

---

# VILA<sup>2</sup>: VILA Augmented VILA

---

**Yunhao Fang<sup>1</sup>\***   **Ligeng Zhu<sup>1</sup>\***   **Yao Lu<sup>1</sup>**,   **Yan Wang<sup>1</sup>**   **Pavlo Molchanov<sup>1</sup>**

**Jang Hyun Cho<sup>2</sup>**   **Marco Pavone<sup>1</sup>**   **Song Han<sup>1,3</sup>**   **Hongxu Yin<sup>1</sup>**

NVIDIA<sup>1</sup>   UT Austin<sup>2</sup>   MIT<sup>3</sup>

## Abstract

Visual language models (VLMs) have rapidly progressed, driven by the success of large language models (LLMs). While model architectures and training infrastructures advance rapidly, data curation remains under-explored. When data quantity and quality become a bottleneck, existing work either directly crawls more raw data from the Internet that does not have a guarantee of data quality or distills from black-box commercial models (*e.g.*, GPT-4V [1] / Gemini [2]) causing the performance upper bounded by that model. In this work, we introduce a novel approach that includes a self-augment step and a specialist-augment step to iteratively improve data quality and model performance. In the self-augment step, a VLM re-captions its own pretraining data to enhance data quality, and then retrains from scratch using this refined dataset to improve model performance. This process can iterate for several rounds. Once self-augmentation saturates, we employ several specialist VLMs finetuned from the self-augmented VLM with domain-specific expertise, to further infuse specialist knowledge into the generalist VLM through task-oriented re-captioning and retraining. With the combined self-augmented and specialist-augmented training, we introduce VILA<sup>2</sup> (*VILA-augmented-VILA*), a VLM family that consistently improves the accuracy on a wide range of tasks over prior art, and achieves new state-of-the-art results on MMMU leaderboard [3] among open-sourced models.

## 1 Introduction

The success of large language models (LLMs) [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15] has offered the cornerstone for cross-modality tasks. Through the alignment of visual encoders to LLMs, visual language models have enabled myriad appealing capabilities to visual tasks, such as instruction following, zero-shot generalization, few-shot in-context learning (ICL), and enhanced world knowledge [16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. The field has progressed rapidly in the past two years, yielding effective alignment training recipes [18, 23, 25] and model architectures [16, 17, 18, 19, 20].

Contrary to the fast-evolving training enhancement, the underlying human-generated datasets and tasks remain simple [26, 27, 28, 29]. Given the costly nature of VLM training, most methods are confined with *coarse-quality large-scale* captioning image-text pairs (pretraining), followed by *fine-grained small-scale* supervised finetuning (SFT). Enhancement of image-text pairs with millions and billions of instances can inevitably impose a huge amount of human effort, and thus not realistic. Recent methods have observed rewarding distillation possibilities from proprietary commercial models like GPT-4V [1] and Gemini [2]. However, the performance is upper bounded by

---

\* Equal contribution.

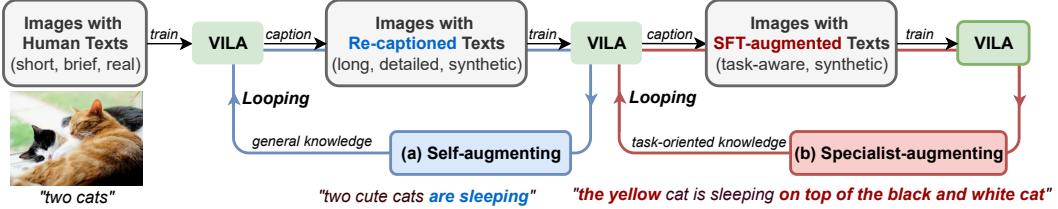


Figure 1: The schematic diagram to train VILA<sup>2</sup>, short for VILA-augmented VILA. We re-formulate visual language model (VLM) training with model in the loop to remedy training data deficiency. We start with validating design options in constructing a self-augmenting loop (Section 2.1) to improve on caption quality of the default training task. After the saturation of this process, we challenge the VLM to generate data conforming to extra SFT-enabled tasks to further VLM learning (Section 2.2). Our new design insights yield off-the-shelf performance boosts to VLMs (Section 3).

these models. In the meantime, studies remain very sparse on how to better utilize VLMs to correct human error and remedy dataset task simplicity for enhanced training.

In this work, we aim to answer “*whether it is possible that the VLM itself can remedy dataset deficiency and enhance its training.*” We delve deep into the potential of using VLM itself to refine and augment the pretraining data and performance in an iterative way. Our new training regime is summarized in Figure 1, which consists of two main steps: a self-augment step and a specialist-augment step. We start with the self-augment loop (Figure 1 (a)) that leverages VLMs to enhance the quality of pretraining data. We demonstrate that synthetic data, combined with the original data, can collaboratively generate stronger models in a bootstrapped loop manner. Intuitively and as we observed, the loops offer performance boosts *for free*, but suffers diminishing returns after 3 rounds. To facilitate additional learning, we reformulate a more challenging task-specific loop (Figure 1 (b)). In this loop, a specialist with a focus on new knowledge or tasks, such as a specially-aware expert, is finetuned from the self-augmented VLM using a limited amount of additional SFT data. The specialist can then recaption a massive amount of pretraining data. Finally, the self-augmented VLM can be retrained on the specialist-recaptioned pretraining data to further boost the performance. We demonstrate the benefits of this loop using three different specialist VLMs with expertise in spatial awareness, OCR, and grounding.

Our novel VLM augmentation training regime progressively improves data quality, *i.e.*, covering enhanced visual semantics and reducing hallucinations, as demonstrated in Figure 1 and more in our experimental section. This offers a direct performance boost to VLMs. We introduce a new family of VILA<sup>2</sup> models, as in VILA-augmented-VILA. VILA<sup>2</sup> outperforms state-of-the-art methods across main benchmarks, all enhanced without bells and whistles via self-bootstrapped training. Now constituting a new state of the art on MMMU benchmark [3] among all open-sourced models, we hope that the insights and release of VILA<sup>2</sup>’s recipe, data, and code can assist with our community for better understanding and usage of synthetic data to train stronger VLMs.

## 2 Methodology

In this paper, we focus on auto-regressive VLMs where image tokens are projected into the textual space and concatenated with text tokens, in line with [16, 2, 25]. This approach is chosen because of its flexibility when handling multi-modal inputs. We follow the widely adapted three-stage training paradigm, *i.e.*, align-pretrain-SFT, to ablate our studies. We start to self-augment VLM training by constructing a bootstrapped loop leveraging VLM’s general captioning capability. After the bootstrapping saturates, we then introduce specialist augmenting exploiting VLM’s skills picked up during SFT across additional visual tasks as specialist feedback to its pretraining stage. We next elaborate on these steps in detail.

### 2.1 Self-augmenting via General Knowledge Enhancement

Existing VLM training largely relies on data from the internet, where the texts are usually brief and short, see Table. 1 where the average number of words is less than 20 for MMC4 [26] and COYO [28]. In addition to brevity, human annotations can also fall short in explaining to LLMs

	MMC4 [26]	COYO [28]	COYO-VILA <sub>1</sub>	COYO-VILA <sub>2</sub>	COYO-VILA <sub>3</sub>	COYO-VILA <sub>4</sub>
Avg #Words VQA <sup>v2</sup>	17.1 ± 25.0 N.A.	11.9 ± 9.0 61.6	101.2 ± 43.0 62.5	117.1 ± 49.1 63.5	126.77 ± 50.10 <b>63.7</b>	125.9 ± 51.2 <u>63.6</u>

Table 1: The average number of words and VQA<sup>v2</sup> evaluation comparison between the original dataset and the re-captioned dataset. Best performance is bolded and second best is underlined. During self-augmentation, the caption lengths increases significantly, thus offering more details and information. The length and VQA score growth plateau after three rounds of self-augmenting.

