A Single Transformer for Scalable Vision-Language Modeling

Anonymous authors
Paper under double-blind review

Abstract

We present SOLO, a single transformer for **S**calable visi**O**n-**L**anguage m**O**deling. Current large vision-language models (LVLMs) such as LLaVA mostly employ heterogeneous architectures that connect pre-trained visual encoders with large language models (LLMs) to facilitate visual recognition and complex reasoning. Although achieving remarkable performance with relatively lightweight training, we identify four primary scalability limitations: (1) The visual capacity is constrained by pre-trained visual encoders, which are typically an order of magnitude smaller than LLMs. (2) The heterogeneous architecture complicates the use of established hardware and software infrastructure. (3) Study of scaling laws on such architecture must consider three separate components — visual encoder, connector, and LLMs, which complicates the analysis. (4) The use of existing visual encoders typically requires following a pre-defined specification of image inputs pre-processing, for example, by reshaping inputs to fixed-resolution square images. This inflexibility can create bottlenecks and impede scalability. A unified single Transformer architecture, like SOLO, effectively addresses these scalability concerns in LVLMs; however, its limited adoption in the modern context likely stems from the absence of reliable training recipes that balance both modalities and ensure stable training for billion-scale models. In this paper, we introduce the first open-source training recipe for developing SOLO, an open-source 7B LVLM with the single Transformer architecture using moderate academic resources (8 x A100 80GB GPUs). The training recipe involves initializing from LLMs, sequential pre-training on ImageNet and web-scale data, and instruction fine-tuning on our curated high-quality datasets. On extensive evaluation, SOLO demonstrates performance comparable to LLaVA-v1.5-7B, particularly excelling in visual mathematical reasoning.

1 Introduction

Large vision-language models (LVLMs) demonstrate remarkable performance on downstream tasks (Li et al., 2023c; Zhu et al., 2023; Liu et al., 2023c; Chen et al., 2023c; Kim & Ji, 2024). They can effectively extract visual information (Wang et al., 2023b) and follow human instructions to generate insightful responses (Li et al., 2023b; Chen et al., 2024d). Two established approaches for vision-language modeling include: (1) Connecting pre-trained visual encoders (Dosovitskiy et al., 2021b; Radford et al., 2021) and large language models (LLMs) (Touvron et al., 2023; Jiang et al., 2023) via a learned projection module that maps the *visual embeddings* to the embedding space of LLMs (Dai et al., 2023; Gao et al., 2023; Liu et al., 2023c), or an intermediate symbolic layer Wang et al. (2024c). (2) Leveraging a pre-trained visual encoder to extract features and aligning feature embeddings with a pre-defined codebook (Esser et al., 2021) to convert each image into a sequence of *discrete visual tokens*, thus enabling LVLMs to process both images and language tokens (Wang et al., 2022b; Peng et al., 2022; Anil et al., 2023; Team, 2024; Diao et al., 2023).

However, despite their effectiveness, these approaches have limitations that make them hard to scale. We consider an architecture scalable when it can demonstrate sustained performance improvement through scaling (*e.g.,* more computational resources and/or data) without hitting any inherent bottleneck in the machine learning system. The bottleneck in prevalent LVLMs is evident across four dimensions (§2.1), primarily due to their reliance on a pre-trained visual encoder:

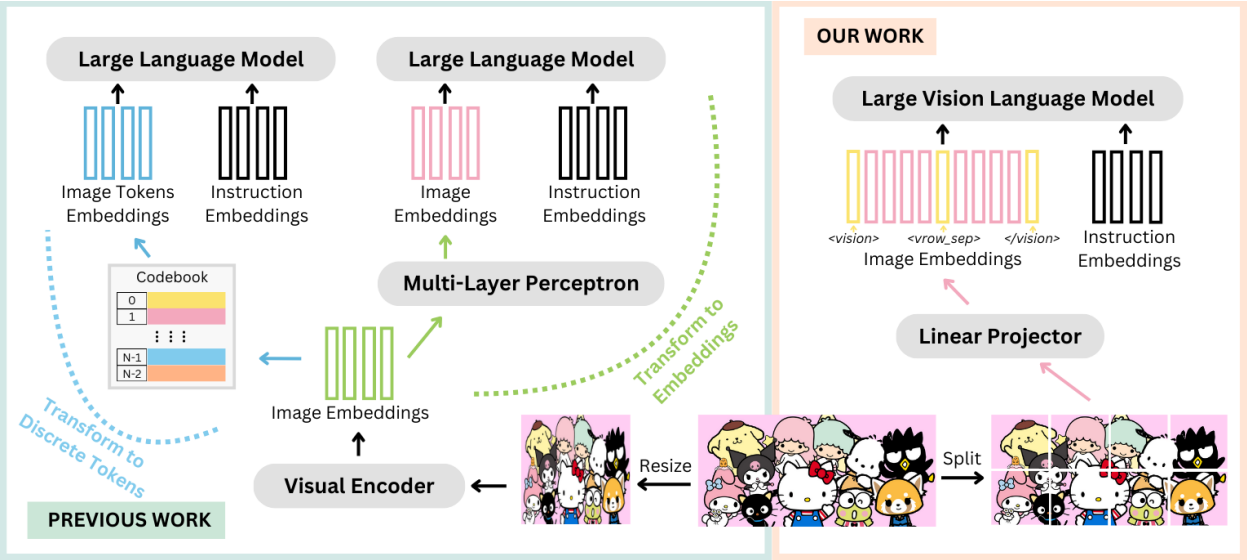


Figure 1: (***Previous work***) The mainstream approaches for vision-language modeling rely on pre-trained visual encoders for visual feature extraction, which exhibits scalability limitations. (***Our work***) We advocate for a unified transformer architecture that processes both images and text, employing a simple linear projection to directly handle raw image pixels. `<vision>`, `</vision>`, and `<vrow_sep>` are special tokens designed explicitly for visual modality encoding.

(1) **Constrained visual capabilities:** The visual capacities of a pre-trained vision encoder are largely pre-determined and limited by the data distribution and volume used during pre-training. Due to the significantly smaller size of visual encoders—approximately over ten times smaller than LLMs—they can be a performance bottleneck in solving complex vision-language tasks.

(2) **Challenges in large-scale training and deployment:** The heterogeneous architecture of LVLMs with vision encoders poses challenges in adapting existing training and inference frameworks and utilizing existing hardware, which is mostly tailored and optimized for single Transformer architectures.

(3) **Multiple components complicate the scaling analysis:** The analysis of scaling laws, which are crucial for the development of foundation models, is complicated by the necessity to consider the size of several distinct components independently: the visual encoder, the connector, and the LLMs.

(4) **Limited image pre-processing flexibility:** Most vision encoders pre-define a specification on *how* image inputs should be pre-processed. For example, the widely used visual backbones, such as CLIP-ViT-336 (Radford et al., 2021), require a square image input with a resolution of 336×336. The inflexibility of image pre-processing can cause bottlenecks that hinder scalability.

To address these limitations, we present `SOLO`, which employs a **single Transformer architecture for unified and end-to-end vision-language modeling**. `SOLO` accepts both raw image patches (in *pixels*) and texts as inputs, without using a separate pre-trained vision encoder (Fig. 1). This simplifies the model design and enhances the scalability and adaptability of the LVLM architecture. By simplifying from multi-component LVLM to a single Transformer model, this architecture is unconstrained on the capabilities of visual encoders, easier to train and deploy using existing hardware and software, allows more straightforward scaling law analysis, and can easily scale to image data with diverse resolutions and aspect ratio. `SOLO`, with a 7-billion parameter count, is initialized from Mistral LLM v0.1 (Jiang et al., 2023) and leverages its extensive pre-trained knowledge.

This modeling strategy is inspired by the foundational modeling framework of VisualBERT (Li et al., 2019) and industry efforts to scale unified LVLMs to the billion-scale (Bavishi et al., 2023). Despite the simplicity and scalability, its limited contemporary adoption can be attributed to the lack of reliable training recipes, as balancing vision and language modalities in unified LVLMs often leads to training divergence. This paper details the first open-source recipe for developing scalable unified LVLMs, using modest academic computational resources, specifically 8 NVIDIA A100 80GB GPUs (§3). Our training recipe

involves initializing with pre-trained LLMs, sequential pre-training on ImageNet and web-scale datasets, and instruction fine-tuning on our curated high-quality data mixture.

While still lags behind recent state-of-the-art LVLMs on evaluation benchmarks, SOLO exhibits performance on par with LLaVA-v1.5-7B (§4) and the variant LLaVA-7B* (§6), which is created following our training recipe in the controlled setting. In particular, SOLO distinguishes itself in the domain of visual mathematical reasoning. Further scalability analysis reveals better scaling behaviors of SOLO and the scalability of our flexible image preprocessing pipeline (§6.2). In addition, through comprehensive ablation studies, we validate the design choices of our training recipe. Our empirical results confirm that the sequential pre-training on ImageNet and web-scale datasets and instruction fine-tuning on our carefully curated data mixture are both essential for the training of such single Transformer LVLMs (§5). Interestingly, we find that after removing the first stage of pre-training on ImageNet, the LVLM will produce outputs of drastically different quality while exhibiting similar image-conditioned language modeling loss (§5.1, Fig. 3).

