also known as BLIP-3), a framework for developing Large Multimodal Models (LMMs). The framework comprises meticulously curated datasets, a training recipe, model architectures, and a resulting suite of LMMs. xGen-MM, short for xGen-MultiModal, expands the Salesforce xGen initiative on foundation AI models. Our models undergo rigorous evaluation across a range of tasks, including both single and multi-image benchmarks. Our pre-trained base model exhibits strong in-context learning capabilities and the instruction-tuned model demonstrates competitive performance among open-source LMMs with similar model sizes. In addition, we introduce a safety-tuned model with DPO, aiming to mitigate harmful behaviors such as hallucinations and improve safety. We open-source our models, curated large-scale datasets, and our fine-tuning codebase to facilitate further advancements in LMM research. Associated resources will be available on our project page above.

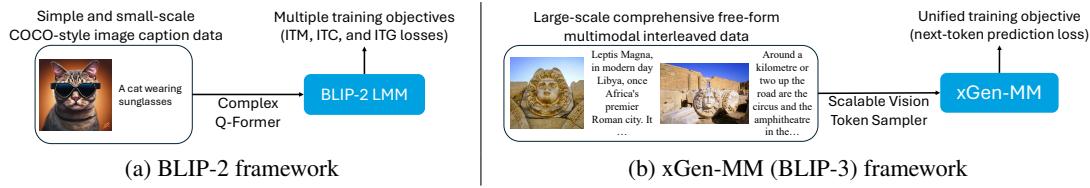


Figure 1: **We introduce xGen-MM (BLIP-3)**, a framework (b) for developing Large Multimodal Models (LMMs). Our framework improves upon BLIP-2 (a) [1] by (1) increasing the richness, scale, and diversity of training data, (2) replacing the Q-Former layers with a more scalable vision token sampler, and (3) simplifying the training process via the unification of the training objectives to a single loss at every training stage. The resulting suite of LMMs can perform various visual language tasks and achieve competitive performance across benchmarks.

1 Introduction

Large Multimodal Models (LMMs) have attracted significant attention with their potential applications and emergent capabilities. Recent advancements in both proprietary models [2–5] and open-source LMMs [6, 1, 7–11] highlight the rapid progress and growing interest in this field. However, despite these advancements,

there is still a gap between open-source models and proprietary ones in terms of access to open weights, training recipes, and curated datasets. Such limitations hinder the open-source communities from replicating, understanding, and improving LMMs.

Recent works have demonstrated that large-scale and high-quality data are essential for training robust LMMs [8–12]. BLIP-2 [1] was one of the pioneering efforts in exploring LMMs, which leveraged synthetic data to achieve impressive results at the time (Figure 1 (a)). However, the data used in BLIP-2 lacks the scale, quality, and diversity required to reach competitive performance compared to more modern LMMs nowadays. In addition, BLIP-2 employs an intricate Q-Former [1] architecture to bridge the vision and language modalities, coupled with a suite of complex training objectives (ITM, ITC, and ITG losses), both of which pose obstacles for larger-scale training. Moreover, BLIP-2 supports only single-image input, whereas interleaved multimodal data formats are the most natural form of multimodal data [13].

In response to these challenges, we introduce xGen-MM (BLIP-3) (Figure 1 (b)), a new framework designed to scale up LMM training by utilizing an ensemble of multimodal interleaved datasets, curated caption datasets, and other publicly available datasets [14–17]. xGen-MM, short for xGen-MultiModal, further expands our previous generative AI initiatives and corresponding foundation models for text xGen [18], code generation codeGen [19, 20], function calling APIGen [21], among others. In xGen-MM (BLIP-3), as illustrated in Figure 2, we streamline the model architecture by replacing the Q-Former [1] with a more scalable vision token sampler (specifically, a perceiver resampler [22]) and simplifying the training objectives to focus solely on the auto-regressive loss of text tokens in a multimodal context. Our primary focus is on dataset curation and scaling up the training data. Recently, our BLIP-3 team introduced two large-scale, high-quality datasets: MINT-1T [12], a trillion-token scale interleaved dataset; and BLIP3-KALE, a knowledge-augmented high-quality dense captions dataset. In this technical report, we introduce two additional specialized datasets: BLIP3-OCR-200M, a large-scale dataset with dense OCR annotations; and BLIP3-GROUNDING-50M, a large-scale visual grounding dataset.

In addition to these datasets, we are committed to open-sourcing the series of models developed in this work, including the base, instruction-tuned, and DPO models. Along with the model release, we also provide code for easy fine-tuning of our base model on custom datasets. By making these resources publicly available, we aim to make LMM research and development more accessible to the community, and we encourage researchers and practitioners to use our models and datasets to understand and further explore the potential and emergent abilities of LMMs.

2 Related Work

Recent advancements in Large Multimodal Models (LMMs) have explored two main architectural approaches: the cross-attention style [22, 23] and the decoder-only style. The cross-attention approach, exemplified by models like Flamingo [22, 23] and Llama 3.1 [5], integrates vision and language modalities through a complex attention mechanism to enable deep multimodal understanding. Another approach is the decoder-only architecture [1, 7, 8, 24–36], which we adopt in xGen-MM (BLIP-3), offers a more streamlined solution. This approach connects pre-trained language models to visual inputs using lightweight connectors, simplifying the integration process while maintaining robust multimodal capabilities. The effectiveness of this architecture is evident in models such as MM1 [9], VILA [10], LLaVA [8], phi3-vision [37], and Otter [38].

Training methodologies for LMMs typically follow one of the two strategies. The first one uses a light pre-training procedure and heavily relies on visual instruction tuning, as seen in the LLaVA series [8, 29]. Extensive research has been conducted on creating effective instruction-tuning data for a variety of tasks [32, 39–43]. The second strategy involves extensive pre-training on large-scale, diverse datasets, followed by visual instruction fine-tuning. This approach, employed by models like MM1 and Idefics2 [11], infuses broad knowledge into the model, which is then fine-tuned to align with human-like interactions and safety standards. While MM1 [9] provides extensive ablations and studies on the recipes aimed at improving LMMs, it releases limited resources for practitioners to adopt the model (*e.g.*, MM1 models and datasets