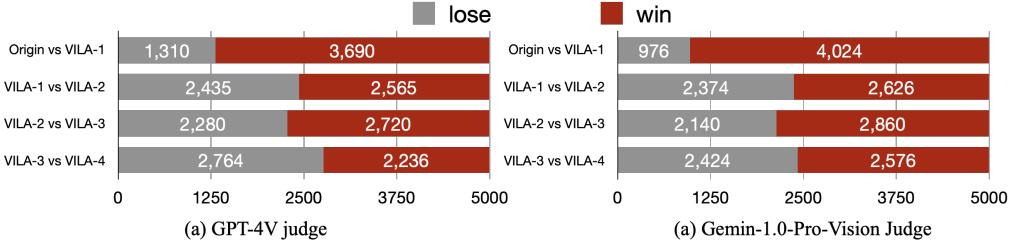


Figure 2: LLM judgement for captions from VILA<sub>i</sub>, *i* indicating the self-augmenting round. Evaluations are based on 5,000 sampled data from Coyo [28]. Both GPT-4V [1] and Gemini-1.0-pro [2] prefer for VILA<sup>2</sup> augmented texts and captions from later rounds got higher score.

the versatile semantics an image presents. As another example, Figure. 4 indicates that an original COYO caption only describes the person riding on the street while lacking a description of clothing and surroundings. Previous studies have either assigned humans to write dense captions or relied on commercial propriety APIs for image descriptions. The first option can be labor-intensive and costly, while the second may inherit model biases, upper bound the model performance, and also potentially introduce copyright issues. Such detailed information, however, can be highly beneficial for VLM training, as we show later.

Rather than distilling proprietary models or relying on manual laboring, we aim to use *VILA to generate better captions for VILA’s pre-training*, exploiting the power of the already-intelligent VILA within intermediate training stages to conduct laborious relabelling effort.

Beginning with the original dataset, we first train the initial version of VILA, referred to as VILA<sub>0</sub> in subsequent experiments. Next, we use VILA<sub>0</sub> to re-caption VILA’s pre-training datasets. With appropriate prompt choice and conversation template, VILA<sub>0</sub> is able to generate *long* and *detailed* captions. Then the augmented datasets, consisting of real images from the internet and synthetic texts from VILA<sub>0</sub>, are then re-used to train the next round of VILA, named VILA<sub>1</sub>. This self-augmenting process can be repeated several rounds before convergence, as shown in Figure 1 (a).

### 2.1.1 Prompts and Template Design

We also augment prompt styles in the self-augmentation loop to diversify caption styles. This approach leads to immediate performance improvements. The key idea is to advance the knowledge embedded in new captions and to avoid repetitive generations.

To validate design choices, we conducted an in-depth study on prompt choices as follows, where <img> indicates the location where image features will be inserted, and discuss our findings.

- Prompt Simple: <img> Describe the image briefly.
- Prompt Long-v1:<img> Describe the image in details.
- Prompt Long-v2:<img> Elaborate on the visual and narrative elements of the image in detail.
- Prompt Long-v3: <img> Instead of describing the imaginary content, only describing the content one can determine confidently from the image. Do not describe the contents by itemizing them in list form. Minimize aesthetic descriptions as much as possible.

	Avg #words	VQA <sup>v2</sup>	GQA	SQA	VQA <sup>T</sup>	POPE	LLaVA <sup>W</sup>	MM-Vet	MMMU
Baseline	17.1	79.6	62.4	68.4	61.6	84.2	68.4	34.5	33.8
<i>Prompt Ablation for Self-Augmenting</i>									
Self-augment Iter1 - Simple	90.4	79.4	63.0	68.7	62.4	<b>87.0</b>	68.3	34.5	33.1
Self-augment Iter1 - Long v1	94.8	80.0	62.7	<b>71.1</b>	<u>62.2</u>	84.0	71.7	34.5	34.4
Self-augment Iter1 - Long v2	105.4	<b>80.1</b>	63.2	70.7	62.7	84.6	71.7	34.9	34.7
Self-augment Iter1 - Long v3	102.4	<b>80.1</b>	63.4	71.0	62.9	85.0	71.4	34.4	34.7
<i>Conversation Template Ablation for Self-Augmenting</i>									
Mixed - re-caption text only	101.2	79.6	62.5	<b>71.1</b>	62.3	81.0	71.8	34.2	34.1
Mixed - concatenated	127.3	80.0	<b>63.2</b>	71.0	<b>62.5</b>	<u>85.0</u>	<u>72.2</u>	<b>34.8</b>	<b>35.8</b>

Table 2: Comparison with different prompts and training templates when self-augmenting for one round. The best and second-best results are highlighted with **bold** and underline respectively. The results show that prompts design are critical for self-augmenting. Re-captioning the dataset with naive prompt "*Describe the image briefly*" does not improve while designed prompt can significantly boost the mode performance.

**Brief and Short Re-captioning is NOT Helpful.** We begin with a straightforward prompt asking VLMs to *briefly* describe the image. Despite these brief recaptions being significantly longer than the original texts (90 vs. 17 words), there is no notable improvement in VLM benchmarks, as shown in Table. 1. In fact, metrics even deteriorate in benchmarks such as Science-Image and MMMU. This decline is due to the lack of details and the potential hallucinations introduced during recaption.

Next, we re-design the prompt to encourage VLMs to provide a more detailed description of visual narrative elements in images. We also referenced the prompt template from ShareGPT-4V [30] to ensure the descriptions are accurate and precise. Our experiments demonstrate that using three different long prompts improves the quality of recaptioning and boosts performance in benchmarks, with a clear ranking among them, detailed in Table. 1. Therefore, we leverage a mixture of these prompts by randomly selecting from versions v1 to v3.

**Keeping Origin Human Text is Important.** We compare different conversation templates in Table. 1, where the first only uses real human data and *concatenated* adapts both human and synthetic description. We conduct experiments to train with both choices and find that using self-augmented data leads to improvement on major benchmarks such as LLaVA-Bench, Science, TextVQA, and MME. However, we notice there is still a decline in several metrics, prompting us to concatenate both the origin and re-captioned texts in order to best preserve information. This template consistently improves all VLM metrics (Table. 1).

With the proper prompt choice and template design, self-augmenting can continuously the caption quality. Shown in Table. 1, the synthetic captions become much *longer* and more *informative* than real human texts (reflecting on VQAv2 score). We further follow LLMJudge [11] to compare caption quality from different rounds. Both Gemini-1.0-Pro and GPT-4V prefer VILA<sup>2</sup> generated captions (higher winrate).

## 2.2 Surpassing the Limit with Specialist VLM Augmentation

We observe that the improvement from self-augmentation starts to saturate after a few rounds of bootstrapping, with all free lunches having been squeezed (Table. 3). We hypothesize that this shortcoming stems from the monotonic task of *general* descriptive captioning, which is also heavily influenced by language modeling priors.

To further facilitate the focus of visual features and enable effective annotation, we propose incorporating extra *task-specific* knowledge into the generalist VLM to create several specialist VLMs. Each specialist model is finetuned with data that requires a deeper understanding of image components and semantics, *e.g.*, spatial relations, localization, and OCR.

It might seem as if we require a separate specialist model for each task whereas we show that quick finetuning of VLM during the SFT stage towards downstream tasks already suffices for this role. A bootstrapped loop can then transfer such specialist knowledge from small-scale SFT data onto a large number of pretraining images. In this manner, we jointly exploit the synergy of abundant image amounts in pretraining and specialist knowledge acquired during SFT.

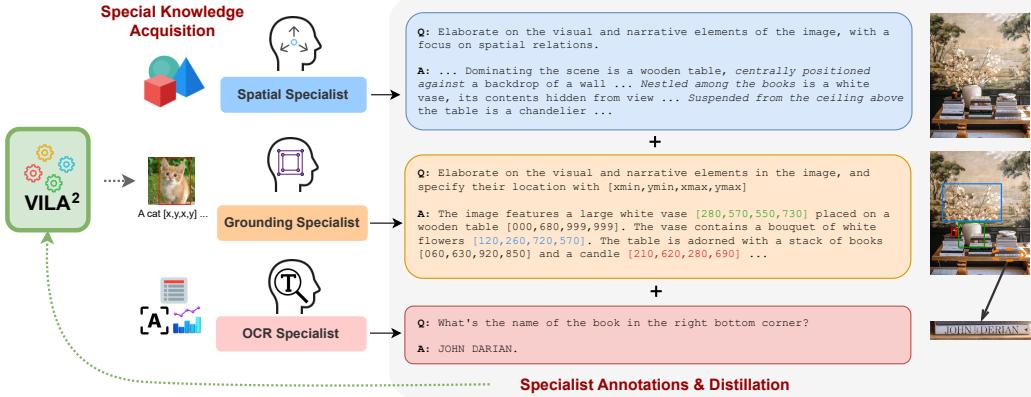


Figure 3: VILA<sup>2</sup> specialist VLM augmentation overview. We start by gathering task-specific knowledge to create task-specialist VLMs. These specialist VLMs annotate images with task-oriented prompts. The generated question-answering pairs are then utilized to re-train the next iteration of VILA<sup>2</sup>.