2 Tackling Scalability Limitations via Integrated Architectures

2.1 Scalability Limitations in Existing LVLMs

The scalability constraints of existing LVLMs are currently articulated from four critical perspectives that limit their efficiency in utilizing expanded computational resources and larger datasets due to the bottlenecks in the system design:

Fixed and Constrained Visual Capabilities The fixed nature of visual encoders severely limits the adaptability of LVLMs to novel visual data distribution and more complex vision-language tasks since these encoders are trained on specific distributions and training objectives. Current approaches address this issue by continuing the training of visual encoders (Bai et al., 2023) or by integrating features derived from various visual encoders (Lin et al., 2023). Nonetheless, the scope of data used for continued pre-training is substantially less than that used initially, which only marginally enhances encoder adaptability, and employing multiple encoders complicates the process of image feature extraction, thereby impeding the scalability of LVLMs. Moreover, the smaller scale of visual encoders compared to LLMs frequently results in the visual understanding component becoming a bottleneck. Consequently, visual representation learning is limited to the smaller visual encoders, hindering the full utilization of LLM capabilities in existing LVLMs.

Challenges in Large-Scale Training and Deployment The heterogeneous architecture with multiple components complicates the implementation of machine learning systems for large-scale training and deployments. (1) **Training challenge**: At large training scale (*i.e.,* multi-node clusters), it is necessary to distribute not only the Transformer-based LLMs but also the vision model and the MLP connector across multiple devices, employing techniques such as tensor and pipeline parallelism. Thus, prevalent LVLMs cannot directly use existing industry-grade training frameworks optimized for the Transformer architecture (Shoeybi et al., 2019; Cano et al., 2023), thus necessitating the development of new tensor-sharding mechanisms. In addition, AI alignment typically employs algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017), which necessitate simultaneously maintaining multiple models (*e.g.,* reward and critic models) in GPU memory and cause difficulty in the algorithm implementations for heterogeneous architectures. (2) **Deployment**: The heterogeneous architecture complicates the deployment process due to similar model and tensor sharding challenges described above. Consequently, this hampers the large-scale services of existing LVLMs. Moreover, existing specialized AI chips (Techcrunch) and inference libraries, such as vLLM (Kwon et al., 2023) and MLC-LLM (team, 2023), are mostly optimized for Transformer architectures, presenting significant challenges in the deployment of these models on end devices.

Multiple Components Complicate the Scaling Analysis The complexity introduced by the multiple components of LVLMs is a significant barrier to understanding and improving these systems. Each component —the visual encoder, the connector, and the language models—operates with its own parameters and training strategies (Radford et al., 2021; 2019; Brown et al., 2020a), which can lead to a lack of cohesion in the overall model behavior. Scaling laws are crucial for guiding the development of large foundational models by

forecasting the performance of a target model using data from several sampled models that are significantly smaller in sizes (Kaplan et al., 2020; Bahri et al., 2021). However, applying these approaches to existing LVLMs requires simultaneous consideration and scaling of various components, hereby increasing complexity.

Limited Image Pre-Processing Flexibility The strict requirements for image pre-processing imposed by the specifications of visual encoders may create bottlenecks that hinder the scalability. For instance, the requirement of a consistent input resolution can make it difficult to process images that are naturally high-resolution or have non-standard aspect ratios without compromising on the quality or representational fidelity of the input. Current mitigation strategies involve splitting the original image into multiple sub-images, independently extracting visual features from each sub-image using pre-trained visual encoders and subsequently aggregating the representation embeddings (Xu et al., 2024a; Liu et al., 2024a; Dong et al., 2024). However, these approaches seem ad-hoc and can be suboptimal as the visual backbone is not pre-trained to handle these inputs, potentially impacting the effective handling of these high-resolution images.

2.2 Unified Vision-Language Modeling with Integrated Architectures

We revisit the foundational modeling framework of VisualBERT (Li et al., 2019), initially proposed in the early stages of research on pre-trained vision-language models. The key idea is to use one single Transformer, initialized from BERT (Devlin et al., 2018) in VisualBERT, to uniformly process the image patches and language tokens. Fuyu-8B exemplifies the industry's effort to scale this modeling approach (Li et al., 2019) to billion-scale models (Bavishi et al., 2023). However, the limited widespread implementation of this unified architecture may be due to the lack of an established training recipe, as only the pre-trained model is released by Bavishi et al. (2023) without training details. Training such unified LVLMs presents significant challenges in balancing the two modalities and maintaining stable training, for which clear solutions are currently lacking. In this paper, we present SOLO with full details of its unified and integrated architecture design and training recipe.

3 SOLO: Scalable Vision-Language Modeling

SOLO consolidates image and language capabilities into a single model, enables data-driven determination of visual representations and parameter allocation across visual and language modalities, simplifies the scaling laws analysis, and allows it to handle high-resolution images and those with uncommon aspect ratios flexibly. For large-scale training (§3.2), SOLO also seamlessly integrates with established software frameworks for large-scale Transformer pre-training (Shoeybi et al., 2019).

3.1 Model Architecture

The architecture of SOLO is shown in Fig. 1, which diverges from earlier models primarily in the extraction of visual features. Instead of resizing the image into a fixed resolution adapted to the pre-trained image encoders, SOLO keeps their original resolutions and aspect ratios. The feature extraction involves splitting the image into patches with a pre-defined size. Through a trainable linear projection, these raw image patches (in *pixels*) are transformed to obtain continuous embeddings that represent the visual features of the images. Thus, we can integrate image and language processing within a single model. We maintain a list of special tokens designed explicitly for visual modality encoding: `<vision>` and `</vision>` tokens mark the beginning and end of a span of image patches respectively; `<vrow_sep>` acts as a row separator within the image patches and helps the model distinguish between different rows of image patches, aiding in structured visual understanding.

```python
PATCH_SIZE = 32
MAX_RESOLUTION = 1024  # 32 x 32

def get_resize_output_image_size(image_size):
    l1, l2 = image_size
    if l2 <= l1:
        short, long = l2, l1
    else:
        short, long = l1, l2

    requested_new_long = min(
        int(long / PATCH_SIZE + 1) * PATCH_SIZE,
        MAX_RESOLUTION
    )
    new_long = requested_new_long
    new_short = int(new_long * short / long)
    new_short = int(new_short / PATCH_SIZE + 1)
        * PATCH_SIZE

    if l2 <= l1:
        return new_long, new_short
    else:
        return new_short, new_long
```

Figure 2: The input image resize algorithm to maintain the aspect ratio.

Table 1: Summary of datasets used in the three stages of pre-training. Each image patch counts as a vision token. Number of *instances* is calculated after packing them into sequences with 32K length.

Training Stage	Dataset	#Instances	#Image	#Token	#Text Tokens	#Vision Tokens
Stage-1	ImageNet21K (Ridnik et al., 2021b)	74,283	13,151,276	2,423,203,108	212,745,573	2,210,457,535
	SlimPajama, *subset* (Soboleva et al., 2023)	120,839	0	4,340,877,587	4,340,877,587	0
	Total	195,122	-	6,764,080,695	(67.32%) 4,553,623,160	(32.68%) 2,210,457,535
Stage-2	Capfusion (subset) (Yu et al., 2024)	204,978	23,681,864	6,664,351,863	1,172,726,505	1,172,726,505
	Websight (Laurençon et al., 2024)	71,579	1,922,671	2,300,945,215	1,087,060,511	1,213,884,704
	CC3M (Sharma et al., 2018b)	32,760	2,331,439	1,064,477,314	76,092,147	988,385,167
	Detailed Captions (lz)	6,225	368,767	202,016,770	44,788,200	157,228,570
	LLaVAR (Zhang et al., 2023b)	3,602	422,315	117,448,784	31,390,556	86,058,228
	DVQA (Kafle et al., 2018)	2,917	200,000	94,853,796	55,653,796	39,200,000
	OCR-VQA (Mishra et al., 2019)	1,593	165,746	51,920,705	21,161,018	30,759,687
	FigureQA (Kahou et al., 2017)	1,526	100,000	49,586,305	24,803,256	24,783,049
	SlimPajama, *a different subset* (Soboleva et al., 2023)	120,385	0	4,300,998,161	4,300,998,161	0
	Total	445,565	-	14,846,598,913	(45.90%) 6,814,674,150	(54.10%) 8,031,924,763
Stage-3	ALLaVA-LAION (Chen et al., 2024a)	13,725	438,992	442,509,490	176,660,898	265,848,592
	ALLaVA-VLFLAN (Xu et al., 2024b)	4,469	207,549	144,577,377	77,835,919	66,741,458
	LLaVAR (Zhang et al., 2023b)	3,602	422,315	117,448,784	31,390,556	86,058,228
	DVQA (Kafle et al., 2018)	2,917	200,000	94,853,796	55,653,796	39,200,000
	FigureQA (Kahou et al., 2017)	1,526	100,000	49,586,305	24,803,256	24,783,049
	SlimPajama, *a different subset* (Soboleva et al., 2023)	12,085	0	430,688,442	430,688,442	0
	Total	38,324	-	1,279,664,194	(62.28%) 797,032,867	(37.72%) 482,631,327

Formally, we define the patch size P and the maximal resolution M. For an image of dimension size (W, H), it is resized to (W', H') to ensure divisibility by P. Fig. 2 details the resizing process, which adjusts the W and H to the nearest multiples of P while preserving the original aspect ratio to the extent possible and complying with the constraints imposed by M. Subsequently, the image is divided into N patches, where $N = (W'/P) \times (H'/P)$, each with dimensions $P \times P \times 3$. A trainable linear projector then maps each patch from a flattened $P \times P \times 3$ vector to an output dimension compatible with the embedding space of LLMs, extracting N embeddings as the image's feature representation. These visual embeddings, along with special visual modality tokens and embeddings of the text tokens, are concatenated and processed through a single Transformer, facilitating unified vision-language modeling. Notably, compared to prevalent LVLMs, this modeling strategy facilitates a much earlier fusion of visual and language modalities, allowing LVLMs to extract relevant information conditioned on the given instructions. In our implementation, we initialize `SOLO` from the Mistral-7B-v0.1 base LLM. The max resolution M of processed images is set as 1024. The patch size P is set as 32.

