Building and better understanding vision-language models: insights and future directions

Hugo Laurençon[*] Andrés Marafioti Victor Sanh Léo Tronchon[*]

Hugging Face

Abstract

The field of vision-language models (VLMs), which take images and texts as inputs and output texts, is rapidly evolving and has yet to reach consensus on several key aspects of the development pipeline, including data, architecture, and training methods. This paper can be seen as a tutorial for building a VLM. We begin by providing a comprehensive overview of the current state-of-the-art approaches, highlighting the strengths and weaknesses of each, addressing the major challenges in the field, and suggesting promising research directions for underexplored areas. We then walk through the practical steps to build Idefics3-8B, a powerful VLM that significantly outperforms its predecessor Idefics2-8B, while being trained efficiently, exclusively on open datasets, and using a straightforward pipeline. These steps include the creation of Docmatix, a dataset for improving document understanding capabilities, which is 240 times larger than previously available datasets. We release the model along with the datasets created for its training.

1 Introduction

Vision-language models (VLMs), that take images and texts as inputs and output texts, are highly effective in various applications such as document and figure understanding (Hu et al., 2024), solving visual mathematical problems (Gao et al., 2023), or converting webpage screenshots into code (Laurençon et al., 2024). The advancement of powerful open large language models (Touvron et al., 2023; Jiang et al., 2023; Team et al., 2024) and vision encoders (Zhai et al., 2023; Sun et al., 2023; Radford et al., 2021) allows researchers to build upon these unimodal pre-trained models to create advanced VLMs that solve these tasks with increasing accuracy (Dai et al., 2023; Liu et al., 2023; Bai et al., 2023; Lin et al., 2023; Li et al., 2023; Wang et al., 2023).

Despite advancements in the field, the literature highlights a variety of divergent design choices across key aspects of the development pipeline, indicating a lack of consensus. For instance, while many recent models (Koh et al., 2023; Li et al., 2023; Liu et al., 2023) have chosen to concatenate the sequence of image hidden states with the sequence of text embeddings before feeding it as input to the language model, the Llama 3-V model (Dubey et al., 2024) use interleaved Transformer-based cross-attentions to fuse the visual information into the LLM, similar to Flamingo (Alayrac et al., 2022). These different core choices in VLM development, often not ablated or justified in research papers, make it challenging to distinguish which decisions impact model performance and assess the compute and data efficiency trade-offs associated with each method.

In this paper, we begin by guiding the reader through the main research questions in the field, offering a detailed overview of the latest VLM approaches to address these challenges, along with the strengths and weaknesses of each. Specifically, we focus on (a) the various architectures used to connect pre-trained language models with vision encoders, (b) the different types of data employed in VLM training, their utility, and the typical stage at which they are introduced, (c) the training methods for

[*] Equal contribution

VLMs, which are often divided into multiple stages for efficiency and stability, and (d) the challenges encountered in model evaluation. We propose future research directions, particularly around data, to enhance model performance.

Building on this overview, we then walk through the practical steps for building Idefics3-8B[2], a powerful VLM trained efficiently, using only open datasets and a straightforward pipeline. Idefics3-8B significantly outperforms its predecessor, Idefics2-8B, particularly in document understanding tasks, with a 13.7-point improvement on DocVQA (Mathew et al., 2021). To especially boost the capabilities on this task, we created the Docmatix[3] dataset, which includes 2.4 million images and 9.5 million QA pairs derived from 1.3 million PDF documents—a 240-fold increase in scale compared to previous open datasets. We release our model alongside the datasets used for its training.

2 Analyzing architectural choices in VLMs

2.1 Connecting unimodal pre-trained models

Since the introduction of Frozen (Tsimpoukelli et al., 2021) and Flamingo (Alayrac et al., 2022), most VLMs have been built on top of unimodal pre-trained backbones, a language model and/or a vision encoder, rather than training entirely new models from scratch (Koh et al., 2023; Li et al., 2023; Liu et al., 2023). The availability of powerful open-source LLMs (Dubey et al., 2024; Jiang et al., 2023; Team et al., 2024) and image encoders (Zhai et al., 2023; Sun et al., 2023; Radford et al., 2021), which are increasingly expensive to train, enables researchers to leverage these models to create high-performing VLMs at a reduced cost (Dai et al., 2023; Koh et al., 2023; Liu et al., 2023; Vallaeys et al., 2024). These two pre-trained models are usually connected with either a cross-attention or a self-attention architecture.

2.1.1 Cross-attention architecture

The cross-attention architecture is introduced in Flamingo (Alayrac et al., 2022). The image hidden states encoded by the vision backbone are used to condition the frozen language model using freshly initialized cross-attention layers that are interleaved between the pretrained language model layers. The keys and values in these layers are obtained from the vision features, while the queries are derived from the language inputs. In practice, a cross-attention block is inserted after every four Transformer blocks in the LLM, adding newly initialized parameters equivalent to roughly 1/4th of the LLM's size. This significant increase in parameters enhances the model's expressivity, allowing it to achieve strong performance without unfreezing the LLM during training, thereby preserving the pre-trained LLM's performance on text-only tasks.

Flamingo

Idefics1 (Laurençon et al., 2023) and OpenFlamingo (Awadalla et al., 2023) are open replications of Flamingo. More recently, Llama 3-V (Dubey et al., 2024) also adopted this approach to adapt Llama 3 to multimodality.

2.1.2 Self-attention architecture

In the self-attention architecture (or fully-autoregressive architecture), introduced in FROMAGe (Koh et al., 2023) and BLIP2 (Li et al., 2023), the output of the vision encoder is treated as tokens and concatenated to the sequence of text tokens. The entire sequence is then passed as input to the language model. The sequence of visual tokens can be optionally pooled into a shorter sequence, making the model more efficient both during the training and at inference. We refer to the layers that map the vision-hidden space to the text-hidden space as modality projection layers. Figure 1 highlights the different components of the self-attention architecture.

Most recent VLMs have now adopted this design, including Llava (Liu et al., 2023), Qwen-VL (Bai et al., 2023), DeepSeek-VL (Lu et al., 2024), SPHINX (Lin et al., 2023), VILA (Lin et al., 2023), MiniGemini (Li et al., 2024), Monkey (Li et al., 2023), MM1 (McKinzie et al., 2024), Idefics2 (Laurençon et al., 2024), MiniCPM-V (Yao et al., 2024), InternLM (Dong et al., 2024) or InternVL (Chen et al., 2024).

[2] `https://huggingface.co/HuggingFaceM4/Idefics3-8B-Llama3`

[3] `https://huggingface.co/datasets/HuggingFaceM4/Docmatix`

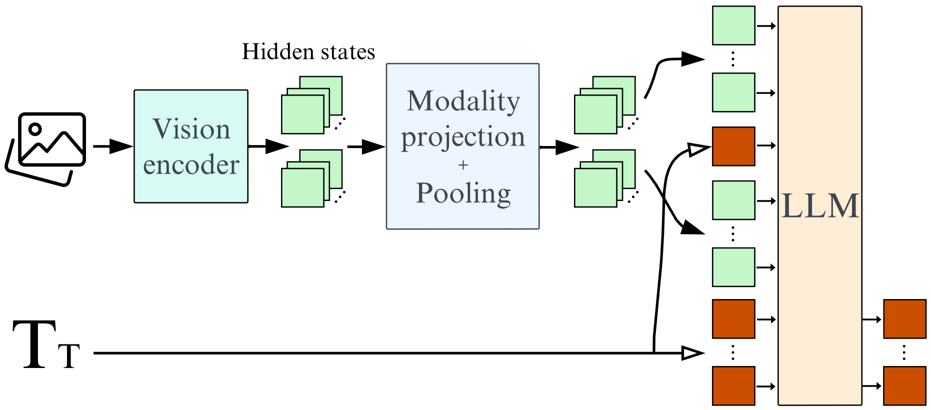


Figure 1: From Laurençon et al. (2024). The self-attention, or fully-autoregressive, architecture: Input images are processed by the Vision encoder. The resulting visual features are mapped (and optionally pooled) to the LLM input space to get the visual tokens. They are concatenated (and potentially interleaved) with the input sequence of text embeddings (green and red column). The concatenated sequence is fed to the language model (LLM), which predicts the text tokens output.

2.1.3 Which architecture performs best?

The performance comparison between these two main types of architectures was explored in Laurençon et al. (2024). The pre-trained unimodal models are Mistral-7B (Jiang et al., 2023) for the LLM and SigLIP-SO400M (Zhai et al., 2023) for the vision encoder. The model with the self-attention architecture has a total of 8.3B parameters, including 740M newly initialized, while the model with the cross-attention architecture has a total of 10B parameters, including 2.5B newly initialized. The authors demonstrate that the cross-attention architecture significantly outperforms when the backbones are kept frozen during training. However, when parts of the vision encoder and language model are trained with LoRA (Hu et al., 2022), adding an extra 200M trainable parameters distributed across both models, the cross-attention architecture performs worse despite having more parameters overall.

Nonetheless, this study did not evaluate the performance of the VLMs on text-only benchmarks. Intuitively, when parts of the language model are unfrozen during training, we need to incorporate data from the LLM training data mixture into the VLM training data to maintain performance on text-only benchmarks.

2.1.4 Impact of the pre-trained backbones on performance

Various studies find that the performance of each standalone unimodal pre-trained backbone correlates with the performance of the resulting VLM. For instance, in (Laurençon et al., 2024), the authors demonstrate that replacing the language model from LLaMA-1-7B (Touvron et al., 2023) (35.1% on MMLU (Hendrycks et al., 2021)) with Mistral-7B (Jiang et al., 2023) (60.1% on MMLU) leads to a substantial improvement across benchmarks. Analogously, replacing CLIP-ViT-H (Radford et al., 2021) (78.0% on ImageNet (Deng et al., 2009)) with SigLIP-SO400M (Zhai et al., 2023) (83.2% on ImageNet), also leads to a substantial performance improvement across all benchmarks, without changing the total number of parameters of the VLM.

Because vision encoders are often trained on different datasets and optimized for various tasks, some models, like SPHINX (Lin et al., 2023), combine representations from multiple encoders, such as DINOv2 (Oquab et al., 2023) and CLIP (Radford et al., 2021), to create a richer sequence of visual embeddings, though this comes at the expense of computational efficiency.

Recent research has heavily focused on improving open language models (Touvron et al., 2023; Dubey et al., 2024; Team et al., 2024; Jiang et al., 2023; Zheng et al., 2024; Conover et al., 2023; Mehta et al., 2024; Abdin et al., 2024; Hu et al., 2024; DeepSeek-AI et al., 2024; Bai et al., 2023). In contrast, few open-vision encoders have been released, with SigLIP-SO400M standing out due to its favorable performance-to-parameter ratio with only 400M parameters. This suggests a need for extensively trained open-source vision encoders at scale.

2.2 Examining the other architectural choices

2.2.1 Is a vision encoder really necessary?

Instead of employing a vision encoder, Fuyu (Bavishi et al., 2023) feeds image patches directly into the language model after applying a simple linear projection to adjust the dimensions. This architecture offers two main advantages: it is independent of another pre-trained model and preserves all the information from the original image. The latter point is crucial since the original image details might be necessary for accurately responding to the prompt. On the other hand, a pre-trained vision encoder transforms an image into a representation that is independent of the user's prompt. As a result, vision encoders aim to capture as much information as possible and can still miss details pertinent to the prompt. VisFocus (Abramovich et al., 2024) attempts to address this drawback by incorporating the user's prompt into the vision encoder. However, this approach is less natural in interleaved image-text conversations, where prompts may refer back to previous questions.

Despite these advantages, this architecture has not yet demonstrated superior performance. Fuyu scores significantly lower on benchmarks compared to the best models of similar size released around the same time. PaliGemma (Beyer et al., 2024) also experimented with this approach and reported a notable drop in performance compared to using a pre-trained vision encoder. The authors suggest that bypassing a vision encoder pre-trained on billions of images could lead to longer training times to achieve similar performance.

Furthermore, handling image representation within the language model might decrease its performance on text-only benchmarks. Even if this approach outperformed others on multimodal benchmarks, most VLMs are still not evaluated on text-only benchmarks, making it unclear whether omitting a vision encoder affects text benchmark performance.

Finally, this approach has not been tested yet with an efficient pooling strategy that does not significantly reduce information by operating directly on raw pixels. Looking ahead, for tasks like video understanding or extension to other modalities, it will be important to develop an architecture that can efficiently reduce the number of visual tokens passed to the language model to maintain a reasonable sequence length.

2.2.2 How should we connect the vision encoder to the language model?

Many models, such as FROMAGe (Koh et al., 2023) and LLaVA (Liu et al., 2023), use a simple linear layer between the vision encoder and the LLM, ensuring that all encoded visual information is retained since no pooling strategy is applied. However, this approach results in a long sequence of visual tokens, making training and inference less efficient. To address this, Qwen-VL (Bai et al., 2023) reduces the number of visual tokens by using a single-layer cross-attention module between a group of embeddings and the image hidden states. Similarly, Idefics2 (Laurençon et al., 2024) employs a cross-attention module within a perceiver resampler (Jaegle et al., 2021; Alayrac et al., 2022), demonstrating that the number of visual tokens can be compressed to as few as 64 (divided by 77) while maintaining performance for most tasks, except those that require extensive OCR capabilities. InternLM-XComposer2-4KHD (Dong et al., 2024) also shows that increasing the number of visual tokens per image is primarily necessary for benchmarks focused on OCR tasks, such as InfoVQA (Mathew et al., 2022) and DocVQA (Mathew et al., 2021).