### 2.2.1 Acquiring Specialized Knowledge

In this paper, we focus on three challenging tasks that require direct visual understanding: *spatial relations understanding*, *Grounded narration*, and *OCR*, as summarized in Figure 3. We next elaborate on the process of specialist construction as follows:

**Spatial Specialist.** To explicitly challenge the model to acquire additional spatial awareness, we curated *SpatialRelationQA*, a dataset containing 1 million conversations about spatial relations within scenes. Our dataset is built on LV3D, a comprehensive collection of both indoor and outdoor 3D datasets from Cube-LLM [31] that is designed to enhance the understanding of 3D spatial relations requiring both perceptual and grounding skills.

We formulated a three-step process to create the QA pairs. First, for each cleaned 3D scene, we iterated through all 3D bounding boxes and randomly sampled from five object-object relations (*closest*, *in front of*, *behind*, *left*, *right*), with six object-camera relations (*close*, *far*, *closest*, *farthest*, *left*, *right*). Next, we checked if any remaining bounding boxes matched these sampled relations. Finally when a match was found, we randomly selected question templates to construct the QA pairs, incorporating instances, their projected 2D bounding boxes, and relations. A sample question can be constructed as: “Where is the *chair closest to the table* [ $x_{left}, y_{top}, x_{right}, y_{bottom}$ ] in the image?”, with answer being the target bounding box.

**Grounding Specialist.** To also enhance knowledge towards grounding awareness we exploited grounded narration, a highly visual-centric task that requires VLMs to generate detailed captions to accurately locate major visual elements using 2D bounding boxes, as shown in Figure 3. This approach provides dense supervision and allows us to verify if VLMs hallucinate. To develop the VILA<sup>2</sup> grounding specialist, we used image-grounded caption pairs from the 20M GRIT dataset [32]. We first filtered out bounding boxes covering more than 70% of the image area, as many images in GRIT are book or album covers not closely related to the captions. Next, we removed images containing more than three instances of the same category to reduce complexity and decrease noise in generation orders. This process yielded 4M higher-quality instances for grounding specialist training, which we further split into two subsets: *Grounding-Short* 3M and *Grounding-Long* 838K for a two-stage finetuning.

**OCR Specialist.** Finally we focus on OCR strength enrichment. We utilize a diverse set of images featuring textual content, such as tables, charts, and documents, to develop an OCR specialist. Each image is annotated with QA pairs that focus on text recognition (e.g., *Q: What is the title of the book?*), comprehension (e.g., *Q: Which bar has the largest value?*), and reasoning (e.g., *Q: What is the main idea of the quote from Albert Camus?*). Dataset details are provided in appendix A.2.

The three specialist are then applied to the final augmentation stage. To this end we use a new set of task-oriented prompts to activate the specialists’ knowledge and improve their instruction-following ability by narrowing focus. Specifically, we prompt the spatial relations understanding specialist with evenly sampled templates of “<image> Elaborate on the visual and narrative elements of the image in detail, with a focus on spatial relations.” and “<image> Can you explain the content of the image and their spatial relations in detail?” during the specialist augmentation stage. Similarly, the grounding specialist generated captions with bounding boxes for the major visual focus, and the OCR specialist identified most textual content in the images. The responses from these different specialist VLMs are then curated as QA pairs and appended to the original captions of COYO images for the next pertaining iteration of VILA<sup>2</sup>. More details can be found in appendix A.1.

### 3 Experiments

**Model Architecture.** We follow the architecture from VILA [25], where a multi-modal large model consists of three key components: an *LLM* for auto-regressive generation, a *visual encoder* for extracting visual features, and an image-text *projector* to align the visual and text modalities.

We use Llama2-7B [9] for exploratory experiments to address the question “*To what extent can a VLM self-bootstrap?*”. Then we switch to our previous SOTA settings with Llama3-8B-Instruct [33] and Yi-34B [14] when compared to other methods. For visual encoders, we use SigLIP [34] for LLaMA-series models and InternViT-6B [35] for the Yi-34B model. For projection layers, we follow LLaVA [16, 36] to adapt simple linear layers for bridging image and text modalities. At the same time, we introduce a  $4 \times$  downsampling of visual tokens by concatenating  $2 \times 2$  neighboring patches along the channel dimension and using a simple MLP for the downsampling process.

**Training Strategies.** We conduct VILA<sup>2</sup> training following widely used three-stage settings.

1. *Projector Initialization.* The language models and ViT are separately pre-trained, while the projector is randomly initialized. To initially align the feature space between the visual and text modalities, we utilize the LLaVA align dataset [16].
2. *Vision-Language Pre-training.* We then pre-train the model (LLM and the projector) on the visual language corpus. We consider interleaved image text corpus (*e.g.*, MMC4 [26]) and image-text pairs (*e.g.*, COYO [28]). We apply our VILA<sup>2</sup> for the pre-training data and the augmented data would used in this stage to replace original COYO captions.
3. *Visual Instruction-tuning.* Finally, we perform instruction tuning of the pre-trained model on visual language instruction datasets. The blending details is attached in the appendix.

Without specifically mentioned, our experiments are conducted with 128 GPUs and a global batch size of 1024. We employ AdamW optimizer with learning rate  $\{10^{-3}, 5 \times 10^{-5}, 2 \times 10^{-5}\}$  for aforementioned three stages respectively. Each stage is trained with one epoch with a 3% warmup strategy. No weight decay is applied. In some self-/specialist augmented training, VILA<sup>2</sup> may involve extra stage to further improve. Please refer to Section 2.2 and Appendix for more details.

**Data.** Our pretraining stage consists of 6M sampled MMC4 [26], 25M sampled Coyo [28], and the full ShareGPT4v [30]. To ensure a fair comparison, we only replace the text captions during our experiments while keeping all image sources the same. We use two SFT data blends for different purposes: a smaller blend of 1.8M samples for exploratory experiments in Table 3–Table 4; a larger blend of 5.9M samples augmented from VILA’s training receipt\*, for state-of-the-art experiments in Table 5–Table 6. Detailed SFT recipe can be found in the appendix.

**Specialist Setups.** We integrated specialty data with high-quality image captioning datasets and diverse instruction finetuning datasets, ensuring the models retain their narrative and instruction-following abilities while acquiring task-specific knowledge.

1. **Spatial Specialists.** We continued training the specialist from the stage 2, ALLaVA caption [37], and GPT-4V caption from ShareGPT4V [30].

---

\*[https://github.com/NVlabs/VILA/tree/main/data\\_prepare](https://github.com/NVlabs/VILA/tree/main/data_prepare)

**Raw caption:** Boris Johnson riding on the street.

**Self-Augmented Round-1:** In the image, a man is seen **riding a bicycle** on a city street. He is dressed in a **blue suit** and tie, giving him a professional appearance. The bicycle he's riding has a **basket attached to it**, suggesting that he might be using it for transportation or leisurely rides around the city. The background of the image reveals a lively urban scene. There are buildings visible, indicating that this is likely an **urban area**. A group of people can also be seen walking on the sidewalk, going about their day. Adding a touch of color to the scene is a **yellow flag** with red and **black stripes** hanging from one of the buildings. This could possibly indicate the presence of a business or organization associated with the building. Overall, the image captures a moment of everyday life **in a bustling city**, with the man on his bicycle adding a unique element to the scene.



**Self-Augmented Round-2:** The image captures a moment of **surprise and excitement** on the face of a man, who is **riding a bicycle**. He's dressed in a **gray jacket and blue tie**, giving him an air of professionalism. His blonde hair stands out against his dark jacket, adding to the contrasting elements in the scene. The background provides context to the setting. A brick building with a **yellow awning** suggests it could be a commercial area or perhaps a residential neighborhood. People can be seen walking by, going about their day, oblivious to the man's unexpected ride. A notable detail is the sign for "**Bacardi**" **visible** in the background. This could indicate that the location is near a Bacardi establishment, possibly a bar or restaurant. The sign adds an element of intrigue to the scene, hinting at stories untold beyond what we see.

