

NVLM: Open Frontier-Class Multimodal LLMs

Wenliang Dai* Nayeon Lee* Boxin Wang* Zhuolin Yang*
 Zihan Liu Jon Barker Tuomas Rintamaki Mohammad Shoeybi Bryan Catanzaro
 Wei Ping*,†

NVIDIA

* Equal contributions, ordered alphabetically
 {wdai, nayeonl, boxinw, zhuoliny, wping}@nvidia.com

† Leads the effort.

Abstract

We introduce NVLM 1.0,¹ a family of frontier-class multimodal large language models (LLMs) that achieve state-of-the-art results on vision-language tasks, rivaling the leading proprietary models (e.g., GPT-4o) and open-access models (e.g., Llama 3-V 405B and InternVL 2). Remarkably, NVLM 1.0 shows improved text-only performance over its LLM backbone after multimodal training.

In terms of **model design**, we perform a comprehensive comparison between decoder-only multimodal LLMs (e.g., LLaVA) and cross-attention-based models (e.g., Flamingo). Based on the strengths and weaknesses of both approaches, we propose a novel architecture that enhances both training efficiency and multimodal reasoning capabilities. Furthermore, we introduce a 1-D tile-tagging design for tile-based dynamic high-resolution images, which significantly boosts performance on multimodal reasoning and OCR-related tasks. Regarding **training data**, we meticulously curate and provide detailed information on our multimodal pretraining and supervised fine-tuning datasets. Our findings indicate that dataset quality and task diversity are more important than scale, even during the pretraining phase, across all architectures. Notably, we develop **production-grade multimodality** for the NVLM-1.0 models, enabling them to excel in vision-language tasks while maintaining and even improving text-only performance compared to their LLM backbones. To achieve this, we craft and integrate a high-quality text-only dataset into multimodal training, alongside a substantial amount of multimodal math and reasoning data, leading to enhanced math and coding capabilities across modalities.

To advance research in the field, we release the model weights at <https://huggingface.co/nvidia/NVLM-D-72B> and will open-source the training code for the community soon. For more information, please visit our project page at <https://research.nvidia.com/labs/adlr/NVLM-1>.

¹ NVIDIA Vision Language Model.

Contents

1	Introduction	3
2	Qualitative Study	5
3	Preliminaries	5
3.1	Essential Building Blocks	5
3.2	Architectural Designs	6
3.3	High-Resolution Inputs	7
4	NVLM: Models and Training Methods	7
4.1	Shared Vision Pathway	7
4.2	NVLM-D: Decoder-only Model	8
4.3	NVLM-X: X-attention Model	9
4.4	NVLM-H: Hybrid Model	11
4.5	Model Configurations and Training Method	11
5	Training Data	12
5.1	Multimodal Pretraining Data	12
5.2	Multimodal SFT Data	13
5.3	Text-only SFT Data	14
6	Results	14
6.1	Benchmarks	14
6.2	Baseline Models	15
6.3	Main Results	15
6.4	Text-only Performance	16
6.5	Frozen versus Unfrozen LLM during Mutimodal SFT	17
7	Conclusion	18
8	Acknowledgement	18
A	Qualitative Examples from the NVLM-1.0-D 72B Model	27
B	Training Hyperparameters	29
C	Perceiver Resampler in Flamingo Impacts OCR Performance	29
D	Evaluation Details of AI2D	30
E	Data Formats and ChatML Tamplate	30
F	Unoptimized Results Using Yi-34B as the Backbone LLM	32

(a)

User Explain why this meme is funny.

NVLM-1.0

The humor in this image comes from the visual representation of the common phrase "the abstract and the paper." The phrase is often used to describe the difference between the summary of a research paper and the full paper itself. The image shows a large, fierce-looking lynx on the left, labeled "the abstract," and a small, domestic cat on the right, labeled "the paper." The contrast between the two animals is humorous because it visually represents the idea that the summary of a paper (the lynx) can be much more impressive or intimidating than the actual content of the paper (the cat).

(b)

User What is the difference between the left, middle and right object in the image?

NVLM-1.0

The left object in the image is a CPU, the middle object is a GPU, and the right object is a TPU. The main difference between these three objects is their architecture and purpose. CPUs are designed for general-purpose computing and have a few cores optimized for sequential processing. GPUs, on the other hand, have many cores and are designed for parallel processing, making them ideal for tasks such as graphics rendering and deep learning. TPUs are specialized chips designed by Google specifically for deep learning workloads, with a focus on high performance and energy efficiency.

(c)

User You are a helpful driving assistant. In this scene, which lane should I choose and why?

NVLM-1.0

In this scene, you should choose the right lane. The signs indicate that the left two lanes are closed, while the right lane is open for buses and RVs. This suggests that the right lane is the only available option for vehicles, including buses and RVs.

(d)

User Write code based on the provided pseudo code.

NVLM-1.0

The pseudo code in the image describes the selection sort algorithm. The code for the same is as follows:

```
def selection_sort(arr):
    n = len(arr)
    for i in range(n):
        min_index = i
        for j in range(i+1, n):
            if arr[j] < arr[min_index]:
                min_index = j
        arr[i], arr[min_index] = arr[min_index], arr[i]
    return arr
```

(e)

Food	Number of people who ate that food	Number of people who ate the food and got sick
Cold chicken	86	34
Potato salad	54	38
Egg sandwiches	76	40
Fruit pie and cream	32	12
Cheese	48	12

User According to the table, explain which food is the most likely cause of the outbreak of food poisoning?

NVLM-1.0

The table shows the number of people who ate a certain food and the number of people who ate the food and got sick. The food that is the most likely cause of the outbreak of food poisoning is the one with the highest percentage of people who ate it and got sick.

To calculate the percentage, divide the number of people who ate the food and got sick by the total number of people who ate the food, and then multiply by 100.

For example, for cold chicken: $(34/86) \times 100 = 39.5\%$

For potato salad: $(38/54) \times 100 = 70.4\%$

For egg sandwiches: $(40/76) \times 100 = 52.6\%$

For fruit pie and cream: $(12/32) \times 100 = 37.5\%$

For cheese: $(12/48) \times 100 = 25\%$

Therefore, the food that is the most likely cause of the outbreak of food poisoning is potato salad, with 70.4% of people who ate it getting sick.

(f)

User What percentage of market share does NVIDIA have for data center GPUs in 2023?

NVLM-1.0

NVIDIA has 92% of the market share for data center GPUs in 2023.

User Provide short description of middle pie chart.

NVLM-1.0

The middle pie chart shows the market share of leading vendors in the Generative AI market for the year 2023. The chart shows that OpenAI has the largest market share at 39%, followed by Microsoft at 30%. AWS and Google have 8% and 7% market share respectively. The remaining 16% of the market is held by other vendors.

Figure 1: Qualitative examples generated by our NVLM-D_{1.0}72B model. We demonstrate diverse capabilities of our model, including chart and table understanding, OCR, localization, knowledge-grounded image description, humorous meme understanding, scene understanding, math reasoning and coding capabilities. For more examples, refer to Appendix A.

Benchmark	Open-access				Proprietary			
	NVLM-D _{1.0} 72B	*Llama 3-V 70B	*Llama 3-V 405B	InternVL2- Llama3-76B	GPT-4V	GPT-4o	Claude 3.5 Sonnet	Gemini 1.5 Pro (Aug 2024)
MMMU (College-level multidisciplinary VQA)	59.7	60.6	64.5	55.2	56.8	69.1	68.3	62.2
MathVista (Math problems in visual context)	65.2	-	-	65.5	49.9	63.8	67.7	63.9
OCR Bench (A comprehensive OCR evaluation)	853	-	-	839	645	736	788	754
AI2D (Diagram understanding)	94.2	93.0	94.1	94.8	78.2	94.2	94.7	94.4
ChartQA (Chart understanding)	86.0	83.2	85.8	88.4	78.5	85.7	90.8	87.2
DocVQA (Document understanding)	92.6	92.2	92.6	94.1	88.4	92.8	95.2	93.1
TextVQA (Scene text comprehension)	82.1	83.4	84.8	84.4	78.0	-	-	78.7
RealWorldQA (Visual perception)	69.7	-	-	72.2	61.4	-	-	70.4
VQAv2 (Natural image understanding)	85.4	79.1	80.2	-	77.2	-	-	80.2
Text (MMLU, GSMBK, Math, HumanEval)	84.1 <small>+ 4.3 ↗</small>	81.8 <small>0</small>	86.7 <small>0</small>	69.9 <small>- 6.7 ↘</small>	-	-	87.0	82.1

Figure 2: NVLM 1.0 *versus* leading proprietary and open-access multimodal LLMs. Note that the model weights for *Llama 3-V [82] have not been released as of the time of this report. The results demonstrate that NVLM 1.0 achieves performance on par with leading models across both vision-language and text-only tasks. Additionally, we compare multimodal LLM to its backbone LLM on text-only tasks. Llama 3-V 70B and 405B show no degradation in text-only tasks, as their LLM backbones are frozen during multimodal training. In contrast, our NVLM-D_{1.0} 72B model demonstrates significant improvements over its text backbone on text-only math and coding benchmarks, with average accuracy increasing by 4.3 points after multimodal training. See Table 7 and Table 8 for full details.

1 Introduction

Large language models (LLMs) [11] have laid the foundation for the rapid progress in AI recently. Since the introduction of ChatGPT [104], LLMs have revolutionized the text domain and are becoming universal task solvers for natural language processing, math and coding problems. Simultaneously, multimodal LLMs (MLLMs) [4; 107], which bridge the physical world with language models, have gained significant traction. The release of GPT-4V [107] has sparked a competitive race in the development of proprietary multimodal LLMs for vision-language intelligence [35; 108; 5; 6; 153; 154; 122]. However, the model architectures, training data, and methods used to build these proprietary models remain undisclosed, preventing the research community from building upon them.

A notable feature of leading proprietary multimodal LLMs is their exceptional performance on both multimodal and text-only tasks, a quality we refer to as *production-grade multimodality* [108; 35; 36]. For example, GPT-4o is a single neural network trained end-to-end on text and images, achieving state-of-the-art results in both text-only and vision-language tasks [110]. This unified approach simplifies deployment by eliminating the need to route different input modalities to separate LLMs, offering users a seamless experience for switching between modalities without losing text or multimodal context.

The community has made significant progress in advancing the capabilities of open-access multimodal LLMs [26; 79; 18; 71; 139]. Notable families of open models include BLIP [66; 67; 26], LLaVA [79; 78; 80; 65], InternVL [19; 18; 111], and Llama 3-V [82]. The most common architectures used to build these multimodal LLMs are the decoder-only architecture (e.g., LLaVA [79] and InternVL [18]), which processes image tokens within the LLM self-attention layers, and the cross-attention-based architecture (e.g., Flamingo [4] and Llama 3-V [82]), which handles image tokens through LLM cross-attention layers.

However, the previous studies of multimodal LLMs have several limitations:

- In contrast to the convergence of model architectures to build LLM in the text domain [11; 133; 105; 5; 35], i.e., the decoder-only transformer [143], existing multimodal LLM architectures (e.g., decoder-only vs. cross-attention models) have not been studied and compared in an apples-to-apples manner. There is no information regarding the architectures of proprietary models. Furthermore, studies on open-access models differ in their choice of LLM backbones, vision encoders, and, most importantly, training data, making direct comparisons challenging. For these reasons, IDEFICS-80B, an open-access reproduction of Flamingo [62] based on LLaMA-65B [140], is perceived as significantly lagging behind LLaVA-1.5-13B [78], which is based on Vicuna-13B [21], in VQA tasks.
- Model designs that handle high-resolution image input (e.g., dynamic high-resolution [160; 80; 30; 18]) significantly boost performance on OCR-related tasks (e.g., OCRCBench [81]), but sometimes show reduced accuracy on reasoning-related tasks (e.g., MMMU [166]) compared to their low-resolution counterparts.
- Although open-access multimodal LLMs achieve impressive benchmark results on vision-language tasks, we observe a significant degradation in text-only performance (see Table 8), unlike leading proprietary models (e.g., GPT-4o). The only work that provides substantial technical details addressing this issue is Llama 3-V [82], which freezes the LLM parameters and trains only the cross-attention layers. However, these models have not yet been made publicly available.