3.2 Training Recipe

We describe our approach for training unified billion-scale LVLMs, including pre-training (§3.2.1) and instruction fine-tuning (§3.2.2). For both stages, we optimize exclusively the language modeling loss on natural language tokens, without optimizing loss on image patches and special image tokens (*e.g.,* `<vision>`). We substantiate the essential ingredients in our recipe in §5.

3.2.1 Pre-Training

We introduce a three-stage pre-training curriculum that progressively enhances the visual capabilities of LVLMs while preserving their fundamental language capabilities. We present datasets and their statistics for each stage in Tab. 1.

Stage-1 ImageNet Pre-Training for Initialization We leverage ImageNet21K (Ridnik et al., 2021a), encompassing a broad spectrum of fine-grained visual categories, for the initial pre-training stage. In this process, we train `SOLO` to predict *only* fine-grained labels in natural language tokens (class name of images, *e.g.,* "golden retriever") conditioned on the image patches, thereby developing visual representations that initialize subsequent pre-training runs. In §5, we demonstrate the critical role of this stage in training unified LVLMs: when this stage is removed, the LVLM pre-trained on web-scale data from stage 2 (*e.g.,* captioning) failed to generate meaningful captions (Fig. 4).

Stage-2 Pre-Training on Web-Scale Data ImageNet21K, composed chiefly of visual concept data annotated by humans, faces scalability constraints in both knowledge breadth and data volume. In Stage-2, we scale up the pre-training data to encompass web-scale data, primarily consisting of image-caption pairs from sources like Capfusion (Yu et al., 2024) and CC3M (Sharma et al., 2018a). Additionally, we include synthetically generated web pages with associated HTML code from Websight (Laurençon et al., 2024) to improve OCR performance, and we also include a small set of supervised datasets to improve the data diversity. In this stage, the language modeling loss is applied uniformly across all language tokens, encompassing captions, HTML code, and questions and responses within the supervised datasets.

Stage-3 Annealing Following MiniCPM (Hu et al., 2024a), we perform a final annealing stage to conclude the pre-training. In this stage, we incorporate a limited selection of supervised datasets–either down-sampled or omitted from the instruction fine-tuning dataset mixture (*e.g.,* ALLaVA, Chen et al. 2024a)–to prime `SOLO` for the subsequent instruction fine-tuning stage. The primary purpose of this stage is to transition `SOLO` from a noisy web data to being trained on high-quality data mixtures.

Balancing Text and Vision Capability Through Language Corpus Blending Initiating with a base LLM and performing full-parameters training necessitates carefully preserving its inherent language comprehension abilities while performing image representation learning since most real-world vision-language tasks require text-only capabilities such as instruction comprehension and complex reasoning. At each stage of `SOLO`'s pre-training, we mix in a non-trivial proportion of text-only pre-training data (SlimPajama, Soboleva et al. 2023) to maintain the text capability. We present more empirical results on how data mixture affects image and text loss trade-offs in §F.

Pre-Training Infrastructure We modify the standard Megatron-LLM (Cano et al., 2023) to support arbitrary image patch inputs. We use one node with 8 NVIDIA A100 80G GPU for pre-training. We use 2-way tensor parallelism (Shoeybi et al., 2019) and 4-way data parallelism for training. Following Shoeybi et al. (2019), we adopt distributed optimizer to shard optimizer states across different GPUs for memory efficiency. We obtain a training throughput of 20k tokens per second.

Training Hyperparameter We use a global batch size of 128 examples (*i.e.,* 4M tokens) and each pre-training example is packed to $32,768$ tokens. We adopt a learning rate of 5e-5 with cosine decay to a minimum learning rate of 5e-6 and warm up for 200 steps. We use weight decay of 0.1. For training efficiency, we pack shorter sequences into one longer sequence and re-adjust the attention mask to make sure tokens from different examples cannot attend to each other. The training process consists of 1525 steps in Stage 1, 3480 steps in Stage 2, and 300 steps in Stage 3.

3.2.2 Instruction Fine-Tuning

Dataset Curation We meticulously select a diverse range of supervised datasets to perform instruction fine-tuning, aiming to enhance their performance across various domains of vision-language tasks. Our dataset selection strategy is based on the empirical analysis derived in Laurençon et al. (2024); Lin et al. (2024); Lu et al. (2024), and is mainly driven by the objective to cover a comprehensive range of data types, including language-only data, detailed image captions, scientific documents, tables, documents, charts, OCR and text-rich images, and general visual question-answering (VQA) tasks. In §A, we present datasets and their statistics in Tab. 3 and more details regarding data curation.

Implementation Details To efficiently conduct experiments with various data mixtures for iterative testing, we utilize DeepSpeed (Rasley et al., 2020), as implemented in Accelerate (Gugger et al., 2022), for instruction fine-tuning. The global batch size is configured at 640, with a weight decay parameter of 0.1. We train for 1 epoch with a maximum learning rate of 1e-5, which follows a linear warm-up phase and transitions to a cosine decay schedule.

Table 2: The main experimental results of SOLO. We compare with LVLMs with diverse capabilities, released at different times and categorized into three levels. SOLO aligns with the second level of prior LVLMs advancements focusing on LLaVA-Style modeling, and distinguishes itself in visual mathematical reasoning. (C) denotes the visual encoders from CLIP.

Level	Model	Visual	Language	MMStar	MME	SEED	ScienceQA	MathVista	AI2D
Level-1	OpenFlamingo v2	(C) ViT-L/14	MPT-7B	26.9	607.2	28.8	44.8	18.6	31.7
	MiniGPT-4-v2	EVA-G	Llama2-13B	21.3	968.4	29.4	54.7	23.1	30.5
	VisualGLM	EVA-CLIP	ChatGLM-6B	5.9	738.1	47.0	56.1	21.9	41.2
	InstructBLIP	EVA-G	Vicuna-7B	32.7	1391.4	44.5	54.1	24.4	40.6
	LLaVA-v1-7b	(C) ViT-L/14	Llama-7B	27.1	1075.5	50.4	61.8	25.2	48.3
Level-2	LLaVA-v1.5 7b	(C) ViT-L/14	Vicuna-V1.5-7B	33.1	1808.4	65.8	69.2	25.6	55.5
	mPLUG-OWL v2	(C) ViT-L/14	Llama2-7B	34.8	1786.4	64.5	69.5	25.4	55.7
	XComposer	EVA-G	InternLM-7B	6.9	1874.2	66.1	89.8	29.8	56.9
	MiniGPM-V	SigLIP-400M	MiniCPM-2.4B	38.6	1650.2	65.6	77.0	30.6	56.3
Level-3	Monkey	ViT-BigHuge	Qwen-7B	37.0	1759.9	64.3	72.1	33.5	62.5
	LLaVA-Next	(C) ViT-L/14	Mistral-7B	38.4	1821.2	72.4	73.0	34.6	69.0
	MiniCPM-v2	SigLIP-400M	MiniCPM-2.4B	39.1	1808.2	67.1	80.7	39.8	62.9
	DeepSeek-VL	Hybrid	DeepSeek-7B	40.5	1765.4	70.1	80.9	36.9	65.3
Ours	SOLO	Mistral-7B		35.5	1260.0	64.4	73.3	34.4	61.4

4 Comparison to Existing LVLMs

4.1 Model

We select various open-source LVLMs for comparison to better understand the capabilities of SOLO. Based on the release time and capabilities of LVLMs, we select 3 groups of LVLMs to better understand the current development phase of SOLO. Level-1 LVLMs represent the pioneering generation, which initiate the integration of visual encoders with pre-trained LLMs, with releases prior to October 2023. Level-2 LVLMs, released before early 2024, typically feature a more refined selection of instruction fine-tuning data to enhance performance. Level-3 marks the state-of-the-art (SoTA) LVLMs, released within the last five months, incorporating advanced training recipes, superior LLM backbones, and support for high-resolution images.

- **Level-1**: (1) OpenFlamingo v2 (Awadalla et al., 2023), (2) MiniGPT4 v2 (Chen et al., 2023a), (3) VisualGLM (Du et al., 2022), (4) InstructBLIP (Dai et al., 2023), (5) LLaVA v1 (Liu et al., 2023a).
- **Level-2**: (6) LLaVA v1.5 (Liu et al., 2024a), (7) mPLUG-Owl v2 (Ye et al., 2024), (8) InternLM-XComposer (Zhang et al., 2023a), (9) MiniCPM-v1 (Hu et al., 2023),
- **Level-3**: (10) Monkey (Li et al., 2024b). (11) LLaVA-NEXT (Liu et al., 2024b), (12) MiniCPM-v2 (Hu et al., 2024b), (13) DeepSeek-VL (Lu et al., 2024).