Despite the efficiency of the perceiver resampler, its use has been challenged in several papers, which suggest leveraging the 2D structure of images more effectively. For instance, HoneyBee (Cha et al., 2024) introduces the C-Abstractor, which reintroduces 2D positional embeddings to the visual features, followed by ResNet blocks (Xie et al., 2017). In mPLUG-DocOwl-1.5 (Hu et al., 2024), the H-Reducer is introduced, using convolutions to divide the number of image hidden states by 4. InternVL (Chen et al., 2024) also achieves a fourfold compression using a simple pixel shuffle strategy. Recently, MiniCPM-V 2.6 (Yao et al., 2024), like Idefics2, chose the perceiver resampler with 64 learnable embeddings but enhanced it by adding 2D positional embeddings.

2.2.3 The image-splitting strategy: a trick to increase the number of visual tokens

Introduced in UReader (Ye et al., 2023) and SPHINX (Lin et al., 2023), the image splitting strategy involves dividing an original image into multiple sub-images, each of which is encoded separately by the vision encoder. The number of tiles can be fixed, such as consistently using four crops per image, or it can vary depending on the image's original resolution, with the image split every N pixels, for example.

When the number of tiles is based on the original resolution, the model is trained with varying numbers of visual tokens. This approach is particularly advantageous during inference: for simpler tasks, fewer visual tokens are needed, saving computational resources, while more computing can be allocated by increasing the image resolution for tasks that require intensive OCR. This flexibility is highly beneficial for models designed to excel both at reasoning on a single image with high computational resources and at processing videos with many frames while maintaining a reasonable sequence length by using a lower resolution for each frame.

Most vision encoders are designed for relatively low, fixed image resolutions and are not well-suited for processing large images. The image-splitting strategy addresses this by enabling the use of off-the-shelf pre-trained vision encoders at their original resolution, simply by feeding multiple smaller sub-images to the encoder instead of the original large image. Since the vision encoder's weights are shared across each sub-image, this approach also enhances training efficiency.

However, since the tiles of an image are not independent, encoding each one separately can be suboptimal and may result in a loss of global context. To address this, the current strategy involves adding the downscaled original image to the list of tiles, resizing it to match the resolution supported by the vision encoder. While this helps retain some of the overall context, it's not a perfect solution, as the reduced resolution of the original image makes it difficult to capture finer details and its resolution depends on the original image's resolution.

Can we do better than the image-splitting strategy? An alternative to the image-splitting strategy and a promising direction for future research is to develop a vision encoder that can natively process images of varying resolutions, including very large ones, without changing the original aspect ratios, potentially incorporating a mechanism for handling long-context efficiently. This model could be trained efficiently using the Patch'n'Pack (Dehghani et al., 2023) strategyate a different number of visual tokens per image based on the original resolution, enabling the entire image to be encoded directly without the need to crop it into multiple sub-images.

3 Training methods and datasets for VLMs

Training VLMs typically occurs in multiple stages, primarily due to (a) the limited availability of high-quality data at scale, (b) memory constraints for efficient training, and (c) stability issues. During these stages, progressively higher-quality data is introduced, the maximum image resolution is gradually increased, and more model parts are unfrozen. Figure 2 illustrates the key stages of training and the types of datasets used at each stage. As discussed in the previous section, the process begins with two unimodal pre-trained backbones: a language model and a vision encoder.

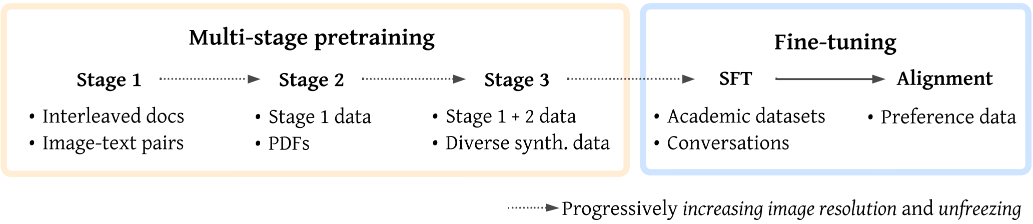
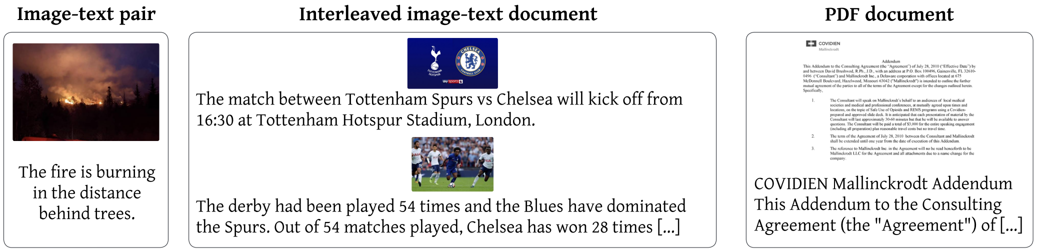


Figure 2: The different stages of training and the types of datasets used.



The fire is burning in the distance behind trees.

The match between Tottenham Spurs vs Chelsea will kick off from 16:30 at Tottenham Hotspur Stadium, London.

The derby had been played 54 times and the Blues have dominated the Spurs. Out of 54 matches played, Chelsea has won 28 times [...]

COVIDIEN Mallinckrodt Addendum This Addendum to the Consulting Agreement (the "Agreement") of [...]

Figure 3: Types of examples used during the pre-training of VLMs. (a) An image-text pair from LAION COCO, (b) an interleaved image-text document from OBELICS, (c) a PDF document from OCR-IDL.

3.1 Multi-stage pre-training

The primary goal of pre-training is to align the backbone models and train the newly initialized parameters in the model. This is achieved using large-scale datasets to expose the VLM to a wide variety of examples to build extensive knowledge and improve robustness against out-of-domain data. To preserve the initial performance of the LLM, some models, like VILA (Lin et al., 2023) and LLaVA-NeXT (Liu et al., 2024), begin training by freezing the backbone models and focusing solely on the newly initialized parameters (the connector) until a satisfactory performance level is achieved. Afterward, the vision encoder and/or the language model can be gradually unfrozen. If instabilities arise, or if there's a need to enhance the model's expressivity while adding more regularization than full unfreezing, a LoRA (Hu et al., 2022) approach can be effective even during the pre-training phase (Laurençon et al., 2024).

To efficiently train on a large number of images, the image resolution is typically kept low at the start of training and gradually increased over time. Once the resolution is sufficiently high, datasets containing large images, such as PDFs, can be incorporated into the training data.

In the following paragraphs, we will discuss the various types of data typically used during this process. Examples of the most common ones are illustrated in Figure 3.

Image-text pairs Image-text pair datasets are generally created by crawling the web, downloading images, and extracting the corresponding alt-text from the original HTML files. Due to the ease of collecting these raw image-text pairs and their effectiveness in establishing strong alignment between images and text, many large-scale datasets have been created, such as LAION (Schuhmann et al., 2022) with 5B images, COYO (Byeon et al., 2022) with 700M images, and DataComp (Gadre et al., 2024) with 12.8B images.

However, the alt-texts in these datasets are often noisy, ungrammatical, or too brief, making training challenging. Recent approaches have achieved better results by using synthetic re-captioning, where the same images from the original datasets are re-captioned using another model (McKinzie et al., 2024; Betker et al., 2023; Laurençon et al., 2024). For example, LAION COCO (Schuhmann et al., 2022) re-captioned 600 million images from LAION-5B using an ensemble of BLIP (Li et al., 2022) and two CLIP models (Radford et al., 2021). Similarly, VeCap (Lai et al., 2023) combines the original alt-text with a synthetically generated caption from LLaVA (Liu et al., 2023) to create a dataset of 300 million samples.

While these efforts have mainly focused on generating high-quality captions for given images, less attention has been paid to the initial selection of "good" images, which remains a promising area of research. This is important given the high proportion of web images that may not be useful for VLM training (e.g., logos, icons, portraits of non-public figures). Synth2 (Sharifzadeh et al., 2024) addresses this by reversing the usual process, starting with LLM-generated captions and then using a Text-to-Image model to generate corresponding images. Furthermore, studies such as SNIP-Dedup Webster et al. (2023) and SemDeDup (Abbas et al., 2023) have shown that by applying image deduplication, it is possible to train on just half of the LAION dataset with only a minimal reduction in performance compared to using the full dataset.

Interleaved image-text documents Training on interleaved image-text documents, also called web documents, was first introduced in Flamingo (Alayrac et al., 2022) using the proprietary M3W dataset. OBELICS (Laurençon et al., 2023) is an open-source dataset of interleaved image-text documents, containing 141 million documents and 353 million images. This dataset was constructed from HTML files obtained from Common Crawl dumps, which were carefully filtered. The resulting documents maintain the original linearity of images and texts as they appeared on the websites, while removing spam and ads.

The authors highlight several advantages of using web documents in the training data mix: (a) it enhances in-context learning abilities, (b) it improves the model's ability to understand an arbitrary number of images interleaved with text, and (c) it exposes the model to a much wider distribution of texts than what is available in standard image-text pair datasets. This aligns with findings from MM1 (McKinzie et al., 2024), which showed that interleaved data is instrumental for few-shot and text-only performance. OBELICS has been used in the training of various VLMs, including MM1 (McKinzie et al., 2024), Idefics2 (Laurençon et al., 2024), and BLIP-3 (Xue et al., 2024). Recently, the scale of these datasets has been significantly expanded, with MINT-1T (Awadalla et al., 2024) growing to 1T documents and 3.4B images, and OmniCorpus (Li et al., 2024) reaching 2.2B documents and 8.6B images.

Model-based filtering on educational content, similar to the approach in Phi-3 (Abdin et al., 2024) and FineWeb-Edu (Penedo et al., 2024), remains unexplored for these multimodal datasets and could likely offer significant improvements.

PDF documents Two primary datasets for PDF documents paired with their text transcriptions are OCR-IDL (Biten et al., 2022) and PDFA[4]. OCR-IDL includes 26M pages of industry documents, while the English-only filtered version of PDFA contains 18M pages sourced from Common Crawl, offering greater diversity than OCR-IDL. Both datasets were created using OCR extraction tools to obtain corresponding texts and their locations within the documents, which can be linearized into a full document transcription. Idefics2 (Laurençon et al., 2024) used these datasets directly during pre-training, an approach also adopted at scale in Llama 3-V (Dubey et al., 2024) to enhance performance on document understanding tasks.

Synthetic data sh foundational skills such as (a) image captioning, (b) handling an arbitrary number of images interleaved with diverse texts, and (c) text transcription, all of which are essential for tackling more complex tasks. These datasets are abundant, as they are primarily built by crawling the web, ensuring a broad distribution of texts and images, enhancing robustness against rare examples. However, these datasets fall short in addressing many of the tasks that users typically require, such as document understanding or visual math reasoning, which are significantly more challenging. Relying on generalization or the limited examples in current fine-tuning datasets to master these tasks is not ideal.

In the training of LLMs, synthetic data has proven to be highly effective (Zheng et al., 2024; Gunasekar et al., 2023; Liu et al., 2024; Dubey et al., 2024). Given the recent advancements in VLMs, which now solve many real-world examples with high accuracy, creating and training on large-scale synthetic datasets is a logical step. These datasets can be tailored to include examples that closely resemble the tasks users will likely request, making them more relevant than the data used in earlier training stages.

The main categories of synthetic data that could be used are outlined below.

Image captioning The leading dataset for images paired with detailed captions is PixelProse (Singla et al., 2024). This dataset, built using images from CC12M (Changpinyo et al., 2021), CommonPool (Gadre et al., 2024), and RedCaps (Desai et al., 2021), contains captions generated by Gemini 1.0 Pro (Team et al., 2023). Despite being smaller in scale with 17M images, PixelProse offers richer descriptions and uses a stronger model for caption generation, making it an improvement over LAION COCO. Future improvements could include a more diverse, filtered, and deduplicated set of images, better models to reduce potential hallucinations in the generations, and various prompts for stylistic diversity. A similar dataset, ShareGPT-4o[5], re-captions images using GPT-4o to obtain 57K examples.

[4] `https://huggingface.co/datasets/pixparse/pdfa-eng-wds`

[5] `https://huggingface.co/datasets/OpenGVLab/ShareGPT-4o`

Real-world visual question answering Datasets in this category contain QA pairs about real-world images, covering topics like identifying people or objects, understanding subtle scenes, counting, color identification, or spatial positioning. The leading dataset in this area is LNQA[6], with 300K images sourced from Localized Narratives (Pont-Tuset et al., 2020) and 1.5M QA pairs.

Text reading in natural images In LLAvAR (Zhang et al., 2023), the authors use OCR tools to extract text from real-world images in the LAION-5B dataset (Schuhmann et al., 2022), resulting in 420K samples. Similar approaches are seen in MiniCPM-V (Yao et al., 2024) and Llama 3-V (Dubey et al., 2024). The key advantage of these datasets is their scalability and the unique distribution of text in natural images compared to PDF documents, which enhances the model's ability to tackle tasks like TextVQA (Singh et al., 2019).