To address these limitations, we introduce NVLM-1.0, a family of frontier multimodal LLMs (see Figure 2 for a comparison with leading models) featuring three distinct architectures: *i*) NVLM-D, a Decoder-only architecture, *ii*) NVLM-X, a cross (**X**)-attention-based architecture, and *iii*) NVLM-H, a novel **Hybrid** architecture. Trained on the same curated data blend, all three architectures achieve state-of-the-art performance, rivaling leading proprietary and open-access models, while offering practitioners flexible and feature-rich model options. Specifically, **we make the following contributions:**

1. **Model architecture:** We compare the pros and cons of the decoder-only and the cross-attention-based models using the same LLM backbones, vision encoder, and well-curated training data. Our findings show that the cross-attention-based NVLM-X offers superior computational efficiency when handling high-resolution images, whereas the decoder-only NVLM-D provides unified multimodal reasoning and achieves higher accuracy in OCR-related tasks. Building on these insights, we propose NVLM-H, a novel hybrid architecture that excels in multimodal reasoning while also delivering improved computational efficiency for high-resolution images.
2. **High-resolution:** To achieve strong accuracy on both OCR-related tasks (e.g., OCRCBench [81]) and multimodal reasoning tasks (e.g., MMMU [166]), we propose a tile-tagging design for the dynamic tiling of high-resolution image inputs. Through comprehensive ablation studies, we find that adding a text-based 1-D tile tag before the image tokens of the corresponding tile in the decoder achieves the best accuracy.
3. **Training data:** We meticulously collect and provide detailed information on our multimodal pretraining and supervised fine-tuning (SFT) datasets, which will support and benefit future research. In the dataset selection and filtering process, we find that *the data quality and task diversity are more important than the scale, even during the pretraining stage*. Furthermore, previous studies have shown that abundant and diverse multimodal pretraining data is crucial for the success of cross-attention-based models, such as Flamingo [4]. In this work, we found that such pretraining data can also significantly improve the performance of decoder-only models, like LLaVA [78], even with a simplified design that involves training only an MLP projection layer during pretraining. For the curation of SFT data, we collected a much larger set of task-oriented datasets compared to previous studies [18].
4. **Production-grade multimodality:** We develop *production-grade multimodality* for NVLM models, enabling them to excel in both vision-language tasks (e.g., multimodal reasoning, OCR, natural image understanding) and text-only tasks (e.g., multidisciplinary knowledge reasoning, coding, and math). To maintain text-only performance during multimodal training, we investigate two approaches: *i*) For the cross-attention-based NVLM-X, we find that freezing the LLM’s parameters and training only the cross-attention layers [4] during both the pretraining and SFT stages works reasonably well, with a moderate performance trade-off on vision-language tasks.

ii) We curate a high-quality text-only dataset and integrate it into the multimodal SFT stage, effectively preserving text-only performance with no degradation, and even achieving noticeable improvements on text-only math and coding benchmarks after multimodal training across all NVLM models. We attribute this to the superb quality of text-only data and the significant amount of multimodal math data (e.g., geometry) incorporated into multimodal SFT blend, which improves NVLM’s reasoning capabilities, regardless of modality.

We organize the rest of this paper as follows. In § 2, we present a qualitative study of our model’s capabilities, showcasing generated samples. In § 3, we introduce the preliminaries of multimodal LLMs and discuss related work. In § 4, we present the NVLM-1.0 model family, followed by details on the training data in § 5. We introduce the evaluation benchmarks and report results in § 6. We conclude the paper in § 7.

2 Qualitative Study

We conduct a qualitative analysis of NVLM-1.0 with diverse images and instructions. As illustrated in Figure 1, NVLM-1.0 can handle diverse types of images including memes in Figure 1 (a), object-centric images in Figure 1 (b), real-world scene images in Figure 1 (c), hand-written pseudo code in Figure 1 (d), table in Figure 1 (e), and charts in Figure 1 (f).

Our NVLM-D_{1.0}-72B demonstrates versatile capabilities in various multimodal tasks by jointly utilizing OCR, reasoning, localization, common sense, world knowledge, and coding ability. For instance, our model can understand the humor behind the “abstract vs. paper” meme in Figure 1 (a) by performing OCR to recognize the text labels for each image and using reasoning to grasp why juxtaposing “the abstract” — labeled with a fierce-looking lynx — and “the paper” — labeled with a domestic cat — is humorous. NVLM accurately performs localization to effectively answer location-sensitive questions, such as “What is the difference between the left, middle, and right objects in the image?” in Figure 1 (b). NVLM is capable of performing mathematical reasoning and coding based on visual information, such as tables and handwritten pseudocode, as illustrated in Figure 1 (d) and (e). For more examples, refer to Appendix A or our project site: <https://nvlm-project.github.io/>.

3 Preliminaries

Vision language models [120; 9; 149; 4; 146; 174; 17; 145; 169] build the connection between the visual world and open text domain. Among these works, the multimodal LLMs augmented from pretrained large language models (LLMs) [4; 7; 158; 66; 67; 26; 79; 78; 80; 65; 19; 18; 111; 71; 147; 8; 15; 139; 159] have become visual assistants and universal task solvers for various vision-language tasks, including image / video captioning [72; 157], visual understanding and reasoning [166], chart and diagram-related QA [93], math reasoning in visual context [87], and optical character recognition (OCR) [81].

3.1 Essential Building Blocks

Multimodal LLM typically consists of two indispensable components: large language model (LLM) and vision encoder.

Large Language Model A multimodal LLM typically builds upon a text-only LLM for initialization. While there are exceptions where multimodal LLMs are pretrained from scratch using multimodal data [35; 3], these approaches, though conceptually compelling, lack clear evidence of superior performance in vision-language tasks compared to multimodal LLMs built on a text-only LLM.

Instruction-tuned LLMs [150; 23; 115] serve as universal task solvers in the text domain, as they can follow user-provided instructions to address a variety of tasks. As a result, it is common to build multimodal LLMs on instruction-tuned LLMs rather than base LLMs in previous studies, [65; 18; 71; 139; 82], as the instruction-following capability is essential for solving a wide range of vision-language tasks. Various instruction-tuned LLMs have been used to build multimodal LLMs in different study, including Vicuna-1.5 [21], LLaMA-2-Chat [141], Mistral 7B [50], Yi-34B [161], Llama3-Instruct [82], and Qwen2-Instruct [119]. In this work, we use Qwen2-72B-Instruct [119]

as the default text-only LLM backbone. We also employ Nous-Hermes-2-Yi-34B [102] for ablation study and faster experimentation.

Vision Encoder. Multimodal LLMs [e.g., 4; 65; 67; 19] typically leverage pretrained vision encoders (e.g., CLIP [120]) to extract visual features from input images or video frames, with only a very few exceptions [3]. These vision encoders [120; 47; 19; 167; 27] are often trained on large-scale, diverse, and noisy text-image pairs sourced from the web [124; 12; 33]. This allows for large-scale training and enhances the generalization needed to effectively process visual input in unseen domains. The other types of datasets, such as those used for optical character recognition (OCR) [19] and image segmentation [58], are also incorporated to enhance the specific capabilities of vision encoders. In this study, we use InternViT-6B [19] as the default vision encoder due to its strength. We keep this vision encoder frozen at all stages of training, as this simplifies the training process while still delivering strong results.

3.2 Architectural Designs

There are various architectural designs for constructing multimodal LLMs (MLLMs) using existing LLMs and vision encoders [4; 66; 79; 147; 8]. We discuss the two most common architectures.

Decoder-only MLLMs. Decoder-only architectures are popular mainly for their simplicity and unified handling of all modalities by aligning other modality tokens into the text token embedding space. It also facilitates the extension to generating other modalities [35; 108]. The notable examples of decoder-only multimodal LLMs include LLaVA [79; 78; 80; 65], InternVL [19; 18; 111], and Cambrian-1 [139]. In these models, image tokens from the vision encoder are projected into the text-embedding space via a projector module, e.g., position-wise multi-layer perceptron (MLP), and then directly fed into the decoder-only LLM, just like the text tokens. Some variants, such as Qwen-VL [8], utilize more advanced modules, e.g., *Perceiver* [48], to down-sample the image tokens before they are fed into the LLM.

Training decoder-only multimodal LLMs typically involves two stages: *pretraining* and *supervised fine-tuning* (SFT). At the start of pretraining, the randomly initialized MLP or projector module needs to be trained while keeping the LLM frozen to avoid disrupting the LLM’s weights [79; 80]. Related work has also shown cases where both the projector and vision encoder are jointly trained during the pretraining stage [18; 8]. Due to the limited capacity of the MLP or projector module, the LLM need to be unfrozen during multimodal supervised fine-tuning (SFT) to achieve good performance on vision-language tasks [71]. The vision encoder is typically kept frozen during the SFT stage. There are some exceptions, though, where the entire multimodal LLM is trained end-to-end [65], usually with smaller vision encoder [167].

Cross-attention-based MLLMs. Cross-attention-based architectures are similar to encoder-decoder transformer models for machine translation [143], where the text decoder processes flattened image tokens via cross-attention layers, treating them as if they were a foreign language. One of the early successful cross(X)-attention architectures is Flamingo [4], which is built on frozen pretrained LLMs [42] and often serves as the starting point for many studies on this type of model. The Flamingo model has two sets of trainable modules: *i*) a *perceiver resampler* [48] positioned after the frozen vision encoder [120], which is designed to down-sample the vision encoder output to a specified size of representations, and *ii*) the *gated x-attention layers* interleaved with frozen LLM layers, which read output representations from the perceiver resampler. In contrast, our NVLM-1.0-X and the concurrent Llama 3-V [82] models utilize only gated cross-attention layers to process image tokens and do not include the Perceiver module.

The Flamingo model was trained in two stages: 1) pretraining with a large (and possibly noisy) set of image-text pairs or interleaved image-text data, and 2) supervised fine-tuning (SFT) with high-quality data. It always freezes self-attention layers in LLM decoder and only trains cross-attention layers and perceiver during both pretraining and supervised fine-tuning (SFT) to maintain text-only performance. At inference time, the gate of the X -attention layers can be turned ON for multimodal tasks and OFF for text-only tasks. Thanks to the frozen LLM and gated X -attention designs, the text-only performance is guaranteed not to degrade after multimodal training. The follow-up work includes IDEFICS [62] and OpenFlamingo [7], which are open-source reproductions of Flamingo.

In contrast to decoder-only models, cross-attention-based MLLMs are generally considered more complex to implement. This complexity arises from the introduction of additional modules, the need for proper cross-attention masking in interleaved image-text settings, and the significantly heavier pretraining data requirements [4; 62; 176]. However, a notable advantage of the X -attention-based architecture is its computational efficiency, as it does not require unrolling all image tokens in the LLM decoder, which typically results in long sequences during both training and inference, especially for high-resolution images. See §4.3 for further study.

3.3 High-Resolution Inputs

Properly handling high-resolution images is crucial for achieving state-of-the-art results in many OCR-related tasks. However, vision encoders are typically trained with static resolution of 224^2 or 336^2 pixels for efficiency [120; 100], when the image patch size per token is usually 14^2 or 16^2 . For example, feeding a 224^2 image to ViT-L/14 (patch size 14^2) results in $(\frac{224}{14})^2 = 256$ tokens. There are specialized vision encoders that can directly handle static high-resolution images. For instance, the SAM encoder [58], designed for image segmentation, can process images of 1024^2 pixels with a ViT-L/16 backbone (16^2 pixels per patch), producing a 4096-token output. This can be costly, especially when training datasets and downstream tasks contain a mix of low-resolution and high-resolution images.

The dynamic high-resolution mechanism [160; 80; 30; 18] has been proposed to address the waste of compute in such scenarios. For example, given a ViT-L/14 vision encoder trained on low-resolution images (e.g., 224^2), a high-resolution image (e.g., 896×672) is divided into tiles based on the aspect ratio and resolution of the input image ($\frac{896}{224} \times \frac{672}{224} = 12$ tiles in this case). Each tile is independently fed into the ViT-L/14, producing 256 tokens per tile and 3072 tokens in total. Meanwhile, it only produces 512 tokens for an input image with 448×224 resolution. This dynamic approach is particularly well-suited for multimodal LLMs, which need to handle different types of tasks with varying image resolutions.

4 NVLM: Models and Training Methods

In this section, we introduce NVLM-1.0, a family of frontier-class multimodal LLMs featuring three architectures: *i*) Decoder-only NVLM-D, *ii*) Cross (X)-attention based NVLM-X, and *iii*) NVLM-H with Hybrid architecture. Figure 3 illustrates these architectures. We will begin by detailing the vision pathway shared by all NVLM models.

4.1 Shared Vision Pathway

Several studies have compared various vision encoders in multimodal LLMs, suggesting that unfreezing and combining multiple smaller vision encoders offer advantages [139]. In this work, we employ a single, large, and powerful vision encoder, InternViT-6B-448px-V1-5 [113; 18], as the default for all three architectures, keeping it frozen throughout all training stages. It processes images at a fixed resolution of 448^2 , generating 1,024 output tokens.

We use the similar dynamic high-resolution (DHR) approach outlined in Chen et al. [18]. See the left part of Figure 3 for an illustration. We allow a maximum of 6 tiles at training. Thus, the predefined aspect ratios are: {1:1, 1:2, 1:3, 1:4, 1:5, 1:6, 2:1, 2:2, 2:3, 3:1, 3:2, 4:1, 5:1, 6:1}, encompassing all possible combinations of aspect ratios formed by 1 to 6 tiles. For each input image, we dynamically match it to a predefined aspect ratio and divide it into 1 to 6 tiles, each corresponding to 448×448 pixels, based on the image’s resolution. We include a thumbnail tile, which is a scaled-down version of the entire image to capture the global context. Each tile is then fed into InternViT-6B-448px-V1-5 [113], generating 1,024 tokens. We apply a *downsampling* operation to reduce the 1,024 image tokens to 256, reducing the processing overhead for the LLM. This operation groups four neighboring image tokens into one by concatenating them along the channel dimension, a.k.a. pixel shuffle [18]. See Figure 4 for a detailed illustration of this process.

This dynamic high-resolution (DHR) design significantly improves performance on OCR-related tasks [18; 30], but sometimes results in degraded results on reasoning-related tasks [166] when all image tokens from the tiles are simply concatenated and fed directly into the LLM. We will address this issue across the three architectures, respectively.