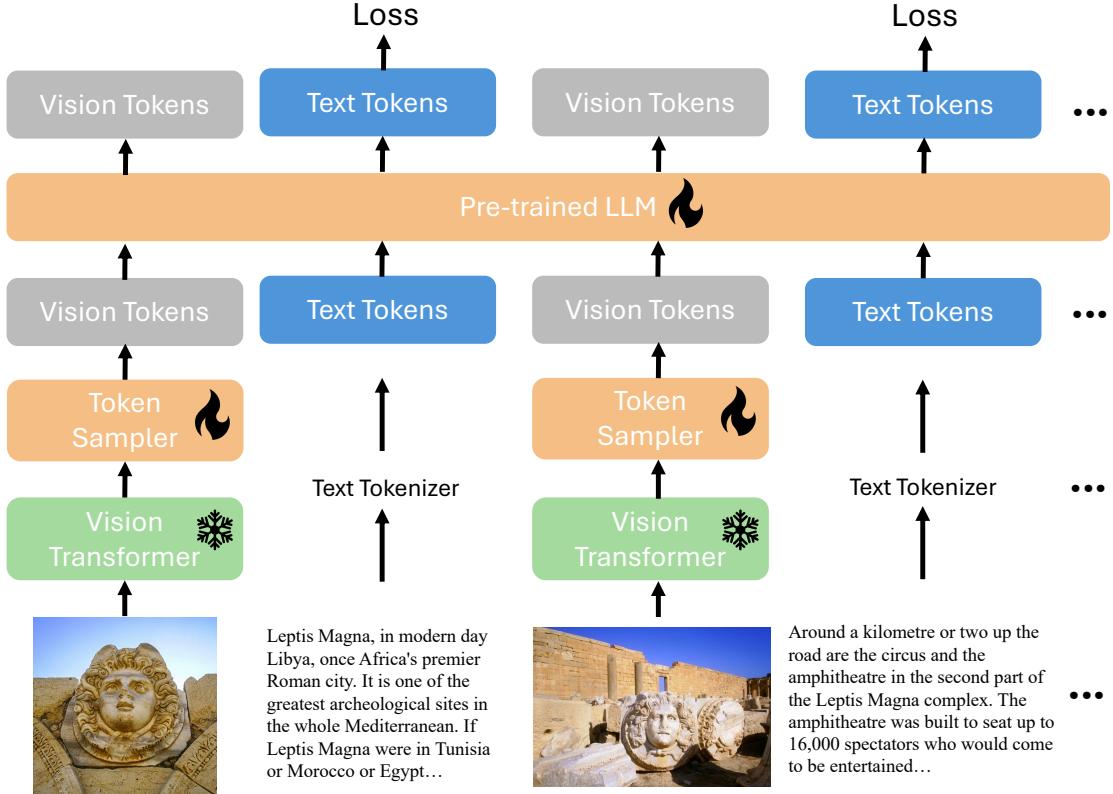


Figure 2: **Overview of the xGen-MM (BLIP-3) framework.** Free-form interleaved images and texts from the ensembled interleaved and caption datasets are input into the framework, with each modality undergoing a separate tokenization process to be fed into the pre-trained LLM in natural order. A standard auto-regressive loss is then applied to the text tokens. The Vision Transformer is kept frozen during training, while all other parameters, including the token sampler and the pre-trained LLM, are trained.

are close-sourced). Idefics2 [11] is a more recent open-source work that open-sources a suite of models along with detailed training strategies and data recipes, but Idefics2 mostly resuses existing datasets in their experiments without contributing new ones.

In this work, we present xGen-MM (BLIP-3). Unlike previous works, xGen-MM (BLIP-3) is an open-source family of a series of models, data recipes, fine-tuning code, and two large-scale foundational multimodal datasets, which we hope will enable and advance future research in this area.

3 Model Architecture

Architecture Overview. As illustrated in Figure 2, the xGen-MM (BLIP-3) framework adopts an architecture consisting of a ViT [44, 45], a vision token sampler (perceiver resampler [22]) to downsample the image embeddings, and a pre-trained Large Language Model (phi3-mini [37]). The input to the model can be free-form multimodal interleaved texts and vision tokens from the diverse multimodal data sources we ensemble.

Any-Resolution Vision Token Sampling. As proved effective in recent LMMs [46–48], we adopt a dynamic high-resolution (*i.e.*, “any-resolution”) image encoding strategy at the fine-tuning and post-training stages. We enable higher-resolution image understanding with image patch-wise encoding. The patch-wise

encoding preserves the resolution of the original image as much as possible by splitting a single image into multiple patches and encoding them separately. Following the previous convention, we concatenate the encoded image patches with a downsized original image that provides global information.

In the VL connector, we use a perceiver resampler to downsample the vision tokens. With any-resolution image encoding, we perform the downsampling for each image patch (including the downsized original image) independently. The downsampled vision tokens are then concatenated together and sent to the LLM. With the downsampling in our VL connector, we can reduce the sequence length of vision tokens by a factor of five or more depending on the number of query tokens in the perceiver resampler. We provide ablation studies on different token sampling strategies in Section 7.2.

4 Training

Pre-training. The pre-training objective is to predict the next text token across the dataset mixture we pre-train on. Overall, the resulting base model xGen-MM-Phi3-mini-base-r is pre-trained for about 100 billion multimodal tokens from the ensembled dataset, and our pre-training resolution is 384x384 pixels, which aligns with SigLIP [45].

Supervised Fine-tuning (SFT). We further fine-tune our pre-trained models on instruction-following examples to make them better understand and follow user queries. At the fine-tuning stage, we use a collection of publicly available instruction-following datasets [11, 29, 49]. We adopt the any-resolution vision token sampling strategy to allow a better understanding of images of higher resolutions such as text-rich document data. We introduce the technical details for the fine-tuning stage in the following sections.

Interleaved Multi-Image Supervised Fine-tuning. We conduct a second-stage fine-tuning on the instruction fine-tuned model on a mixture of multi-image and single-image instructions-following samples. The goal for this second-stage fine-tuning is to enhance the model’s ability to comprehend interleaved image-text input, which is helpful for multimodal in-context learning, multi-image question answering, and many more practical use cases. For the multi-image fine-tuning, we also adopt the any-resolution vision token sampling strategy same as in the previous SFT stage.

Post-training. Finally, we perform two stages of post-training to improve the model’s helpfulness while mitigating harmful qualities such as hallucination and toxicity. We first perform direct preference optimization (DPO [50]) to improve the model’s helpfulness and visual faithfulness. We then perform safety fine-tuning to improve the model’s harmlessness. We quantitatively demonstrate Pareto gains in model harmlessness and helpfulness after our post-training.

5 Data

5.1 Pre-training Data Recipe

As indicated in Figure 3, in xGen-MM (BLIP-3), we pre-train on an ensemble of diverse multimodal datasets with the indicated sampling ratios.

Interleaved Dataset Mixture. We combine MINT-1T (including its HTML, PDF, and ArXiv subsets) with OBELICS (HTML only) to create a more diverse and comprehensive dataset mixture that covers a broader range of domains.