Text transcription The leading dataset for text transcription is PDFA, mentioned above. However, linearizing texts coherently from bounding boxes can be challenging, and math equations are often inaccurately transcribed or omitted, an area where models like Nougat (Blecher et al., 2023) excel. Additionally, figures and tables are often poorly transcribed by OCR tools. A better strategy for text transcription would involve combining a traditional OCR tool, a document-specialized model like Nougat, and a robust VLM to judge, refine, and merge the outputs of these models.

Document understanding Understanding documents from images is complex, making the generation of quality synthetic QA pairs challenging even for advanced VLMs. However, accurate text transcriptions from document images can be obtained with OCR tools, and text-only LLMs are performant at generating QA pairs from these transcriptions. This approach was used to create the dataset Docmatix, introduced in detail later in this paper, which includes 1.3M documents up to 4 pages long and 9.5M QA pairs. Enhancements could involve generating more diverse questions, such as summarizing a paragraph, and employing a strong VLM to filter out erroneous generated QA pairs.

Chart understanding ChartGemma (Masry et al., 2024) uses Gemini 1.5 Flash (Reid et al., 2024) to generate 160K QA pairs for chart analysis, covering a range of questions like summarizing insights, converting charts to Markdown tables, and assessing the validity of stated facts based on the chart.

Table understanding A dataset for table understanding can be created by either using a strong VLM with table images taken from the web, or by synthetically generating tables with an LLM, rendering them to images, and generating QA pairs with the LLM. However, to our knowledge, there is currently no large-scale open-source synthetic dataset available for this task.

Reasoning with chain-of-thought In Meteor (Lee et al., 2024), the authors developed a proprietary dataset to enable a model to answer complex questions using a chain-of-thought strategy (Wei et al., 2022). They began by collecting challenging QA pairs from academic datasets, where the answers were provided without explanations. Then, they employed Claude 3 Haiku (Anthropic, 2024) to generate detailed and comprehensive rationales for these answers. These rationales were finally filtered by GPT-4V (Achiam et al., 2023) to ensure quality, resulting in a final set of 1.1M question-rationale-answer triples.

Visual mathematical reasoning Even the most advanced VLMs currently struggle with complex mathematical reasoning and geometry tasks. Generating synthetic data directly from a teacher model is problematic because the teacher often fails to provide correct answers. Instead, datasets like Geo170K (Gao et al., 2023) and MAVIS-Instruct (Zhang et al., 2024) are created by augmenting small and accurate academic mathematical datasets using an LLM. In AlphaGeometry (Trinh et al., 2024), the authors train a model exclusively on synthetically generated geometric problems, enabling it to solve olympiad-level challenges effectively.

Converting web screenshots into HTML code To develop models capable of efficiently converting web screenshots into functional HTML code, WebSight (Laurençon et al., 2024) introduced a fully synthetic dataset comprising 2M pairs of HTML code and their corresponding screenshots. The HTML and TailWind CSS code were generated using DeepSeek-Coder (Guo et al., 2024), merged into a single file, and then filtered and rendered to obtain

[6] `https://huggingface.co/datasets/vikhyatk/lnqa`

the web screenshot. Instead of relying on a general LLM coder, further improvements could be achieved by using a specialist LLM fine-tuned specifically for HTML and CSS generation, enabling the creation of more diverse and visually appealing websites. In InternLM-XComposer-2.5 (Zhang et al., 2024), in addition to the WebSight dataset, the authors built a proprietary dataset that includes HTML and CSS files from The Stack v2 (Lozhkov et al., 2024) which were heavily filtered to remove external links and irrelevant content. This approach benefits from more diverse websites in the dataset, though it may introduce challenges with potentially noisy, lengthy, or difficult-to-learn examples.

Locating objects in an image Determining the exact positions of objects within an image by generating bounding boxes around them is useful for various applications, such as enabling a VLM to navigate the web by selecting where to click based on positional output. In BLIP3-GROUNDING-50M (Xue et al., 2024), large-scale grounding datasets are created by using a diverse set of images, where objects and their locations are identified using open-world image tagging and object detection models.

3.2 Fine-tuning

Similar to the approach commonly used with LLMs (Touvron et al., 2023), fine-tuning is typically done in two stages: supervised fine-tuning (SFT) followed by an alignment phase.

Which datasets should be used for the SFT? The literature offers many high-quality datasets containing diverse images and covering a wide range of tasks. They are often annotated by humans, ensuring accurate QA pairs. Although most of them are relatively small individually, when combined, they provide a sufficient number of examples for an effective SFT.
Inspired by previous work on LLMs (Wei et al., 2022; Sanh et al., 2022), InstructBLIP (Dai et al., 2023) and M3IT (Li et al., 2023) were among the first to introduce curated mixtures of academic datasets for fine-tuning VLMs. Building on these efforts, The Cauldron (Laurençon et al., 2024) introduced a collection of 50 high-quality datasets covering a broad range of tasks, including general visual question answering, counting, captioning, text transcription, document understanding, chart/figure analysis, table understanding, visual reasoning, geometry, spotting differences between two images, and converting screenshots into functional code. Each dataset in this compilation is formatted into a standardized question/answer format, and when multiple QA pairs exist per image, they are combined into a multi-turn conversation. However, a drawback of academic datasets is that their answers tend to be concise, which may lead the model to generate similarly brief responses, which are often less preferred by users. A potential solution is to use an LLM to expand and rephrase the answers, as in M3IT (Li et al., 2023) and Llava 3-V (Dubey et al., 2024).

Alignment phase There are several reasons to include an alignment stage following supervised fine-tuning. The first objective is to align the model's output with human preferences, making it more intuitive and better at following complex instructions. Additionally, as demonstrated in RLHF-V (Yu et al., 2024), this stage effectively reduces hallucinations, where the model might describe objects or details not actually present in the image. It also enhances model safety by minimizing the risk of generating harmful content. It also may further improve overall model performance.
RLAIF-V (Yu et al., 2024) provides a dataset of 80K preference pairs, used in the training of MiniCPM-V 2.5 (Yao et al., 2024). VLFeedback (Li et al., 2023) offers 380K comparison pairs, where model responses sampled from 12 VLMs are ranked by GPT-4V (Achiam et al., 2023). Similarly, SPA-VL (Zhang et al., 2024) generates 100K preference pairs through a comparable approach. DPO (Rafailov et al., 2024) is then commonly applied to these datasets during the alignment phase.

4 Challenges in evaluating VLMs

4.1 Open-ended and multiple-choice benchmarks

The earliest and most popular multimodal benchmarks, such as VQAv2 (Goyal et al., 2017), OKVQA (Marino et al., 2019), TextVQA (Singh et al., 2019), and COCO Captioning (Lin et al., 2014), are mainly open-ended. These benchmarks rely on specific ground-truth answers for each question, so even minor variations in the model's responses can lead to a score marked as incorrect.
This method of evaluation tends to favor models that produce answers closely aligned with the

benchmark's expected format or writing style. For example, VQAv2, which assesses general real-world image understanding, typically expects short answers, often just one or two words. Even when the evaluation prompt clearly specifies this format, models like Gemini 1.0 Ultra (Team et al., 2023) and GPT-4V (Achiam et al., 2023) achieve scores of 77.8 and 77.2, respectively. These scores are notably lower than those of much smaller models that include a small portion of VQAv2 in their fine-tuning data: MM1-3B-Chat (McKinzie et al., 2024) reaches 82.0, and moondream2 achieves 79.4 with only 1.9B parameters. This discrepancy highlights the challenge of evaluating different models without letting the benchmark's template influence the results.

One potential way to mitigate this bias is to perform few-shot evaluations, although this approach is less effective than training on the benchmark training set, and is not currently used for evaluating instruct models.

However, the level of ambiguity in these evaluations can vary by benchmark. For instance, TextVQA and DocVQA (Mathew et al., 2021) require the model to read and extract text directly from an image without rephrasing it, which reduces ambiguity. In MathVista (Lu et al., 2024), where answers are always numerical, each question is paired with specific instructions, such as indicating whether the answer should be an integer or a float rounded to two decimal places.

Recently proposed, the LAVE metric (Mañas et al., 2024) consists of asking an LLM to evaluate whether the response generated by the VLM is correct, given the ground truth and the specific question, thereby reducing the template problem.

Another way to reduce ambiguity is to use benchmarks that include multiple-choice questions (MCQs), where the model selects the correct option by choosing the corresponding letter. Many recent benchmarks have adopted this approach, such as MMMU (Yue et al., 2024), MMStar (Chen et al., 2024), and MMBench (Liu et al., 2023).

4.2 Challenges in model evaluation during the pre-training stage

There is a significant discrepancy between the performance of VLMs at the pre-training stage versus after fine-tuning. For instance, Idefics2-base (Laurençon et al., 2024) scores 57.9 on TextVQA (Singh et al., 2019) using 8 in-context examples and less than 55 on DocVQA (Mathew et al., 2021) during pre-training. However, after fine-tuning, it achieves 70.4 on TextVQA and 67.3 on DocVQA in a zero-shot setting, without employing the image-splitting strategy. As noted earlier, these open-ended tasks are less influenced by the specific template expected by the benchmark.

One reason for this gap is that the model only starts learning the specific task of visual question answering (beyond just image captioning or text transcription) during the fine-tuning stage—unless a third pre-training stage is conducted using large synthetic VQA datasets, as described in Figure 2, which offer examples more aligned with the ones present in benchmarks.

When instruction data is omitted during pre-training, more complex tasks like document understanding may perform poorly, and the impact of development choices in the VLM may only become evident after fine-tuning, leading to a delayed feedback loop. This delay can make pre-training ablations misleading. For example, in Idefics2, the authors found no noticeable improvements during pre-training when using 128 visual tokens instead of 64 with their architecture. While this held true for most tasks, the benefit of using more visual tokens per image became apparent in OCR tasks after fine-tuning with the image-splitting strategy. Therefore, to obtain more accurate insights during pre-training ablations, we recommend incorporating instruction data into the data mixture.

4.3 Risk of contamination and overoptimization in some benchmarks

Some benchmarks are derived from the validation or test sets of existing academic datasets. For instance, MathVista (Lu et al., 2024), a leading benchmark for evaluating reasoning and math capabilities, shows signs of potential contamination. We found that at least 6.6% of the questions include images from the training sets of academic datasets often used in supervised fine-tuning, and 2.2% feature both an image and a question that is identical or highly similar.

Additionally, this benchmark often includes questions that are especially difficult to answer unless the model has encountered them during training. For example, we find that at least 6.1% of the questions in MathVista ask variations of the question, `"What is the age gap between these two people in the image?"`. Variants of this question are also abundant on KVQA (Shah et al., 2019). Therefore, models incorporating in their fine-tuning data will have an advantage for MathVista.

Ultimately, benchmarks should be used to measure model performance, not as a training objective. Fine-tuning on similar examples can boost scores, but it provides little evidence for the model's ability to generalize to real-world scenarios. Thus, we encourage researchers to exclude images used in the benchmarks they evaluate from their supervised fine-tuning data.

5 Idefics3: adapting Llama 3 to multimodality

In this section, we detail the construction of Idefics3, a VLM based on Llama 3.1 (Dubey et al., 2024) and SigLIP-SO400M (Zhai et al., 2023). First, we begin by preparing the dataset used for training.

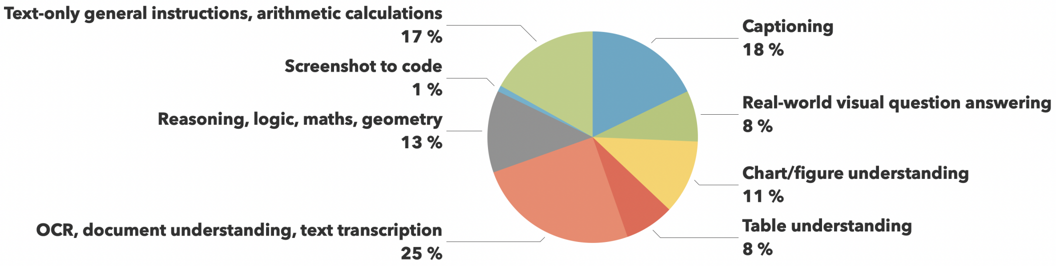
5.1 Dataset preparation

Our approach mainly takes the datasets used in the training of Idefics2 (Laurençon et al., 2024) while also adding complementary datasets for supervised fine-tuning to expand the range of tasks covered. These datasets are detailed below.

5.1.1 Extending The Cauldron

As previously mentioned, The Cauldron (Laurençon et al., 2024) is a collection of 50 high-quality datasets from existing literature. We have expanded this collection by adding 6 more datasets: Cord-v2[7] for training models to output information in JSON format, LNQA for large-scale real-world visual question answering, ShareGPT-4o and IIW-400 (Garg et al., 2024) for generating detailed captions, Geo170K (Gao et al., 2023) for tasks involving geometry, and Docmatix for document understanding.