**Self-Augmented Round-3:** The image captures a **lively scene on a bustling street**. At the **center of the frame**, a man is energetically pedaling a **white bicycle**. He's dressed in a **black jacket and blue tie**, adding a touch of formality to his otherwise casual attire. His **mouth is open wide** as if he's shouting or laughing, perhaps expressing excitement or surprise. The background of the image provides context to the location - it appears to be a busy city street with buildings and people going about their day. The photo seems to have been taken during the day under **clear skies**, casting bright light that illuminates the entire scene. Despite the hustle and bustle, there's a sense of joy and freedom conveyed by the man riding the bike, making this an interesting snapshot of city life.

**Specialist-Augmented.** In the image, a man is the **central figure**, riding a bicycle on a street. He is dressed in a **gray jacket and a blue tie**, giving him a professional appearance. The bicycle he's riding has a **blue sign on the front**, although the text on the sign is **not discernible** from the image. The **man is not alone; he is accompanied by a group of people** who are **walking behind him**. Their exact number is hard to determine from the image, but they appear to be a small crowd. The setting of the image is a street, with a building visible in the background. The building's details are not clear, but it adds context to the scene. The man on the bicycle is **facing** towards the right side of the image, suggesting that he might be moving in that direction. **The people walking behind him** are also **facing the same direction**, indicating that they might be following the man on the bicycle. Overall, the image captures a moment of everyday life, with the man on the bicycle as the main focus, and the people and the building providing the backdrop. The image does not contain any text. The relative positions of the objects suggest a dynamic scene, with the man on the bicycle leading the way for the **people behind him**.

Figure 4: VILA<sup>2</sup> continuously enhances caption quality over self-augmenting. The sample is from the COYO [28]. We mark facts in **green**, hallucinations in **red**, and spatial information in **blue**.

2. **Grounding Specialist.** We split the cleaned 4M grounded narration into *Grounding-Short* 3M and *Grounding-Long* 838K for a two-stage training process. In stage 3, we combined *Grounding-Short* 3M with ALLaVA caption [37] to adapt to new tasks of grounded narration while maintaining the narrative ability. In stage 4, we combine *Grounding-Long* 838K with Shikra GPT-4 [38], Visual7W [39], LLaVA-SFT, and 100K GPT-4V captions from ShareGPT4V to sustain both narrative and instruction following capacities.
3. **OCR Specialist.** We trained our OCR specialist with various internet datasets focused on text recognition, understanding, and reasoning, including LLaVA-SFT, TextOCR-GPT4V [40], SynthDoG-En [41], OCRVQA [42], TextCaps [43], ArxivQA [44], DocVQA [45], AI2D [46], ChartQA [47], LLaVAR [48] and 35 OCR-related datasets from The Cauldron [49].

**Evaluation.** We ablate our models in the following common VLM benchmarks. Note that some metrics are shortened due to space limits. VQA<sup>v2</sup> [50]; GQA [51]; SQA: ScienceQA [52]; VQA<sup>T</sup>: TextVQA [53]; POPE [54]; MMB: MMBench [55]; MMB<sup>CN</sup>: MMBench-Chinese [55]; SEED: SEED-Bench [56]; LLaVA<sup>W</sup>: LLaVA-Bench (In-the-Wild) [16]; MM-Vet [57]; MMMU [3].

### 3.1 Self-Augmentation Results

Our goal is to "augment" existing pretraining datasets by rewriting captions with dense and informative texts. Instead of relying on human labor or black-box APIs, we use VILA to generate these captions. Therefore, the enriched caption can help develop better VILAs based on which VILA can also feedback to further enhance the captions for the training dataset.

**VILA<sup>2</sup> Enriches Dataset Text Quality.** The main VLM's performance boost comes from the caption quality improvement. As demonstrated in Table. 1, the average number of words grows quickly

	VQA <sup>v2</sup>	GQA	SQA	VQAT	POPE	LLaVA <sup>W</sup>	MM-Vet	MMMU
VILA <sub>0</sub> - Baseline	79.6	62.4	68.4	61.6	84.2	68.4	34.5	33.8
VILA <sub>1</sub>	80.0	63.2	71.0	62.5	84.6	72.2	34.8	<u>35.8</u>
VILA <sub>2</sub>	80.8	<b>63.5</b>	71.5	63.5	84.7	71.2	34.9	<u>35.2</u>
VILA <sub>3</sub>	80.7	<b>63.5</b>	71.5	63.7	84.5	<b>72.3</b>	35.5	35.5
VILA <sub>4</sub>	80.7	63.4	71.2	63.6	<b>85.0</b>	<b>72.3</b>	35.5	35.0
VILA <sub>3</sub> +Spatial Specialist	<b>81.1</b>	62.8	<b>72.9</b>	<b>65.0</b>	<b>85.0</b>	71.4	<b>37.1</b>	<b>36.8</b>

Table 3: Self-augmenting can consistently enhance the performance of model training. For VILA<sub>1–4</sub>, the best and second-best results are highlighted in **bold** and underline, respectively. With each iteration, VLM improves the quality of the pretraining dataset’s captions. These improved descriptions lead to progressively better performance when training subsequent VLMs. Although the effects of self-augmentation plateau after three rounds, they can be further improved by our specialist.

after self-augmenting and saturates around round-3 and round-4. This is consistent with the trend observed in the benchmark side. In round-1, caption length increases significantly from 12 to 101, while in round-3, the increase is only from 117 to 126. The increase in caption length is no longer significant after round-1 but we still observe consistent improvement on VLM benchmarks. We hypothesize that self-augmentation after round-1 primarily improves caption quality, providing more accurate details and reducing hallucination. The results are visualized in Figure 4. The origin caption for the provided example is brief and short, only describing Boris’ riding action and no other information. After self-augmenting once, the caption becomes longer and much more informative, including descriptions of dressing and surrounding environments. Several new hallucinations are also added to the text (*e.g.*, suit color, basket not showing in the image). While self-augmenting iterates, the hallucinations gradually get removed from captions. For example, the no-showing basket and wrongly-read “Barcardi” text are no longer mentioned in round-3. VILA<sub>3</sub> only describe visual elements that are confident with.

**VILA<sup>2</sup> Bootstraps VILA’s Performance.** We follow the same pre-training + SFT process as VILA [25] and sample 5% data from the pre-training phase to ablate. The images are from the existing COYO [28] and MMC4 [26] and in each loop, we use the models trained last round to generate new captions for half of the sampled COYO images. MMC4 is not re-captioned because of its interleaved feature. Other settings are kept the same. We compare VILA<sub>i</sub> from different looping steps on common VLM benchmarks. We notice that self-augmented data help improves the model performance across different iterations: VILA<sub>i+1</sub> is consistently better than VILA<sub>i</sub> and the looping progressively boosts the performance (VILA<sub>1–4</sub> in Table 3). The self-augmenting technique is consistently useful until three rounds. VILA<sub>4</sub> reaches saturation and no longer bring consistent improvement of VILA<sub>3</sub>.

**Surpassing the Limit with VILA<sup>2</sup> Specialist.** The “self-augmentation then training” cycle is yielding a diminishing return and is likely to reach saturation within three iterations, as illustrated in Tab. 3. However, by incorporating external spatial relations data and fine-tuning a spatial specialist VLM, we can extend this process beyond its current limits. Training a specialist VLM with domain-specific knowledge and using task-specific prompts during the re-captioning process can effectively enrich the resulting captions with expertise knowledge, thereby enhancing benchmark performance.

We show that VILA<sub>3</sub> *spatial relations understanding* specialist represents an extra focus on describing the spatial information in Fig. 4. Caption augmented by the specialist (the last example) will keep the most visible details, and provide more information about spatial relations compared to the “Round-4” caption. This additional information encompasses not only the object-to-object relations but also the localization and clear pose of the major focus, which is not present in the *SpatialRelationQA* dataset. We hypothesize that this improvement might result from the compositional combination of knowledge in other parts of the training data, as a synergy result from both the new task learning and the self-bootstrapping procedure. The effectiveness of these enriched captions is demonstrated in Table 3. Here, we integrate an *additional* 5% of COYO data re-captioned by the VILA<sub>3</sub> specialist into the pre-training stage. Following the same SFT stage, we observe notable improvements in 5 out of the 8 benchmarks.