1. **MINT-1T** [12] is a trillion token scale multimodal interleaved dataset, containing data sources from HTML, PDF, and ArXiv. As evidenced by MM1 [9] and Idefics2 [11], such multimodal interleaved

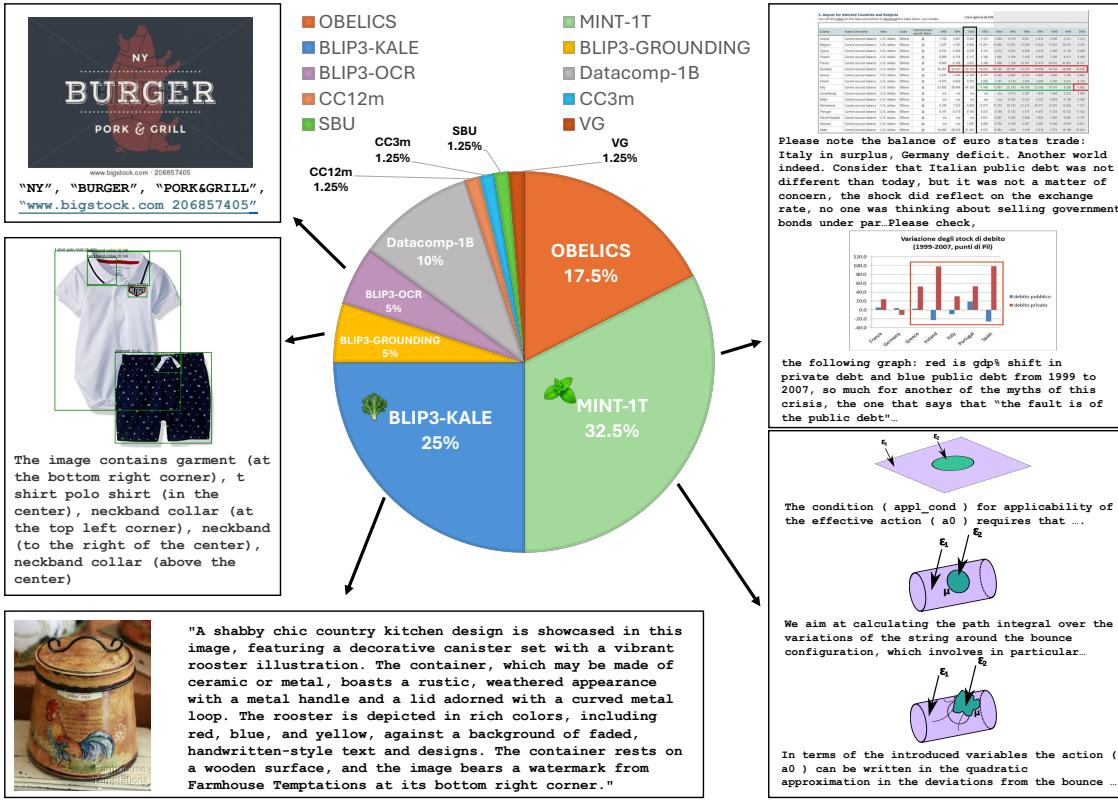


Figure 3: Overview of xGen-MM (BLIP-3) Pre-training Datasets.

datasets are essential for scaling up large multimodal model training and enabling fundamental capabilities like multimodal in-context learning. Notably, different from the OBELICS [11], MINT-1T has three subsets from different sources: the HTML subset, the PDF subset, and the ArXiv subset. In xGen-MM (BLIP-3), these three subsets are mixed in a 7:5:1 ratio.

2. **OBELICS** [11] is another large-scale multimodal interleaved dataset constructed from HTML documents solely. It differs slightly in domain coverage from MINT-1T due to the specific preprocessing steps adopted.

Caption Dataset Mixture. We integrate a diverse range of caption datasets, with the specifics outlined in the following details.

1. **BLIP3-KALE** is a large-scale curated high-quality caption dataset. Details will be discussed in another paper, and the dataset will be made public very soon.
2. **BLIP3-OCR-200M** is a curated large-scale OCR dataset to address the limitations of current large multimodal models in handling text-rich images like documents and charts, as traditional image-text datasets often lack adequate OCR annotations. To enhance text comprehension abilities, we use the OCR engine PaddleOCR [51] to annotate images with OCR-specific annotations. Overall, we curate a dataset of 200 million high-resolution images from Datacomp-1B[17]. For each image, we create captions with OCR data by identifying and extracting textual elements using the off-the-shelf OCR engine [51]. Text segments in a caption like "... text ..." are modified to include OCR information as "... text (ocr_info) ...", where ocr_info contains bounding box coordinates for the extracted

text, specifying its exact position within the image in the format "<bbox> x_1, y_1, x_2, y_2 </bbox>". We have multiple granularities of OCR information, including with and without bounding box data. In our work, we only utilize textual information without bounding box data, which has proven to be effective. Note that, in xGen-MM (BLIP-3), we preprocess the captions to remove filler texts such as "the text", which we find improves OCR-related benchmark performance. We hypothesize that this is because such filler text is relatively easy to predict, potentially diluting the loss associated with OCR-relevant tokens. Nonetheless, incorporating bounding box information could further enhance performance, and we encourage researchers in the community to explore this potential.

3. **BLIP3-GROUNDING-50M** is a curated large-scale grounding dataset to enhance the ability to ground semantic concepts in visual features, which is crucial for tasks like object detection, semantic segmentation, and understanding referring expressions [52] (e.g., "the object to the left of the dog"). We curate a dataset of 50 million images from Datacomp-1B [17]. For each image, we identify objects and their location information using one of the state-of-the-art open-world image tagging [53] and object detection models [54]. Objects mentioned in a caption like "... *object* ..." are modified to include grounding information as "... *object* (*grounding_info*) ...", where *grounding_info* contains bounding box information in one of three formats, each capturing a different granularity of localization: (1) <bbox> x_1, y_1, x_2, y_2 </bbox>, (2) "starts at (x_1, y_1) and extends up to (x_2, y_2) ", or (3) "top-left corner of the image".
4. **Other Public Datasets Mixture:** We also include other publicly available datasets such as uncurated Datacomp-1B [17] image-text pairs, CC12M [14], CC3M [14], VG [15], and SBU [16].

5.2 Supervised Fine-tuning Data Recipe

The datasets used in the fine-tuning stage are from public datasets of different domains. We sample data with various domain focuses including multi-modal conversation [29], image captioning [55, 56], visual question answering [57–60], chart/document understanding [61–64], science and math [65, 66]. In addition to the multi-modal image-text data, we also mix in pure text instruction following data [67, 68] during visual instruction tuning. Ultimately, we collect a mixture of 1 million publicly available instruction-tuning samples, on which we fine-tune our model for one epoch.

The multi-image instruction tuning stage starts with a model fine-tuned on single-image samples. We use a mixture of public multi-image/interleaved image-text instruction data [69, 70]. To prevent the model from deteriorating on single-image capabilities, we reuse a subset of single-image datasets used in the previous fine-tuning stage and mix them into the multi-image training data.

5.3 Post-training Data Recipe

Improving Truthfulness by Direct Preference Optimization. We employ VLFeedback [71], a synthetically annotated multimodal preference dataset that uses off-the-shelf VLMs to generate responses to a diverse mix of multimodal instructions that are then scored by GPT4-V [2] along three axes – helpfulness, visual faithfulness, and ethics. The dataset contains 80k such instructions from which we construct preference data by marking as preferred (and dispreferred) the response with the highest (and lowest) average score across models and filtering out examples with low-scoring preferred responses. We thus generate 62.6k preference examples.