In Table 1, we present the statistics of the datasets included in The Cauldron and the text-only instruction datasets used for the supervised fine-tuning. For each dataset, we give the number of different images it contains, the number of question-answer pairs, the total number of tokens for the answers in the question-answer pairs, and the selected percentage of answer tokens it represents in our final mixture after upsampling or downsampling.



Dataset	# images	# QA pairs	# tokens	% mix
Captioning				
ShareGPT-4o [8]	57,259	57,259	39,696,010	13.03%
LNarratives (Pont-Tuset et al., 2020)	507,444	507,444	21,328,731	1.40%
TextCaps (Sidorov et al., 2020)	21,953	21,953	389,658	1.28%
VisText (Tang et al., 2023)	7,057	9,969	1,245,485	1.23%
IIW-400 (Garg et al., 2024)	400	400	103,024	0.68%
Screen2Words (Wang et al., 2021)	15,730	15,743	143,103	0.23%
Real-world visual question answering				
LNQA [9]	302,780	1,520,942	21,107,241	3.46%

[7] https://huggingface.co/datasets/naver-clova-ix/cord-v2

[8] https://huggingface.co/datasets/OpenGVLab/ShareGPT-4o

[9] https://huggingface.co/datasets/vikhyatk/lnqa

VQAv2 (Goyal et al., 2017)	82,772	443,757	1,595,929	2.10%
COCO-QA (Ren et al., 2015)	46,287	78,736	286,982	0.94%
Visual7W (Zhu et al., 2016)	14,366	69,817	279,268	0.92%
OK-VQA (Marino et al., 2019)	8,998	9,009	38,853	0.26%
VSR (Liu et al., 2023)	2,157	3,354	10,062	0.13%
OCR, document understanding, text transcription				
Docmatix[10] (ours)	1,273,215	9,488,888	392,302,612	10.31%
RenderedText[11]	999,000	999,000	27,207,774	7.15%
DocVQA (Mathew et al., 2021)	10,189	39,463	337,829	2.22%
TextVQA (Singh et al., 2019)	21,953	34,602	181,918	1.19%
Cord-v2 [12]	800	800	178,388	1.17%
ST-VQA (Biten et al., 2019)	17,247	23,121	127,846	0.84%
OCR-VQA (Mishra et al., 2019)	165,746	801,579	6,073,824	0.60%
VisualMRC (Tanaka et al., 2021)	3,027	11,988	168,828	0.55%
IAM (Marti and Bunke, 2002)	5,663	5,663	144,216	0.47%
InfoVQA (Mathew et al., 2022)	2,118	10,074	61,048	0.40%
Diagram image-to-text[13]	300	300	22,196	0.07%
Chart/figure understanding				
Chart2Text (Obeid and Hoque, 2020)	26,985	30,242	2,852,827	4.38%
DVQA (Kafle et al., 2018)	200,000	2,325,316	8,346,234	4.27%
ChartQA (Masry et al., 2022)	18,271	28,299	185,835	1.90%
PlotQA (Methani et al., 2020)	157,070	20,249,479	8478299.278	0.65%
FigureQA (Kahou et al., 2017)	100,000	1,327,368	3,982,104	0.61%
MapQA (Chang et al., 2022)	37,417	483,416	6,470,485	0.33%
Table understanding				
TabMWP (Lu et al., 2023)	22,729	23,059	1,948,166	1.60%
TAT-QA (Zhu et al., 2021)	2,199	13,215	283,776	1.40%
HiTab (Cheng et al., 2022)	2,500	7,782	351,299	1.15%
MultiHiertt (Zhao et al., 2022)	7,619	7,830	267,615	0.88%
FinQA (Chen et al., 2021)	5,276	6,251	242,561	0.64%
WikiSQL (Zhong et al., 2017)	74,989	86,202	9,680,673	0.64%
SQA (Iyyer et al., 2017)	8,514	34,141	1,894,824	0.62%
TQA (Kembhavi et al., 2017)	1,496	6,501	26,004	0.34%
WTQ (Pasupat and Liang, 2015)	38,246	44,096	6,677,013	0.33%
Reasoning, logic, maths, geometry				
Geo170K (Gao et al., 2023)	9,067	177,457	17,971,088	2.95%
GeomVerse (Kazemi et al., 2024)	9,303	9,339	2,489,459	2.45%
CLEVR-Math (Lindström and al, 2022)	70,000	788,650	3,184,656	2.09%
CLEVR (Johnson et al., 2017)	70,000	699,989	2,396,781	0.79%
A-OKVQA (Schwenk et al., 2022)	16,539	17,056	236,492	0.78%
IconQA (Lu et al., 2021)	27,315	29,859	112,969	0.74%
AI2D (Kembhavi et al., 2016)	3,099	9,708	38,832	0.51%
NLVR2 (Suhr et al., 2019)	50,426	86,373	259,119	0.43%
RAVEN (Zhang et al., 2019)	42,000	42,000	105,081	0.43%
TallyQA (Acharya et al., 2019)	98,680	183,986	738,254	0.36%
Spot the diff (Jhamtani et al., 2018)	8,566	9,524	221,477	0.36%
GSD (Li et al., 2023)	70,939	141,869	4,637,229	0.30%
ScienceQA (Lu et al., 2022)	4,985	6,218	24,872	0.16%
Inter-GPs (Lu et al., 2021)	1,451	2,101	8,404	0.11%
HatefulMemes (Kiela et al., 2020)	8,500	8,500	25,500	0.08%

[10] `https://huggingface.co/datasets/HuggingFaceM4/Docmatix`

[11] `https://huggingface.co/datasets/wendlerc/RenderedText`

[12] `https://huggingface.co/datasets/naver-clova-ix/cord-v2`

[13] `https://huggingface.co/datasets/Kamizuru00/diagram_image_to_text`

Screenshot to code				
WebSight (Laurençon et al., 2024)	500,000	500,000	276,743,299	0.91%
DaTikz (Belouadi et al., 2023)	47,974	48,296	59,556,252	0.02%
Text-only general instructions, math problems, arithmetic calculations				
OpenHermes-2.5 (Teknium, 2023)	0	1,006,223	248,553,747	8.16%
MetaMathQA (Yu et al., 2024)	0	395,000	74,328,255	2.44%
AtlasMathSets[14]	0	17,807,579	455,411,624	2.24%
MathInstruct (Yue et al., 2024)	0	261,781	45,393,559	1.49%
OrcaMath (Mitra et al., 2024)	0	200,031	63,780,702	1.05%
Goat (Liu and Low, 2023)	0	1,746,300	167,695,693	0.55%
LIMA (Zhou et al., 2023)	0	1,052	633,867	0.52%
Dolly (Conover et al., 2023)	0	14,972	1,329,999	0.44%
CamelAIMath (Li et al., 2023)	0	49,744	21,873,629	0.04%

Table 1: The statistics of datasets used for instruction fine-tuning. # tokens is the total number of tokens for each dataset for the answers only. % mix is our selected percentage of answer tokens for each dataset in the final mixture.

5.1.2 Enhancing document understanding capabilities with Docmatix

Document understanding is a critical business application for VLMs. Yet, only a few open-source datasets are available for boosting the performance of models in this area, and they typically include only a limited number of examples. For instance, DocVQA (Mathew et al., 2021) offers 10K images and 40K QA pairs, InfographicVQA (Mathew et al., 2022) contains 2K images and 10K QA pairs, and VisualMRC (Tanaka et al., 2021) provides 3K images and 12K QA pairs.

Moreover, generating high-quality synthetic data for this task is relatively straightforward if we reframe the problem as one of LLM-based data generation rather than relying solely on VLMs. Standard OCR tools can accurately extract text from PDF documents, and an LLM can then be used to generate QA pairs based on this text.
These motivations lead us to build a large-scale document understanding dataset.

We begin with the text transcriptions from the English PDFA dataset and use Phi-3-small (Abdin et al., 2024) to generate QA pairs. To ensure diverse outputs, we employ five different prompts. To maintain dataset quality, we filter the results, discarding 15% of QA pairs flagged as incorrect. This is done by using regular expressions to detect code and removing answers containing the keyword "unanswerable." Figure 4 shows an overview of our dataset creation pipeline.

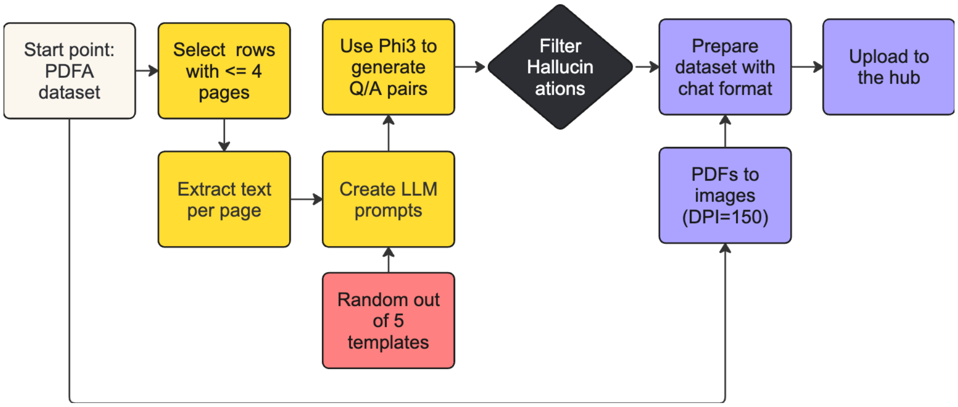

Start point: PDFA dataset
Select rows with <= 4 pages
Use Phi3 to generate Q/A pairs
Filter Hallucin ations
Prepare dataset with chat format
Upload to the hub
Extract text per page
Create LLM prompts
PDFs to images (DPI=150)
Random out of 5 templates

Figure 4: Overview of the pipeline used for the creation of Docmatix.

[14] https://huggingface.co/datasets/AtlasUnified/atlas-math-sets

The resulting dataset, Docmatix [15], includes 2.4M images and 9.5M QA pairs derived from 1.3M PDF documents, representing a 240-fold increase in scale compared to previous open datasets.

Model / Training data	Model size	DocVQA (ANLS)
Florence-2 / DocVQA	700M	60.1
Florence-2 / Docmatix	700M	71.4
Idefics2 / General mixture	8B	74.0

Table 2: Ablation on the importance of Docmatix to improve the performance on document understanding tasks.

To assess Docmatix's effectiveness, we conduct ablation studies using the Florence-2 (Xiao et al., 2024) model. We train two versions of the model: one trained over multiple epochs on the DocVQA dataset, and another trained for a single epoch on a subset of Docmatix (20% of images and 4% of QA pairs), followed by an epoch on DocVQA to ensure proper format for evaluation. The results, shown in Figure 2, are significant: training on this small portion of Docmatix leads to a nearly 20% relative improvement. Additionally, the specialist 0.7B Florence-2 model performs only 5% worse than the much larger 8B Idefics2 (Laurençon et al., 2024) generalist model.

Since Docmatix was made publicly available prior to this paper, it has already been used to enhance the performance of the moondream2 model[16], which achieved a 103% improvement on DocVQA compared to its previous version.

5.2 Building Idefics3

5.2.1 Architecture and training methods

Following Idefics2 (Laurençon et al., 2024), we use SigLIP-SO400M (Zhai et al., 2023) for the vision encoder, and swap the language model for Llama 3.1 instruct (Dubey et al., 2024), as it significantly outperforms Mistral-7B (Jiang et al., 2023).

For the connector between these backbones, Idefics2 uses a perceiver resampler to encode each image up to 980x980 pixels into 64 visual tokens. With Idefics3, we aim to enhance OCR capabilities. To address the bottleneck for OCR tasks of having too few visual tokens per image, we replace the perceiver resampler with a simple pixel shuffle strategy. This method, as in InternVL-1.5 (Chen et al., 2024), acts as a pooling technique that reduces the number of image hidden states by a factor of 4, encoding each image up to 364x364 pixels into 169 visual tokens.

During both training and inference, we follow the image-splitting strategy, where the original image is divided into a matrix of tiles of 364x364 pixels. The number of rows and columns in this matrix depends on the resolution of the original image. The vision encoder processes each tile separately, resulting in a sequence of visual tokens.
However, because images have a 2D structure, and the number of rows and columns in the tile matrix varies for each image, linearizing these visual tokens into a single sequence can cause the model to lose the information about the original arrangement of tiles, making it difficult to reconstruct their positions accurately. To address this issue, we follow the common practice of inserting a text token '\n' after each row of tiles, and of appending the original image, downscaled to 364x364 pixels, to the sequence of tiles to provide the model with the complete image in its entirety (Lin et al., 2023; Dong et al., 2024). Additionally, as in mPLUG-DocOwl-1.5 (Hu et al., 2024), we prepend each tile with the textual tokens '<row_x_col_y>', where x and y indicate the tile's position in the matrix.

Details of the Idefics3 training process are summarized in Table 3. The training involves three stages of pre-training followed by supervised fine-tuning.

In the first pre-training stage, the model's backbones remain frozen to preserve their performance while learning the newly initialized parameters. We gradually increase the maximum image resolution from 364^2 to 1820^2. From the second stage onward, we efficiently train the backbones using DoRA (Liu et al., 2024), a variant of LoRA (Hu et al., 2022), and introduce larger images into the training data. The final pre-training stage focuses on training with large synthetic datasets.