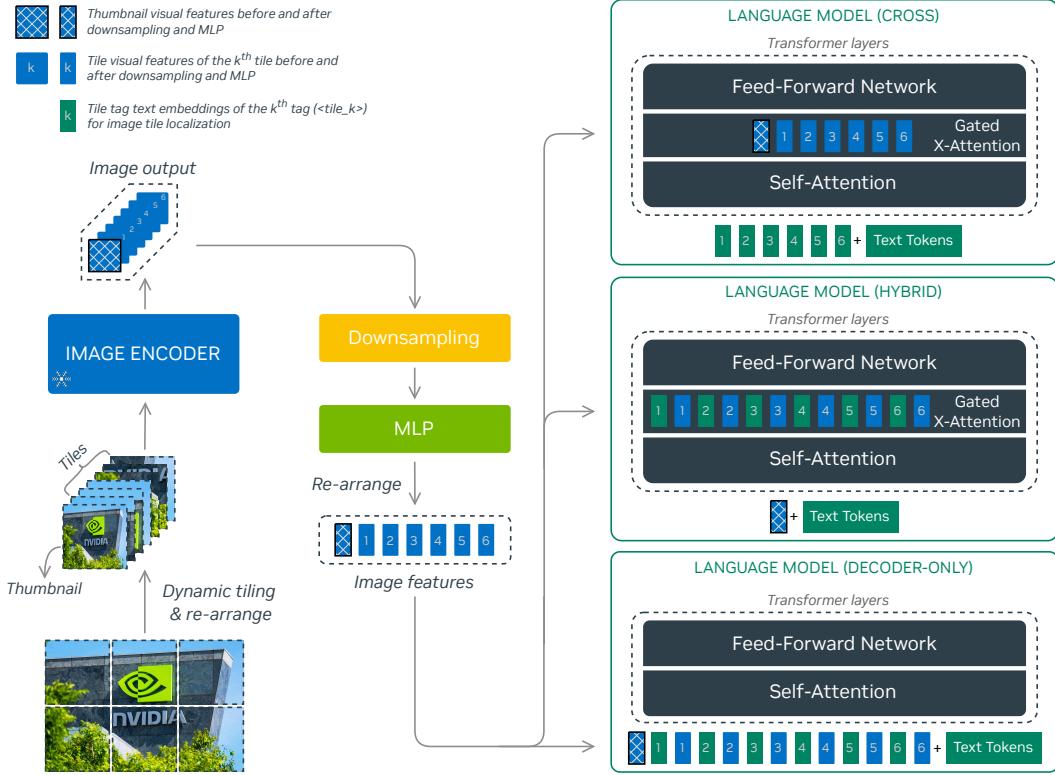


Figure 3: NVLM-1.0 offers three architectural options: the cross-attention-based NVLM-X (top), the hybrid NVLM-H (middle), and the decoder-only NVLM-D (bottom). The dynamic high-resolution vision pathway is shared by all three models. However, different architectures process the image features from thumbnails and regular local tiles in distinct ways.

4.2 NVLM-D: Decoder-only Model

Similar to previous decoder-only multimodal LLMs [79; 18], NVLM-D model connects the pretrained vision encoder to the LLM using a 2-layer MLP as the projector or modality-alignment module.

Training NVLM-D involves two stages: pretraining and supervised fine-tuning (SFT). The MLP is randomly initialized and needs to undergo pretraining first, with both the vision encoder and LLM backbone kept frozen. In our early exploration, we found that joint pretraining of the MLP projector and vision encoder is beneficial when the vision encoder is relatively weak (e.g., ViT-L/14 [100]) and the pretraining datasets are sufficiently diverse. However, after upgrading to the more powerful InternViT-6B-448px-V1-5 [113], the performance gains became marginal. Consequently, we opt to keep the vision encoder frozen during pretraining for the sake of simplicity. During the SFT stage, both the MLP projector and LLM are trained to learn new vision-language tasks with novel instructions, while the vision encoder remains frozen. However, a less frequently discussed point in decoder-only MLLM literature is that leaving the LLM unfrozen during multimodal SFT training often results in significant degradation in text-only performance. Our NVLM-D model effectively maintains text-only performance by incorporating a high-quality text-only SFT dataset. The model configuration and training details for NVLM-D models are in § 4.5.

Tile Tag for Dynamic High-Resolution. As illustrated in Figure 3, the LLM backbone needs to process the flattened image tokens from all dynamic high-resolution tiles, including an additional thumbnail tile. Directly concatenating flattened tokens without delimiters could confuse the LLM, as LLM lacks prior knowledge of the dynamic tiling process. To address this, we insert a text-based tile tag in the input sequence to signal the start of a tile and the position of this tile within the whole tiling structure. After the tile tag, we append the flattened 256 image tokens of the tile. Note that our design differs from previous work [30], which globally flattens the image tokens from different tiles



Figure 4: Dynamic tiling of high-resolution input images. Each tile is encoded separately by InternViT-6B, producing 1,024 tokens, which are downsampled to 256 tokens using a pixel shuffle operation.

Table 1: Ablation study of tile tag formats for dynamic high-resolution (DHR) using the decoder-only **NVLM-D** with Yi-34B as the backbone LLM. All models are trained for 20K iterations with batch size 128 without checkpoint selection to ensure a straightforward comparison.

Tile tag format	MMMU (val)	MathVista	AI2D (test)	ChartQA	DocVQA	TextVQA	OCRBench
Low-resolution (448 ²)	50.9	46.1	67.0	64.8	52.9	78.2	622
DHR + No tag	50.0	51.7	79.9	76.1	80.2	78.4	728
DHR + 2-D grid tag	51.1	52.8	81.7	81.1	86.7	79.4	787
DHR + 2-D bbox tag	50.3	50.6	81.2	80.8	86.7	79.7	791
DHR + 1-D tag	52.0	53.8	82.1	81.1	87.4	79.9	806

and inserts a newline symbol at the end of each row of tokens. We observe improved results with our approach, particularly as we scale up the model size and training data.

We introduce three different tile tags, and perform an ablation study on NVLM-D with Yi-34B [102] as the LLM backbone using the following variants of tile tags:

- a) No tag: Simple concatenation without tile tag, which is the design of InternVL-1.5 [18].
- b) 1-D flattened tile tag: <tile_1>, <tile_2>, ..., <tile_6>, <tile_global>.
- c) 2-D grid tag: <tile_x0_y0>, <tile_x1_y0>, ..., <tile_xW_yH>, <tile_global>, where the {i : j} of <tile_xi_yj> can be in {1:1, 1:2, 1:3, 1:4, 1:5, 1:6, 2:1, 2:2, 2:3, 3:1, 3:2, 4:1, 5:1, 6:1}.
- d) 2-D bounding-box tag: <box>(x₀, y₀), (x₁, y₁)</box>, ..., <box>(x_W, y_H), (x_{W+1}, y_{H+1})</box>, where the (x_i, y_j), (x_{i+1}, y_{j+1}) are the (left, top), (right, bottom) coordinates of that particular title within the whole high-resolution image.

From Table 1, we can observe that: 1) The vanilla dynamic high-resolution method (DHR + No tag) significantly improves performance across all benchmarks, except for MMMU (50.0 vs. 50.9), compared to its low-resolution counterpart. It is worth mentioning that previous DHR methods [30; 18] also exhibit lower MMMU accuracy compared to their low-resolution counterparts. 2) Inserting all types of tile tags into the LLM decoder significantly outperforms simple concatenation with no tags. In particular, we find that the introduction of tile tag greatly improves the performance on OCR-related tasks, including ChartQA [93], DocVQA [81] and OCRBench [81]. 3) 1-D tile tag <tile_k> performs generally better than other tags. We hypothesize that although the 1-D tile tag does not tell 2-D information (e.g., 2×3 vs. 3×2), it offers better generalization at test time. Importantly, this tile tag design for dynamic high-resolution also offers moderate improvement on math and multidisciplinary reasoning tasks, including MathVista [87] and MMMU [166].

4.3 NVLM-X: X-attention Model

NVLM-X employs gated cross-attention to process image tokens and differs from the Flamingo model [4] in two key ways:

- During our initial exploration, we found that while the *perceiver resampler* is beneficial for natural image captioning, it negatively impacts dense OCR tasks, such as transcribing text from scanned documents (see Appendix C for further details). The primary reason is that the cross-attention to latent array in the Perceiver [48] mixes the input image tokens, potentially disrupting the spatial relationships between image patches, which are crucial for document OCR. Based on this observation, our NVLM-X architecture does not use a perceiver resampler; instead, it relies solely on cross-attention to read image tokens directly from the vision encoder.

Table 2: Ablation study of using tile tag $\langle \text{tile_k} \rangle$ for dynamic high-resolution (DHR) using the cross-attention-based NVLM-X with Yi-34B as the backbone LLM. All models are trained for 10K iterations with batch size 512 without checkpoint selection to ensure a straightforward comparison.

Tile-tag format	MMMU (val)	MathVista	AI2D (test)	ChartQA	DocVQA	TextVQA	OCRBench
Low-resolution (448^2)	53.2	57.4	69.8	65.6	62.3	63.2	612
DHR + No tag	53.0	57.6	78.5	73.2	75.4	75.2	682
DHR + 1-D tag	54.1	59.6	80.7	79.2	80.9	78.6	744

Table 3: Training throughput (samples per second) of NVLM-X, NVLM-D, and NVLM-H with Yi-34B as the backbone LLM. We use 128 H100 GPUs during supervised fine-tuning with unfrozen LLMs. All three models are implemented in Megatron-LM with tensor parallelism set to 8 [129]. Sequence lengths in the LLM decoder are set with 1,024 tokens for text, 256 image tokens for the thumbnail tile, and 256×6 for the 6 regular tiles.

Models	Batch size	# of H100 GPUs	Sequence length in LLM decoder	# of Tiles	Elapsed time (ms) per iteration	Throughput samples / sec
NVLM-X 34B	256	128	1,024	6+1	5,063	50.6
NVLM-D 34B	256	128	$1,024 + 256 \times 7 = 2,816$	6+1	8,885	28.8
NVLM-H 34B	256	128	$1,024 + 256 = 1,280$	6+1	7,071	36.2

- Freezing the LLM during the multimodal SFT stage compromises performance on vision-language tasks, as the multimodal LLM needs to quickly adapt to new tasks and novel instructions that were not encountered during text-only instruction tuning. We illustrate this observation in Table 9 in § 6. Thus, we unfreeze the LLM backbone of NVLM-X during multimodal SFT and blend in a high-quality text-only SFT dataset to maintain strong text-only performance. Note that this also differs from Llama 3-V [82], which freezes the LLM during multimodal training.

The model configuration and training details for NVLM-X models can be found in § 4.5.

Tile Tag for Dynamic High-Resolution. NVLM-X uses the same dynamic high-resolution approach as NVLM-D to obtain image tokens from a global thumbnail tile and regular tiles. As illustrated in Figure 3, NVLM-X employs gated X -attention to process the flattened image tokens for each tile, rather than feeding them directly into the LLM decoder. Similar to the design used in NVLM-D, we insert a sequence of text-based tile tags $\langle \text{tile_1} \rangle \dots \langle \text{tile_k} \rangle$ in the LLM decoder, while allowing each tag $\langle \text{tile_k} \rangle$ to only attend to its corresponding image tokens by properly configuring the X -attention mask. This approach ensures that the LLM is better informed about the tiling structure without needing to infer it from the content of thumbnail tile and regular tiles.

In Table 2, we present an ablation study of NVLM-X with Yi-34B LLM backbone using low-resolution 448^2 input, dynamic high-resolution (DHR) without tile tags and with 1-D $\langle \text{tile_k} \rangle$ tags. We find that: 1) The vanilla dynamic high-resolution approach (DHR + No tag) significantly outperforms its low-resolution counterpart across all benchmarks, except MMMU (53.0 vs. 53.2). 2) Adding tile tags further improves performance across all benchmarks, including multimodal reasoning (MMMU: 54.1 vs. 53.0, MathVista: 59.6 vs. 57.6) and OCR-related tasks.

Decoder-only vs. X-attention. The pros and cons of cross-attention-based NVLM-X and decoder-only NVLM-D can be summarized in the following. *i) Parameter efficiency:* NVLM-D has fewer parameters than NVLM-X, as the latter has the newly introduced gated cross-attention layers. The number of additional parameters becomes significant as the model scales up. For instance, Llama 3-V 405B added 100B parameters to the text-only Llama-3.1-405B-Instruct [82]. *ii) Training efficiency:* NVLM-X enables more efficient processing of high-resolution images by eliminating the need to unroll all image tokens on the LLM decoder side. See Table 3 for a comparison of training *throughput* between 34B NVLM-D and NVLM-X models. Note that the decoder-only NVLM-D requires much longer sequence lengths, as all image tokens are concatenated and fed into the LLM decoder, leading to higher GPU memory consumption and lower training throughput. *iii) Multimodal reasoning:* NVLM-D performs unified processing of all tokens from different modalities, enabling joint multimodal reasoning at the LLM decoder. However, the long sequence of tokens for high-resolution images (e.g., $256 \times 7 = 1792$ tokens) may still make reasoning challenging, even with the assistance of tile tags.