Each LVLM may have multiple variants based on different LLM sizes and architectures. If possible, we opt for the variant equipped with a 7B Mistral LLM. For the remaining LVLMs, we select the variant whose configuration most closely aligns with our specifications (Mistral-7B-LLM). We directly present the evaluation results of existing LVLMs from the leaderboard (OpenCompass) when available, to ensure a fair comparison.

4.2 Benchmarks

We select a wide range of benchmarks, encompassing both general vision-language tasks and specific task-oriented datasets, for evaluation and analysis. For general vision-language capability evaluation, we choose MMStar (Chen et al., 2024b), MME (Fu et al., 2024), and SEED-Bench (Li et al., 2024a). Specifically, MMStar measures elite vision-indispensable capabilities, MME measures both the perception and cognition capabilities, and SEED-Bench covers 12 evaluation dimensions covering various aspects of LVLMs capabilities. For scientific document understanding, we choose AI2D (Kembhavi et al., 2016) and ScienceQA (Lu et al., 2022a). For visual mathematical reasoning, we choose MathVista (Lu et al., 2023). We adopt VLMEvalKit (Contributors, 2023) to perform the unified evaluation.

4.3 Results

The experimental results are shown in Tab. 2. We find that SOLO significantly outperforms Level-1 LVLMs and also performs comparably to Level-2 LVLMs, despite slightly underperforming Level-3 LVLMs. Furthermore,

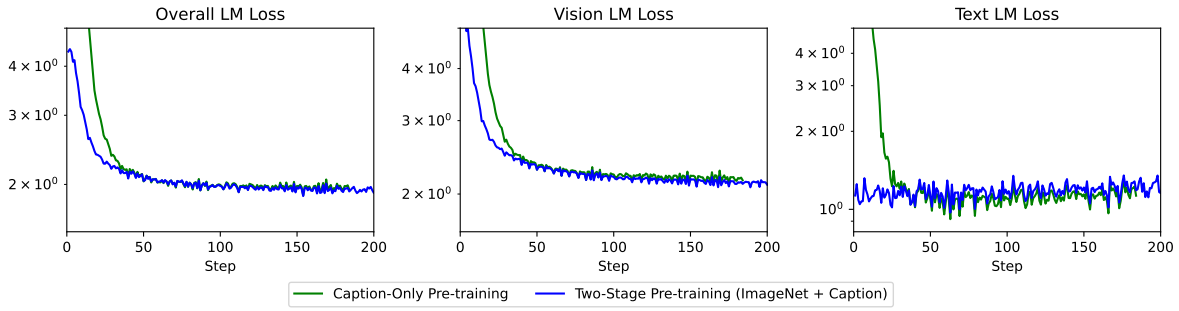


Figure 3: Image captioning loss using two differently initialized checkpoints: (**1**) caption-only pre-training (green) initialized from the LLM; (**2**) two-stage pre-training (blue) initialized from the Stage-1 ImageNet pre-trained LVLM.

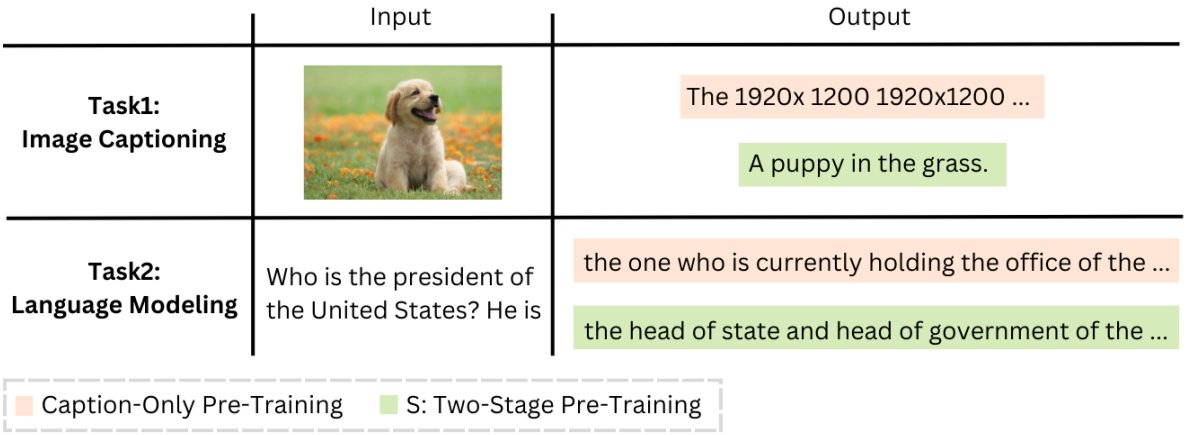


Figure 4: Qualitative analysis of caption-only pre-training and SOLO's two-stage pre-training. Comparisons are made on two checkpoints with comparable vision-language modeling loss (*i.e.,* 2.1). Specifically, we select the caption-only checkpoint at pre-training step 150, and SOLO at step 100.

SOLO excels in task-oriented benchmarks, especially in areas requiring scientific knowledge and mathematical reasoning, due to its successful integration of image representation and complex reasoning within a single unified model. Overall, while SOLO does not yet meet the SoTA performance of the leading LVLMs (Level-3) within the prevalent multi-component LVLM framework, it marks a substantial progression in unified vision-language modeling. This establishes SOLO as a viable candidate for future developments aimed at closing the performance gap with SoTA LVLMs, while more flexibility and scalability by avoiding issues in prior architectures (§2.1).

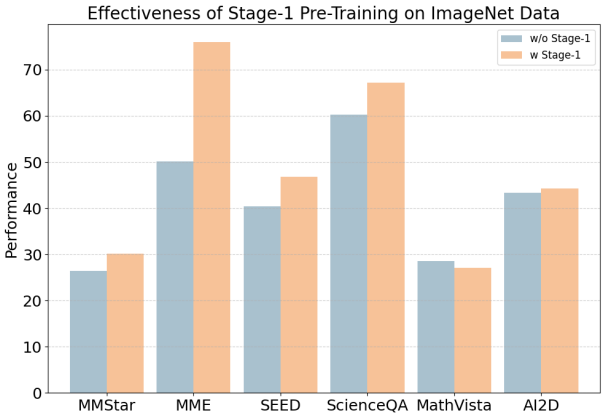
5 Validating Key Ingredients in Our Recipe

5.1 LVLMs Generate Meaningless Captions without Stage-1 Pre-training

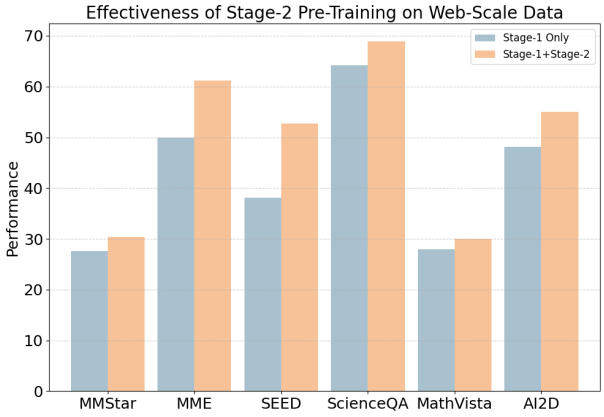
We assess the necessity of Stage-1 pre-training by comparing the Stage-2 LVLM checkpoints *with* and *without* undergoing Stage-1 ImageNet pre-training. In Fig. 3, we observe that these two variants overall achieve similar pre-training loss curves on vision-language modeling and (text) language modeling.

Select Checkpoints for Comparison We select two checkpoints for comparison: one using caption-only pre-training (Stage-2 only) and the other utilizing SOLO's two-stage pre-training, both of which achieve an equivalent vision-language modeling loss of 2.1.

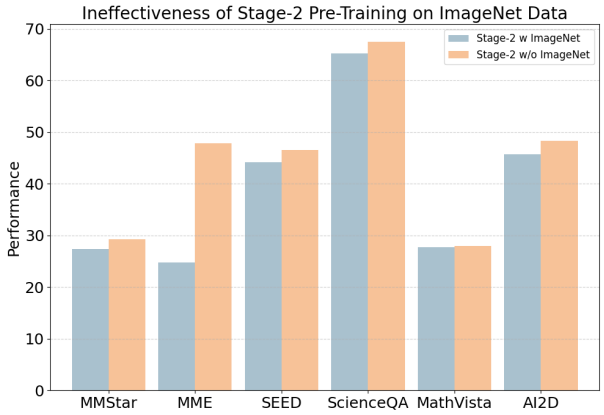
Qualitative Comparison We *randomly* select one example for qualitative analysis (in Fig. 4). Despite the equivalent loss of the selected checkpoints in Fig. 3, we find that without ImageNet pre-training (Stage-2 only), the model generates irrelevant and meaningless image captions, indicates a training divergence.



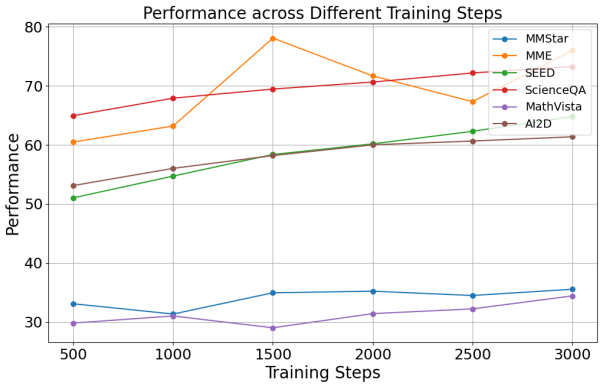
(a) The effectiveness of Stage-1 pre-training on ImageNet data for initialization.