During the supervised fine-tuning phase, we apply NEFTune noise (Jain et al., 2024) to the inputs and calculate the loss only on the answer tokens. The learning rate is kept constant during the first two

[15] https://huggingface.co/datasets/HuggingFaceM4/Docmatix

[16] https://huggingface.co/vikhyatk/moondream2

	Pre-training			**SFT**
	Stage 1	*Stage 2*	*Stage 3*	
Number of steps	1000	3000	1500	5000
Learning rate (max, min)	(10^{-4}, 10^{-4})	(10^{-4}, 10^{-4})	(10^{-4}, 0)	(5x10^{-5}, 0)
Batch size	1024			
Sequence length	10K			
Max image resolution	364^2 $364^2 \rightarrow 728^2$ $728^2 \rightarrow 1092^2$ $1092^2 \rightarrow 1456^2$ $1456^2 \rightarrow 1820^2$	1820^2	1820^2	1820^2
Backbones training	Frozen	LoRA	LoRA	LoRA
Data	• OBELICS • LAION COCO	• OBELICS • LAION COCO • PDFA	• PDFA • Docmatix • Websight • LNQA • PixelProse • ChartGemma	• The Cauldron

Table 3: The different training stages of Idefics3, along with the parameters and datasets used.

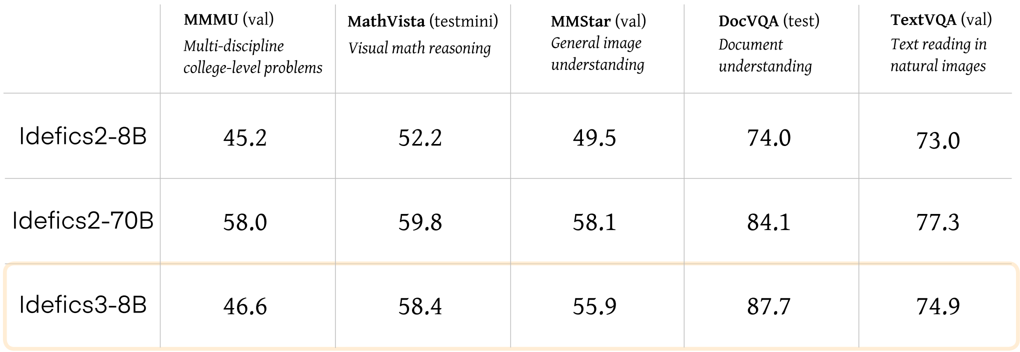
pre-training stages but is linearly decayed to zero during the final pre-training stage and supervised fine-tuning. The entire training process, including restarts, is completed in 5 days on 32 H100 nodes.

Opportunities for improvement There are several straightforward opportunities for improvement. First, although we did not encounter instabilities when fully unfreezing the backbones, we opt for a LoRA approach to enhance training efficiency. However, we believe that carefully executed full unfreezing can lead to better performance. Additionally, during the first two pre-training stages, the loss function is far from converging, but we move to the next stage to reduce computational costs. In stage 3 of pre-training, only a fraction of the examples available in the chosen datasets are used, again to reduce computational demands. Further significant improvements can be achieved by creating and incorporating the synthetic datasets mentioned in Section 3.1 into the stage 3 data mixture.

5.2.2 Evaluation

We evaluate Idefics3 on commonly adopted and challenging benchmarks: MMMU (Yue et al., 2024) for multidiscipline college-level problems, MathVista (Lu et al., 2024) for visual mathematical reasoning, MMStar (Chen et al., 2024) for general image understanding, DocVQA (Mathew et al., 2021) for document understanding, and TextVQA (Singh et al., 2019) for text reading on natural images. For Idefics3, we evaluate the benchmarks by resizing all images so that the longest side is 4x364 pixels. The exception is DocVQA, which has larger images, where we resize them to 5x364, matching the maximum resolution used during training. For Idefics2-70B, we resize the longest side of all images to 1960 pixels, the maximum resolution seen during its training. The prompts used for the evaluations are provided in Section A.1.1.

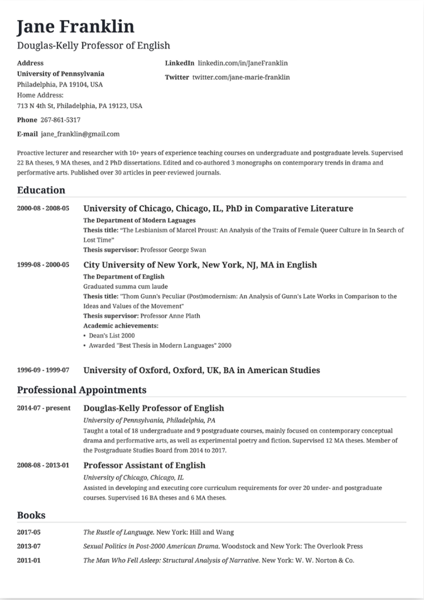
Figure 5 presents the results of Idefics3 against Idefics2-70B and Idefics2-8B (Laurençon et al., 2024). The detailed performance of Idefics3 across each category of MMMU is present in Table 4. Compared to Idefics2, Idefics3 benefits from having more visual tokens per image, a third stage of pre-training on large high-quality synthetic datasets, and an improved language model backbone. Despite being trained less extensively during the first two pre-training stages, these enhancements led to significant improvements across various tasks, particularly in document understanding tasks, with a boost of 13.7 points on DocVQA. However, the large gap of 11.4 points between Idefics2-70B and Idefics3-8B on MMMU indicates that scale is necessary for this benchmark to encapsulate sufficient knowledge into the model's weights.



	MMMU (val) *Multi-discipline college-level problems*	**MathVista** (testmini) *Visual math reasoning*	**MMStar** (val) *General image understanding*	**DocVQA** (test) *Document understanding*	**TextVQA** (val) *Text reading in natural images*
Idefics2-8B	45.2	52.2	49.5	74.0	73.0
Idefics2-70B	58.0	59.8	58.1	84.1	77.3
Idefics3-8B	46.6	58.4	55.9	87.7	74.9

Figure 5: Performance of Idefics3 against Idefics2-8B and Idefics2-70B. The evaluations are done in zero shot and without any chain-of-thought prompting.
(Benchmark, Split, Metric): (MMMU, val, MMMU score), (MathVista, testmini, MMMU score), (MMStar, val, acc.), (DocVQA, test, ANLS score), (TextVQA, val, VQA acc.).

User prompt:
Extract the information in this CV.



Jane Franklin

Douglas-Kelly Professor of English

Address
University of Pennsylvania
Philadelphia, PA 19104, USA
Home Address:
713 N 4th St, Philadelphia, PA 19123, USA

Phone 267-861-5317

E-mail jane_franklin@gmail.com

LinkedIn linkedin.com/in/JaneFranklin

Twitter twitter.com/jane-marie-franklin

Proactive lecturer and researcher with 10+ years of experience teaching courses on undergraduate and postgraduate levels. Supervised 22 BA theses, 9 MA theses, and 2 PhD dissertations. Edited and co-authored 3 monographs on contemporary trends in drama and performative arts. Published over 30 articles in peer-reviewed journals.

Education

2000-08 - 2008-05 **University of Chicago, Chicago, IL, PhD in Comparative Literature**
The Department of Modern Laguages
Thesis title: "The Lesbianism of Marcel Proust: An Analysis of the Traits of Female Queer Culture in In Search of Lost Time"
Thesis supervisor: Professor George Swan

1999-08 - 2000-05 **City University of New York, New York, NJ, MA in English**
The Department of English
Graduated summa cum laude
Thesis title: "Thom Gunn's Peculiar (Post)modernism: An Analysis of Gunn's Late Works in Comparison to the Ideas and Values of the Movement"
Thesis supervisor: Professor Anne Plath
Academic achievements:
- Dean's List 2000
- Awarded "Best Thesis in Modern Languages" 2000

1996-09 - 1999-07 **University of Oxford, Oxford, UK, BA in American Studies**

Professional Appointments

2014-07 - present **Douglas-Kelly Professor of English**
University of Pennsylvania, Philadelphia, PA
Taught a total of 18 undergraduate and 9 postgraduate courses, mainly focused on contemporary conceptual drama and performative arts, as well as experimental poetry and fiction. Supervised 12 MA theses. Member of the Postgraduate Studies Board from 2014 to 2017.

2008-08 - 2013-01 **Professor Assistant of English**
University of Chicago, Chicago, IL
Assisted in developing and executing core curriculum requirements for over 20 under- and postgraduate courses. Supervised 16 BA theses and 6 MA theses.

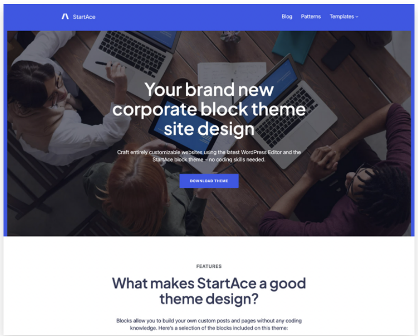
Books

2017-05	*The Rustle of Language.* New York: Hill and Wang
2013-07	*Sexual Politics in Post-2000 American Drama.* Woodstock and New York: The Overlook Press
2011-01	*The Man Who Fell Asleep: Structural Analysis of Narrative.* New York: W. W. Norton & Co.

Idefics3 output:

```
1. Personal Information:
  - Name: Jane Franklin
  - Title: Douglas-Kelly Professor of English
  - Address: University of Pennsylvania, Philadelphia, PA 19104, USA
  - Home Address: 713 N 4th St, Philadelphia, PA 19123, USA
  - Phone: 267-861-5317
  - E-mail: jane_franklin@gmail.com
2. Professional Experience:
  - 2014-07 - present: Douglas-Kelly Professor of English, University of Pennsylvania, Philadelphia, PA
    - Taught a total of 18 undergraduate and 9 postgraduate courses, mainly focused on contemporary conceptual drama and performance arts, as well as experimental poetry and fiction.
    - Supervised 12 MA theses.
    - Member of the Postgraduate Studies Board from 2014 to 2017.
  - 2008-08 - 2013-01: Professor Assistant of English, University of Chicago, Chicago, IL
    - Assisted in developing and executing core curriculum requirements for over 20 undergraduate and postgraduate courses.
    - Supervised 16 BA theses and 6 MA theses.
3. Education:
  - 2000-08 - 2008-05: University of Chicago, Chicago, IL, PhD in Comparative Literature
    - The Department of Modern Languages
    - Thesis title: "The Lesbianism of Marcel Proust: An Analysis of the Traits of Female Queer Culture in Search of Lost Time"
    - Thesis supervisor: Professor George Swan
  - 1999-08 - 2000-05: City University of New York, New York, NJ, MA in English
    - The Department of English
    - Graduated summa cum laude
    - Thesis title: "Thomas Gunn's Peculiar (Post)Modernism: An Analysis of Gunn's Late Works in Comparison to the Ideas and Values of the Movement"
    - Thesis supervisor: Professor Anne Plath
    - Academic achievements:
      - Dean's List 2000
      - Awarded "Best Thesis in Modern Languages" 2000
  - 1996-09 - 1999-07: University of Oxford, Oxford, UK, BA in American Studies
4. Books:
  - 2017-05: "The Rustle of Language," New York: Hill and Wang
  - 2013-07: "Sexual Politics in Post-2000 American Drama," Woodstock and New York: The Overlook Press
  - 2011-01: "The Man Who Fell, Asleep: Structural Analysis of Narrative," New York: W. W. Norton & Co.
```

User prompt:
Transform this website screenshot into HTML code



StartAce

Blog Patterns Templates

Your brand new corporate block theme site design

Craft entirely customizable websites using the latest WordPress Editor and the StartAce block theme – no coding skills needed.

DOWNLOAD THEME

FEATURES

What makes StartAce a good theme design?

Blocks allow you to build your own custom posts and pages without any coding knowledge. Here's a selection of the blocks included on this theme:

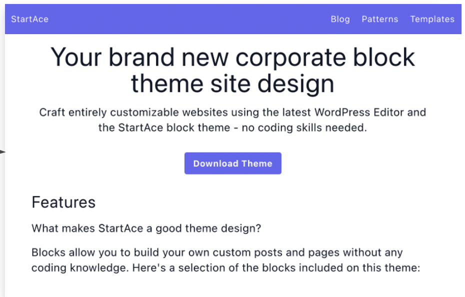
Assistant prefix:
<html>

Idefics3 output:

```
<html>
<link href="https://cdn.jsdelivr.net/npm/tailwindcss@2.2.19/dist/tailwind.min.css"
rel="stylesheet">
<body class="font-sans antialiased text-gray-900 leading-normal tracking-wider bg-cover
bg-center" style="background-image:
url('https://source.unsplash.com/random/1600x900/?web');">
```

[...] (output truncated)

Website rendering



StartAce

Blog Patterns Templates

Your brand new corporate block theme site design

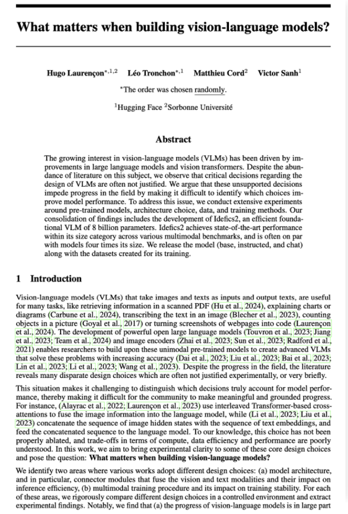
Craft entirely customizable websites using the latest WordPress Editor and the StartAce block theme - no coding skills needed.