We perform 1 epoch of DPO on the combined preference dataset while updating a subset (2.5%) of LLM backbone weights using low-rank adaptation (LoRA [72]). Also, following recent work [50], we generate an additional set of responses that capture the model's *intrinsic* hallucinations, by performing a second step of DPO per-iteration against the models' output to a *noised* version of the input image and original query, which we treat as an additional dispreferred response.

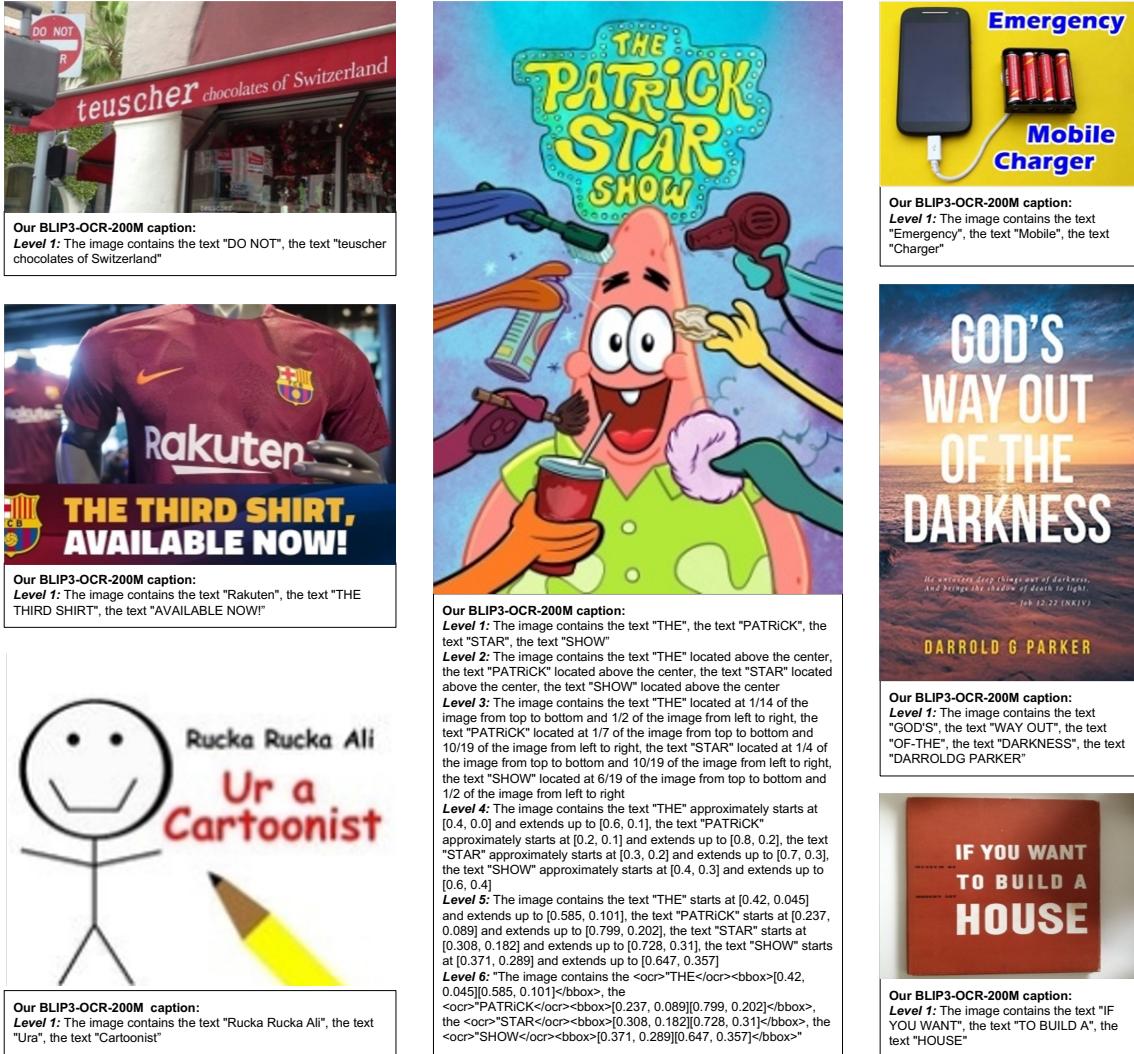


Figure 4: **Samples from BLIP3-OCR-200M.** Six levels of OCR information granularity are extracted, with and without bounding box data. Note that in xGen-MM (BLIP-3), OCR-related captions are preprocessed to remove filler phrases like 'the text,' resulting in improved OCR benchmark performance.

Improving Harmlessness by Safety Fine-tuning. Next, we perform 3 epochs of safety fine-tuning on the train split of the VLGuard [73] dataset, which contains 2k examples of unsafe images and instructions. VLGuard comprises two types of unsafe examples: (1) objectionable images paired with safe instructions and a desirable abstention response, and (2) safe images paired with two types of instruction-response pairs, one safe and another unsafe. The dataset consists of unsafe examples belonging to various subcategories including privacy-violating, risky/sensitive topics (such as politics, sex, and violence), deception, and discrimination. Following the original work, we randomly sample 5k additional examples from the instruction fine-tuning dataset to retain the model's helpfulness without exaggerating its safety behavior. As before, we update a subset (2.5%) of LLM backbone weights using low-rank adaptation.

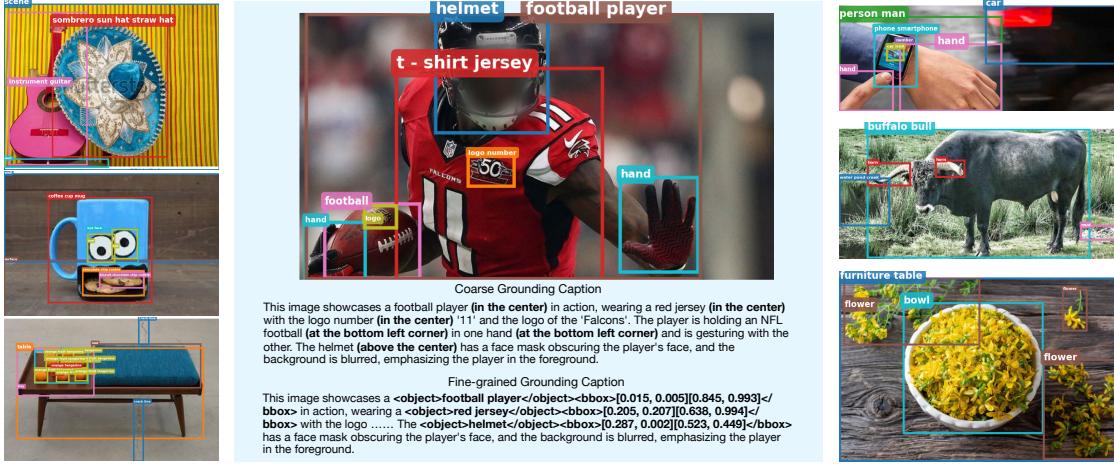


Figure 5: **Samples from BLIP3-GROUNDING-50M.** We introduce a large-scale dataset of images and corresponding captions containing localization information about objects. Furthermore, we release the associated object bounding box data to facilitate the creation of captions with custom templates.