	VQA <sup>v2</sup>	GQA	VQA <sup>T</sup>	POPE	SEED-I	MME	MM-Vet	MMMU (T)
<i>Pretrain Data: 10% MMC4-core+10% COYO-25M+ShareGPT4V-Pretrain</i>								
Original Caption	81.4	63.8	65.2	85.5	70.6	1472.5	34.0	31.8
+ Spatial Specialist	<b>81.9</b> <sub>+0.5</sub>	<b>64.1</b> <sub>+0.3</sub>	<b>66.0</b> <sub>+0.8</sub>	85.9 <sub>+0.4</sub>	71.8 <sub>+1.2</sub>	1476.5 <sub>+4.0</sub>	36.7 <sub>+2.7</sub>	32.5 <sub>+0.7</sub>
+ OCR Specialist	81.8 <sub>+0.4</sub>	64.0 <sub>+0.2</sub>	65.3 <sub>+0.1</sub>	86.4 <sub>+0.9</sub>	<b>72.1</b> <sub>+1.5</sub>	1500.2 <sub>+27.7</sub>	34.3 <sub>+0.3</sub>	32.1 <sub>+0.3</sub>
+ Grounding Specialist	81.8 <sub>+0.4</sub>	64.0 <sub>+0.2</sub>	65.1 <sub>+0.1</sub>	<b>86.7</b> <sub>+1.2</sub>	71.0 <sub>+0.4</sub>	<b>1536.4</b> <sub>+63.9</sub>	<b>37.5</b> <sub>+3.5</sub>	<b>32.6</b> <sub>+0.8</sub>
<i>Pretrain Data: MMC4-core+COYO-25M+ShareGPT4V-Pretrain</i>								
Original Caption	82.2	63.9	66.7	<b>86.5</b>	71.2	1518.2	42.6	33.4
+ All 3 Specialist	<b>83.0</b> <sub>+0.8</sub>	<b>64.7</b> <sub>+0.8</sub>	<b>70.9</b> <sub>+4.2</sub>	86.4 <sub>-0.1</sub>	<b>74.0</b> <sub>+2.8</sub>	<b>1656.2</b> <sub>+142</sub>	<b>44.7</b> <sub>+2.1</sub>	<b>35.8</b> <sub>+2.4</sub>

Table 4: Effectiveness of the data re-captioned by specialists: We mark the best performance with **bold**. The results in the same block are trained with different pretraining data but the same SFT data. Specialists-annotated data consistently improves on both 10% and 100% pretraining setting.

Method	LLM	Res.	VQA <sup>v2</sup>	GQA	VizWiz	SQA <sup>I</sup>	VQA <sup>T</sup>	MMB	MMB <sup>CN</sup>	SEED	LLaVA <sup>W</sup>	MM-Vet
BLIP-2 [20]	Vicuna-13B	224	41.0	41	19.6	61	42.5	—	—	46.4	38.1	22.4
InstructBLIP [59]	Vicuna-7B	224	—	49.2	34.5	60.5	50.1	36	23.7	53.4	60.9	26.2
InstructBLIP [59]	Vicuna-13B	224	—	49.5	33.4	63.1	50.7	—	—	—	58.2	25.6
Qwen-VL [22]	Qwen-7B	448	78.8	59.3	35.2	67.1	63.8	38.2	7.4	56.3	—	—
Qwen-VL-Chat [22]	Qwen-7B	448	78.2	57.5	38.9	68.2	61.5	60.6	56.7	58.2	—	—
LLaVA-1.5 [60]	Vicuna-1.5-7B	336	78.5	62.0	50.0	66.8	58.2	64.3	58.3	58.6	63.4	30.5
LLaVA-1.5 [60]	Vicuna-1.5-13B	336	80.0	63.3	53.6	71.6	61.3	67.7	63.6	61.6	70.7	35.4
VILA-7B [25]	Llama 2-7B	336	79.9	62.3	57.8	68.2	64.4	68.9	61.7	61.1	69.7	34.9
VILA-13B [25]	Llama 2-13B	336	80.8	63.3	60.6	73.7	66.6	70.3	<u>64.3</u>	<u>62.8</u>	73.0	38.8
LLaVA-NeXT-8B [36]	Llama 3-8B	672	—	<b>65.2</b>	—	72.8	64.6	72.1	—	—	80.1	—
Cambrian-1-8B [61]	Llama 3-8B	1024	—	64.6	—	80.4	71.7	75.9	—	—	—	—
Mini-Gemini-HD-8B [62]	Llama 3-8B	1536	—	64.5	—	75.1	70.2	72.7	—	—	—	—
VILA <sup>2</sup> -8B (ours)	Llama 3-8B	384	<b>82.9</b>	64.1	<b>64.3</b>	<b>87.6</b>	<b>73.4</b>	<b>76.6</b>	<b>71.7</b>	<b>66.1</b>	<b>86.6</b>	<b>50.0</b>

Table 5: Comparison with state-of-the-art methods on 10 visual-language benchmarks. Our models consistently improve VILA under a head-to-head comparison, using the same prompts and the same base LLM, showing the effectiveness of enhanced pretraining data quality. We mark the best performance **bold** and the second-best underlined.

### 3.2 Specialist Augmentation Results

We next transition to our previous state-of-the-art settings to explore the significance of *scaling-up* our VILA<sup>2</sup> approach with more pretraining data and larger models to fully study the benefits of specialists. Following our VILA-1.5 release practice<sup>†</sup>, we use S2 [58] with Llama 3-8B-Instruct to enhance the visual encoder by scaling on scales rather than increasing input resolution. Additionally, we utilize a larger visual encoder, the Intern ViT-6B, which has a comparable number of parameters to the Yi-34B model to ensure balanced training.

We demonstrate the effectiveness of each specialist on a ~10% subset of the entire pretraining data blend and the 1.8M SFT blend, as shown in Table 4. Generally, specialist VLMs show overall improvements across most VQA benchmarks. We then combine annotations from all three specialists into multi-round QA pairs for each image and re-train VILA. This synergy among the specialists proves highly effective, as scaling-up to the full pretraining set yields significant improvements.

### 3.3 Benchmark Comparison to Prior Work

We now perform a comprehensive comparison to prior work over 10 major benchmarks and summarize results in Table 5 and Table 6. Note that we used a total of 25 million COYO data sampled from the 700 million with the highest CLIP score. We augment the original short real labels with multi-round QA pairs annotated by three specialists, all from 8B models. For 40B models, we continue to train from the stage 2 checkpoints with a mix of 3.75 M recaptioned COYO and a 200K caption dataset [37]. We observed the improvements in quality is consistent and can scale to 40B VILA checkpoints. The detailed training recipes of our 8B and 40B checkpoints are included in the Appendix and will be released jointly with the code base.

Remarkably we observed the enhanced self-augmentation and specialist augmentation training recipes, backed by enhanced and refined datasets, support VILA<sup>2</sup> to further push the performance boundary

<sup>†</sup><https://github.com/NVlabs/VILA>

Method	Overall	Art & Design	Business	Science	Health	Human & Social	Tech. & Eng.
GPT-4V [23]	55.7	65.3	64.3	48.4	63.5	76.3	41.7
SenseChat-V [63]	50.3	62.7	44.1	42.3	55.7	74.7	43.5
<b>VILA<sup>2</sup>-40B (ours)</b>	<b>47.9</b>	<b>62.0</b>	<b>42.3</b>	<b>38.5</b>	<b>51.9</b>	<b>71.9</b>	<b>42.3</b>
Qwen-VL-MAX [15]	46.8	64.2	39.8	36.3	52.5	70.4	40.7
InternVL-Chat-V1.2 [35]	46.2	62.5	37.6	37.9	49.7	70.1	40.8
LLaVA-1.6 [36]	44.7	58.6	39.9	36.0	51.2	70.2	36.3
Marco-VL-Plus*	44.3	57.4	34.7	38.5	48.7	72.2	36.7
Yi-VL [14]	41.6	56.1	33.3	32.9	45.9	66.5	36.0
Qwen-VL-PLUS [15]	40.8	59.9	34.5	32.8	43.7	65.5	32.9
Marco-VL-Plus*	40.4	56.5	31.0	31.0	46.9	66.5	33.8
Weito-VL-1.0*	38.4	56.6	30.5	31.1	38.4	59.0	34.2
<b>VILA<sup>2</sup>-8B (ours)</b>	<b>38.3</b>	<b>54.3</b>	<b>32.0</b>	<b>29.3</b>	<b>39.7</b>	<b>56.8</b>	<b>34.4</b>
InfiMM-Zephyr [64]	35.5	50.0	29.6	28.2	37.5	54.6	31.1
SVIT [65]	34.1	48.9	28.0	26.8	35.5	50.9	28.8
Emu2-Chat [66]	34.1	50.6	27.7	28.0	32.4	50.3	31.3
BLIP-2 FLAN-T5-XXL [20]	34.0	49.2	28.6	27.3	33.7	51.5	30.4
InstructBLIP-T5-XXL [59]	33.8	48.5	30.6	27.6	33.6	49.8	29.4
LLaVA-1.5 [16]	33.6	49.8	28.2	25.9	34.9	54.7	28.3