Download Theme

Features

What makes StartAce a good theme design?

Blocks allow you to build your own custom posts and pages without any coding knowledge. Here's a selection of the blocks included on this theme:



What matters when building vision-language models?

Hugo Laurençon[*,1,2] Léo Tronchon[*,1] Matthieu Cord[2] Victor Sanh[1]

[*]The order was chosen randomly.

[1]Hugging Face [2]Sorbonne Université

Abstract

The growing interest in vision-language models (VLMs) has been driven by improvements in large language models and vision transformers. Despite the abundance of literature on this subject, we observe that critical decisions regarding the design of VLMs are often not justified. We argue that these unsupported decisions impede progress in the field by making it difficult to identify which choices improve model performance. To address this issue, we conduct extensive experiments around pre-trained models, architecture choice, data, and training methods. Our consolidation of findings includes the development of Idefics2, an efficient foundational VLM of 8 billion parameters. Idefics2 achieves state-of-the-art performance within its size category across various multimodal benchmarks, and is often on par with models four times its size. We release the model (base, instructed, and chat) along with the datasets created for its training.

1 Introduction

Vision-language models (VLMs) that take images and texts as inputs and output texts, are useful for many tasks, like retrieving information in a scanned PDF (Hu et al., 2024), explaining charts or diagrams (Carbune et al., 2024), transcribing the text in an image (Blecher et al., 2023), counting objects in a picture (Goyal et al., 2017) or turning screenshots of webpages into code (Laurençon et al., 2024). The development of powerful open large language models (Touvron et al., 2023; Jiang et al., 2023; Team et al., 2024) and image encoders (Zhai et al., 2023; Sun et al., 2023; Radford et al., 2021) enables researchers to build upon these unimodal pre-trained models to create advanced VLMs that solve these problems with increasing accuracy (Dai et al., 2023; Liu et al., 2023; Bai et al., 2023; Lin et al., 2023; Li et al., 2023; Wang et al., 2023). Despite the progress in the field, the literature reveals many disparate design choices which are often not justified experimentally, or very briefly.

This situation makes it challenging to distinguish which decisions truly account for model performance, thereby making it difficult for the community to make meaningful and grounded progress. For instance, (Alayrac et al., 2022; Laurençon et al., 2023) use interleaved Transformer-based cross-attentions to fuse the image information into the language model, while (Li et al., 2023; Liu et al., 2023) concatenate the sequence of image hidden states with the sequence of text embeddings, and feed the concatenated sequence to the language model. To our knowledge, this choice has not been properly ablated, and trade-offs in terms of compute, data efficiency and performance are poorly understood. In this work, we aim to bring experimental clarity to some of these core design choices and pose the question: **What matters when building vision-language models?**

We identify two areas where various works adopt different design choices: (a) model architecture, and in particular, connector modules that fuse the vision and text modalities and their impact on inference efficiency, (b) multimodal training procedure and its impact on training stability. For each of these areas, we rigorously compare different design choices in a controlled environment and extract experimental findings. Notably, we find that (a) the progress of vision-language models is in large part

User prompt:
Give a summary of this paper.

Idefics3 output:

TLDR of the paper: The paper discusses the challenges in building vision-language models (VLMs) and proposes a method to address these challenges. The authors argue that the design of VLMs are often not justified experimentally and that there is a lack of understanding of the impact of different design choices on model performance. They conduct extensive experiments around pre-trained models, architecture choice, data, and training methods to identify the most important factors that affect VLM performance. The authors also release a model that achieves state-of-the-art performance within its size category across various multimodal benchmarks.

Figure 6: Generation of Idefics3 on various examples.

Qualitative evaluation In Figure 6, we present examples where Idefics3 is able to accurately extract information from a CV, generate HTML code to recreate a website from a screenshot, and summarize a research paper given a screenshot. However, since the model was mainly trained on short answers during supervised fine-tuning, and did not benefit from an alignment phase, we observe that it can sometimes struggle to follow instructions for more challenging prompts. Nonetheless, we find that adding a brief prefix to the assistant's response allows the user to easily shape the generated output as desired.

6 Conclusion

In this paper, we provided a comprehensive tutorial on building vision-language models (VLMs), emphasizing the importance of architecture, data, and training methods in the development pipeline. Through an in-depth analysis of current state-of-the-art approaches, we highlighted the strengths and weaknesses of various design choices and suggested potential research directions for improving the models. We then detailed the practical steps taken to build Idefics3-8B, a VLM that demonstrates significant improvements in document understanding tasks, particularly through the use of the introduced Docmatix dataset. By releasing both the model and the datasets, we aim to contribute to develop the next generation of responsible and open VLMs.

Acknowledgement

We thank Amy Sartran for helping to integrate Idefics3 into the Transformers library (Wolf et al., 2020), Merve Noyan for building a demo for the model, and Leandro von Werra for reviewing this paper.

References

Abbas, A., K. Tirumala, D. Simig, S. Ganguli, and A. S. Morcos (2023). Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

Abdin, M., S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, et al. (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Abramovich, O., N. Nayman, S. Fogel, I. Lavi, R. Litman, S. Tsiper, R. Tichauer, S. Appalaraju, S. Mazor, and R. Manmatha (2024). Visfocus: Prompt-guided vision encoders for ocr-free dense document understanding. *arXiv preprint arXiv:2407.12594*.

Acharya, M., K. Kafle, and C. Kanan (2019). Tallyqa: Answering complex counting questions. In *AAAI*.

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agrawal, H., K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson (2019, October). nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Alayrac, J.-B., J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. a. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan (2022). Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 23716–23736. Curran Associates, Inc.

Anthropic, A. (2024). The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card 1*.

Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh (2015). VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Awadalla, A., I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, et al. (2023). Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Awadalla, A., L. Xue, O. Lo, M. Shu, H. Lee, E. K. Guha, M. Jordan, S. Shen, M. Awadalla, S. Savarese, et al. (2024). Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *arXiv preprint arXiv:2406.11271*.

Bach, S., V. Sanh, Z. X. Yong, A. Webson, C. Raffel, N. V. Nayak, A. Sharma, T. Kim, M. S. Bari, T. Fevry, Z. Alyafeai, M. Dey, A. Santilli, Z. Sun, S. Ben-david, C. Xu, G. Chhablani, H. Wang, J. Fries, M. Al-shaibani, S. Sharma, U. Thakker, K. Almubarak, X. Tang, D. Radev, M. T.-j. Jiang, and A. Rush (2022, May). PromptSource: An integrated development environment and repository for natural language prompts. In V. Basile, Z. Kozareva, and S. Stajner (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Dublin, Ireland, pp. 93–104. Association for Computational Linguistics.

Bai, J., S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, J., S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou (2023). Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Bavishi, R., E. Elsen, C. Hawthorne, M. Nye, A. Odena, A. Somani, and S. Taşırlar (2023). Introducing our multimodal models.

Belouadi, J., A. Lauscher, and S. Eger (2023). Automatikz: Text-guided synthesis of scientific vector graphics with tikz. *arXiv preprint arXiv:2310.00367*.

Betker, J., G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al. (2023). Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf 2*(3), 8.

Beyer, L., A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, et al. (2024). Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Biten, A. F., R. Tito, L. Gomez, E. Valveny, and D. Karatzas (2022). Ocr-idl: Ocr annotations for industry document library dataset. In *European Conference on Computer Vision*, pp. 241–252. Springer.

Biten, A. F., R. Tito, A. Mafla, L. Gomez, M. Rusiñol, C. Jawahar, E. Valveny, and D. Karatzas (2019). Scene text visual question answering. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4290–4300.

Blecher, L., G. Cucurull, T. Scialom, and R. Stojnic (2023). Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 1877–1901. Curran Associates, Inc.

Byeon, M., B. Park, H. Kim, S. Lee, W. Baek, and S. Kim (2022). Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`.

Cai, Z., M. Cao, H. Chen, K. Chen, K. Chen, X. Chen, X. Chen, Z. Chen, Z. Chen, P. Chu, et al. (2024). Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Carbune, V., H. Mansoor, F. Liu, R. Aralikatte, G. Baechler, J. Chen, and A. Sharma (2024). Chart-based reasoning: Transferring capabilities from llms to vlms. *arXiv preprint arXiv:2403.12596*.

Cha, J., W. Kang, J. Mun, and B. Roh (2024). Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13817–13827.

Chang, S., D. Palzer, J. Li, E. Fosler-Lussier, and N. Xiao (2022). MapQA: A dataset for question answering on choropleth maps. In *NeurIPS 2022 First Table Representation Workshop*.

Changpinyo, S., P. Sharma, N. Ding, and R. Soricut (2021). Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Chen, L., J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin (2023). Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Chen, L., J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, et al. (2024). Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Chen, X., J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, C. R. Ruiz, S. Goodman, X. Wang, Y. Tay, et al. (2023). Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.

Chen, X. and X. Wang (2022). Pali: Scaling language-image learning in 100+ languages. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Chen, X., X. Wang, L. Beyer, A. Kolesnikov, J. Wu, P. Voigtlaender, B. Mustafa, S. Goodman, I. Alabdulmohsin, P. Padlewski, et al. (2023). Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*.

Chen, Z., W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, and W. Y. Wang (2021, November). FinQA: A dataset of numerical reasoning over financial data. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, pp. 3697–3711. Association for Computational Linguistics.

Chen, Z., W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. (2024). How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Cheng, Z., H. Dong, Z. Wang, R. Jia, J. Guo, Y. Gao, S. Han, J.-G. Lou, and D. Zhang (2022, May). HiTab: A hierarchical table dataset for question answering and natural language generation. In S. Muresan, P. Nakov, and A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, pp. 1094–1110. Association for Computational Linguistics.

Chu, X., L. Qiao, X. Zhang, S. Xu, F. Wei, Y. Yang, X. Sun, Y. Hu, X. Lin, B. Zhang, et al. (2024). Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*.

Conover, M., M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, and R. Xin (2023). Free dolly: Introducing the world's first truly open instruction-tuned llm. `https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm`. Accessed: 2023-06-30.

Dai, W., J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi (2023). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Darcet, T., M. Oquab, J. Mairal, and P. Bojanowski (2024). Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*.

DeepSeek-AI, A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Xu, H. Yang, H. Zhang, H. Ding, H. Xin, H. Gao, H. Li, H. Qu, J. L. Cai, J. Liang, J. Guo, J. Ni, J. Li, J. Chen, J. Yuan, J. Qiu, J. Song, K. Dong, K. Gao, K. Guan, L. Wang, L. Zhang, L. Xu, L. Xia, L. Zhao, L. Zhang, M. Li, M. Wang, M. Zhang, M. Zhang, M. Tang, M. Li, N. Tian, P. Huang, P. Wang, P. Zhang, Q. Zhu, Q. Chen, Q. Du, R. J. Chen, R. L. Jin, R. Ge, R. Pan, R. Xu, R. Chen, S. S. Li, S. Lu, S. Zhou, S. Chen, S. Wu, S. Ye, S. Ma, S. Wang, S. Zhou, S. Yu, S. Zhou, S. Zheng, T. Wang, T. Pei, T. Yuan, T. Sun, W. L. Xiao, W. Zeng, W. An, W. Liu, W. Liang, W. Gao, W. Zhang, X. Q. Li, X. Jin, X. Wang, X. Bi, X. Liu, X. Wang, X. Shen, X. Chen, X. Chen, X. Nie, X. Sun, X. Wang, X. Liu, X. Xie, X. Yu, X. Song, X. Zhou, X. Yang, X. Lu, X. Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zhao, Y. Sun, Y. Li, Y. Wang, Y. Zheng, Y. Zhang,

Y. Xiong, Y. Zhao, Y. He, Y. Tang, Y. Piao, Y. Dong, Y. Tan, Y. Liu, Y. Wang, Y. Guo, Y. Zhu, Y. Wang, Y. Zou, Y. Zha, Y. Ma, Y. Yan, Y. You, Y. Liu, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Huang, Z. Zhang, Z. Xie, Z. Hao, Z. Shao, Z. Wen, Z. Xu, Z. Zhang, Z. Li, Z. Wang, Z. Gu, Z. Li, and Z. Xie (2024). Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Dehghani, M., J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, R. Jenatton, L. Beyer, M. Tschannen, A. Arnab, X. Wang, C. Riquelme Ruiz, M. Minderer, J. Puigcerver, U. Evci, M. Kumar, S. V. Steenkiste, G. F. Elsayed, A. Mahendran, F. Yu, A. Oliver, F. Huot, J. Bastings, M. Collier, A. A. Gritsenko, V. Birodkar, C. N. Vasconcelos, Y. Tay, T. Mensink, A. Kolesnikov, F. Pavetic, D. Tran, T. Kipf, M. Lucic, X. Zhai, D. Keysers, J. J. Harmsen, and N. Houlsby (2023, 23–29 Jul). Scaling vision transformers to 22 billion parameters. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning*, Volume 202 of *Proceedings of Machine Learning Research*, pp. 7480–7512. PMLR.