4.4 NVLM-H: Hybrid Model

Drawing inspiration from the comparison of NVLM-X and NVLM-D, we propose NVLM-H, a novel hybrid architecture that combines the best of both approaches. As illustrated in Figure 3, NVLM-H separates the processing of image tokens into two paths. The thumbnail image tokens are fed into the LLM alongside text tokens and processed by self-attention layers, enabling joint multimodal reasoning. Simultaneously, a dynamic number of regular tiles are processed through gated cross-attention, enabling the model to capture finer image details. This approach enhances high-resolution capability compared to NVLM-X while significantly improving computational efficiency compared to NVLM-D. Table 3 demonstrates that NVLM-H has higher throughput than NVLM-D in training.

Tile Tag for Dynamic High-Resolution. NVLM-H utilizes the same 1-D flattened tile tag $\langle \text{tile_k} \rangle$ introduced in §4.2 for NVLM-D. The primary distinction lies in the processing location. As shown in Figure 3, text embeddings of $\langle \text{tile_k} \rangle$ are integrated into the gated cross-attention layers alongside visual embeddings. This approach is effective because the text and visual embeddings are well-aligned during pre-training, enabling the model to seamlessly interpret tile tags within the cross-attention mechanism. Consistent with the results in Table 1 and Table 2, adding tile tags enhances NVLM-H’s performance on OCR-related tasks compared to no tagging.

4.5 Model Configurations and Training Method

We provide the model configurations and training details for all NVLM-1.0 models below.

Backbone LLMs and Vision Encoder. For the NVLM-D, NVLM-X, and NVLM-H 72B models, we use Qwen2-72B-Instruct [119] as the backbone LLM. For computational reasons, we also use the smaller Nous-Hermes-2-Yi-34B [102] as the LLM backbone for faster ablation studies and experimentation. After finalizing the optimized designs, we shifted our computational resources to improving the NVLM-1.0 72B models. Across all NVLM models, InternViT-6B-448px-V1-5 [113] serves as the vision encoder.

Modality-Alignment Module. We include the details of modality-alignment modules for three NVLM architectures in the following:

- For NVLM-D models, the LLM and vision encoder are connected by a two-layer MLP to align the modalities, with hidden dimensions of $12800 \rightarrow 20480 \rightarrow 7168$ for 34B model, and $12800 \rightarrow 29568 \rightarrow 8192$ for 72B model. Note that InternViT-6B has a hidden dimension of 3200, which increases to $3200 \times 4 = 12800$ after applying pixel shuffle. Yi-34B [102] has hidden dimension 7168, and Qwen2-72B has hideen dimension 8192.
- For NVLM-X models, the images features are first projected to LLMs’ hidden dimension with a one-layer MLP, $12800 \rightarrow 7168$ for 34B model, and $12800 \rightarrow 8192$ for 72B model. We insert a gated X -attention layer every 6 and 8 LLM self-attention layers, respectively. This results in a total of 10 X -attention layers for both models.
- The NVLM-H 34B and 72B models utilize a two-layer MLP and X -attention layers as the modality-alignment module. The image tokens for both thumbnail and regular tiles are projected through the two-layer MLP, with hidden dimensions of $12800 \rightarrow 20480 \rightarrow 7168$ for the 34B model, and $12800 \rightarrow 29568 \rightarrow 8192$ for the 72B model. The projected thumbnail image tokens are then directly fed into the LLM decoder. The projected image tokens of regular tiles are cross-attended by the X -attention layers. As with NVLM-X, ten gated X -attention layers are inserted for both the 34B and 72B models.

Training Method. We employ a unified training method for all NVLM models. The training process involves two stages: *i) Pretraining*: we freeze both the LLM backbone and vision encoder for all models. We only train the modality-alignment modules, i.e., projector MLP or X -attention layers, using our pretraining dataset detailed in Table 4. For pretraining hyperparameters, one can refer to Table 10 in Appendix B. We find a large batch size of 2048 improves the pretraining with frozen LLMs. *ii) Supervised fine-tuning* (SFT): we keep the vision encoder frozen while training both the LLM and modality-alignment modules with our multimodal SFT datasets detailed in Table 6, along with a text-only SFT dataset. For hyperparameters of SFT, one can refer to Table 11 in Appendix B.

Table 4: Datasets used by NVLM-1.0 at the pretraining stage.

Task	Dataset
Captioning	COCO [72], CC3M [127], SBU [114], LAION-115M (sanitized) [123; 66]
VQA (natural image)	VQAv2 [38], Visual Genome [59]
Chart	DVQA [51]
Document	Docmatix [90]
OCR / Scene-Text	OCR-VQA [98], COCO-Text [144], TextOCR [132], ReCTs [170], RRC-Art [22], RRC-LSVT [134] RCTW [128], synthdog-en [57], pdfa-eng-wds [117]
Math	CLEVR-Math [73]

Table 5: An ablation study comparing the use of our pretraining data in Table 4 with the pretraining data from LLaVA-1.5 [77] using decoder-only NVLM-D with Yi-34B as the backbone LLM. Both models are trained for 20K iterations with batch size 128 without checkpoint selection to ensure a straightforward comparison.

Pretraining data	MMMU (val)	MathVista	AI2D (test)	ChartQA	DocVQA	TextVQA	OCRBench
LLaVA-1.5 data [77]	51.8	48.9	80.5	80.3	85.2	78.9	760
Our pretraining data	52.0	53.8	82.1	81.1	87.4	79.9	806

5 Training Data

In this section, we provide details of the pretraining and supervised fine-tuning (SFT) datasets. These curated training datasets are used across all three architectures in the NVLM family. All datasets are formatted based on the task type and the chat template provided in Appendix E.

5.1 Multimodal Pretraining Data

We curate a diverse, high-quality multimodal pretraining dataset, all sourced from the open-source community. We find that *the quality of the dataset matters more than its scale, even at the pretraining stage*. In early exploration, we experimented with much larger but noisier datasets [e.g., 124; 33], commonly used for training CLIP-style vision encoders. However, we found these unfiltered datasets to be less effective for training both decoder-only and X -attention-based multimodal LLMs, even with a frozen LLM during pretraining. The potential reason could be that noisy text-image data leads to inaccurate alignment between the two modalities. We also experimented with interleaved text-image datasets, including MMC4 [176] and OBELICS [63], and found that they had minimal impact on downstream vision-language tasks, even in few-shot settings, within the state-of-the-art NVLM framework. We hypothesize more careful filtering and recaptioning are needed for such interleaved text-image datasets.

We provide a list of the pretraining datasets in Table 4. These datasets cover different tasks: 1) Captioning. In particular, we use a filtered and recaptioned version of LAION-115M from Li et al. [66]. We perform thorough data scanning and sanitization to ensure the dataset is free of any harmful or inappropriate content. 2) Visual question answering (VQA) on natural image. 3) VQA on chart and scanned document. 4) Math reasoning in a visual context. 5) OCR and scene-text recognition. In addition to large-scale captioning datasets, we find that incorporating large task-oriented datasets during the pretraining stage enhances cross-modal alignment and leads to better final results. We also experimented with blending relatively small task-oriented datasets used in SFT into pretraining. However, this approach caused overfitting on these datasets and impaired the model’s reasoning ability when evaluated on zero-shot benchmarks such as MMMU and MathVista.

Previous work [4] has shown that abundant and diverse pretraining data is crucial for the success of cross-attention-based models. In contrast, decoder-only models, such as LLaVA [78; 80], work well with smaller pretraining datasets [77], which are simply filtered subsets of the captioning datasets including CC3M [127], SBU [114], and LAION-115M [123; 66]. In contrast, our findings demonstrate that the diverse pretraining data shown in Table 4 can still significantly enhance the performance of decoder-only multimodal LLMs, even in state-of-the-art settings with highly curated SFT datasets. We conducted an ablation study comparing our pretraining data with LLaVA-1.5’s pretraining data [77] for the NVLM-D with Yi-34B as LLM backbone, as shown in Table 5. The pretrained models are then fine-tuned on the same high-quality SFT dataset in Table 6. One can see

Table 6: Datasets used by NVLM-1.0 at supervised-fine-tuning (SFT).

Task	Dataset
Captioning	COCO [72], TextCaps [130], ShareGPT-4o [61]
VQA (natural image)	VQAv2 [38], Visual Genome [59], TallyQA [2], Visual7W [177], Vizwiz [39]
General Knowledge	OK-VQA [91], A-OKVQA [125]
Visual Reasoning	GQA [45], Super-CLEVR [69], Raven [168], VSR [74]
Chart & Diagram	DVQA [51], PlotQA [97], MMC-Instruction [76], ChartQA [93], InfographicVQA [96] FigureQA [52], IconQA [84], Chart2Text [103], Diagram Image2Text [53]
Table	WikiTableQuestions [116], RobuT(WTQ), WikiSQL, SQA) [173], HiTab [20]
Document	DocVQA [95], Docmatix [90], DUDE [142], VisualMRC [135], TAT-DQA[175] UReader IE [160], UReader KG [160], UReader QA [160],
OCR / Screen / Scene-Text	OCR-VQA [98], TextVQA [131], ST-VQA [10], ScreenQA [43], SlideQA [136], PDF-VQA [29] VQA-CD [89], VQAonBD [1], POIE [60], SROIE [44], ORAND [28], EST-VQA [148] FUNSD [49], SQuAD(rendering) [121], WordArt [155], IAM [92], IIIT5K [46], HME100K [164] synthdog-en [57], Bentham QA [94], HW-SQuAD [94], WebSight [64], ChromeWriting [152] K12 Printing [65], COCO-Text [144], TextOCR [132], ReCTS [170], pdfa-eng-wds [117]
Math	CLEVR-Math [73], GeoQA+ [13], Geometry3K [83], TabMWP [86], GSM8K(rendering) [25] MetaMathQA(rendering) [162], MAVIS Data Engine [171], MAVIS Manual Collection [171] Geo170K Align [34], Geo170K QA [34], GeoMVerse [54], GEOS [126], UniGeo [14]
Science	AI2D [55], ScienceQA [85], TQA [56], ArXivQA [68], textbook data
Visual Instruction-Tuning	LRV-Instruction [75], LLaVA-158K [79], LLaVAR [172]
Text-only SFT	SlimOrca [70], ShareGPT [138], EvolInstruct [156], GPTeacher [137], AlpacaGPT4 [118], UltraInteract [163], OrcaMathWordProblems [99], MathInstruct [165], MetaMath [162], GlaiveCodeAssistant [37], Magicoder [151], WizardCoder [88].

that our diverse pretraining data provide consistent improvements across all benchmarks, in particular a significant improvement in math reasoning and OCR-related tasks, as we add these types of data in pretraining.

5.2 Multimodal SFT Data

We collected a diverse, high-quality, task-oriented SFT dataset to enhance NVLM’s capabilites on a wide range of visoin-language tasks. A detailed list of SFT datasets is provided in Table 6. In addition to high-quality datasets with short captions, such as COCO [72] and TextCaps [130], we also include ShareGPT-4o [61], which provides detailed image descriptions. Additionally, we have included several VQA datasets based on natural images [38], with a focus on object layout [59], counting [2], object-level grounding [177], mobile phone photo with varying quality [39], visual reasoning [45; 69; 168; 74], and knowledge-based VQA [91; 125]. The ability to understand charts, diagrams, tables, document images is a critical real-world application of multimodal LLMs. To enhance this capability, we have incorporated a diverse set of datasets (e.g., DVQA [51], PlotQA [97], WikiTableQuestions [116], DocVQA [95]). OCR is a fundamental capability of multimodal LLMs, as it is directly related to performance on tasks involving scene text [131], screenshots [43], charts [93], tables [20], document images [95], and handwritten text. As a result, we have incorporated a substantial amount of OCR-related datasets in our SFT blend. Another important capability is mathematical reasoning within a visual context. To enhance this, we have incorporated many multimodal math reasoning datasets listed in Table 6. Interestingly, the abundant multimodal math data not only leads to significant improvements in vision-language tasks like MathVista [87], but also results in substantial gains on text-only math benchmarks, including GSM8K [24] and MATH [41].

Following previous leading open-source work [e.g., 18; 65], we incorporate the training splits of datasets including ChartQA [93], DocVQA [95], VQAv2 [38], TextVQA [131] and AI2D [55] into the SFT blend. Their test sets are used as evaluation benchmarks in Section 6, meaning they are not evaluated in a zero-shot setting. **We want to emphasize that we did not apply data augmentation to the training splits of these benchmark datasets. While such techniques certainly improve benchmark results, they contribute minimal value to the model’s out-of-domain generalization.** Note that it is unknown whether the proprietary multimodal LLMs are being evaluated on these benchmarks in a zero-shot or fine-tuning setting, as no information is provided regarding their training

datasets. We hypothesize that it is a fine-tuning setting, based on observed accuracy gaps between the training and test sets for some proprietary models; however, this is not conclusive.

5.3 Text-only SFT Data

We curated a high-quality text-only SFT dataset and incorporated it into the multimodal fine-tuning stage, effectively preserving the LLM backbone’s text-only performance and preventing catastrophic forgetting. Previous leading open-access multimodal LLMs [18; 65] also include text-only SFT datasets but still show significant performance degradation on text-only benchmarks (see Table 8 for details). The key difference between our recipe and theirs lies in the quality of the data.