(b) The effectiveness of Stage-2 pre-training on web-scale data for knowledge breadth and data volume.



(c) The ineffectiveness of including ImageNet data in Stage-2 pre-training.



(d) The performance across training steps on the fine-tuning dataset.

Figure 5: The evaluation performance of various ablations to validate key ingredients of our recipe. The MME scores are normalized for better illustration.

Quantitative Comparison We perform a quantitative comparison on the same two checkpoints by training them on the instruction fine-tuning data mixture for 800 steps (see Fig. 5a). Compared to the two-stage pre-trained SOLO, we observe a performance degradation across multiple benchmarks on the checkpoint *without* Stage-1 pre-training, further validating the importance of the first stage.

Discussion We hypothesize that discrepancies between a model's vision and language capabilities can lead to the observed behaviors. Specifically, when there is a significant imbalance—such as with the Mistral 7B model, which possesses advanced language abilities but lacks vision understanding—the model may reduce loss by replicating caption patterns, including redundant text tokens irrelevant to the visual content. For instance, in a caption like "This is a dog", the essential element is "dog". Focusing solely on minimizing language modeling loss without a robust initialized vision representation may lead the model to favor generic phrases like "This is a" over the more discriminative "dog". This is because the former includes more tokens, disproportionately influencing the overall language modeling loss. Pre-training on ImageNet at Stage 1, which emphasizes predicting only the "dog" token, helps the model develop a solid visual representation, effectively narrowing the gap between vision and language capabilities and mitigating this issue. In addition, the results indicate that pre-training loss on vision-language data does not reliably indicate the performance of LVLMs. Detailed analysis are provided in §D, which also demonstrates that training loss on the instruction fine-tuning data mixture is similarly unreliable as an indicator.

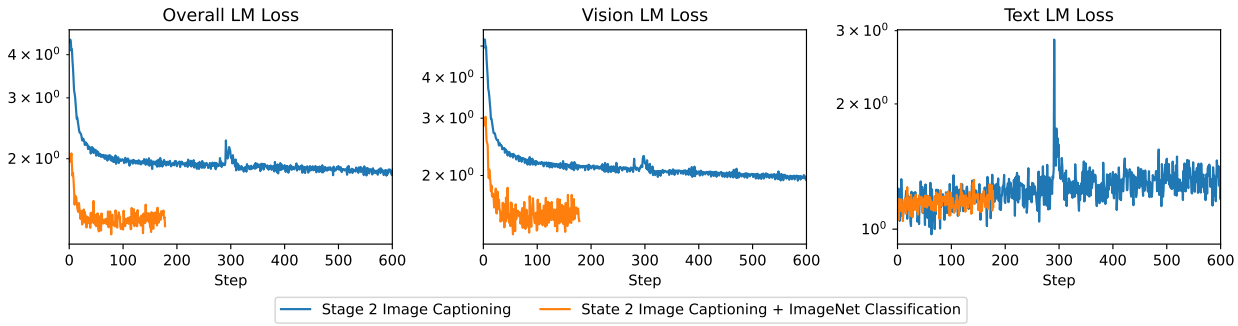


Figure 6: Stage-2 language modeling loss when trained on (**1**) Image captioning objective only (blue); (**2**) Image captioning objective *and* image classification objective used in Stage-1 (orange).

5.2 Stage-2 Pre-Training on Web-Scale Data

Stage-2 Pre-Training Improves Performance on Top of Stage-1 We verify the effectiveness of Stage-2 pre-training on web-scale data by comparing the performance of two LVLMs. Each model is fine-tuned for 800 steps using the same instruction fine-tuning data mixture but initialized differently—one from the end of Stage 1 and the other from Stage 2. In Fig. 5b, we observe significant improvement on all evaluation datasets after pre-training on web-scale data, showing the substantial advantages of Stage 2 pre-training compared to solely using ImageNet data (Stage 1 only).

Combining ImageNet and Captioning at Stage-2 Hurts Performance In addition, it is pertinent to ask whether ImageNet21K data can be combined with web-scale data for Stage 2. We include an ablation with SOLO trained on ImageNet21K and the web-scale data included in the second stage. Fig. 6 illustrates the training curves for comparison. The results suggest that while ImageNet pre-training effectively establishes an initial visual representation, it may not be optimal for subsequent Stage-2 pre-training on web-scale data, as it potentially impedes the optimization of vision-language modeling on image captions (*i.e.,* vision language modeling loss stop improving). This discrepancy may arise from the divergence in image classification and captioning capabilities; the former is emphasized in the first stage. This two-stage approach aligns with the principles of continual curriculum learning, where the model must maintain proficiency in familiar tasks while integrating new ones. This conclusion is also supported by our evaluation of downstream tasks (see Fig. 5c). We train two different checkpoints with and without ImageNet21K data in the second stage on the instruction fine-tuning data mixture (§3.2.2) for 800 steps. Note that we select two checkpoints with comparable vision-language modeling losses for analysis. The results indicate that incorporating ImageNet21K data in the second stage may detrimentally impact overall performance by inhibiting adaptation to and learning from web-scale data.

5.3 Performance Boost via Instruction Fine-Tuning

We evaluate the performance of SOLO on different training steps throughout the instruction fine-tuning stage (see Fig. 5d). The results indicate a consistent improvement in SOLO's performance with prolonged training on the fine-tuning dataset, although the MME scores exhibit some fluctuations. This outcome contrasts with the findings of Liu et al. (2024a), where the performance quickly plateaus upon training with a limited subset of the fine-tuning dataset. This illustrates the increased scalability of SOLO during the instruction fine-tuning stage, suggesting that acquiring additional high-quality supervised datasets for fine-tuning could consistently enhance performance.

5.4 Additional Validation Experiments

We present the results to demonstrate the effectiveness of Stage-3 annealing in §B and to validate the curated data mixture for instruction fine-tuning in §C. We also find that balancing vision and language capabilities during pre-training is challenging on a 7B scale, and present our analysis in §F.

6 Further Analysis

6.1 Controlled Analysis of Fuyu and LLaVA

We conduct a controlled analysis to compare SOLO with LLaVA and Fuyu (see Fig. 7). SOLO with our training recipe consistently outperforms Fuyu-8B, which adopts the same unified modeling strategy, across all evaluation benchmarks. To facilitate a controlled comparison with LLaVA, we develop LLaVA-7B*, which integrates CLIP-ViT-336 and Mistral-7B-base-v0.1, utilizing our specific training procedure and data. The results reveal that LLaVA-7B* achieves performance similar with LLaVA-v1.5-7B (Liu et al., 2024a), indicating that our training recipe, which utilizes large-scale datasets and extensive training, may not significantly impact LLaVA-style LVLMs. Notably, while LLaVA-7B* excels in general visual-language tasks, SOLO demonstrates superior capabilities in visual mathematical reasoning, with overall performance being similar.

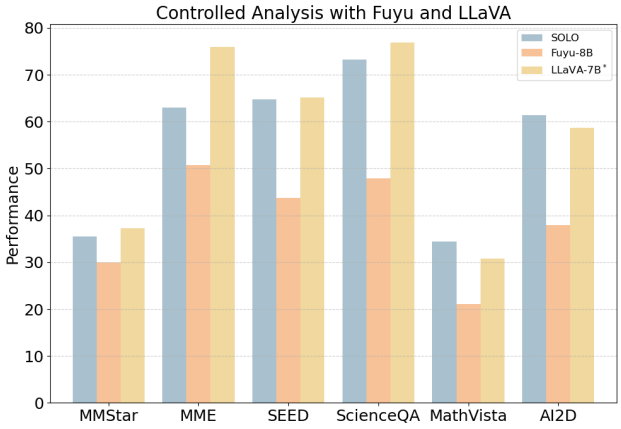


Figure 7: The controlled analysis of Fuyu-8B and LLaVA-7B.

6.2 Scalability Analysis

SOLO Shows Better Scaling Behaviors Compared to LLaVA We perform instruction fine-tuning on pre-trained LLaVA obtained in §6.1 and compare its scaling behaviors with SOLO by measuring the performance improvement per training token (see Fig. 8). For evaluation purposes, we fine-tune both models with 50 steps as the starting points as the pre-trained models are ineffective at following instructions. We measure performance improvement per token every 500 steps and average these to obtain the metric. We observe that SOLO demonstrates better scaling behaviors compared to LLaVA across all evaluation benchmarks since it can more effectively transform the increasing training tokens into performance gains.

SOLO Demonstrates Improved Performance when Scaling up Image Resolution We train SOLO on different image resolutions during the instruction fine-tuning stage for 1,000 steps due to the compute limits (see Fig. 9). The image resolution during inference matches that used in the instruction fine-tuning stage. We find that SOLO continues to improve the performance with increasing image resolution, especially for the visual mathematical reasoning task. In addition, there is no significant difference in performance between the 1024-square resolution and the adapted resolution with 1024 as the maximum used in SOLO. This demonstrates the efficiency and scalability of our flexible image pre-processing pipeline.