6 Experiments

6.1 Pre-training

Few-shot Evaluation. After the pre-training stage, we evaluate our pre-trained model on classic captioning and VQA tasks, in comparison with previous models that support few-shot learning multi-modal evaluation. We present zero-shot and few-shot (4- and 8-shots) results, as shown in Table 1. Overall, our model achieves competitive multimodal in-context learning performance with comparable-sized LMMs¹. For the OCR tasks (TextCaps and TextVQA) and VQA-v2, it significantly outperforms MM1-3B and even larger models such as Idefics-9B [13] and MM1-7B [9]. On all benchmarks, increasing the number of shots can improve the performance, demonstrating the model’s ability to adapt to in-context distributions.

6.2 Supervised Fine-tuning

We evaluate our model on a comprehensive suite of multi-modal (image-text) benchmarks, assessing the model’s ability from multiple perspectives. Our evaluation covers general VQA benchmarks [74–78], visual perception [49], domain knowledge [79, 80], OCR ability [57, 81], and hallucination [82, 83]. For models fine-tuned on interleaved multi-image datasets, we also evaluate their performance on common multi-image benchmarks [69, 84–86].

Single-Image Evaluation. In Table 2, we compare with models in comparable sizes (< 5B), including both closed-source [9] and SoTA open-source models [10, 37]. We report individual benchmark scores along with two average scores: “Avg.(all)” is simply the average over all benchmarks, and “Avg.(perc.)” is the average score over benchmarks that focus on general VQA and visual perceptions. xGen-MM-instruct outperforms previous baselines on both general VQA and visual perception benchmarks. In addition, we

¹For few-shot evaluation, the zero-shot results are used mainly as a reference for their corresponding few-shot numbers, and they can be sensitive to prompts. As also mentioned in [9], they are mostly indicative of how well the pre-training distribution matches the associated evaluation task format. In pre-training evaluation, we mainly care for few-shot performance, which is robust to prompt templates.

Model	Shot	Visual Question Answering			Captioning		
		VQAv2	TextVQA	OKVQA	COCO	NoCaps	TextCaps
<i>< 5B Model Comparisons</i>							
Flamingo-3B [22]	0	49.2	30.1	41.2	73.0	–	–
	4	53.2	32.7	43.3	85.0	–	–
	8	55.4	32.4	44.6	90.6	–	–
MM1-3B [9]	0	46.2	29.4	26.1	73.5	55.6	63.3
	4	57.9	45.3	44.6	112.3	99.7	84.1
	8	63.6	44.6	48.4	114.6	104.7	88.8
xGen-MM-base (4B)	0	43.1	34.0	28.0	67.2	82.6	69.5
	4	66.3	54.2	48.9	107.6	100.8	89.9
	8	66.9	55.3	50.1	109.8	104.6	94.0
<i>Larger Models for Reference</i>							
Flamingo-9B [22]	8	58.0	33.6	50.0	99.0	–	–
Idefics-9B [13]	8	56.4	27.5	47.7	97.0	86.8	63.2
MM1-7B [9]	8	63.6	46.3	51.4	116.3	106.6	88.2
Idefics2-8B [11]	8	70.3	57.9	54.6	116.0	–	–

Table 1: **Few-shot Pretraining Evaluation.** Following [9], we randomly sample demonstrations from the training set as few-shot examples. We report CIDEr score for captioning and accuracy for VQA.

find that xGen-MM-instruct-interleave, although further fine-tuned on multi-image data, maintains good performance on single-image benchmarks and has the highest overall scores.

Model (Size)	SEED -IMG	SEED v2	MMB (dev)	MM Star	MME (norm)	CVB -2D	CVB -3D	RealW QA	MMMU (val)	Math Vista	Sci QA	POPE	Text VQA	Avg. (all)	Avg. (perc.)
<i>Closed-source models</i>															
GPT-4V	72.0	–	80.8	49.7	63.3	64.3	73.8	56.5	53.8	48.2	82.1	75.4	–	–	–
MM1-3B-Chat (3B)	68.8	–	67.8	–	62.9	–	–	–	33.9	–	–	87.4	–	–	–
<i>Open-source models</i>															
HPT-1.5-edge (4B)	72.3	–	74.6	45.8	–	–	–	–	42.6	45.1	85.4	91.0	–	–	–
VILA-1.5-3B (3B)	67.9	–	63.4	–	–	–	–	–	33.3	–	69.0	85.9	–	–	–
VILA-1.5-3B* (3B)	67.9	51.9	62.4	40.3	58.5	50.1	60.3	53.3	34.1	30.6	68.9	86.9	58.1	55.6	59.1
phi-3-vision (4B)	–	–	80.5	–	–	–	–	–	–	44.5	90.8	85.8	70.9	–	–
phi-3-vision* (4B)	71.0	52.7	74.2	47.9	55.3	60.7	68.2	59.1	46.1	45.1	90.2	83.5	73.3	63.6	63.6
xGen-MM-inst. (4B)	71.8	<u>53.9</u>	<u>76</u>	46.7	<u>63.8</u>	<u>66.2</u>	75.4	61.6	<u>42.8</u>	39.2	85.6	87.0	<u>72.0</u>	<u>64.8</u>	<u>66.9</u>
xGen-MM-inst.-interleave (4B)	<u>72.2</u>	55.5	76.8	48.1	64.4	69.3	<u>72.3</u>	<u>60.5</u>	41.1	<u>39.6</u>	<u>88.3</u>	87.0	71.0	65.1	67.3

Table 2: **Evaluation on single-image benchmarks.** phi-3-vision* and VILA-1.5-3B* are tested with our evaluation code² for a fair comparison. We also include the GPT-4V (gpt-4-1106-preview) performance (provided by the evaluation codebase) as a reference in the first row.

Multi-Image Evaluation. In Table 3, we compare xGen-MM-instruct with xGen-MM-instruct-interleave on multi-image benchmarks. Although the former is fine-tuned from xGen-MM-base which can comprehend interleaved image-text data, it performs poorly on multi-image benchmarks. We suspect it is because solely fine-tuning on single-image data hurts such ability. With multi-image SFT, we see significantly improved scores. In addition, we also evaluate Xgen-MM-interleave on single-image benchmarks (See Table 2) and find that it maintains good performance on all benchmarks with the highest overall scores.