Our text-only SFT dataset is built on top of open-source SFT datasets. We collect SFT datasets from general categories, including ShareGPT [21; 138], SlimOrca [70; 101], EvolInstruct [156], GPTeacher [137], AlpacaGPT4 [118], and UltraInteract [163]. Additionally, we collect datasets from math category, including OrcaMathWordProblems [99], MathInstruct [165], MetaMath [162], and from code category, including Magicoder [151], WizardCoder [88], and GlaiveCodeAssistant [37]. After that, we leverage OpenAI models, GPT-4o [108] and GPT-4o-mini [109], to further refine the responses of the prompts from these datasets to enhance the quality of our SFT dataset. Finally, we conduct data decontamination to make sure our dataset does not contain the prompts from all benchmark test datasets.

6 Results

In this section, we present a comprehensive evaluation of the NVLM-1.0 model family across a wide range of benchmarks to assess their multimodal capabilities, comparing them to other leading open-access and proprietary multimodal LLMs. Additionally, we evaluate the NVLM-1.0 models and other top open-access multimodal LLMs on key text-only benchmarks, demonstrating either no degradation or even improvements in the text-only performance of the NVLM-1.0 models, in sharp contrast to the significant degradation observed in other open-access multimodal LLMs.

6.1 Benchmarks

We first introduce the vision-language and text-only benchmarks used in this work. Following previous frontier-class multimodal LLMs [e.g., 108; 6; 111], we evaluate NVLM on nine vision-language benchmarks, focusing on multimodal reasoning, math reasoning in visual context, natural image understanding, scene-text reading, chart understanding, document understanding, real-world perception, and OCR capabilities:

- ◊ MMMU [166] is one of the most popular multimodal reasoning benchmarks, covering multi-disciplinary college-level problems. We do evaluations on both validation and test sets.
- ◊ MathVista [87] is a math reasoning benchmark that covers a variety of mathematical problems, e.g., geometry, function plot, table/chart related arithmetic, in visual contexts. We perform the evaluation on its testmini set.
- ◊ VQAv2 [38] is a natural image understanding benchmark. We evaluate NVLM models on the test-dev set.
- ◊ AI2D [55] is a multimodal reasoning dataset with Grade School Science diagrams. We evaluate the test set using two evaluation settings from VLMEvalKit [31] (see Appendix 10 for examples). In the first setting (“test”), the text in the image is replaced with letter options from the answer choices. In the second setting (“test_no_mask”), the text in the image is replaced with both the letter option and the corresponding value of the answer choices, which we refer to as *no_mask*. Note that the first setting is used as the default metric unless *no_mask* is explicitly stated.
- ◊ TextVQA [131] is a scene-text reading benchmark that includes various text-reading problems from natural images. We conduct evaluation on its validation set.
- ◊ ChartQA [93] is a chart understanding benchmark that involves visual and logical reasoning. We perform evaluation on its test set.
- ◊ DocVQA [95] is dataset for VQA on document images. We do evaluation on its test set.

- ◊ RealWordQA [153] is a benchmark focused on physical world perception and understanding.
- ◊ OCRBench [81] is a comprehensive benchmark created to evaluate the OCR capabilities of multimodal LLMs. It consists of five components: text recognition in images, scene text-centric VQA, document-oriented VQA, key information extraction, and handwritten mathematical expression recognition.

To assess the degradation of text-only performance during multimodal training, we evaluate and compare the multimodal LLMs against their corresponding text-only LLM backbones across four key benchmarks, focusing on multidisciplinary knowledge reasoning, math reasoning, and coding capabilities:

- MMLU [40] is a multidisciplinary benchmark that covers 57 subjects, including elementary mathematics, U.S. history, computer science, law, and more
- GSM8K [24] is a benchmark consisting of grade school math word problems.
- MATH [41] is a math reasoning benchmark that covers math problems ranging across 5 levels of difficulty and 7 sub-disciplines.
- HumanEval [16] is a coding benchmark that measures functional correctness for synthesizing programs from docstrings.

6.2 Baseline Models

We compare our models to leading proprietary and open-access multimodal LLMs. The state-of-the-art (SOTA) proprietary models include GPT-4o [108], Claude 3.5 [6], Gemini Pro 1.5 [36], and Grok-2 [154]. The SOTA open-access models include InternVL-2-Llama3-76B [112], InternVL 2-Pro [111], LLaVA-OneVision 72B [65], Llama 3-V 70B and 405B [82]. Note that the model weights of top-performing InternVL 2-Pro (size unspecified) and Llama 3-V have not yet been made open-access. We optimize and evaluate the following NVLM-1.0 models: *i*) decoder-only NVLM-D_{1.0} 72B, which process image tokens within the LLM decoder, *ii*) cross-attention-based NVLM-X_{1.0} 72B, which handle image tokens through X -attention layers, and *iii*) hybrid NVLM-H_{1.0} 72B, which process global thumbnail image tokens using self-attention layers in the LLM decoder, and regular tile image tokens using X -attention layers.

Given limited computational resources and the goal of building a frontier-class multimodal LLM, we used the smaller 34B models for faster ablation studies and iterations, without focusing on careful checkpoint selection or hyperparameter optimization. We include the unoptimized 34B results in Appendix F for reference purpose. Note that, although our 34B results are not optimized, the 34B NVLM models still significantly outperform other models, including VILA-1.5 40B [71] and Cambrian-1 34B [139].

6.3 Main Results

The main results are presented in Table 7, which includes outcomes from nine vision-language benchmarks and four text-only benchmarks. Our NVLM-1.0 72B models rival the leading proprietary models (e.g., GPT-4o) and open-access models, including LLaMA 3V (not yet publicly available) and InternVL 2. Specifically, the following observations can be drawn from Table 7:

- NVLM-D_{1.0} 72B achieves the highest scores on OCRBench (853) and VQAv2 (85.4) among all leading proprietary and open-access models. Its MMMU score (59.7) also significantly surpasses all leading open-access models at the time of this report’s publication, including LLaVA-OneVision 72B (56.8) [65] and InternVL-2-Llama3-76B (55.2) [112]. On AI2D, TextVQA, ChartQA, and DocVQA, it performs only slightly worse than the best-performing InternVL-2-Llama3-76B, matches very strong GPT-4o [108], and significantly outperforms other leading open-access models, including Cambrian-1 [139] and LLaVA-OneVision 72B [65].
- NVLM-H_{1.0} 72B achieves the highest MMMU (Val) score (60.2) among all multimodal LLMs that are open-access at the time of this report’s publication. It also achieves the best MathVista score (66.6) within NVLM-1.0 family, which already outperforms many very strong models including GPT-4o [108], Gemini Pro 1.5 (Aug 2024) [36], InternVL-2-Pro [111]. This demonstrate its superb multimodal reasoning capability.

Table 7: Evaluation on vision-language and text-only benchmarks. For vision-language benchmarks, all baseline model results are sourced from official reports and the benchmark hosts. For open multimodal LLMs, we list the models that were open-access at the time of this report’s publication and mark * for models not yet open-access. We **highlight** the highest score for each benchmark in both the proprietary and open-access categories. **Text-only Avg.** 4 represents the average accuracy degradation (–) or improvement (+) of the multimodal LLM compared to its backbone LLM on text-only benchmarks after multimodal training, measured across four key benchmarks: MMLU, GSM8K, MATH, and HumanEval (see Table 8 for full results).

Tasks	MMMU test / val	MathVista testmini	VQAv2 test-dev	AI2D test / no_mask	TextVQA val	ChartQA test	DocVQA test	Real- WorldQA	OCR- Bench	Text-only Avg. 4
Proprietary										
GPT-4V [107]	56.1 / 56.8	49.9	77.2	78.2	78.0	78.5	88.4	61.4	645	-
GPT-4-Turbo [106]	- / 63.1	58.1	-	89.4	-	78.1	87.2	-	678	-
GPT-4o [108]	- / 69.1	63.8	-	94.2	-	85.7	92.8	-	736	-
Claude 3 Sonnet [5]	- / 53.1	47.9	-	88.7	-	81.1	89.5	51.9	646	-
Claude 3 Opus [5]	- / 59.4	50.5	-	88.1	-	80.8	89.3	49.8	694	-
Claude 3.5 Sonnet [6]	- / 68.3	67.7	-	94.7	-	90.8	95.2	-	788	-
Gemini Pro 1.0 [35]	- / 47.9	45.2	71.2	73.9	74.6	74.1	88.1	-	659	-
Gemini Ultra 1.0 [35]	- / 59.4	53.0	77.8	79.5	82.3	80.8	90.9	-	-	-
Gemini Pro 1.5 [36]	- / 58.5	52.1	80.2	80.3	73.5	81.3	86.5	67.5	-	-
Gemini Pro 1.5 (Aug 2024)	- / 62.2	63.9	80.2	94.4	78.7	87.2	93.1	70.4	754	-
Grok-1.5V [153]	- / 53.6	52.8	-	88.3	78.1	76.1	85.6	68.7	-	-
Grok-2 [154]	- / 66.1	69.0	-	-	-	-	93.6	-	-	-
<i>Others</i>										
Qwen-VL-MAX	46.8 / 51.4	51.0	78.8	79.3	79.5	79.8	93.1	-	723	-
Adept Fuyu-Heavy [3]	- / 48.3	-	77.8	81.2	-	75.4	-	-	-	-
Open-access										
LLaVA-Next 34B [80]	44.7 / 51.1	46.5	-	-	69.5	-	-	-	574	-
VILA-1.5 40B [71]	46.9 / 51.9	-	84.3	-	-	-	-	-	-	-6.9
Cambrian-1 34B [139]	- / 49.7	53.2	-	79.7	76.7	75.6	75.5	67.8	600	-
LLaVA-OneVision 72B [65]	- / 56.8	67.5	-	85.6	-	83.7	91.3	-	-	-6.3
InternVL-1.2 40B [19]	- / 51.6	47.7	-	79.0	72.5	68.0	57.7	67.5	569	-
InternVL-1.5 26B [18]	- / 45.2	53.5	-	80.7	80.6	83.8	90.9	66.0	724	-
InternVL-2 40B [111]	- / 53.9	63.7	-	87.1	83.0	86.2	93.9	71.8	837	-
InternVL-2-Llama3-76B	- / 55.2	65.5	-	87.6 / 94.8	84.4	88.4	94.1	72.2	839	-6.7
*InternVL-2-Pro [111]	- / 58.9	66.3	-	87.3 / 96.0	-	87.1	95.1	-	837	-
*Llama 3-V 70B [32]	- / 60.6	-	79.1	93.0	83.4	83.2	92.2	-	-	0
*Llama 3-V 405B [32]	- / 64.5	-	80.2	94.1	84.8	85.8	92.6	-	-	0
NVLM-D _{1.0} 72B	54.6 / 59.7	65.2	85.4	85.2 / 94.2	82.1	86.0	92.6	69.7	853	+ 4.3
NVLM-X _{1.0} 72B	53.6 / 57.4	64.6	85.2	84.2 / 93.6	80.2	82.9	82.9	66.1	828	+ 2.5
NVLM-H _{1.0} 72B	53.0 / 60.2	66.6	85.2	83.8 / 93.3	80.3	83.3	83.1	66.0	831	+ 2.7

- NVLM-X_{1.0} 72B also achieves frontier-class results and stands as the best-in-class cross-attention-based multimodal LLMs, rivaling the yet-to-be-released Llama 3-V 70B [82]. One notable advantage of NVLM-X_{1.0} is its significantly faster training and inference speeds compared to its decoder-only counterpart, as demonstrated in Table 3.
- Open-access multimodal LLMs, such as LLaVA-OneVision 72B and InternVL-2-Llama3-76B, show significant performance degradation on text-only tasks after multimodal training. In contrast, our NVLM-1.0 models exhibit even improved text-only performance, thanks to the inclusion of high-quality text-only SFT data. This demonstrates that unfreezing the LLM backbone during multimodal SFT does not compromise text performance, as long as high-quality text alignment data is incorporated.

6.4 Text-only Performance

We present detailed results of text-only performance for our NVLM models, along with leading proprietary and open-access multimodal LLMs, in Table 8. It can be observed that all open-access models experience a significant drop in accuracy compared to their LLM backbones. For instance, VILA-1.5 40B sees a notable decrease of 6.9 points, from 54.8 to 47.9. Similarly, the average accuracy of LLaVA-OneVision 72B and InternVL-2-Llama3-76B drops by 6.3 and 6.9 points, respectively. Llama 3-V experiences no degradation in text-only performance because the LLM is frozen during multimodal training. However, as we will demonstrate in § 6.5, this frozen LLM strategy may lead to an unnecessary trade-off in vision-language performance.

By incorporating high-quality text SFT data, both the NVLM-1.0 72B models achieve higher average accuracy than their respective LLM backbone, Qwen2-72B-Instruct [119]. For example, the average accuracy across four benchmarks increases from 79.8 to 84.1 for the NVLM-D_{1.0} 72B model. It is particularly interesting that the math capabilities of the NVLM-1.0 models improve significantly

Table 8: Evaluation on text benchmarks: MMLU, GSM8K, MATH and HumanEval. For leading proprietary models, information about potential text performance degradation during multimodal training has not been disclosed. The model weights of *LLaMA 3-V had not been released at the time of this report. Text-only Avg. 4 represents the average accuracy degradation (–) or improvement (+) of the multimodal LLM compared to its backbone LLM on text benchmarks after multimodal training.