7 Related Work

Model Architecture Existing research advances the development of LVLMs capable of addressing diverse tasks via a unified interface that can directly generate natural language, thus avoiding task-specific modifications (Wang et al., 2021; 2022a; Li et al., 2023c). Utilizing pre-trained LLMs (Brown et al., 2020b; Bubeck et al., 2023) as the language component paired with pre-trained visual encoders (Radford et al., 2021; Dosovitskiy et al., 2021a), recent approaches further enhance the instruction-following, user-friendly responses generation, and complex reasoning ability of LVLMs (Liu et al., 2023c; Zhu et al., 2023; Dai et al., 2023; Alayrac et al., 2022; Li et al., 2023a; Ye et al., 2024). Concurrently, Wang et al. (2022b); Peng et al. (2022); Anil et al. (2023); Team (2024); Ge et al. (2023) propose to further learn a codebook in the initial stage to discretize the continuous embeddings extracted by visual encoders into a sequence of image tokens. These approaches enable a uniform vision-language modeling strategy for image and language tokens. However, the dependence on pre-trained visual encoders restricts the scalability of LVLMs. In this study, we address this challenge by readopting the conventional vision-language modeling approach that utilizes a single

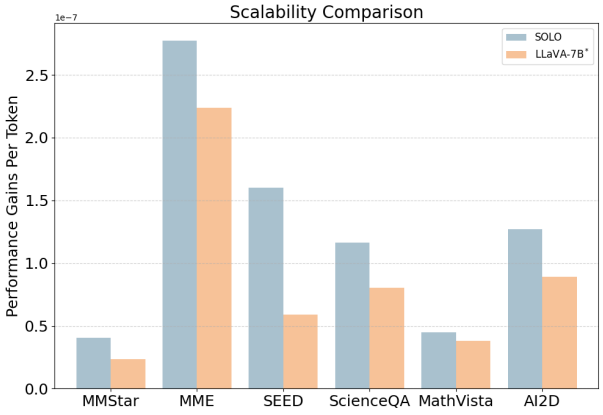


Figure 8: We compare the scaling behaviors of SOLO and LLaVA by measuring the improvement on benchmark performance per token.

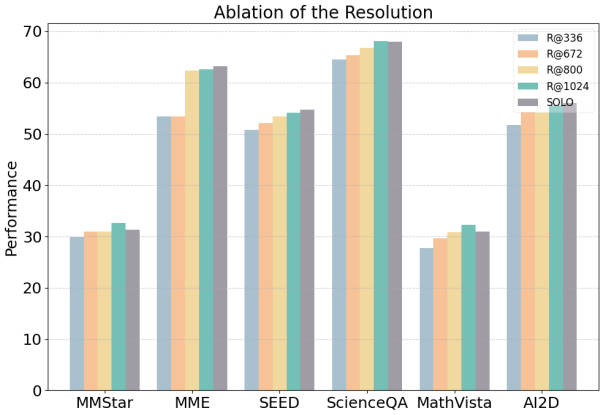


Figure 9: The performance of SOLO when trained and tested on different resolutions of images. R@X denotes the resolution of X.

Transformer for both image and text processing (Li et al., 2019). Furthermore, while Bavishi et al. (2023) extend this approach to billion-scale models, they do not disclose the specifics of their training processes. We address this gap by offering reproducible training recipes, complete with publicly released code, for scalable vision-language modeling on a 7-billion LVLM.

Training Data Typically, LVLMs leverage extensive image-caption pair datasets (Lin et al., 2014a; Schuhmann et al., 2021; 2022; Yu et al., 2024; Chen et al., 2023b) to train a projector or a codebook that map continuous image features into the embedding space of LLMs, thereby aligning the two modalities (Li et al., 2023c; Gong et al., 2023; Zeng et al., 2023; Sun et al., 2023a). Furthermore, large-scale vision-language instruction tuning datasets (Su et al., 2023; Wei et al., 2023; Liu et al., 2023b; Gong et al., 2023; Gao et al., 2023; Li et al., 2023a) and feedback datasets (Li et al., 2023d; Sun et al., 2023b; Chen et al., 2024c; Zhang et al., 2024b) are utilized to further boost the fundamental capabilities of LVLMs and align LVLMs with human preferences, ensuring their ability to comprehend instructions and generate responses that are user-friendly. In this work, we propose a recipe that encompasses the selection of pre-training and instruction fine-tuning datasets, along with corresponding multi-stage paradigms, to facilitate the training of billion-scale LVLMs of a single Transformer architecture. Related work about the LVLMs evaluation benchmarks is in §E.

8 Conclusion

This work revisits the simple vision-language modeling framework with a single Transformer. We argue that this approach effectively mitigates the scalability limitations inherent in prevailing models. With academic resources, we build SOLO, a 7B LVLM initialized from the Mistral LLM. We detail the training recipe and conduct extensive analysis and evaluation to validate the ingredients in our recipe. Experimental results show that SOLO demonstrates performance comparable to LLaVA-v1.5, supporting the continued investigation into this unified vision-language modeling approach for improved scalability.

Limitations and Broader Impact Statement

The investigation into large-scale vision-language modeling using a unified transformer architecture remains nascent, with our model not yet reaching optimal performance across diverse benchmarks. Continued advancements in the direction of unified LVLMs for scalable vision-language modeling are anticipated. However, although developing LVLMs with strong capabilities brings significant advancements in AI, it also poses potential negative impacts. One concern is the risk of misuse, where the model could be employed for malicious purposes, such as generating misleading content that could manipulate public opinion or deceive individuals. Additionally, the model may inadvertently exacerbate biases present in the training data, leading to unfair or discriminatory outcomes in decision-making processes.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pp. 8076–8084, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:2204.14198, 2022.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. CoRR, abs/2312.11805, 2023. doi: 10.48550/ARXIV.2312.11805. URL https://doi.org/10.48550/arXiv.2312.11805.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pp. 2425–2433, 2015.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390, 2023.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. arXiv preprint arXiv:2102.06701, 2021.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL https://www.adept.ai/blog/fuyu-8b.

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, pp. 4291–4301, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020a.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 2020b.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. CoRR, 2023.

Alejandro Hernández Cano, Matteo Pagliardini, Andreas Köpf, Kyle Matoba, Amirkeivan Mohtashami, Olivia Simin Fan, Axel Marmet, Deniz Bayazit, Igor Krawczuk, Zeming Chen, Francesco Salvi, Antoine Bosselut, and Martin Jaggi. epfllm megatron-lm, 2023. URL https://github.com/epfLLM/Megatron-LLM.

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 a lite vision-language model. arXiv preprint arXiv:2402.11684, 2024a.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478, 2023a.

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, 2023b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? arXiv preprint arXiv:2403.20330, 2024b.

Yangyi Chen, Xingyao Wang, Manling Li, Derek Hoiem, and Heng Ji. Vistruct: Visual structural knowledge extraction via curriculum guided code-vision representation. In Proc. The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP2023), 2023c.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14239–14250, 2024c.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Measuring and improving chain-of-thought reasoning in vision-language models. In Proc. 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL2024), 2024d.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. CoRR, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Shizhe Diao, Wangchunshu Zhou, Xinsong Zhang, and Jiawei Wang. Write and paint: Generative vision-language models are unified modal learners. In The Eleventh International Conference on Learning Representations, 2023.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. arXiv preprint arXiv:2305.14233, 2023.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, 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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs], 2021a.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In Iclr, 2021b.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 320–335, 2022.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective. arXiv preprint arXiv:2403.15796, 2024.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pp. 12873–12883. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01268. URL https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html.

Francis Ferraro, Nasrin Mostafazadeh, Lucy Vanderwende, Jacob Devlin, Michel Galley, Margaret Mitchell, et al. A survey of current datasets for vision and language research. arXiv preprint arXiv:1506.06833, 2015.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

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

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010, 2023.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. arXiv preprint arXiv:2307.08041, 2023.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. arXiv preprint arXiv:2305.04790, 2023.

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, pp. 6904–6913, 2017.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300, 2020.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, et al. Large multilingual models pivot zero-shot multimodal learning across languages. arXiv preprint arXiv:2308.12038, 2023.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kai Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies. CoRR, abs/2404.06395, 2024a. doi: 10.48550/ARXIV.2404.06395. URL https://doi.org/10.48550/arXiv.2404.06395.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. arXiv preprint arXiv:2404.06395, 2024b.

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

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. arXiv preprint arXiv:2310.06825, 2023.

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, pp. 5648–5656, 2018.

Kushal Kafle, Robik Shrestha, and Christopher Kanan. Challenges and prospects in vision and language research. Frontiers in Artificial Intelligence, 2:28, 2019.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. arXiv preprint arXiv:1710.07300, 2017.

Yasindu Kamizuru. Diagram image to text. URL https://huggingface.co/datasets/Kamizuru00/diagram_image_to_text.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

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, pp. 235–251. Springer, 2016.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 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, pp. 4999–5007, 2017.

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

Jeonghwan Kim and Heng Ji. Finer: Investigating and enhancing fine-grained visual concept recognition in large vision language models. In arxiv, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683, 2017.

LAION. Laion-gpt-4v. URL https://huggingface.co/datasets/laion/gpt4v-dataset.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? arXiv preprint arXiv:2405.02246, 2024.

Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset, 2024.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726, 2023a.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. CoRR, abs/2311.17092, 2023b. doi: 10.48550/ARXIV.2311.17092. URL https://doi.org/10.48550/arXiv.2311.17092.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13299–13308, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. CoRR, 2023c.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. arXiv preprint arXiv:2312.10665, 2023d.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019.

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, 2023e.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26763–26773, 2024b.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, 2014a.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. Springer, 2014b.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575, 2023.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. Transactions of the Association for Computational Linguistics, 11:635–651, 2023a.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. arXiv preprint arXiv:2306.14565, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. CoRR, 2023c.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26296–26306, 2024a.