Dehghani, M., B. Mustafa, J. Djolonga, J. Heek, M. Minderer, M. Caron, A. P. Steiner, J. Puigcerver, R. Geirhos, I. Alabdulmohsin, A. Oliver, P. Padlewski, A. A. Gritsenko, M. Lucic, and N. Houlsby (2023). Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

Desai, K., G. Kaul, Z. Aysola, and J. Johnson (2021). Redcaps: Web-curated image-text data created by the people, for the people. In J. Vanschoren and S. Yeung (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Volume 1. Curran.

Dong, X., P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, S. Zhang, H. Duan, W. Zhang, Y. Li, et al. (2024). Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*.

Driess, D., F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence (2023). Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Duan, H., J. Yang, Y. Qiao, X. Fang, L. Chen, Y. Liu, X. Dong, Y. Zang, P. Zhang, J. Wang, et al. (2024). Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*.

Dubey, A., A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gadre, S. Y., G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, et al. (2024). Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems 36*.

Gao, J., R. Pi, J. Zhang, J. Ye, W. Zhong, Y. Wang, L. Hong, J. Han, H. Xu, Z. Li, et al. (2023). G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Gao, P., R. Zhang, C. Liu, L. Qiu, S. Huang, W. Lin, S. Zhao, S. Geng, Z. Lin, P. Jin, et al. (2024). Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*.

Garg, R., A. Burns, B. K. Ayan, Y. Bitton, C. Montgomery, Y. Onoe, A. Bunner, R. Krishna, J. Baldridge, and R. Soricut (2024). Imageinwords: Unlocking hyper-detailed image descriptions. *arXiv preprint arXiv:2405.02793*.

Goyal, Y., T. Khot, D. Summers-Stay, D. Batra, and D. Parikh (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334.

Gunasekar, S., Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, et al. (2023). Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Guo, D., Q. Zhu, D. Yang, Z. Xie, K. Dong, W. Zhang, G. Chen, X. Bi, Y. Wu, Y. Li, et al. (2024). Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

He, X., Y. Zhang, L. Mou, E. Xing, and P. Xie (2020). Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Hendrycks, D., C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Hong, W., W. Wang, Q. Lv, J. Xu, W. Yu, J. Ji, Y. Wang, Z. Wang, Y. Dong, M. Ding, et al. (2023). Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*.

Hu, A., H. Xu, J. Ye, M. Yan, L. Zhang, B. Zhang, C. Li, J. Zhang, Q. Jin, F. Huang, et al. (2024). mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*.

Hu, E. J., yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hu, S., Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao, et al. (2024). Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Huang, S., L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, and F. Wei (2023). Language is not all you need: Aligning perception with language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hudson, D. A. and C. D. Manning (2019). 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)*, pp. 6693–6702.

Iyyer, M., W.-t. Yih, and M.-W. Chang (2017, July). Search-based neural structured learning for sequential question answering. In R. Barzilay and M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, pp. 1821–1831. Association for Computational Linguistics.

Jaegle, A., F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira (2021, 18–24 Jul). Perceiver: General perception with iterative attention. In M. Meila and T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning*, Volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR.

Jain, N., P. yeh Chiang, Y. Wen, J. Kirchenbauer, H.-M. Chu, G. Somepalli, B. R. Bartoldson, B. Kailkhura, A. Schwarzschild, A. Saha, M. Goldblum, J. Geiping, and T. Goldstein (2024). NEFTune: Noisy embeddings improve instruction finetuning. In *The Twelfth International Conference on Learning Representations*.

Jhamtani, H. et al. (2018, October-November). Learning to describe differences between pairs of similar images. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 4024–4034. Association for Computational Linguistics.

Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jiang, D., X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen (2024). Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Johnson, J., B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1997.

Kafle, K., S. Cohen, B. Price, and C. Kanan (2018). Dvqa: Understanding data visualizations via question answering. In *CVPR*.

Kahou, S. E., V. Michalski, A. Atkinson, Á. Kádár, A. Trischler, and Y. Bengio (2017). Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Karamcheti, S., S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh (2024). Prismatic vlms: Investigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*.

Kazemi, M., H. Alvari, A. Anand, J. Wu, X. Chen, and R. Soricut (2024). Geomverse: A systematic evaluation of large models for geometric reasoning. In *Synthetic Data for Computer Vision Workshop @ CVPR 2024*.

Kembhavi, A., M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi (2016). A diagram is worth a dozen images. In B. Leibe, J. Matas, N. Sebe, and M. Welling (Eds.), *Computer Vision – ECCV 2016*, Cham, pp. 235–251. Springer International Publishing.

Kembhavi, A., M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi (2017). Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5376–5384.

Kiela, D., H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 2611–2624. Curran Associates, Inc.

Kingma, D. and J. Ba (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA.

Koh, J. Y., R. Salakhutdinov, and D. Fried (2023). Grounding language models to images for multimodal inputs and outputs.

Lai, Z., H. Zhang, W. Wu, H. Bai, A. Timofeev, X. Du, Z. Gan, J. Shan, C.-N. Chuah, Y. Yang, et al. (2023). From scarcity to efficiency: Improving clip training via visual-enriched captions. *arXiv preprint arXiv:2310.07699*.

Lau, J., S. Gayen, A. Ben Abacha, and D. Demner-Fushman (2018, 11). A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data 5*, 180251.

Laurençon, H., L. Saulnier, T. Wang, C. Akiki, A. Villanova del Moral, T. Le Scao, L. Von Werra, C. Mou, E. González Ponferrada, H. Nguyen, J. Frohberg, M. Šaško, Q. Lhoest, A. McMillan-Major, G. Dupont, S. Biderman, A. Rogers, L. Ben allal, F. De Toni, G. Pistilli, O. Nguyen, S. Nikpoor, M. Masoud, P. Colombo, J. de la Rosa, P. Villegas, T. Thrush, S. Longpre, S. Nagel, L. Weber, M. Muñoz, J. Zhu, D. Van Strien, Z. Alyafeai, K. Almubarak, M. C. Vu, I. Gonzalez-Dios, A. Soroa, K. Lo, M. Dey, P. Ortiz Suarez, A. Gokaslan, S. Bose, D. Adelani, L. Phan, H. Tran, I. Yu, S. Pai, J. Chim, V. Lepercq, S. Ilic, M. Mitchell, S. A. Luccioni, and Y. Jernite (2022). The bigscience roots corpus: A 1.6tb composite multilingual dataset. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 31809–31826. Curran Associates, Inc.

Laurençon, H., L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh (2023). 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*.

Laurençon, H., L. Tronchon, M. Cord, and V. Sanh (2024). What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Laurençon, H., L. Tronchon, and V. Sanh (2024). Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029*.

Lee, B.-K., C. W. Kim, B. Park, and Y. M. Ro (2024). Meteor: Mamba-based traversal of rationale for large language and vision models. *arXiv preprint arXiv:2405.15574*.

Lee, B.-K., B. Park, C. W. Kim, and Y. M. Ro (2024). Moai: Mixture of all intelligence for large language and vision models. *arXiv preprint arXiv:2403.07508*.

Lee, K., M. Joshi, I. Turc, H. Hu, F. Liu, J. Eisenschlos, U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova (2023). Pix2struct: screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Li, B., R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan (2023). Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Li, B., Y. Zhang, L. Chen, J. Wang, F. Pu, J. Yang, C. Li, and Z. Liu (2023). Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Li, B., Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li (2024). Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, G., H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem (2023). CAMEL: Communicative agents for ”mind” exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Li, J., D. Li, S. Savarese, and S. Hoi (2023). Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Li, J., D. Li, C. Xiong, and S. Hoi (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Li, L., Z. Xie, M. Li, S. Chen, P. Wang, L. Chen, Y. Yang, B. Wang, and L. Kong (2023). Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.

Li, L., Y. Yin, S. Li, L. Chen, P. Wang, S. Ren, M. Li, Y. Yang, J. Xu, X. Sun, et al. (2023). M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.

Li, Q., Z. Chen, W. Wang, W. Wang, S. Ye, Z. Jin, G. Chen, Y. He, Z. Gao, E. Cui, et al. (2024). Omnicorpus: An unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*.

Li, Y., Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen (2023, December). Evaluating object hallucination in large vision-language models. In H. Bouamor, J. Pino, and K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 292–305. Association for Computational Linguistics.

Li, Y., Y. Zhang, C. Wang, Z. Zhong, Y. Chen, R. Chu, S. Liu, and J. Jia (2024). Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Li, Z., B. Yang, Q. Liu, Z. Ma, S. Zhang, J. Yang, Y. Sun, Y. Liu, and X. Bai (2023). Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*.

Lin, B., Z. Tang, Y. Ye, J. Cui, B. Zhu, P. Jin, J. Zhang, M. Ning, and L. Yuan (2024). Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Lin, J., H. Yin, W. Ping, Y. Lu, P. Molchanov, A. Tao, H. Mao, J. Kautz, M. Shoeybi, and S. Han (2023). Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*.

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014). Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014*, Cham, pp. 740–755. Springer International Publishing.

Lin, Z., C. Liu, R. Zhang, P. Gao, L. Qiu, H. Xiao, H. Qiu, C. Lin, W. Shao, K. Chen, et al. (2023). Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*.

Lindström, A. D. and al (2022). Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*.

Liu, F., G. Emerson, and N. Collier (2023). Visual spatial reasoning. *Transactions of the Association for Computational Linguistics 11*, 635–651.

Liu, H., C. Li, Y. Li, and Y. J. Lee (2023). Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Liu, H., C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee (2024, January). Llava-next: Improved reasoning, ocr, and world knowledge.

Liu, H., C. Li, Q. Wu, and Y. J. Lee (2023). Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Liu, H., Q. You, X. Han, Y. Wang, B. Zhai, Y. Liu, Y. Tao, H. Huang, R. He, and H. Yang (2024). Infimm-hd: A leap forward in high-resolution multimodal understanding. *arXiv preprint arXiv:2403.01487*.

Liu, R., J. Wei, F. Liu, C. Si, Y. Zhang, J. Rao, S. Zheng, D. Peng, D. Yang, D. Zhou, and A. M. Dai (2024). Best practices and lessons learned on synthetic data for language models. *CoRR abs/2404.07503*.

Liu, S.-Y., C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen (2024). Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.

Liu, T. and B. K. H. Low (2023). Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks. *arXiv preprint arXiv:2305.14201*.

Liu, Y., H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. (2023). Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Loshchilov, I. and F. Hutter (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lozhkov, A., R. Li, L. B. Allal, F. Cassano, J. Lamy-Poirier, N. Tazi, A. Tang, D. Pykhtar, J. Liu, Y. Wei, et al. (2024). Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*.

Lu, H., W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, Y. Sun, et al. (2024). Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Lu, J., C. Clark, S. Lee, Z. Zhang, S. Khosla, R. Marten, D. Hoiem, and A. Kembhavi (2023). Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*.

Lu, P., H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao (2024). Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Lu, P., R. Gong, S. Jiang, L. Qiu, S. Huang, X. Liang, and S.-C. Zhu (2021). Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Lu, P., S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan (2022). 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 (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 2507–2521. Curran Associates, Inc.

Lu, P., L. Qiu, K.-W. Chang, Y. N. Wu, S.-C. Zhu, T. Rajpurohit, P. Clark, and A. Kalyan (2023). Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*.

Lu, P., L. Qiu, J. Chen, T. Xia, Y. Zhao, W. Zhang, Z. Yu, X. Liang, and S.-C. Zhu (2021). Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

Mañas, O., B. Krojer, and A. Agrawal (2024). Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 38, pp. 4171–4179.

Mañas, O., P. Rodriguez Lopez, S. Ahmadi, A. Nematzadeh, Y. Goyal, and A. Agrawal (2023, May). MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In A. Vlachos and I. Augenstein (Eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia, pp. 2523–2548. Association for Computational Linguistics.

Marino, K., M. Rastegari, A. Farhadi, and R. Mottaghi (2019). Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Marti, U.-V. and H. Bunke (2002, 11). The iam-database: An english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition 5*, 39–46.

Masry, A., D. Long, J. Q. Tan, S. Joty, and E. Hoque (2022, May). ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland, pp. 2263–2279. Association for Computational Linguistics.

Masry, A., M. Thakkar, A. Bajaj, A. Kartha, E. Hoque, and S. Joty (2024). Chartgemma: Visual instruction-tuning for chart reasoning in the wild. *arXiv preprint arXiv:2407.04172*.

Mathew, M., V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. V. Jawahar (2022). Infographicvqa. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591.

Mathew, M., D. Karatzas, and C. V. Jawahar (2021). Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208.

McKinzie, B., Z. Gan, J.-P. Fauconnier, S. Dodge, B. Zhang, P. Dufter, D. Shah, X. Du, F. Peng, F. Weers, et al. (2024). Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*.

Mehta, S., M. H. Sekhavat, Q. Cao, M. Horton, Y. Jin, C. Sun, I. Mirzadeh, M. Najibi, D. Belenko, P. Zatloukal, et al. (2024). Openelm: An efficient language model family with open-source training and inference framework. *arXiv preprint arXiv:2404.14619*.