²<https://github.com/open-compass/VLMEvalKit>

Model	BLINK	QBench-2	Mantis-eval
GPT-4V	51.1	73.4	62.7
VILA-1.5-3B* (3B)	39.8	51.7	41.9
xGen-MM-inst. (4B)	46.6	52.4	42.4
xGen-MM-inst.-interleave (4B)	49.7	75.1	56.7

Table 3: **Evaluation on multi-image benchmarks.** VILA-1.5-3B* results are obtained using the same evaluation code as our models. We include the GPT-4V performance as a reference in the first row.

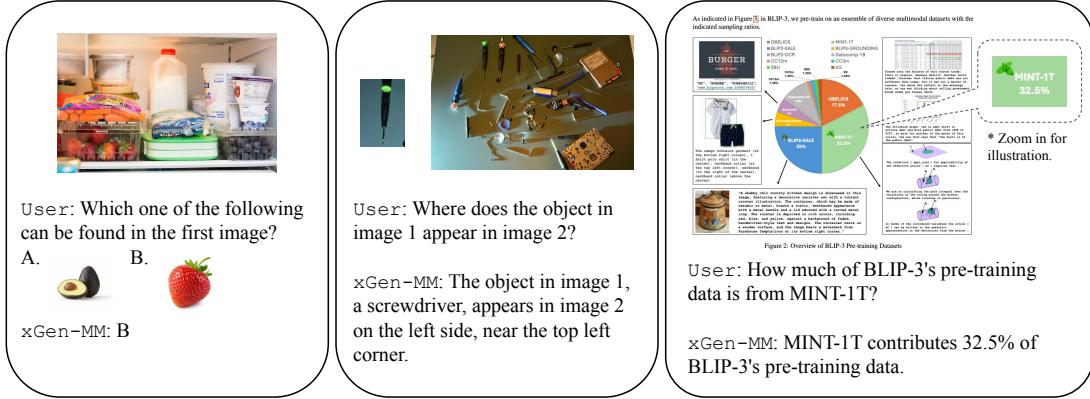


Figure 6: **Example model outputs of xGen-MM-instruct-interleave.** The model is capable of understanding interleaved image-text input and user queries about multiple images while maintaining the performance on single-image QAs.

6.3 Post-training

Table 4 summarizes the results of two post-training strategies for xGen-MM-instruct. We measure safety performance by ASR% (attack success rate) on the VLGuard test split and hallucination performance using HallusionBench [82] (accuracy on image-context reasoning) and POPE [83] (average F1 score on binary entity presence questions). To ensure post-training doesn't compromise helpfulness, we report performance on a few comprehension benchmarks as a control.

DPO enhances truthfulness by improving hallucination benchmarks (Row 2), while safety finetuning significantly reduces ASR (Row 3). Helpfulness is also improved slightly, as shown by control benchmarks. The final model, xGen-MM-dpo, includes both improvements.

Method	Safety		Hallucination		Helpfulness (Control)			
	VLGuard (↓)	HalBench (↑)	POPE (↑)	SEED-IMG (↑)	MMB-dev (↑)	MME (↑)	MMStar (↑)	
xGen-MM-inst. (4B)	56.6	56.3	87.0	71.8	76.0	63.8	46.7	
+ DPO	54.9	57.1	87.0	71.9	76.4	63.0	47.1	
+ Safety FT	5.2	56.6	86.8	72.1	76.4	64.4	47.1	

Table 4: **Post-training results.** We report results on safety and hallucination benchmarks after post-training, as well as on four helpfulness benchmarks as a control. Post-training improves harmlessness without compromising helpfulness.

7 Ablation Studies

7.1 Pre-training Ablation

Scaling Pre-training Data. We perform an ablation study to explore the relation between the amount of pre-training data and the pre-train evaluation metrics, by varying the data scale from 2B multimodal tokens to 100B multimodal tokens. The data recipe we used here is a mixture of image caption datasets and multimodal interleaved data. As shown in Figure 7, we find that scaling up the number of multimodal tokens from 2B to 60B leads to substantial gain for image-text (COCO-Caps) and OCR (Text-Caps, TextVQA) tasks, and further increasing the data size to 100B has moderate additional benefit in terms of few-shot evaluation metrics.

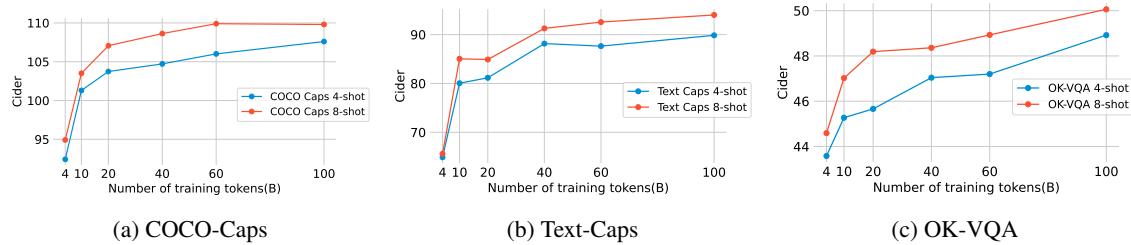


Figure 7: Few-shot performance given different sizes of pretraining data.

Pre-training Data Recipe. We discuss the impact of different data recipes for pre-training. Specifically, we perform ablation studies on top of a base data recipe: use Obelics [13] as the multimodal interleaved data source while keeping the caption datasets mixture the same. We also consider two other recipes (1) using MINT-1T [12] as interleaved data replacement, and (2) mixing additional pure text-only instruction-tuning data as a pre-train dataset. As shown in Table 5, we see a performance improvement using MINT-1T for image-text alignment (COCO-Caps) and OCR (Text-Caps, TextVQA), with a slight performance drop on OK-VQA, which is a knowledge-intensive task. We also find that adding text data can help attain the performance on OK-VQA that relies more on LLM capacity.

Data	Text-VQA	OK-VQA	COCO-Caps	Text-Caps
Obelics	41.1 / 41.9	48.4 / 49.5	107.2 / 109.4	78.2 / 79.9
MINT-1T	41.0 / 42.0	46.5 / 48.2	109.3 / 111.4	80.3 / 82.0
MINT-1T + text data	42.1 / 42.6	48.3 / 49.7	108.0 / 110.2	77.0 / 79.9

Table 5: Few-shot (4-shot / 8-shot) performance given different data recipes.

Visual Backbones. We also explore if different visual backbones have an impact on the performance of vision-language tasks. We compare two types of visual encoders, DFN and SigLIP. Empirically, we find SigLIP provides better visual representations that boost performance on OCR tasks.

Visual Backbone	Text-VQA	OK-VQA	COCO-Caps	Text-Caps
DFN	41.1 / 41.8	48.4 / 49.5	107.2 / 109.4	78.2 / 79.9
SigLIP	49.1 / 50.5	48.4 / 48.9	108.7 / 110.2	84.7 / 88.6

Table 6: Few-shot (4-shot / 8-shot) performance given different visual backbones.

Number of Visual Tokens. Another ablation is to study the impact of different numbers of visual tokens, i.e., input image tokens fed into the language model. We find that reducing the number of visual tokens from 128 to 64 can still attain similar performance, as shown in Table 7. This makes it possible for models to take in more visual images given a fixed context window.