Table 6: Comparison with state-of-the-art methods on the MMMU dataset. \*: model on leader-board with unidentified reference. The models are ranked by overall scores, including both proprietary and open-sourced models. We highlight our results with color green. VILA<sup>2</sup> achieves SOTA performance in the open source category.

of VILA [25] by noticeable margins across almost all benchmarks, consistent with the ablated performance boosts we observed in previous analysis of Table 3. Moreover, VILA<sup>2</sup> now constitutes a SOTA performance on the main MMMU [3] test dataset leaderboard across all open-sourced models, without the usage of proprietary datasets and only based on publicly available datasets.

## 4 Related Work

**Large language models (LLMs).** The LLMs arguably began with the introduction of transformers [67] and subsequent efforts to scale them. For the decoder-only paradigm, OpenAI introduced the Generative Pre-trained Transformer (GPT) models [68], [6], from GPT-2 (1.5B parameters) to GPT-4 [23] (1.76T), demonstrating that scaling parameters along with high-quality data can produce coherent and contextually relevant text across various domains. Transformer-XL [5] extended the context window, enabling better comprehension of longer texts.

Open-sourced LLMs significantly advanced the field including LLaMa-1/2/3 [8, 9, 69], its instruct-finetuned extension Alpaca [10], and its chatbot version Vicuna [11]. Various models have further pushed the boundaries of LLM capabilities, such as Mistral [12] for generating coherent long-form text, Falcon [70] scaling open-sourced LLMs to 180B parameters, Yi [14] and DeepSeek [71] with applications in supervised fine-tuning for unlocking multilingual capabilities, DBRX [72] with a fine-grained mixture-of-experts architecture, and Gemini [2] with 10M context window. The scale of models is also rapidly increasing, *e.g.*, the LLaMa-3 [69] at 400B parameters. Our work utilizes LLaMa-2-7B [9], LLaMa-3-8B [69] and Yi-34B [14] as the LLM backbones.

**Visual language models (VLM).** Visual language models have rapidly progressed in recent years. The success mainly comes from pre-training visual and language models on the internet-scale data. CLIP [73] trained a shared feature space of vision and language modalities through contrastive learning on image-caption data. BLIP [74] and BLIP-2 [20] improved CLIP by leveraging noisy pseudo-labels and expanded multi-modal capabilities by aligning the pre-trained encoder to large language models. LLaVA [16] and InstructBLIP [59] demonstrated that jointly training from a collection of diverse datasets as an instruction-following task leads to strong generalization across tasks. Kosmos-2 [32] and PaLI-X [19] largely scaled the pre-training data by pseudo-labeling bounding boxes from performant open-vocabulary object detectors (GLIP [75] and OWL-v2 [76], respectively). They examined that strong perception capabilities such as object detection and OCR translate to better high-level reasoning tasks like visual question-answering (VQA). Our work expands the horizon of data-scaling through our self-augmenting paradigm. Instead of using VLMs directly for data synthesizing, specialist models with domain expertise such as object detection, spatial relation, OCR, and more have a large potential for participating in the data collection pipeline. In this

paper, We explore the potential of labeling the internet-scale image data with both self-augmentation techniques, and also a model specialized in spatial relations.

**Limitations.** Due to source constraints, we concentrate on the design of a self-augmented data curation pipeline and verify the 7B, 8B, and 40B models with 50M data. Larger models and more data can have the potential to lead to better VLM capabilities with self-augmenting abilities. We will address these aspects in future work.

## 5 Conclusions

This work has explored the techniques, insights, and benefits of using VLMs to self-improve its pre-training. We introduced two primary augmentation loops, one leveraging VLM’s general captioning capacities and the other harnessing their strength in specialized visual tasks. We demonstrated the feasibility of three ‘free lunch’ rounds for VLMs through self-bootstrapping, with further improvements via knowledge distillation from specialist VLMs. Our new VILA<sup>2</sup> models demonstrate SOTA performances across a comprehensive set of benchmarks. Fruitful future directions include a deeper delve into the potential synergy between synthetic and real data toward training stronger foundation models.

## References

- [1] OpenAI. Gpt-4v(ision) system card. 2023.
- [2] Google: Rohan Anil Gemini Team and 1344 other authors. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [3] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.
- [4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [5] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [7] OpenAI. Chatgpt: Optimizing language models for dialogue. <https://openai.com/blog/chatgpt>, 2023. Accessed: 2023.
- [8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [9] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [10] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca), 2023.

- [11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023.
- [12] Siddharth Karamcheti, Laurel Orr, Jason Bolton, Tianyi Zhang, Karan Goel, Avanika Narayan, Rishi Bommasani, Deepak Narayanan, Tatsunori Hashimoto, Dan Jurafsky, et al. Mistral—a journey towards reproducible language model training, 2021.
- [13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [14] Yi-34B large language model. <https://huggingface.co/01-ai/Yi-34B>, 2023.
- [15] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. Technical report, Alibaba Group, 2023. <https://arxiv.org/abs/2303.08774>.
- [16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023.
- [17] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [18] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
- [19] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023.
- [20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [21] Fuyu-8B: A multimodal architecture for AI agents. <https://www.adept.ai/blog/fuyu-8b>, 2023.
- [22] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [23] OpenAI. GPT-4 technical report. Technical report, OpenAI, 2023.
- [24] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [25] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2024.
- [26] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.
- [27] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.
- [28] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Sae-hoon Kim. Coyo-700m: Image-text pair dataset. <https://github.com/kakaobrain/coyo-dataset>, 2022.
- [29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

- [30] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.
- [31] Jang Hyun Cho, Boris Ivanovic, Yulong Cao, Edward Schmerling, Yue Wang, Xinshuo Weng, Boyi Li, Yurong You, Philipp Krähenbühl, Yan Wang, et al. Language-image models with 3d understanding. *arXiv preprint arXiv:2405.03685*, 2024.
- [32] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.
- [33] Author(s). Llama 3 model card. Online, 2024. Accessed: 2024-07-18.
- [34] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.
- [35] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.
- [36] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.
- [37] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models, 2024.
- [38] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.
- [39] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.
- [40] Jimmy Carter. Textocr-gpt4v. <https://huggingface.co/datasets/jimmycarter/textocr-gpt4v>, 2024.
- [41] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022.
- [42] Mishra Anand, Shekhar Shashank, Singh Ajeet, Kumar, and Chakraborty Anirban. Ocr-vqa: Visual question answering by reading text in images, 2019.
- [43] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020.
- [44] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024.
- [45] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.
- [46] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.
- [47] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

- [48] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.
- [49] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024.
- [50] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
- [51] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.
- [52] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- [53] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.
- [54] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.
- [55] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.
- [56] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- [57] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
- [58] Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models?, 2024.
- [59] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023.
- [60] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [61] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024.
- [62] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models, 2024.
- [63] Sensetime. SenseChat-Vision webpage. Technical report, sensenova, 2024.
- [64] InfiMM Team. Infimm: Advancing multimodal understanding from flamingo’s legacy through diverse llm integration, 2024.
- [65] Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023.
- [66] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. 2023.