Methani, N., P. Ganguly, M. M. Khapra, and P. Kumar (2020, March). Plotqa: Reasoning over scientific plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Mishra, A., S. Shekhar, A. K. Singh, and A. Chakraborty (2019). Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 947–952.

Mitra, A., H. Khanpour, C. Rosset, and A. Awadallah (2024). Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*.

Obeid, J. and E. Hoque (2020, December). Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In B. Davis, Y. Graham, J. Kelleher, and Y. Sripada (Eds.), *Proceedings of the 13th International Conference on Natural Language Generation*, Dublin, Ireland, pp. 138–147. Association for Computational Linguistics.

Oquab, M., T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Pasupat, P. and P. Liang (2015, July). Compositional semantic parsing on semi-structured tables. In C. Zong and M. Strube (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China, pp. 1470–1480. Association for Computational Linguistics.

Penedo, G., H. Kydlíček, A. Lozhkov, M. Mitchell, C. Raffel, L. Von Werra, T. Wolf, et al. (2024). The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Penedo, G., Q. Malartic, D. Hesslow, R. Cojocaru, H. Alobeidli, A. Cappelli, B. Pannier, E. Almazrouei, and J. Launay (2023). The refinedweb dataset for falcon LLM: Outperforming curated corpora with web data only. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Pont-Tuset, J., J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari (2020). Connecting vision and language with localized narratives. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020*, Cham, pp. 647–664. Springer International Publishing.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Rafailov, R., A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems 36*.

Reid, M., N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Ren, M., R. Kiros, and R. Zemel (2015). Exploring models and data for image question answering. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 28. Curran Associates, Inc.

Sanh, V., A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, and A. M. Rush (2022). Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Schuhmann, C., R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 25278–25294. Curran Associates, Inc.

Schuhmann, C., A. Köpf, R. Vencu, T. Coombes, and R. Beaumont (2022). Laion coco: 600m synthetic captions from laion2b-en. *URL https://laion.ai/blog/laion-coco*.

Schuhmann, C., R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Schwenk, D., A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi (2022). 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*, Berlin, Heidelberg, pp. 146–162. Springer-Verlag.

Shah, S., A. Mishra, N. Yadati, and P. P. Talukdar (2019). Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, Volume 33, pp. 8876–8884.

Sharifzadeh, S., C. Kaplanis, S. Pathak, D. Kumaran, A. Ilic, J. Mitrovic, C. Blundell, and A. Banino (2024). Synth2: Boosting visual-language models with synthetic captions and image embeddings. *arXiv preprint arXiv:2403.07750*.

Sharma, P., N. Ding, S. Goodman, and R. Soricut (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Shayegani, E., Y. Dong, and N. Abu-Ghazaleh (2024). Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The Twelfth International Conference on Learning Representations*.

Shukor, M., C. Dancette, and M. Cord (2023, oct). ep-alm: Efficient perceptual augmentation of language models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA, pp. 21999–22012. IEEE Computer Society.

Sidorov, O., R. Hu, M. Rohrbach, and A. Singh (2020). Textcaps: A dataset for image captioning with reading comprehension. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020*, Cham, pp. 742–758. Springer International Publishing.

Singh, A., R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela (2022). Flava: A foundational language and vision alignment model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15617–15629.

Singh, A., V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, and M. Rohrbach (2019). Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326.

Singla, V., K. Yue, S. Paul, R. Shirkavand, M. Jayawardhana, A. Ganjdanesh, H. Huang, A. Bhatele, G. Somepalli, and T. Goldstein (2024). From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*.

Srinivasan, K., K. Raman, J. Chen, M. Bendersky, and M. Najork (2021). Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, New York, NY, USA, pp. 2443–2449. Association for Computing Machinery.

Suhr, A., S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi (2019, July). A corpus for reasoning about natural language grounded in photographs. In A. Korhonen, D. Traum, and L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 6418–6428. Association for Computational Linguistics.

Sun, Q., Y. Cui, X. Zhang, F. Zhang, Q. Yu, Z. Luo, Y. Wang, Y. Rao, J. Liu, T. Huang, et al. (2023). Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*.

Sun, Q., Y. Fang, L. Wu, X. Wang, and Y. Cao (2023). Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Sun, Z., S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, et al. (2023). Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Tanaka, R., K. Nishida, and S. Yoshida (2021). Visualmrc: Machine reading comprehension on document images. In *AAAI*.

Tang, B. J., A. Boggust, and A. Satyanarayan (2023). VisText: A Benchmark for Semantically Rich Chart Captioning. In *The Annual Meeting of the Association for Computational Linguistics (ACL)*.

Team, G., R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, G., T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Teknium (2023). Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

Thiel, D. (2023). Identifying and eliminating csam in generative ml training data and models.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trinh, T. H., Y. Wu, Q. V. Le, H. He, and T. Luong (2024). Solving olympiad geometry without human demonstrations. *Nature 625*(7995), 476–482.

Tsimpoukelli, M., J. L. Menick, S. Cabi, S. M. A. Eslami, O. Vinyals, and F. Hill (2021). Multimodal few-shot learning with frozen language models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems*, Volume 34, pp. 200–212. Curran Associates, Inc.

Vallaeys, T., M. Shukor, M. Cord, and J. Verbeek (2024). Improved baselines for data-efficient perceptual augmentation of llms. *arXiv preprint arXiv:2403.13499*.

Wadekar, S. N., A. Chaurasia, A. Chadha, and E. Culurciello (2024). The evolution of multimodal model architectures. *arXiv preprint arXiv:2405.17927*.

Wang, B., G. Li, X. Zhou, Z. Chen, T. Grossman, and Y. Li (2021). Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, UIST '21, New York, NY, USA, pp. 498–510. Association for Computing Machinery.

Wang, W., Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, et al. (2023). Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Webster, R., J. Rabin, L. Simon, and F. Jurie (2023). On the de-duplication of laion-2b. *arXiv preprint arXiv:2303.12733*.

Wei, J., M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le (2022). Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems 35*, 24824–24837.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush (2020, October). Transformers: State-of-the-art natural language processing. In Q. Liu and D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, pp. 38–45. Association for Computational Linguistics.

Xiao, B., H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan (2024). Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829.

Xiao, J., Z. Xu, A. Yuille, S. Yan, and B. Wang (2024). Palm2-vadapter: Progressively aligned language model makes a strong vision-language adapter. *arXiv preprint arXiv:2402.10896*.

Xie, S., R. Girshick, P. Dollár, Z. Tu, and K. He (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500.

Xue, L., M. Shu, A. Awadalla, J. Wang, A. Yan, S. Purushwalkam, H. Zhou, V. Prabhu, Y. Dai, M. S. Ryoo, S. Kendre, J. Zhang, C. Qin, S. Zhang, C.-C. Chen, N. Yu, J. Tan, T. M. Awalgaonkar, S. Heinecke, H. Wang, Y. Choi, L. Schmidt, Z. Chen, S. Savarese, J. C. Niebles, C. Xiong, and R. Xu (2024). xgen-mm (blip-3): A family of open large multimodal models.

Yao, Y., T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, Q. Chen, H. Zhou, Z. Zou, H. Zhang, S. Hu, Z. Zheng, J. Zhou, J. Cai, X. Han, G. Zeng, D. Li, Z. Liu, and M. Sun (2024). Minicpm-v: A gpt-4v level mllm on your phone.

Ye, J., A. Hu, H. Xu, Q. Ye, M. Yan, G. Xu, C. Li, J. Tian, Q. Qian, J. Zhang, et al. (2023). Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Young, P., A. Lai, M. Hodosh, and J. Hockenmaier (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics 2*, 67–78.

Yu, L., W. Jiang, H. Shi, J. YU, Z. Liu, Y. Zhang, J. Kwok, Z. Li, A. Weller, and W. Liu (2024). Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*.

Yu, T., Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H.-T. Zheng, M. Sun, et al. (2024). Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816.

Yu, T., H. Zhang, Y. Yao, Y. Dang, D. Chen, X. Lu, G. Cui, T. He, Z. Liu, T.-S. Chua, et al. (2024). Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Yue, X., Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen (2024). Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Yue, X., X. Qu, G. Zhang, Y. Fu, W. Huang, H. Sun, Y. Su, and W. Chen (2024). MAmmoTH: Building math generalist models through hybrid instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Zhai, X., B. Mustafa, A. Kolesnikov, and L. Beyer (2023). Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986.

Zhang, C., F. Gao, B. Jia, Y. Zhu, and S.-C. Zhu (2019). Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, P., X. Dong, Y. Zang, Y. Cao, R. Qian, L. Chen, Q. Guo, H. Duan, B. Wang, L. Ouyang, et al. (2024). Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

Zhang, R., X. Wei, D. Jiang, Y. Zhang, Z. Guo, C. Tong, J. Liu, A. Zhou, B. Wei, S. Zhang, et al. (2024). Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*.

Zhang, X., C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie (2023). Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Zhang, Y., L. Chen, G. Zheng, Y. Gao, R. Zheng, J. Fu, Z. Yin, S. Jin, Y. Qiao, X. Huang, et al. (2024). Spa-vl: A comprehensive safety preference alignment dataset for vision language model. *arXiv preprint arXiv:2406.12030*.

Zhang, Y., R. Zhang, J. Gu, Y. Zhou, N. Lipka, D. Yang, and T. Sun (2023). Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Zhao, Y., Y. Li, C. Li, and R. Zhang (2022, May). MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, pp. 6588–6600. Association for Computational Linguistics.

Zhao, Y., C. Zhao, L. Nan, Z. Qi, W. Zhang, X. Tang, B. Mi, and D. Radev (2023, July). RobuT: A systematic study of table QA robustness against human-annotated adversarial perturbations. In A. Rogers, J. Boyd-Graber, and N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, pp. 6064–6081. Association for Computational Linguistics.

Zheng, L., W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems 36*.

Zhong, V., C. Xiong, and R. Socher (2017). Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Zhou, B., Y. Hu, X. Weng, J. Jia, J. Luo, X. Liu, J. Wu, and L. Huang (2024). Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

Zhou, C., P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. YU, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy (2023). LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhu, F., W. Lei, Y. Huang, C. Wang, S. Zhang, J. Lv, F. Feng, and T.-S. Chua (2021, August). TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, pp. 3277–3287. Association for Computational Linguistics.

Zhu, W., J. Hessel, A. Awadalla, S. Y. Gadre, J. Dodge, A. Fang, Y. Yu, L. Schmidt, W. Y. Wang, and Y. Choi (2023). Multimodal c4: An open, billion-scale corpus of images interleaved with text. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhu, Y., O. Groth, M. Bernstein, and L. Fei-Fei (2016). Visual7W: Grounded Question Answering in Images. In *IEEE Conference on Computer Vision and Pattern Recognition*.

A Appendix

A.1 Evaluation of Idefics3

A.1.1 Prompts used for the evaluation

We evaluate MMStar (Chen et al., 2024) using our default template for multiple-choice questions, as seen during supervised fine-tuning:

```
Question: {question}
Choices:
A. {choice_a}
B. {choice_b}
C. {choice_c}
D. {choice_d}
...
Answer with the letter.
```

We evaluate MMMU (Yue et al., 2024) and MathVista (Lu et al., 2024) using the VLMEvalKit (Duan et al., 2024) library. For the multiple-choice questions in these benchmarks, we also use our default template.

For TextVQA (Singh et al., 2019) and DocVQA (Mathew et al., 2021), we evaluate and train using the prompts from Gemini (Reid et al., 2024).

TextVQA

```
Answer the following question about the image using as few words as possible. Follow these additional
instructions:
-Always answer a binary question with Yes or No.
-When asked what time it is, reply with the time seen in the image.
-Do not put any full stops at the end of the answer.
-Do not put quotation marks around the answer.
-An answer with one or two words is favorable.
-Do not apply common sense knowledge. The answer can be found in the image.
Question: question
```

DocVQA

```
Give a short and terse answer to the following question. Do not paraphrase or reformat the text you
see in the image. Do not include any full stops. Just give the answer without additional explanation.
Question: {question}
```

We use the stop words `Question`, `User`, `<end_of_utterance>` and the EOS token to stop a generation.

A.1.2 Detailed performance on MMMU

The detailed performance of Idefics3 across each category of MMMU (Yue et al., 2024) is present in Table 4.

MMMU category	**Score**
Overall	46.6
Accounting	33.3
Agriculture	56.7
Architecture and Engineering	33.3
Art	56.7
Art Theory	76.7
Basic Medical Science	50.0
Biology	36.7
Chemistry	40.0
Clinical Medicine	53.3
Computer Science	50.0
Design	73.3
Diagnostics and Laboratory Medicine	43.3
Economics	40.0
Electronics	40.0
Energy and Power	36.7
Finance	40.0
Geography	50.0
History	56.7
Literature	80.0
Manage	47.7
Marketing	53.3
Materials	26.7
Math	26.7
Mechanical Engineering	33.3
Music	26.7
Pharmacy	53.3
Physics	26.7
Psychology	53.3
Public Health	46.7
Sociology	56.7
Art & Design	58.3
Business	42.7
Health & Medicine	49.3
Humanities & Social Science	61.7
Science	36.0
Tech & Engineering	39.5

Table 4: Detailed performance of Idefics3 across each category of MMMU. (Yue et al., 2024).