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 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. arXiv preprint arXiv:2305.07895, 2023d.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525, 2024.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. arXiv preprint arXiv:2110.13214, 2021.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In NeurIPS, 2022a.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. arXiv preprint arXiv:2209.14610, 2022b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255, 2023.

lz. Detailed caption. URL https://huggingface.co/datasets/echo840/Detailed_Caption.

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

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.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1697–1706, 2022.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In 2019 international conference on document analysis and recognition (ICDAR), pp. 947–952. IEEE, 2019.

Jason Obeid and Enamul Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. arXiv preprint arXiv:2010.09142, 2020.

OpenCompass. Opencompass. URL https://huggingface.co/spaces/opencompass/open_vlm_leaderboard.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. CoRR, abs/2208.06366, 2022. doi: 10.48550/ARXIV.2208.06366. URL https://doi.org/10.48550/arXiv.2208.06366.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision, pp. 2641–2649, 2015.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, pp. 647–664. Springer, 2020.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, 2021.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3505–3506, 2020.

Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. Advances in neural information processing systems, 28, 2015.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972, 2021a.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021b.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114, 2021.

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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. CoRR, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In European Conference on Computer Vision, pp. 146–162. Springer, 2022.

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 the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2018a.

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 the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2556–2565, 2018b.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053, 2019.

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, pp. 742–758. Springer, 2020.

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, pp. 8317–8326, 2019.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama, 2023. URL https://huggingface.co/datasets/cerebras/SlimPajama-627B.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355, 2023.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222, 2023a.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. arXiv preprint arXiv:2309.14525, 2023b.

Benny J Tang, Angie Boggust, and Arvind Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. arXiv preprint arXiv:2307.05356, 2023.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. CoRR, abs/2405.09818, 2024. doi: 10.48550/ARXIV.2405.09818. URL https://doi.org/10.48550/arXiv.2405.09818.

MLC team. MLC-LLM, 2023. URL https://github.com/mlc-ai/mlc-llm.

Techcrunch. Etched is building an AI chip that only runs one type of model. https://techcrunch.com/2024/06/25/etched-is-building-an-ai-chip-that-only-runs-transformer-models/. Accessed: 2024-07-06.

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.

Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. Multimodal research in vision and language: A review of current and emerging trends. Information Fusion, 77:149–171, 2022.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. arXiv preprint arXiv:2311.07574, 2023a.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. arXiv preprint arXiv:2402.14804, 2024a.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. Pmlr, 2022a.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. CoRR, abs/2208.10442, 2022b. doi: 10.48550/ARXIV.2208.10442. URL https://doi.org/10.48550/arXiv.2208.10442.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better llm agents. arXiv preprint arXiv:2402.01030, 2024b.

Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching video-language foundation models with action knowledge. In Proc. 2023 Conference on Neural Information Processing Systems (NeurIPS2023) [Spotlight Paper], 2023b.

Zhenhailong Wang, Joy Hsu, Xingyao Wang, Kuan-Hao Huang, Manling Li, Jiajun Wu, and Heng Ji. Text-based reasoning about vector graphics. In arxiv, 2024c.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904, 2021.

Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. arXiv preprint arXiv:2308.12067, 2023.

Chris Wendler. Renderedtext. URL https://huggingface.co/datasets/wendlerc/RenderedText.

Shujin Wu, Yi R Fung, Sha Li, Yixin Wan, Kai-Wei Chang, and Heng Ji. Macaroon: Training vision-language models to be your engaged partners. arXiv preprint arXiv:2406.14137, 2024.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. arXiv preprint arXiv:2403.11703, 2024a.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. arXiv preprint arXiv:2402.11690, 2024b.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13040–13051, 2024.

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14022–14032, 2024.

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.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. Advancing llm reasoning generalists with preference trees. arXiv preprint arXiv:2404.02078, 2024.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830, 2019.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. What matters in training a gpt4-style language model with multimodal inputs? arXiv preprint arXiv:2307.02469, 2023.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. arXiv preprint arXiv:2401.13601, 2024a.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. arXiv preprint arXiv:2309.15112, 2023a.

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, 2023b.

Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, Feng Zhao, Tao Gui, and Jing Shao. Spa-vl: A comprehensive safety preference alignment dataset for vision language model, 2024b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. CoRR, 2023.

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, pp. 4995–5004, 2016.

Appendix

A Details of Instruction-tuning Data Curation

The curated instruction fine-tuning data mixture is shown in Tab. 3. Each category is chosen to address specific challenges and capabilities of SOLO. For instance, datasets like UltraInteract-SFT (Yuan et al., 2024) and CodeAct-General (Wang et al., 2024b) enable the refinement of language processing and reasoning abilities, while visually rich datasets such as LVIS-Instruct4V (Wang et al., 2023a) and Localized Narratives (Pont-Tuset et al., 2020) enhance the model's basic image understanding and recognition abilities. Scientific document datasets like TQA (Kembhavi et al., 2017) are included to bolster the model's ability to parse and reason with academic visual information. Furthermore, OCR and text-heavy image datasets such as TextCaps (Sidorov et al., 2020) and OCR-VQA (Mishra et al., 2019) provide a crucial source for the model's ability to interpret text within complex images. By selecting datasets with a broad range of complexities, sizes, and focuses, we ensure a robust fine-tuning process that prepares SOLO to handle real-world applications effectively, reflecting a deep and detailed understanding of both vision and language data. Additionally, we conduct a thorough manual inspection and comparisons of the fine-tuning datasets, employing random sampling techniques on some datasets such as DVQA (Kafle et al., 2018) and FigureQA (Kahou et al., 2017) to guarantee diversity and prevent data imbalance.

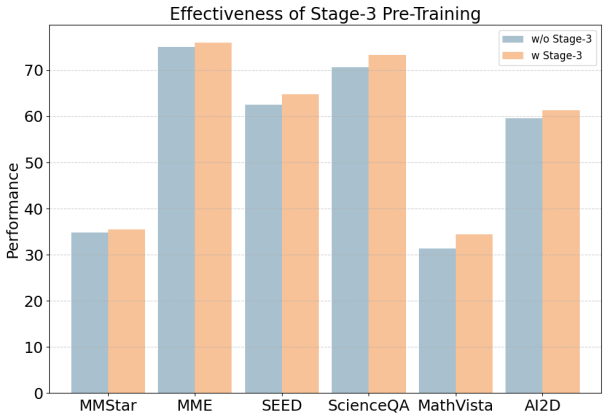
B Stage-3 Annealing

We directly perform instruction fine-tuning on the pre-trained SOLO finished at Stage-2 to understand the effect of the annealing stage. The results shown in Fig. 10a indicate that introducing an annealing stage to conclude pre-training can slightly promote the performance across all evaluation benchmarks.

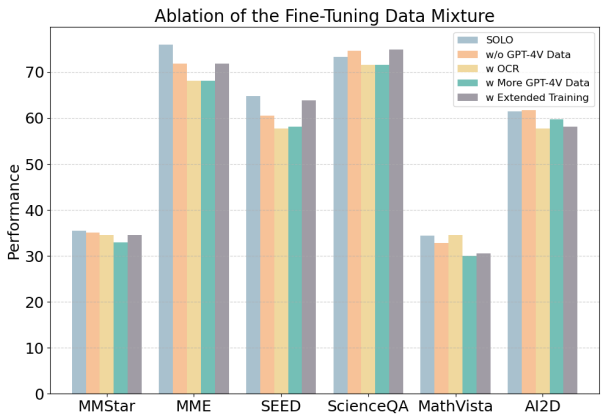
C Effectiveness of Curated Data Mixture for Instruction Fine-Tuning

We conduct an ablation study to validate the curated data mixture for instruction fine-tuning. The ablations included are: (1) Without GPT-4V Data: All data generated by GPT-4V, including detailed captions and instructional fine-tuning samples, is excluded from the fine-tuning mixture. (2) With Additional OCR Data: Additional OCR data from LLaVAR is incorporated into the fine-tuning mixture to enhance OCR capabilities, which are crucial for tasks like require extract text information from charts. (3) With More GPT-4V Data: Data from GPT-4V used in the third stage of pre-training is added to the fine-tuning mixture. (4) Extended Training Duration: SOLO is trained for an additional epoch to investigate the effects of prolonged training.

The results are presented in Fig. 10b. Our analysis indicates that incorporating additional OCR data does not significantly enhance performance in scientific document comprehension or visual mathematical reasoning tasks, which rely extensively on visual text understanding. This lack of improvement can be attributed to the discrepancy between general OCR data and the specific demands of scientific charts. Identifying effective methods for collecting OCR data pertinent to scientific chart comprehension remains a critical area for future research. Furthermore, incorporating this OCR data seems to adversely affect overall visual-language capabilities, as demonstrated by general benchmarks. Regarding the use of GPT-4V data, our findings suggest that a measured inclusion during the pre-training annealing stage enhances performance (see §B), whereas excessive incorporation during fine-tuning can hurt the overall performance. We also find that SOLO exhibits minimal performance decline when trained solely on existing supervised datasets, excluding all data generated by GPT-4V. This demonstrates that GPT-4V data is not essential for enhancing the core capabilities of SOLO. The results overall justify the choice of datasets included in the supervised fine-tuning data mixture. However, although we observe a continual performance improvement across training steps within one epoch (see Fig. 5d), prolonged training on repetitive samples could lead to overfitting and decreased performance. This suggests that while extended exposure to diverse training data generally enhances model performance, overfitting remains a critical challenge when models are exposed repeatedly to a limited data subset. Overall, the ablation study proves the effectiveness of our curated data mixture.