Tasks	Backbone LLM	MMLU	GSM8K	MATH	HumanEval	Avg. Accuracy	Text-only Avg. 4
Proprietary							
GPT-4o [108]	N/A	88.7	-	76.6	90.2	-	unknown
Gemini Pro 1.5 (Aug 2024) [36]	N/A	85.9	90.8	67.7	84.1	82.1	unknown
Claude 3.5 Sonnet [6]	N/A	88.7	96.4	71.1	92.0	87.0	unknown
Open LLM							
(a) Nous-Hermes-2-Yi-34B [102]	N/A	75.5	78.6	21.8	43.3	54.8	N/A
(b) Qwen2-72B-Instruct [119]	N/A	82.3	91.1	59.7	86.0	79.8	N/A
(c) Llama-3-70B-Instruct [32]	N/A	82.0	93.0	51.0	81.7	76.6	N/A
(d) Llama-3.1-70B-Instruct [32]	N/A	83.6	95.1	68.0	80.5	81.8	N/A
(e) Llama-3.1-405B-Instruct [32]	N/A	87.3	96.8	73.8	89.0	86.7	N/A
Open Multimodal LLM							
VILA-1.5 40B [71]	(a)	73.3	67.5	16.8	34.1	47.9	– 6.9
LLaVA-OneVision 72B [80]	(b)	80.6	89.9	49.2	74.4	73.5	– 6.3
InternVL-2-Llama3-76B [111]	(c)	78.5	87.1	42.5	71.3	69.9	– 6.7
*Llama 3-V 70B [32]	(d)	83.6	95.1	68.0	80.5	81.8	0
*Llama 3-V 405B [32]	(e)	87.3	96.8	73.8	89.0	86.7	0
NVLM-D 1.0 72B	(b)	82.0	92.9	73.1	88.4	84.1	+ 4.3
NVLM-X 1.0 72B	(b)	81.4	91.8	70.6	85.2	82.3	+ 2.5
NVLM-H 1.0 72B	(b)	80.4	91.5	71.4	86.6	82.5	+ 2.7

Table 9: Impact on vision-language performance with a frozen vs. unfrozen LLM backbone during multimodal SFT for the cross-attention model NVLM-X.

Tasks	MMMU test / val	MathVista testmini	VQAv2 test-dev	AI2D test	TextVQA val	CharQA test	DocVQA test	RealWorld-QA	OCR-Bench
NVLM-X 34B (frozen)	43.2 / 51.6	51.8	83.8	72.4	72.4	74.4	73.2	63.4	696
NVLM-X 34B	47.2 / 54.0	59.2	84.5	79.6	78.2	79.4	79.2	64.8	802
NVLM-X 72B (frozen)	50.6 / 54.4	60.6	85.3	76.2	76.2	76.2	76.4	65.3	722
NVLM-X 1.0 72B	53.6 / 57.4	64.6	85.2	84.2	80.2	82.9	82.9	66.1	828

compared to its text-only backbone. We attribute this to the high-quality text-only SFT data and the substantial amount of multimodal math data included in our training blend, which enhances math reasoning skills overall, regardless of the modality.

6.5 Frozen versus Unfrozen LLM during Mutimodal SFT

In this subsection, we compare two methods for maintaining text-only performance in the cross-attention-based NVLM-X: *i*) Freezing the LLM during multimodal SFT training, which ensures no degradation in text performance due to the gated X -attention layers, and *ii*) our default approach, which incorporates a high-quality text-only dataset during multimodal SFT training. It is important to note that freezing the LLM for decoder-only multimodal model during SFT leads to poor results on vision-language tasks (as demonstrated in a similar study by [71]), due to the very limited capacity of the MLP projector module.

In Table 9, it can be seen that freezing the LLM yields reasonably good results. Notably, accuracy scales well as the model size increases from 34B to 72B, reaffirming findings from the original Flamingo study, which also froze the LLM during multimodal training. However, compared to the unfrozen setting, freezing the LLM still results in a moderate performance drop on vision-language tasks. For example, NVLM-X_{1.0} 72B with a frozen LLM performs only comparably to the smaller NVLM-X-34B, showing slight improvements on multimodal reasoning tasks (MMMU and MathVista) and natural image understanding tasks (VQAv2 and RealWorldQA), while falling behind on OCR-related tasks (AI2D, TextVQA, ChartQA, DocVQA, and OCRBench).

7 Conclusion

We introduce NVLM-1.0, a family of frontier multimodal large language models that achieve state-of-the-art results on vision-language tasks, rivaling leading multimodal LLMs, without compromising text-only performance during multimodal training. Furthermore, we provide key insights on architecture design, tile-based dynamic high-resolution input, multimodal training data curation, and how to achieve production-grade multimodality with even improved text-only performance after multimodal training. We open-source the model weights and training code to the community. Please refer to the updated information at <https://nvlm-project.github.io/>.

8 Acknowledgement

We would like to express our sincere gratitude to Jared Casper, Peter Dykas, and Mike Chrzanowski for their support with Megatron-LM. Our thanks also go to Lukas Voegtle and Philipp Fischer for the implementation and debugging efforts with their Megatron-Energon dataloader, which is released early. We are also grateful to Hongxu Yin, Timo Roman, Karan Sapra and Valerii Iakovlev for their help and support.

References

- [1] VQAonBD 2023 Dataset, 2023. URL <https://ilocr.iit.ac.in/vqabd/dataset.html>.
- [2] Acharya, M., Kafle, K., and Kanan, C. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, 2019.
- [3] Adept. Adept Fuyu-Heavy: A new multimodal model, 2024.
- [4] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, 2022.
- [5] Anthropic. Introducing the next generation of Claude, 2024.
- [6] Anthropic. Claude 3.5 Sonnet, 2024.
- [7] Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P. W., Ilharco, G., Wortsman, M., and Schmidt, L. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.
- [8] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [9] Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [10] Biten, A. F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., and Karatzas, D. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4291–4301, 2019.
- [11] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *NeurIPS*, 2020.
- [12] Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., and Kim, S. Coyo-700m: Image-text pair dataset. <https://github.com/kakaobrain/coyo-dataset>, 2022.
- [13] Cao, J. and Xiao, J. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1511–1520, 2022.
- [14] Chen, J., Li, T., Qin, J., Lu, P., Lin, L., Chen, C., and Liang, X. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022.
- [15] Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., and Elhoseiny, M. MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

- [16] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- [17] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.
- [18] Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al. How far are we to GTP-4V? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [19] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.
- [20] Cheng, Z., Dong, H., Wang, Z., Jia, R., Guo, J., Gao, Y., Han, S., Lou, J.-G., and Zhang, D. Hitab: A hierarchical table dataset for question answering and natural language generation. *arXiv preprint arXiv:2108.06712*, 2021.
- [21] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- [22] Chng, C. K., Liu, Y., Sun, Y., Ng, C. C., Luo, C., Ni, Z., Fang, C., Zhang, S., Han, J., Ding, E., et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1571–1576. IEEE, 2019.
- [23] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 2024.
- [24] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [25] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>, 2021.
- [26] Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, 2023.
- [27] Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pp. 7480–7512. PMLR, 2023.
- [28] Diem, M., Fiel, S., Kleber, F., Sablatnig, R., Saavedra, J. M., Contreras, D., Barrios, J. M., and Oliveira, L. S. Icfhr 2014 competition on handwritten digit string recognition in challenging datasets (hdsr 2014). In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pp. 779–784. IEEE, 2014.
- [29] Ding, Y., Luo, S., Chung, H., and Han, S. C. Pdf-vqa: A new dataset for real-world vqa on pdf documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 585–601. Springer, 2023.
- [30] Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Zhang, S., Duan, H., Zhang, W., Li, Y., et al. InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD. *arXiv preprint arXiv:2404.06512*, 2024.
- [31] Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., Lin, D., and Chen, K. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models, 2024. URL <https://arxiv.org/abs/2407.11691>.
- [32] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

- [33] Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.
- [34] Gao, J., Pi, R., Zhang, J., Ye, J., Zhong, W., Wang, Y., Hong, L., Han, J., Xu, H., Li, Z., et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.
- [35] Gemini-Team. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [36] Gemini-Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- [37] Glaive-AI. GlaiveCodeAssistant, 2023. URL <https://huggingface.co/datasets/glaiveai/glaive-code-assistant-v2>.
- [38] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.
- [39] Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.
- [40] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- [41] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- [42] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [43] Hsiao, Y.-C., Zubach, F., Wang, M., and Chen, J. ScreenQA: Large-Scale Question-Answer Pairs over Mobile App Screenshots, 2024.
- [44] Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., and Jawahar, C. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520. IEEE, 2019.
- [45] Hudson, D. A. and Manning, C. D. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.
- [46] IIIT. URL <http://cvit.iiit.ac.in/projects/SceneTextUnderstanding/IIIT5K.html>.
- [47] Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. OpenCLIP, 2021.
- [48] Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. Perceiver: General perception with iterative attention. In *ICML*, 2021.
- [49] Jaume, G., Ekenel, H. K., and Thiran, J.-P. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pp. 1–6. IEEE, 2019.
- [50] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- [51] Kafle, K., Price, B., Cohen, S., and Kanan, C. DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5648–5656, 2018.
- [52] Kahou, S. E., Michalski, V., Atkinson, A., Kádár, Á., Trischler, A., and Bengio, Y. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.
- [53] Kamizuru, Y., 2023. URL https://huggingface.co/datasets/Kamizuru00/diagram_image_to_text.

- [54] Kazemi, M., Alvari, H., Anand, A., Wu, J., Chen, X., and Soricut, R. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023.
- [55] Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., and Farhadi, A. A diagram is worth a dozen images. In *ECCV*, 2016.
- [56] Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., and Hajishirzi, H. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, 2017.
- [57] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022.
- [58] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.
- [59] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- [60] Kuang, J., Hua, W., Liang, D., Yang, M., Jiang, D., Ren, B., and Bai, X. Visual information extraction in the wild: practical dataset and end-to-end solution. In *International Conference on Document Analysis and Recognition*, 2023.
- [61] Laboratory, S. A. Sharegpt-4o, 2024. URL <https://sharegpt4o.github.io/>.
- [62] Laurençon, H., van Strien, D., Bekman, S., Tronchon, L., Saulnier, L., Wang, T., Karamcheti, S., Singh, A., Pistilli, G., Jernite, Y., et al. Introducing idefics: An open reproduction of state-of-the-art visual language model, 2023.
- [63] Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A., Kiela, D., et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 2024.
- [64] Laurençon, H., Tronchon, L., and Sanh, V. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029*, 2024.
- [65] Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., and Li, C. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [66] Li, J., Li, D., Xiong, C., and Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, 2022.
- [67] Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.
- [68] Li, L., Wang, Y., Xu, R., Wang, P., Feng, X., Kong, L., and Liu, Q. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024.
- [69] Li, Z., Wang, X., Stengel-Eskin, E., Kortylewski, A., Ma, W., Van Durme, B., and Yuille, A. L. Super-CLEVR: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14963–14973, 2023.
- [70] Lian, W., Wang, G., Goodson, B., Pentland, E., Cook, A., Vong, C., and "Teknium". Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification, 2023. URL <https://huggingface.co/Open-Orca/SlimOrca>.
- [71] Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., and Han, S. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024.
- [72] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV 2014*, 2014.
- [73] Lindström, A. D. and Abraham, S. S. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.

- [74] Liu, F., Emerson, G., and Collier, N. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.
- [75] Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.
- [76] Liu, F., Wang, X., Yao, W., Chen, J., Song, K., Cho, S., Yacoob, Y., and Yu, D. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023.
- [77] Liu, H. Llava-pretrain. <https://huggingface.co/datasets/liuhaojian/LLaVA-Pretrain>, 2023.
- [78] Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2023.
- [79] Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023.
- [80] Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-v1.github.io/blog/2024-01-30-llava-next/>.
- [81] Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X., lin Liu, C., Jin, L., and Bai, X. Ocrbench: On the hidden mystery of ocr in large multimodal models, 2024. URL <https://arxiv.org/abs/2305.07895>.
- [82] Llama-Team. The Llama 3 Herd of Models. *arXiv preprint arXiv:2302.04858*, 2023.
- [83] Lu, P., Gong, R., Jiang, S., Qiu, L., Huang, S., Liang, X., and Zhu, S.-C. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.
- [84] Lu, P., Qiu, L., Chen, J., Xia, T., Zhao, Y., Zhang, W., Yu, Z., Liang, X., and Zhu, S.-C. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.
- [85] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- [86] Lu, P., Qiu, L., Chang, K.-W., Wu, Y. N., Zhu, S.-C., Rajpurohit, T., Clark, P., and Kalyan, A. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022.
- [87] Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.
- [88] Luo, Z., Xu, C., Zhao, P., Sun, Q., Geng, X., Hu, W., Tao, C., Ma, J., Lin, Q., and Jiang, D. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*, 2023.
- [89] Mahamoud, I. S., Coustaty, M., Joseph, A., d’Andecy, V. P., and Ogier, J.-M. Qalayout: Question answering layout based on multimodal attention for visual question answering on corporate document. In *International Workshop on Document Analysis Systems*, pp. 659–673. Springer, 2022.
- [90] Marafioti, A. and Laurendon, H. Docmatix - a huge dataset for document visual question answering, 2024.
- [91] Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.
- [92] Marti, U.-V. and Bunke, H. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5:39–46, 2002.
- [93] Masry, A., Long, D. X., Tan, J. Q., Joty, S., and Hoque, E. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.
- [94] Mathew, M., Gomez, L., Karatzas, D., and Jawahar, C. Asking questions on handwritten document collections. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(3):235–249, 2021.