Visual Token	Text-VQA	OK-VQA	COCO-Caps	Text-Caps
128	41.1 / 41.8	48.4 / 49.5	107.2 / 109.4	78.2 / 79.9
64	41.2 / 42.6	47.6 / 48.3	108.0 / 109.3	79.5 / 81.6

Table 7: Few-shot (4-shot / 8-shot) performance given the different number of visual tokens.

7.2 SFT Ablation

We conduct ablation studies at the instruction fine-tuning stage, focusing on several model design choices and data recipes. The SFT ablation studies are conducted on a simplified SFT data mixture, so the results in this section are not directly comparable to the main results in section 6.2.

Any-Resolution Vision Token Sampling. Our any-resolution strategy differs from previous work [46] in that every group of image embeddings (of the same image patch) is downsampled with a perceiver resampler, which ensures that the number of vision tokens input to the LLM remains relatively small. In this section, we ablate the effectiveness of our any-resolution strategy by comparing it with a “fixed-resolution” baseline and other downsampling designs.

The “fixed-resolution” baseline resizes all images to the default input size of the vision encoder while keeping the original aspect ratios. We also tried another downsampling strategy with the perceiver resampler: Instead of doing downsampling for each patch independently, we consider a “fixed sampling” (denoted as `anyres-fixed-sampling` in Figure 8a). In the fixed sampling, we concatenate the image embeddings from all image patches and then input them as a single sequence to the perceiver resampler to obtain the fixed number of vision tokens for the whole image.

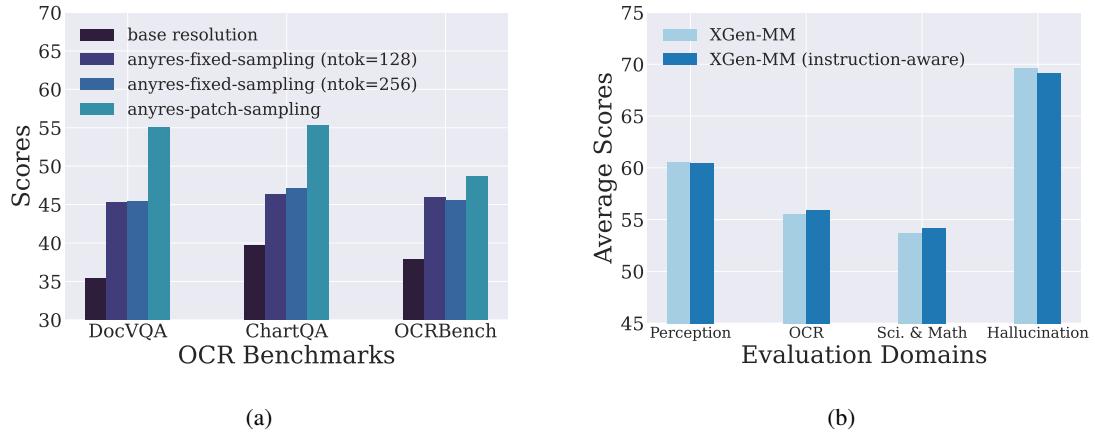


Figure 8: **SFT ablation studies.** (a). Comparison of different vision token sampling strategies on OCR benchmarks. (b). Comparison between our model and its “instruction-aware” alternative. For each evaluation domain in Figure (b), we report the average score on multiple relevant benchmarks.

Our evaluation of this design focuses on text-rich tasks (*e.g.*, document understanding) that would benefit from high-resolution encoding with visual details. From Figure 8a, we can see significant improvements

with our resolution image encoding strategy even with downsampled vision tokens. The fixed sampling strategy, although it shows improvements over the base resolution baseline, is not as good as the patch-wise sampling. We suspect that this may be due to two reasons: (a) With fixed sampling, a vision token sequence that can have as long as over 3,000 embedding tokens will be compressed to 128 tokens, which may be too few to retain the information. (b) The perceiver resampler may not work well with a concatenation of different image embeddings.

Instruction-Aware Vision Token Sampling. InstructBLIP [7] proposes an instruction-aware Q-Former [1] for vision token sampling and shows that it can improve the model performance on some benchmarks. With the perceiver resampler as the VL connector, we can adopt a similar modification to make this process instruction-aware. To make our perceiver resampler “instruction-aware”, we append the text instructions tokens to the query tokens of the perceiver resampler. Unlike Q-Former, there are only cross-attention layers inside the perceiver resampler, so the instruction (text tokens) would interact with both query tokens and image embeddings via cross-attention.

From the comparison in Figure 8b, we do not observe a significant difference between our model and its instruction-aware version on various benchmarks. It could be that our modification to the perceiver resampler can not be identical to the instruction-aware modification made to the Q-Former in Dai et al. [7], and thus the effectiveness differs. Because of the little difference we observe in this ablation study, we keep the original perceiver resampler architecture in our model for simplicity. We leave the further exploration of the “instruction-aware” VL connector to future works.

Text-only SFT data	MMMU (val)	MathVista (mini)	Science QA	MME (norm)	MMStar
Conversation	39.1	37.1	84.8	64.9	46.1
Conversation + Math + Coding	40.9	38.9	81.4	64.8	45.3

Table 8: **The impact of text-only SFT data.** We compare two choices of text-only SFT data used for the image-text SFT data mixture.

Quality of the Text-only Instruction Data. It is a common strategy to train or fine-tune a multi-modal LLM on both multi-modal and pure text data [8, 11]. It is mainly for maintaining the language ability of the fine-tuned model. In this experiment, we study how this pure text subset would affect the performance on multi-modal benchmarks. For the instruction tuning stage, we compare whether the *diversity* of the pure text data would affect the multi-modal performance. For example, how pure-text math data affects a model’s performance on multimodal math benchmarks. In our main experiments, the default collection of pure text instruction data covers diverse domains including conversation [67], math [68, 87], and code [88]. For this ablation study, we substitute these datasets with the same amount of samples that only cover general conversation.

In Table 8, we observe that adding math and coding data, although in pure-text format, can help improve a model on relevant benchmarks like MathVista [80], while has less effects on general VQA benchmarks.

8 Conclusion

We introduce xGen-MM (BLIP-3), a comprehensive framework for training a series of open-source large multimodal models on a curated mixture of large-scale datasets. xGen-MM (BLIP-3) demonstrates emergent abilities such as multimodal in-context learning and achieves impressive results on multimodal benchmarks. By open-sourcing xGen-MM (BLIP-3), our curated datasets, and our SFT fine-tuning codebase, we hope to empower the research community with accessible multimodal foundation models and datasets, allowing practitioners to explore further and advance the potential and emergent abilities of LMMs.