- [67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [68] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [69] meta. Introducing meta llama 3: The most capable openly available llm to date. Technical report, Meta, 2024. <https://ai.meta.com/blog/meta-llama-3/>.
- [70] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.
- [71] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.
- [72] dbrx. Introducing dbrx: A new state-of-the-art open llm. Technical report, Databricks, 2024. <https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-lm>.
- [73] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv: Computer Vision and Pattern Recognition*, 2021.
- [74] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- [75] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.
- [76] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024.
- [77] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.
- [78] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325, 2017.
- [79] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.
- [80] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888, 2021.
- [81] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueling Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.
- [82] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021.
- [83] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

- [84] Jianhao Shen, Ye Yuan, Srbuhi Mirzoyan, Ming Zhang, and Chenguang Wang. Measuring vision-language stem skills of neural models, 2024.
- [85] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019.
- [86] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’21. ACM, July 2021.
- [87] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120, 2022.
- [88] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog, 2017.
- [89] Jack Hessel, Jena D. Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The abduction of sherlock holmes: A dataset for visual abductive reasoning, 2022.
- [90] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-llava: Solving geometric problem with multi-modal large language model, 2023.
- [91] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning, 2024.
- [92] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning, 2024.
- [93] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016.
- [94] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022.
- [95] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. Mammoth: Building math generalist models through hybrid instruction tuning, 2023.
- [96] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world’s first truly open instruction-tuned llm, 2023.
- [97] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023.

## A Appendix / supplemental material

### A.1 Prompts for Specialist-augmentation

We use the following prompts during specialist- augmentation,

- Spatial Relations Understanding Specialist  
" Elaborate on the visual and narrative elements of the image in detail, with a focus on spatial relations."
- Grounded Narration Specialist  
" Elaborate on the visual and narrative elements in the image, and specify their location with [xmin,ymin,xmax,ymax]."
- OCR Specialist  
" Your task is to recognize and describe the text in the image. Please provide a brief description that focuses on the textual content."

### A.2 Specialist Acquisition

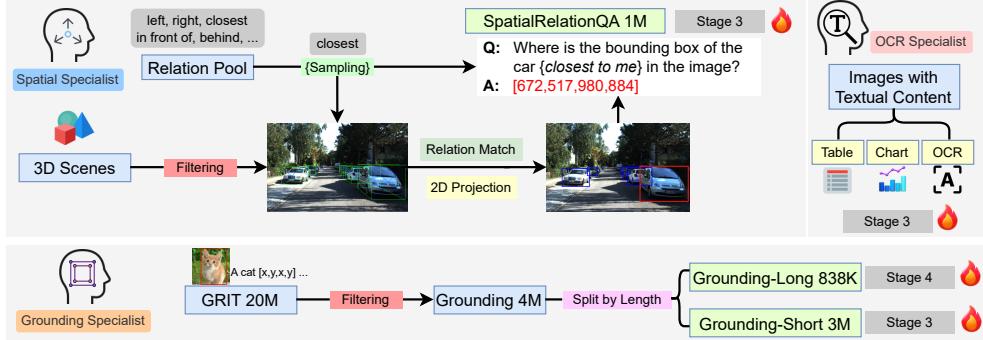


Figure 5: VILA<sup>2</sup> Specialist VLM Acquisition Pipeline. We gather task-specific knowledge from public datasets, followed by filtering noisy samples using rule-based strategies. We then train the specialist VLMs from pretrained checkpoints, employing different data blends and training strategies.

Specialty and expertise can be obtained via gathering existing data from the open-source community, human labeling, and annotating by domain-specific models, e.g. OCR models for text recognition and detection models for bounding box prediction. We also experimented with using open-world detectors like OWLv2[76] to automatically label bounding boxes, VLMs to generate detailed captions, and LLMs such as Llama3-70B-Instruct to merge the information for the grounded narration specialist. However, we found that language models introduced more hallucinations into the merged grounded narration. This is because many different detection labels can share the same meaning and refer to the same instance, making it difficult for the language model to perform the bipartite matching between bounding boxes and their text correspondence.

### A.3 SFT Data

We use two different datasets for our experiments: a 1.8M sample dataset for exploratory experiments and a 5.9M sample dataset for state-of-the-art experiments.

- **1.8M SFT Blend:** This dataset includes samples from the following sources: LLava-SFT, MSR-VTT, TextCaps, Image Paragraph Captioning, CLEVR, NLVR, VisualMRC, ActivityNet-QA, iVQA, MSRVTT-QA, MSVD-QA, DVQA, OCRVQA, ST-VQA, ViQuAE, VQAv2-train, Visual Dialog, GQA-train, FLAN-1M.
- **5.9M SFT Blend:** This dataset comprises all the datasets listed in the following table:

Categories	Datasets
Hybrid	LLaVA-SFT, The Cauldron (subset)
Captioning	MSR-VTT [77], TextCaps, Image Paragraph Captioning [78], LLaVAR, ShareGPT4V-100K
Reasoning	CLEVR [79], NLVR, VisualMRC [80]
Multi-images	ActivityNet-QA [81], iVQA [82], MSRVTT-QA, MSVD-QA [83], STEM-QA [84]
OCR	DVQA, OCRVQA, ST-VQA [85], SynthDoG-en, TextOCR-GPT4V, ArxivQA
World Knowledge	WIT [86]
General VQA	ViQuAE [87], VQAv2-train, Visual Dialog [88], GQA-train [51], ScienceQA-train, SHERLOCK [89], Geo170K [90], MMC-Instruction [91], LRV-Instruction [92], RefCOCO-train [93]
Text-only	FLAN-1M [94], MathInstruct [95], Dolly [96], GSM8K-ScRel-SFT [97]

Table 7: Data mixture for the SFT stage.

#### A.4 Training Detail

We adjust our training strategies akin to varying language model sizes for training cost considerations. We next elaborate on the details.

##### A.4.1 7B & 8B & 13B Models

We divide the entire training process of 7B&8B&13B models into three sub-stages.

- **Stage 1: Alignment Stage.** We train only the multi-modal projector using 595K image-text pairs, as mentioned in LLaVA, to achieve the initial alignment between the two modalities.
- **Stage 2: Pre-training Stage.** We gather a total of 51 million images, consisting of 25 million image-text pairs with the highest CLIP scores from COYO-700M, 25 million images in an interleaved image-text format from the MMC4-Core subset, and 1 million images with detailed captions from ShareGPT4V-Pretrain. During this stage, we unfreeze both the multi-modal projector and the language model to enhance comprehension of the diverse multi-modal training data. **We use the augmented data here to replace the original COYO captions.**
- **Stage 3: Supervised Fine-tuning Stage.** After stage 2, we collect diverse visual question-answer pairs and unfreeze all parameters to fine-tune the model for general-purpose VQA capacities.

##### A.4.2 40B Model

For the VILA<sup>2</sup>-40B model, we skip the cost-intensive stage 2 and train the model with 7.5 million images randomly sampled from the 25 million COYO subset pairing with various caption sources: 2.5 million with original COYO captions, 2.5 million with VILA<sub>3</sub> re-captioned descriptions, and 2.5 million with VILA<sub>3</sub> spatial specialist re-captioned descriptions. Both the multi-modal projector and the language model remain unfrozen. Note that adding interleaved data, such as MMC4, can further boost the performance and we leave this potential to future work.

#### A.5 Hyperparameters

We use a universal batch size of 1024, a cosine decay learning rate schedule, a 0.03 learning rate warmup ratio, no weight decay, and AdamW as the optimizer for stable training, and details are expanded in Table 8. All trainings are conducted with 128 A100 GPUs.

Hyperparameter	Stage 1	Stage 2	Stage 3	Stage 4
batch size	1024	1024	1024	1024
learning rate (lr)	1e-3	5e-5	2e-5	1e-4
lr schedule	cosine	cosine	cosine	cosine
lr warmup ratio	0.03	0.03	0.03	0.03
weight decay	0	0	0	0
epoch	1	1	1	1
optimizer	AdamW	AdamW	AdamW	AdamW
DeepSpeed stage	stage2	stage3	stage3	stage3

Table 8: The detailed training setup for VILA<sup>2</sup> and the hyper-parameters across the training stages.

## A.6 Additional Re-caption Comparisons

We provide additional VILA<sup>2</sup> recaptioned examples from the SAM dataset comparing the baseline captions from alternative methods, *e.g.*, the widely adopted InstructBLIP Flan-T5 XXL [59] and LLaVA-NeXT-34B [36], with captions generated through various rounds of VILA<sup>2</sup>'s intermediate models in Figure 6 through Figure 8. The correct facts are marked with green, hallucinations are marked with red and spatial related information are marked with blue. From these examples, we can see that while the baseline caption is descriptive, it contains more hallucinations. As the re-captioning and training cycle progresses, the model tends to generate captions with more details and fewer hallucinations. Additionally, the spatial specialist introduces more spatial information into the recaptions, as expected.