- [95] Mathew, M., Karatzas, D., and Jawahar, C. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.
- [96] Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., and Jawahar, C. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.
- [97] Methani, N., Ganguly, P., Khapra, M. M., and Kumar, P. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1527–1536, 2020.
- [98] Mishra, A., Shekhar, S., Singh, A. K., and Chakraborty, A. OCR-VQA: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.
- [99] Mitra, A., Khanpour, H., Rosset, C., and Awadallah, A. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*, 2024.
- [100] mlfoundations. An open source implementation of CLIP, 2021. URL https://github.com/mlfoundations/open_clip.
- [101] Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., and Awadallah, A. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.
- [102] Nous. Nous-Hermes-2-Yi-34B, 2024.
- [103] Obeid, J. and Hoque, E. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. *arXiv preprint arXiv:2010.09142*, 2020.
- [104] OpenAI. Introducing ChatGPT, 2022.
- [105] OpenAI. GPT-4, 2023.
- [106] OpenAI. New models and developer products announced at DevDay, 2023.
- [107] OpenAI. GPT-4V(ision) System Card, 2023.
- [108] OpenAI. Hello GPT-4o, 2024.
- [109] OpenAI. GPT-4o mini: advancing cost-efficient intelligence, 2024.
- [110] OpenAI. GPT-4o System Card, 2024.
- [111] OpenGVLab. InternVL2: Better than the Best – Expanding Performance Boundaries of Open-Source Multimodal Models with the Progressive Scaling Strategy, 2024.
- [112] OpenGVLab. InternVL2-Llama3-76B, 2024. URL <https://huggingface.co/OpenGVLab/InternVL2-Llama3-76B>.
- [113] OpenGVLab. InternViT-6B-448px-V1-5, 2024. URL <https://huggingface.co/OpenGVLab/InternViT-6B-448px-V1-5>.
- [114] Ordonez, V., Kulkarni, G., and Berg, T. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.
- [115] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 2022.
- [116] Pasupat, P. and Liang, P. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015.
- [117] PDF-Association, 2024. URL <https://huggingface.co/datasets/pixparse/pdfa-eng-wds>.
- [118] Peng, B., Li, C., He, P., Galley, M., and Gao, J. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.
- [119] Qwen-Team. Qwen2 technical report, 2024.

- [120] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- [121] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
- [122] Reka. Reka Core: Our Frontier Class Multimodal Language Model, 2024.
- [123] Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [124] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.
- [125] Schwenk, D., Khandelwal, A., Clark, C., Marino, K., and Mottaghi, R. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.
- [126] Seo, M., Hajishirzi, H., Farhadi, A., Etzioni, O., and Malcolm, C. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1466–1476, 2015.
- [127] Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.
- [128] Shi, B., Yao, C., Liao, M., Yang, M., Xu, P., Cui, L., Belongie, S., Lu, S., and Bai, X. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th iapr international conference on document analysis and recognition (ICDAR)*, volume 1, pp. 1429–1434. IEEE, 2017.
- [129] Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019. URL <https://github.com/NVIDIA/Megatron-LM>.
- [130] Sidorov, O., Hu, R., Rohrbach, M., and Singh, A. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, 2020.
- [131] Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards VQA models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.
- [132] Singh, A., Pang, G., Toh, M., Huang, J., Galuba, W., and Hassner, T. TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *The Conference on Computer Vision and Pattern Recognition*, 2021.
- [133] Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.
- [134] Sun, Y., Ni, Z., Chng, C.-K., Liu, Y., Luo, C., Ng, C. C., Han, J., Ding, E., Liu, J., Karatzas, D., et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1557–1562. IEEE, 2019.
- [135] Tanaka, R., Nishida, K., and Yoshida, S. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.
- [136] Tanaka, R., Nishida, K., Nishida, K., Hasegawa, T., Saito, I., and Saito, K. SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images. In *AAAI*, 2023.
- [137] Teknium. GPTTeacher-General-Instruct, 2023. URL <https://huggingface.co/datasets/teknium/GPTTeacher-General-Instruct>.
- [138] The-Vicuna-Team. ShareGPT-Vicuna, 2023. URL https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered.
- [139] Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S. C., Yang, J., Yang, S., Iyer, A., Pan, X., et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

- [140] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [141] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [142] Van Landeghem, J., Tito, R., Borchmann, Ł., Pietruszka, M., Joziak, P., Powalski, R., Jurkiewicz, D., Coustaty, M., Anckaert, B., Valveny, E., et al. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19528–19540, 2023.
- [143] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [144] Veit, A., Matera, T., Neumann, L., Matas, J., and Belongie, S. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.
- [145] Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.
- [146] Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.
- [147] Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al. Cogvilm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.
- [148] Wang, X., Liu, Y., Shen, C., Ng, C. C., Luo, C., Jin, L., Chan, C. S., Hengel, A. v. d., and Wang, L. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10126–10135, 2020.
- [149] Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. SimVLM: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022.
- [150] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.
- [151] Wei, Y., Wang, Z., Liu, J., Ding, Y., and Zhang, L. Magicoder: Empowering code generation with oss-instruct. In *Forty-first International Conference on Machine Learning*, 2024.
- [152] Wendler, C., 2023. URL <https://huggingface.co/datasets/wendlerc/RenderedText>.
- [153] xAI. Grok-1.5 vision preview: Connecting the digital and physical worlds with our first multimodal model, 2024.
- [154] xAI. Grok-2 beta release, 2024.
- [155] Xie, X., Fu, L., Zhang, Z., Wang, Z., and Bai, X. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *European conference on computer vision*, pp. 303–321. Springer, 2022.
- [156] Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., Lin, Q., and Jiang, D. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=CfXh93NDgH>.
- [157] Xu, J., Mei, T., Yao, T., and Rui, Y. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.
- [158] Yang, Z., Ping, W., Liu, Z., Korthikanti, V., Nie, W., Huang, D.-A., Fan, L., Yu, Z., Lan, S., Li, B., et al. Re-ViLM: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*, 2023.
- [159] Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

- [160] Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Xu, G., Li, C., Tian, J., Qian, Q., Zhang, J., et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.
- [161] Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.
- [162] Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., and Liu, W. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.
- [163] Yuan, L., Cui, G., Wang, H., Ding, N., Wang, X., Deng, J., Shan, B., Chen, H., Xie, R., Lin, Y., et al. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*, 2024.
- [164] Yuan, Y., Liu, X., Dikubab, W., Liu, H., Ji, Z., Wu, Z., and Bai, X. Syntax-aware network for handwritten mathematical expression recognition. *arXiv preprint arXiv:2203.01601*, 2022.
- [165] Yue, X., Qu, X., Zhang, G., Fu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.
- [166] Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.
- [167] Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.
- [168] Zhang, C., Gao, F., Jia, B., Zhu, Y., and Zhu, S.-C. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5317–5327, 2019.
- [169] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5579–5588, 2021.
- [170] Zhang, R., Zhou, Y., Jiang, Q., Song, Q., Li, N., Zhou, K., Wang, L., Wang, D., Liao, M., Yang, M., et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 1577–1581. IEEE, 2019.
- [171] Zhang, R., Wei, X., Jiang, D., Zhang, Y., Guo, Z., Tong, C., Liu, J., Zhou, A., Wei, B., Zhang, S., et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024.
- [172] Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., and Sun, T. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.
- [173] Zhao, Y., Zhao, C., Nan, L., Qi, Z., Zhang, W., Tang, X., Mi, B., and Radev, D. RobuT: A systematic study of table QA robustness against human-annotated adversarial perturbations. *arXiv preprint arXiv:2306.14321*, 2023.
- [174] Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [175] Zhu, F., Lei, W., Feng, F., Wang, C., Zhang, H., and Chua, T.-S. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4857–4866, 2022.
- [176] Zhu, W., Hessel, J., Awadalla, A., Gadre, S. Y., Dodge, J., Fang, A., Yu, Y., Schmidt, L., Wang, W. Y., and Choi, Y. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.
- [177] Zhu, Y., Groth, O., Bernstein, M., and Fei-Fei, L. Visual7W: Grounded question answering in images. In *CVPR*, 2016.

Appendix

A Qualitative Examples from the NVLM-1.0-D 72B Model

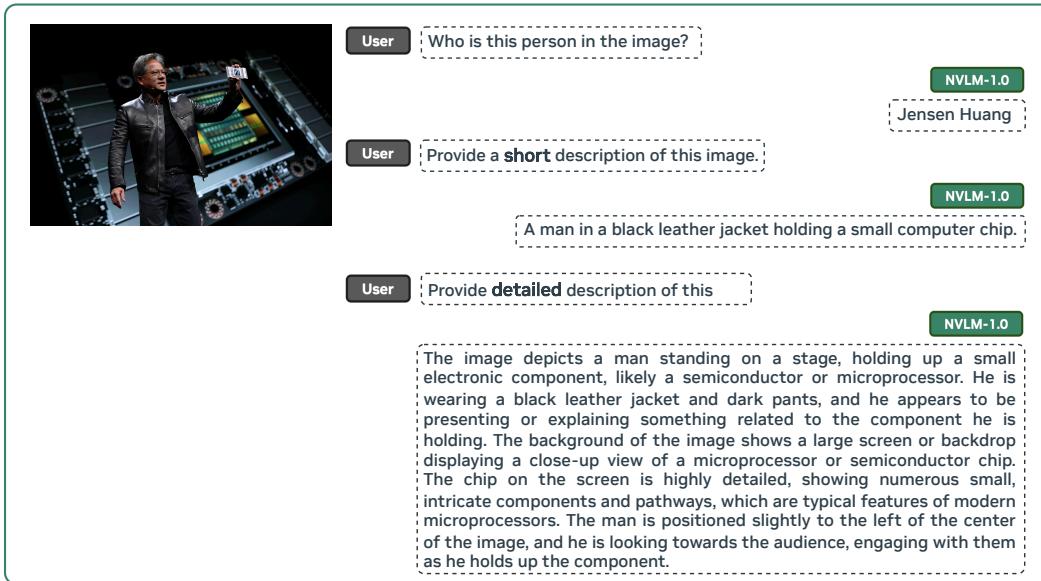


Figure 5: NVLM-1.0-D 72B model demonstrates good instruction-following capability. Based on the instructions, it appropriately controls the target generation's length. It can also generate a very high-quality, detailed description of the provided image.

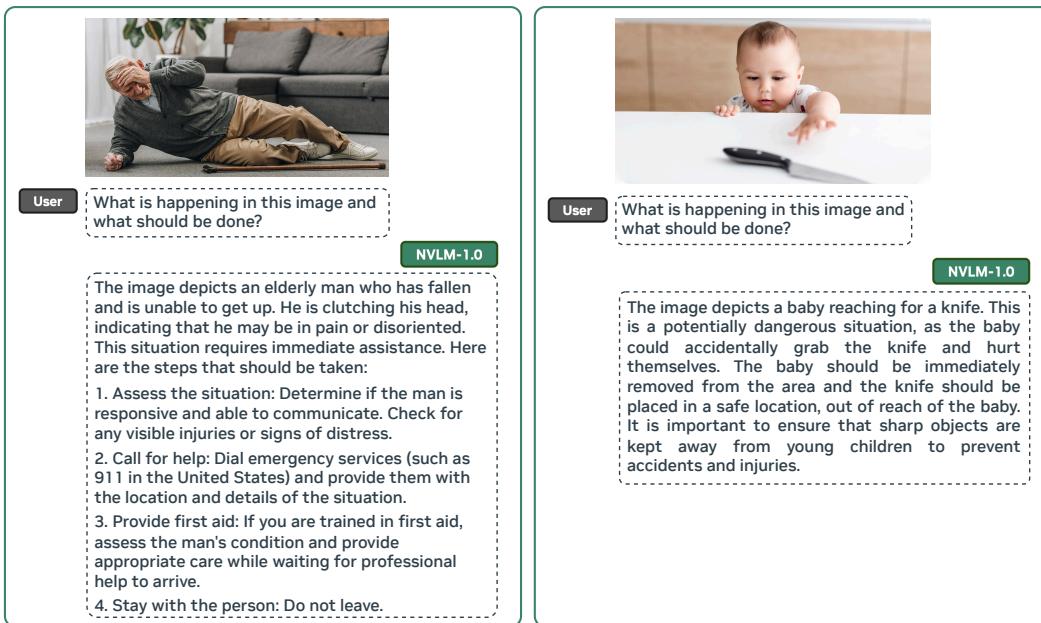


Figure 6: Demonstration of NVLM-1.0-D 72B model's strong scene understanding capability. It has common-sense knowledge to recognize potential dangers or accidents and correctly suggests what immediate actions should be taken.