9 Broader Impact

The xGen-MM (BLIP-3) framework and its suite of Large Multimodal Models (LMMs) have the potential to significantly advance multimodal AI research by providing accessible, open-source resources for the broader community. By facilitating the development and fine-tuning of state-of-the-art LMMs, xGen-MM (BLIP-3) empowers researchers and practitioners across various domains to innovate and apply these models to diverse real-world challenges. Moreover, the integration of safety-tuning protocols within the xGen-MM (BLIP-3) framework helps mitigate ethical risks such as bias and misinformation, promoting the responsible deployment of AI technologies.

10 Acknowledgement

We would like to thank Srinath Meadusani, Lavanya Karanam, Dhaval Dilip Metrani, and Eric Hu for their work on the scientific computation infrastructure, as well as Jason Lee and John Emmons for their efforts in collecting the large-scale text-only SFT datasets used in one of our pre-training ablation studies.

References

- [1] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023.
- [2] OpenAI. Gpt-4v(ision) system card, 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.
- [3] Google. Gemini: A family of highly capable multimodal models, 2023.
- [4] OpenAI. Hello gpt-4o, 2024. URL <https://openai.com/index/hello-gpt-4o/>.
- [5] AI @ Meta Llama Team. The llama 3 herd of models, 2024. URL <https://ai.meta.com/research/publications/the-llama-3-herd-of-models/>.
- [6] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022.
- [7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.
- [8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
- [9] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training, March 2024. URL <http://arxiv.org/abs/2403.09611>. arXiv:2403.09611 [cs].
- [10] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2024. URL <https://arxiv.org/abs/2312.07533>.

- [11] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. URL <https://arxiv.org/abs/2405.02246>.
- [12] Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Eash Kumar Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, et al. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *arXiv preprint arXiv:2406.11271*, 2024.
- [13] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024.
- [14] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.
- [15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- [16] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.
- [17] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.
- [18] Erik Nijkamp, Hiroaki Hayashi Tian Xie, Bo Pang, Congying Xia, Chen Xing, Jesse Vig, Semih Yavuz, Philippe Laban, Ben Krause, Senthil Purushwalkam, Tong Niu, Wojciech Kryscinski, Lidiya Murakhovs'ka, Prafulla Kumar Choubey, Alex Fabbri, Ye Liu, Rui Meng, Lifu Tu, Meghana Bhat, Chien-Sheng Wu, Silvio Savarese, Yingbo Zhou, Shafiq Rayhan Joty, and Caiming Xiong. Long sequence modeling with xgen: A 7b llm trained on 8k input sequence length. ArXiv, 2023. URL <https://arxiv.org/abs/2309.03450>.
- [19] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.
- [20] Erik Nijkamp, Hiroaki Hayashi, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. Codegen2: Lessons for training llms on programming and natural languages. *arXiv preprint arXiv:2305.02309*, 2023.
- [21] Zuxin Liu, Thai Hoang, Jianguo Zhang, Ming Zhu, Tian Lan, Shirley Kokane, Juntao Tan, Weiran Yao, Zhiwei Liu, Yihao Feng, et al. Apigen: Automated pipeline for generating verifiable and diverse function-calling datasets. *arXiv preprint arXiv:2406.18518*, 2024.
- [22] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangoeei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

- [23] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *CoRR*, abs/2308.01390, 2023.
- [24] 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.
- [25] 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.
- [26] 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. *arXiv preprint arXiv:2303.03378*, 2023.
- [27] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [28] 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.
- [29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.
- [30] 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.
- [31] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024.
- [32] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.
- [33] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827, 2024.
- [34] Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024.
- [35] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

- [36] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.
- [37] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, May 2024. URL <http://arxiv.org/abs/2404.14219>. arXiv:2404.14219 [cs].
- [38] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *corr abs/2305.03726* (2023), 2023.
- [39] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.
- [40] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 3, 2023.
- [41] Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023.
- [42] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023.
- [43] 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.
- [44] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.
- [45] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.
- [46] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.

- [47] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, and Yinfei Yang. Ferret-v2: An Improved Baseline for Referring and Grounding with Large Language Models, April 2024. URL <http://arxiv.org/abs/2404.07973>.
- [48] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhui Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD, April 2024. URL <http://arxiv.org/abs/2404.06512>.
- [49] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, June 2024. URL <http://arxiv.org/abs/2406.16860>. arXiv:2406.16860 [cs].
- [50] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.
- [51] <https://github.com/PFCCCLab/PPOCRLabel>. Awesome multilingual ocr toolkits based on paddlepaddle.
- [52] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II* 14, pages 69–85. Springer, 2016.
- [53] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024.
- [54] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- [55] An Yan, Zhengyuan Yang, Junda Wu, Wanrong Zhu, Jianwei Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Julian McAuley, Jianfeng Gao, et al. List items one by one: A new data source and learning paradigm for multimodal llms. *arXiv preprint arXiv:2404.16375*, 2024.
- [56] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.
- [57] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.
- [58] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952, 2019. doi: 10.1109/ICDAR.2019.00156.
- [59] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, page 146–162, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20073-1. doi: 10.1007/978-3-031-20074-8_9. URL https://doi.org/10.1007/978-3-031-20074-8_9.

- [60] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, 2019. doi: 10.1109/CVPR.2019.00686.
- [61] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208, 2021. doi: 10.1109/WACV48630.2021.00225.
- [62] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL <https://aclanthology.org/2022.findings-acl.177>.
- [63] Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018.
- [64] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*. Springer International Publishing, 2016. ISBN 978-3-319-46493-0.
- [65] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, 2017. doi: 10.1109/CVPR.2017.215.
- [66] 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. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/11332b6b6cf4485b84afadb1352d3a9a-Paper-Conference.pdf.
- [67] Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichek Vong, and "Teknium". Openorca: An open dataset of gpt augmented flan reasoning traces. <https://huggingface.co/Open-Orca/OpenOrca>, 2023.
- [68] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math, 2024.
- [69] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. MANTIS: Interleaved Multi-Image Instruction Tuning, May 2024. URL <http://arxiv.org/abs/2405.01483> [cs]. arXiv:2405.01483 [cs].
- [70] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms, 2024. URL <https://arxiv.org/abs/2406.11833>.
- [71] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023.
- [72] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

- [73] Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In *The 41st International Conference on Machine Learning*, 2024.
- [74] 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.
- [75] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.
- [76] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023.
- [77] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.
- [78] ai. Grok-1.5 vision preview, 2024. URL <https://x.ai/blog/grok-1.5v>.
- [79] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *CoRR*, abs/2311.16502, 2023.
- [80] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.
- [81] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Chenglin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models, 2024. URL <https://arxiv.org/abs/2305.07895>.
- [82] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.
- [83] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- [84] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*. OpenReview.net, 2024.
- [85] Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhamo Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding. *CoRR*, abs/2406.09411, 2024.

- [86] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: multimodal large language models can see but not perceive. *CoRR*, abs/2404.12390, 2024.
- [87] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021.
- [88] Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhui Chen, and Xiang Yue. Opencodeinterpreter: Integrating code generation with execution and refinement, 2024. URL <https://arxiv.org/abs/2402.14658>.