<b>InstructBLIP:</b> The image depicts a train traveling along a hillside near a stone tower, which is visible in the background of the image. The train appears to be moving at a steady pace, and it can be seen passing by the tower several times during its journey. There are several people present in the scene, including a man standing next to the train, a woman sitting on the side of the train, and a young boy walking behind the train. All of these individuals appear to be enjoying the scenery and taking in the surrounding environment.	
<b>Self-Augmented Round-1:</b> The image captures a scene of a stone tower standing tall against a backdrop of a clear blue sky dotted with white clouds. The tower, with its red roof and white walls, exudes an air of antiquity. It's perched on a grassy hill, adding to its grandeur. A flagpole rises from the tower, reaching towards the sky. In the foreground, a train engine painted in vibrant shades of yellow and green adds a touch of modernity to the otherwise historical setting. The perspective of the image places the viewer at a lower vantage point, looking up at the tower, enhancing its imposing presence.	
<b>Self-Augmented Round-3:</b> The image captures a scene of a green train parked on a track, with a brick tower standing tall in the background. The tower, constructed from red bricks, features a pointed roof and a flagpole on top, adding a touch of historical charm to the scene. The sky above is a clear blue, dotted with white clouds, creating a serene backdrop for the tower. The perspective of the image is from the ground, looking up at the tower, giving it a sense of grandeur and prominence.	
<b>Specialist-Augmented:</b> The image captures a scene of a green train car, adorned with a yellow stripe, making its way through a verdant landscape. The train car, positioned on the left side of the frame, is moving towards the right, suggesting a journey in progress. On the right side of the image, a stone tower stands tall. It's a two-story structure, topped with a red roof that contrasts with the surrounding greenery. The tower is situated on a grassy hill, adding a sense of depth to the scene. The sky above is a clear blue, dotted with white clouds, providing a serene backdrop to the scene. The train car and the tower are the main subjects of the image, their contrasting colors and shapes creating a visually interesting composition. The train car's movement adds a dynamic element to the otherwise tranquil setting. Overall, the image presents a harmonious blend of man-made structures and natural elements, each contributing to the narrative of the scene.	

Figure 6: More examples of comparison among captions from generation rounds of VILA<sup>2</sup> and the widely used caption baseline of InstructBLIP [59].

**InstructBLIP:** The image depicts a crowd of people watching an old-fashioned car parade on a cobblestone street. There is a white vintage car in the foreground, while several other cars can be seen scattered throughout the scene. Some of the vehicles are placed closer to the center of the road, while others are spread out over a larger area. Overall, the scene offers a sense of nostalgia and celebration with the presence of these classic vehicles.



**Self-Augmented Round-1:** The image captures a vibrant scene on a cobblestone street. Dominating the frame is a white vintage car, its black roof contrasting with its body. The car is parked, its front facing the camera, as if ready to embark on a journey. The license plate, reading "W22-221", is clearly visible, hinting at the car's unique identity. In the background, a yellow car adds a splash of color to the scene, parked further down the street. The cobblestone street, lined with buildings, extends into the distance, creating a sense of depth and perspective. People can be seen walking on the sidewalk, going about their day, oblivious to the vintage charm of the cars. Their presence adds a touch of everyday life to the scene, grounding the image in reality. The image is a blend of the old and the new, the vintage car standing out against the modern cityscape. It's a snapshot of a moment, frozen in time, telling a story of a city where the past and present coexist.

**Self-Augmented Round-3:** The image captures a vibrant scene on a cobblestone street. Dominating the frame is a white vintage car, its black roof contrasting with its body. The car is parked, its front facing the camera, as if ready to embark on a journey. The license plate, reading "W22-221", is clearly visible, hinting at the car's unique identity. In the background, a yellow car adds a splash of color to the scene, parked further down the street. The cobblestone street, lined with buildings, extends into the distance, creating a sense of depth and perspective. People can be seen walking on the sidewalk, going about their day, oblivious to the vintage charm of the cars. Their presence adds a touch of everyday life to the scene, grounding the image in reality. The image is a blend of the old and the new, the vintage car standing out against the modern cityscape. It's a snapshot of a moment, frozen in time, telling a story of a city where the past and present coexist.

**Specialist-Augmented:** The image captures a scene on a city street. Dominating the frame is a white vintage car, its shiny exterior reflecting the light. The car is parked on a brick pavement, its chrome bumper and grille gleaming under the sunlight. The car's design is reminiscent of a bygone era, adding a touch of nostalgia to the modern cityscape. In the background, a yellow car can be seen, its bright color contrasting with the white of the vintage car. The yellow car is parked further down the street, ready to ferry passengers to their destinations. The street itself is lined with buildings, their architecture hinting at the urban setting. A large building with a glass facade stands out in the background, its modern design juxtaposing the vintage charm of the car. The image is taken from a low angle, looking up at the car and the buildings. This perspective gives a sense of grandeur to the scene, making the car and the buildings appear larger and more imposing. Overall, the image presents a blend of old and new, nature and architecture, stillness and motion. It's a snapshot of a moment in a city, where the past and present coexist.

Figure 7: More examples of comparison among captions from generation rounds of VILA<sup>2</sup> and the widely used caption baseline of InstructBLIP [59].

**InstructBLIP-FLAN-T5-XXL:** The image depicts a large swimming pool in the middle of a lush green landscape. There are several lounge chairs placed around the pool, and there is also a group of people lounging near the edge of the pool. A number of trees can be seen surrounding the pool, providing a picturesque backdrop for the scene.



**LLaVA-NeXT-34B:** The image depicts a serene and well-maintained swimming pool area. The pool itself is rectangular with a curved edge on one side, suggesting a design that accommodates both swimming and lounging. The water appears clear and inviting, reflecting the bright sunlight. In the foreground, there are several sun loungers, arranged in a semi-circle around the pool, inviting relaxation and sunbathing. The loungers are empty, which adds to the tranquil atmosphere of the scene. The pool is surrounded by a stone deck, which provides a solid and elegant base for the pool area. The deck's design complements the pool's shape, creating a harmonious visual flow. In the background, there is a large, well-manicured lawn that extends to a line of trees, providing a natural boundary and a sense of privacy. The lawn is lush and green, indicating that the image was likely taken during a warm season. Above the lawn, there is a large, white gazebo with a red roof, which stands out against the greenery. The gazebo's presence suggests a place for social gatherings or perhaps a spot for a quiet read. The overall spatial relationship of the elements in the image is balanced and inviting. The pool is centrally located, drawing the eye, while the sun loungers and gazebo are positioned to offer different experiences for visitors. The lawn and trees provide a sense of openness and connection to nature, while the stone deck and gazebo offer more structured spaces for relaxation. The image does not contain any people, which gives it a peaceful and undisturbed quality. The absence of human activity allows the viewer to focus on the design and layout of the pool area. The sky is clear and blue, suggesting a sunny day, which is ideal for outdoor activities. The sunlight illuminates the scene, casting soft shadows and highlighting the textures of the stone and the smoothness of the water. Overall, the image presents a luxurious and relaxing outdoor space, designed with attention to detail and a focus on comfort and aesthetics. The spatial relationships between the various elements create a harmonious and inviting environment.

**VILA^2-7B:** The image captures a serene outdoor setting, dominated by a large, light blue pool that sits in the center. The pool is encircled by a white wall, which is adorned with a series of small waterfalls, adding a soothing ambiance to the scene. On the right side of the pool, a red umbrella stands out against the natural backdrop, providing a spot of color and shade. The pool area is not devoid of life, with a few people enjoying the tranquil environment. Some are lounging on the poolside, while others are taking a dip in the water. The pool is nestled amidst lush greenery, with trees and bushes enveloping the area, creating a sense of seclusion and tranquility. The sky above is a clear blue, suggesting a bright, sunny day. The overall scene is one of relaxation and leisure, a perfect depiction of a day spent by the pool.

Figure 8: More examples of comparison among captions from generation rounds of VILA<sup>2</sup> and an additional caption baseline of LLaVA-NeXT [36]. LLaVA-NeXT-34B tends to generate longer caption with more hallucinations.