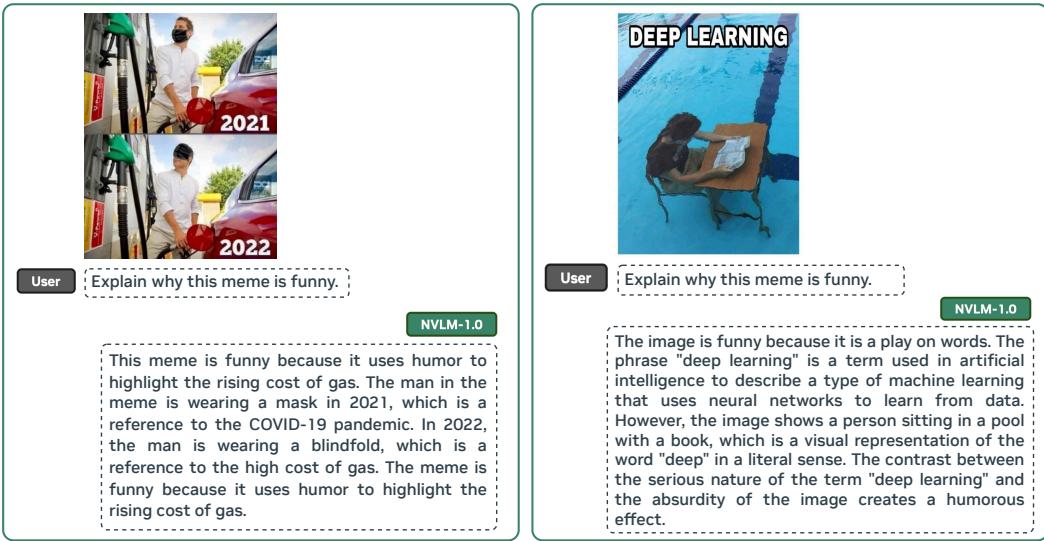


Figure 7: More examples of NVLM-1.0-D 72B model’s ability to understand memes, which is a challenging task that requires an understanding of humor and knowledge of important social trends, context, or events.

A user asks for help solving a math problem: "A square is inscribed in a circle of area 18π square units. Find the length of a side of the square." The model provides a step-by-step solution:

- The area of the circle is given by the formula $A = \pi r^2$, where r is the radius of the circle.
- Given that the area of the circle is 18π square units, we can set up the equation $\pi r^2 = 18\pi$.
- Dividing both sides of the equation by π , we get $r^2 = 18$.
- Taking the square root of both sides, we get $r = \sqrt{18}$.
- Since the square is inscribed in the circle, the diagonal of the square is equal to the diameter of the circle.
- The diameter of the circle is twice the radius, so the diagonal of the square is $2r = 2\sqrt{18}$.
- Using the Pythagorean theorem, we can find the length of a side of the square.
- Let s be the length of a side of the square. Then, by the Pythagorean theorem, we have $s^2 + s^2 = (2\sqrt{18})^2$.
- Simplifying, we get $2s^2 = 72$.
- Dividing both sides by 2, we get $s^2 = 36$.
- Taking the square root of both sides, we get $s = 6$.
- Therefore, the length of a side of the square is 6 units. The answer is: 6

Figure 8: NVLM-1.0-D 72B model can solve math questions by providing step-by-step mathematical reasoning. We render latex equations for readability.

B Training Hyperparameters

We provide the pretraining hyperparameters in Table 10 and the SFT hyperparameters in Table 11.

Table 10: Training hyper-parameters of NVLM models in the pretraining stage.

Hyper-parameters	NVLM-D	NVLM-X	NVLM-H
Trainable weights	MLP	X-attn layers	X-attn layers & MLP
# of gated cross-attention layers	N/A	10	10
Global batch size	2,048	2048	2,048
Max learning rate	1e-4	1e-4	1e-4
Min learning rate	2.5e-5	1e-5	1e-5
Learning rate warmup steps	500	1,000	1,000
Scheduler	cosine	cosine	cosine
Optimizer	AdamW	AdamW	AdamW
Optimizer config	$\beta_1 = 0.9, \beta_2 = 0.95$	$\beta_1 = 0.9, \beta_2 = 0.95$	$\beta_1 = 0.9, \beta_2 = 0.98$
Weight decay	0.1	0.05	0.05
Gradient clipping	10	1.0	1.0
Sequence length in the LLM decoder	512	512	512
Downsampling of visual tokens	1024->256	1024->256	1024->256
# of visual token per tile	256	256	256
# of tiles	1	1	6+1
Tensor parallelism	8	8	8
Pipeline parallelism	1	1	1
# of training steps	20K	20K	20K

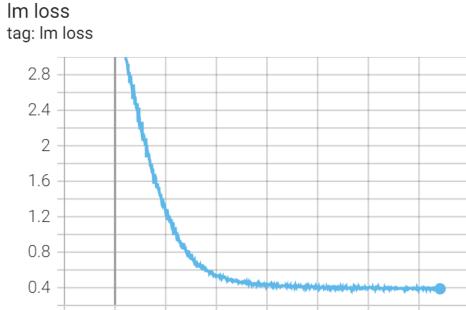
Table 11: Training hyper-parameters of NVLM models in the SFT stage.

Hyper-parameters	NVLM-D	NVLM-X	NVLM-H
Trainable weights	MLP & LLM	X-attn layers & LLM	X-attn layers & MLP & LLM
# of gated cross-attention layers	N/A	10	10
Global batch size	128	512 (34B), 256 (72B)	256
Max learning rate	$2e^{-6}$	$1e^{-5}$	$1e^{-5}$
Min learning rate	$2.5e^{-7}$	$1e^{-6}$ (34B), $1e^{-7}$ (72B)	$1e^{-7}$
Learning rate warmup steps	1,000	500	1,000
Scheduler	cosine	cosine	cosine
Optimizer	AdamW	AdamW	AdamW
Optimizer config	$\beta_1 = 0.9, \beta_2 = 0.98$	$\beta_1 = 0.9, \beta_2 = 0.98$	$\beta_1 = 0.9, \beta_2 = 0.98$
Weight decay	0.1	0.05	0.05
Gradient clipping	10	1.0	1.0
Sequence length in the LLM decoder	3,200	1,024	1,280
Downsampling of visual tokens	1024->256	1024->256	1024->256
# of visual token per tile	256	256	256
# of tiles	6+1	6+1	6+1
Tensor parallelism	8	8	8
Pipeline parallelism	4	1	1
# of training steps	40K	20K	40K

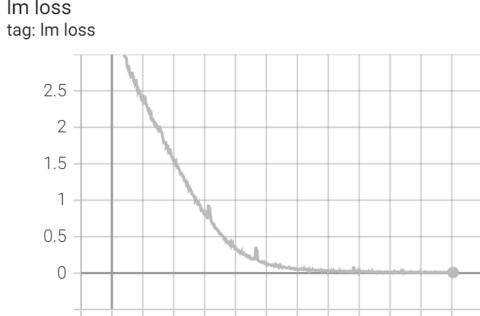
C Perceiver Resampler in Flamingo Impacts OCR Performance

In this study, we utilize a pretrained Flamingo model [4] from Yang et al. [158], built on a 1.3B LLM, and fine-tune it on an internal document OCR dataset consisting of 30K samples. In Figure 9, we observe that the original Flamingo model, incorporating the *perceiver resampler*, struggles to overfit

this OCR dataset. For instance, even after numerous epochs, the training loss remains around 0.4. However, when we remove the *perceiver resampler* and only train the cross-attention layer, the loss decreases to 0 at the same iteration. We hypothesize that the X -attention operation to the *latent array* in Perceiver [48] may shuffle the spatial information among the image patches, making it challenging for the subsequent cross-attention layer to disentangle.



(a) Training loss of Flamingo w/ perceiver.

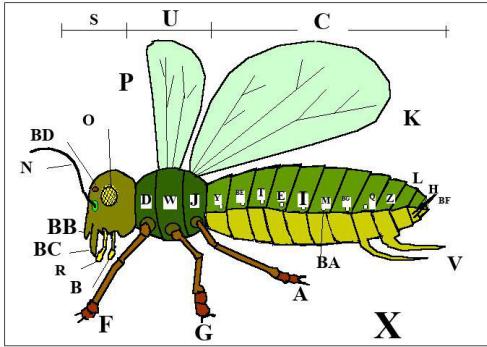


(b) Training loss of “Flamingo” w/o perceiver.

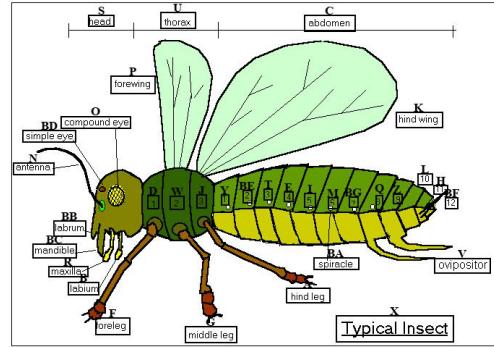
Figure 9: An overfitting experiment on the Flamingo models with and without the perceiver resampler on a document OCR dataset.

D Evaluation Details of AI2D

We provide an illustration of two AI2D evaluation settings in Figure 10a and 10b.



(a) Evaluation Setting 1: With mask



(b) Evaluation Setting 2: With no mask

Figure 10: Illustration of two AI2D evaluation settings adopted from VLMEvalKit using a test sample with the question “Which is the leg closest to the head?”.

E Data Formats and ChatML Template

We provide the examples of training data formats for various tasks in Figure 11 and ChatML template used in SFT in Figure 12.

- Image Captioning**

<Image><image></Image> Give a brief description of the image. A group of airplanes flying through the sky.<|im_end|>
- Grounded OCR**

<Image><image></Image> Detect the text within this section: <box>(743,621), (886,672)</box>. VINEGAR<|im_end|>
- Referring Expression Grounding**

<Image><image></Image> Identify the text "VINEGAR" in the image and provide the coordinates. <box>(743,621),(886,672)</box><|im_end|>
- OCR**

<Image><image></Image> Apply Optical Character Recognition (OCR) to the provided image. <ref>Herald</ref><box>(602,893),(666,933)</box><ref>read </ref><box>(649,851),(688,882)</box><|im_end|>
- Document Transcription**

<Image><image></Image> Extract the text from the provided document. Access PDF Holt Mathematics Lesson 9 4 of the books to browse. The standard Answers book, fiction, history, novel, scientific research, ... (the rest of the doc)<|im_end|>
- Visual Question Answering**

<Image><image></Image> What color is the hydrant? Black and yellow<|im_end|>
- LaTeX Recognition**

<Image><image></Image> Please write out the expression of the formula in the image using LaTeX format. $\backslash\angle A C B = \backslash\angle A ^ { \prime } C B ^ { \prime }$ <|im_end|>

Figure 11: Examples of training formats for various tasks used in pre-training. The emerald colored <image> tag indicates where to insert visual features. The blue colored text represents the ground truth associated with loss.

ChatML Template Example Utilized in SFT

```
<|im_start|>system
Follow the user's instruction and answer questions.<|im_end|>
<|im_start|>user
<Image><image></Image>
What activity are the people engaged in on the green field?<|im_end|>
<|im_start|>assistant
The people are engaged in playing a game of frisbee on the lush green field.<|im_end|>
<|im_start|>user
Is there a person wearing a fairy outfit in the image? If so, what are they doing?<|im_end|>
<|im_start|>assistant
Yes, there is a person wearing a fairy outfit in the image. They are walking in the field, likely
participating in the fun activities along with the other people.<|im_end|>
<|im_start|>user
What can you say about the spectator's location in relation to the houses?<|im_end|>
<|im_start|>assistant
The spectators are watching a match being played in a location close to the houses, possibly
indicating that the field is within a residential area or park.<|im_end|>
```

Figure 12: An example of the ChatML template used in SFT. The emerald colored <image> tag indicates where to insert visual features. The blue colored text represents the ground truth associated with loss.

F Unoptimized Results Using Yi-34B as the Backbone LLM

We only utilized the smaller 34B models for faster ablation studies and iterations, without detailed checkpoint selection or hyperparameter optimization. The unoptimized 34B results are provided in the following Tables for reference. Note that, although our 34B results are not optimized, the 34B NVLM models still significantly outperform other models, including VILA-1.5 40B [71] and Cambrian-1 34B [139].

Table 12: Evaluation of 34B models on vision-language and text-only benchmarks. **Text-only Avg. 4** represents the average accuracy degradation or improvement of the multimodal LLM compared to its backbone text-only LLM after multimodal training, measured across four text benchmarks: MMLU, GSM8K, MATH, and HumanEval (see Table 8 for full results).

Tasks	MMMU test / val	MathVista testmini	VQAv2 test-dev	AI2D test / no_mask	TextVQA val	ChartQA test	DocVQA test	Real- WorldQA	OCR- Bench	Text-only Avg. 4
NVLM-D 34B	48.7 / 52.1	59.9	84.3	82.6 / 93.0	80.0	84.0	89.1	67.3	819	+11.8
NVLM-X 34B	47.2 / 54.0	59.2	84.5	79.6 / 91.2	78.2	79.4	79.2	64.8	802	+11.2
NVLM-H 34B	46.3 / 53.0	59.4	84.8	81.5 / 93.0	79.0	82.0	80.5	66.3	821	+11.6

Table 13: Evaluation of 34B models on text benchmarks: MMLU, GSM8K, MATH and HumanEval.

Tasks	Backbone LLM	MMLU	GSM8K	MATH	HumanEval	Avg. Accuracy	Text-only Avg. 4
(a) Nous-Hermes-2-Yi-34B [102]	N/A	75.5	78.6	21.8	43.3	54.8	N/A
NVLM-D 34B	(a)	73.4	82.3	47.8	62.8	66.6	+11.8
NVLM-X 34B	(a)	73.2	82.2	46.4	62.2	66.0	+11.2
NVLM-H 34B	(a)	73.4	82.7	46.3	63.4	66.4	+11.6