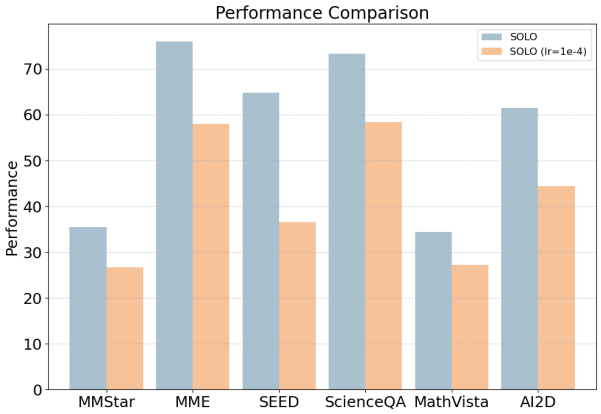
(a) The effectiveness of Stage-3 pre-training in priming SOLO for the instruction fine-tuning stage.



(b) The ablation study of the fine-tuning data mixture.

Figure 10: The evaluation performance of various ablations to validate the effectiveness of Stage-3 pre-training and the fine-tuning data mixture.

(a) The training curves of two variants.



(b) The downstream performance of two variants.

Figure 11: The training curves and downstream performance evaluation of two variants of SOLO. We find that they show significant performance differences although achieving a similar loss on the instruction fine-tuning data mixture.

D (Pre-)training Loss on Vision-Language Data is Not a Reliable Indicator of the Actual Performance.

We find that both the pre-training loss and the instruction fine-tuning loss on the vision-language data are not reliable for the estimation of LVLMs' actual performance. References to support this claim regarding the pre-training loss include observations detailed in Fig. 3, Fig. 4, and Fig. 5a. Despite achieving similar language modeling losses when conditioned on visual inputs, LVLMs exhibit markedly different behaviors and performance across various downstream tasks. This contrasts with findings from pure language modeling, where pre-training loss strongly correlates with various downstream task performance (Du et al., 2024).

We also demonstrate that the loss associated with the instruction fine-tuning data mixture does not reliably indicate task performance. We train a variant of SOLO with a learning rate of 1e-4, deviating from the prescribed rate of 1e-5 in our recipe. We show the training curves (Fig. 11a) and the downstream performance evaluation (Fig. 11b) of these two variants. The two variants exhibit similar training behaviors and losses on

Table 3: Summary of datasets used in the supervised fine-tuning stage.

Category	Dataset	#Sample
Language-Only	CodeAct-General (Wang et al., 2024b)	71K
	UltraInteract-SFT (Yuan et al., 2024)	288K
	UltraChat (Ding et al., 2023)	207K
Detailed Image Caption	LVIS-Instruct4V (Wang et al., 2023a)	223K
	ShareGPT4V (Chen et al., 2023b)	102K
	LAION-GPT4V (LAION)	12K
	Localized Narratives (Pont-Tuset et al., 2020)	200K
	VSR (Liu et al., 2023a)	2K
Scientific Document	TQA (Kembhavi et al., 2017)	2K
	ScienceQA (Lu et al., 2022a)	5K
Table, Document, and Chart	IconQA (Lu et al., 2021)	27K
	TabMWP (Lu et al., 2022b)	23K
	ChartQA (Masry et al., 2022)	18K
	VisText (Tang et al., 2023)	7K
	Chart2Text (Obeid & Hoque, 2020)	27K
	DVQA (Kafle et al., 2018)	20K
	FigureQA (Kahou et al., 2017)	20K
OCR and Text-Rich Images	Diagram Image-to-Text (Kamizuru)	300
	Infographic VQA (Mathew et al., 2022)	2K
	ST-VQA (Biten et al., 2019)	17K
	TextCaps (Sidorov et al., 2020)	22K
	TextVQA (Singh et al., 2019)	22K
	OCR-VQA (Mishra et al., 2019)	17K
	Rendered-Text (Wendler)	10K
General VQA	HatefulMemes (Kiela et al., 2020)	8.5K
	OK-VQA (Marino et al., 2019)	9K
	AOK-VQA (Schwenk et al., 2022)	16.5K
	TallyQA (Acharya et al., 2019)	100K
	Visual7W (Zhu et al., 2016)	14K
	COCO-QA (Ren et al., 2015)	46K
	VQAV2 (Goyal et al., 2017)	82K
	GQA (Hudson & Manning, 2019)	72K

the instruction fine-tuning data mixture, yet they display significant performance disparities in downstream evaluation benchmarks.

Overall, our analysis highlights the need to identify a dependable metric for evaluating LVLMs with the unified architecture in pre-training, particularly for establishing scaling laws in future research.

E Related Work about LVLMs Evaluation Benchmarks

The progress of LVLMs is guided and measured by the continuous development of evaluation benchmarks (Ferraro et al., 2015; Kafle et al., 2019; Gan et al., 2022; Chen et al., 2024d). Initially, evaluation primarily concentrates on fundamental visual-language skills, such as image captioning (Lin et al., 2014b; Plummer et al., 2015), basic visual information recognition (Antol et al., 2015; Goyal et al., 2017), compositional visual understanding (Hudson & Manning, 2019), and knowledge reasoning based on visual information (Marino et al., 2019; Schwenk et al., 2022). Current benchmarks are advancing to encompass more intricate capabilities, requiring LVLMs to perform detailed visual analysis and complex reasoning (Uppal et al., 2022; Zhang et al., 2024a). These benchmarks range from general assessments across various domains and skills (Li et al., 2024a; Fu et al., 2024; Chen et al., 2024b; Yu et al., 2023) to specific tests targeting particular abilities, such as

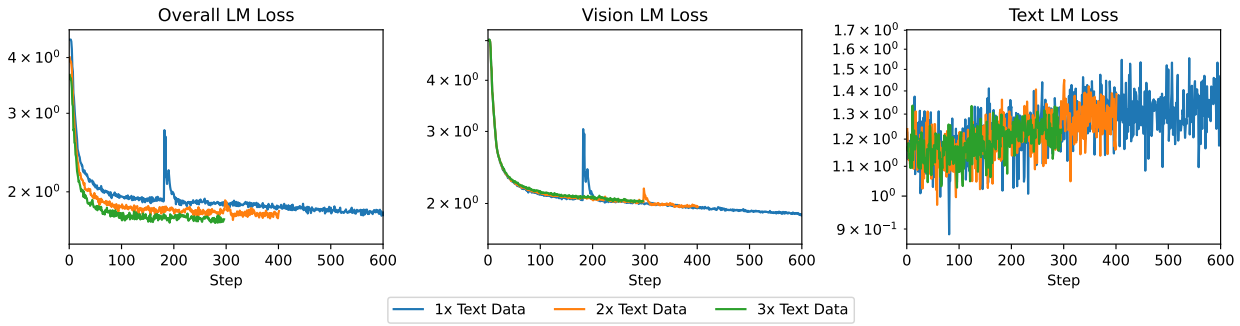


Figure 12: Stage 2 language modeling loss when trained on a mixture with different quantities of text data. 1x reflects the data mixture in Tab. 1, 2x and 3x represent mixtures with 2 or 3 times more text data compared to 1x while keeping the amount of vision data unchanged.

scientific document understanding (Kembhavi et al., 2016; Lu et al., 2022a), mathematical reasoning (Lu et al., 2023; Wang et al., 2024a), multi-discipline understanding and reasoning (Yue et al., 2023; Wu et al., 2024), hallucination (Li et al., 2023e), and OCR ability (Liu et al., 2023d). In this work, we select the advanced general and skill-specific benchmarks for evaluation.

F Balancing Vision and Language Capabilities during Pre-Training is Challenging

We find that on a 7B scale, balancing vision and text capabilities can be challenging. Specifically, we observe that during Stage-2 pre-training, despite the inclusion of text-only pre-training data (§3.2.1) to maintain the language capability of the original LLM, the language modeling loss on the language-only pre-training subset still steadily increases as training continues. In Fig. 12, we introduce a setting where we gradually increase the proportion of text-only data per batch (Tab. 1) and monitor the language modeling loss for text. The results suggest that augmenting text data proportions does not alleviate the rise in language modeling loss, indicating challenges in achieving balanced vision and text capabilities in a 7B-scale model.

To further understand the degradation in language ability of SOLO, we evaluate SOLO on standard LLM evaluation benchmarks, including MMLU (Hendrycks et al., 2020), GSM8k (Cobbe et al., 2021), HellaSwag (Zellers et al., 2019), and RACE (Lai et al., 2017). Our analysis includes comparisons with the backbone LLM of SOLO, specifically Mistral-7B-v0.1-base, as well as Mistral-7B-v0.1-Instruct (see Fig. 13). We observe a decline in language capabilities, particularly in knowledge-intensive benchmarks such as MMLU. There are two potential reasons: (1) Integrating vision capabilities may compromise language performance. (2) The quality of Mistral's pre-training corpus is better than the open-source Slimpajama we employ. Overall, the current results indicate a limitation in the current version of SOLO, as effective performance in real-world vision-language tasks often necessitates strong foundational language capabilities, including knowledge and reasoning. Thus, we plan to maintain the language capabilities of SOLO in the upcoming version by enriching the pre-training dataset with a higher-quality text corpus and increasing the proportion of text data.

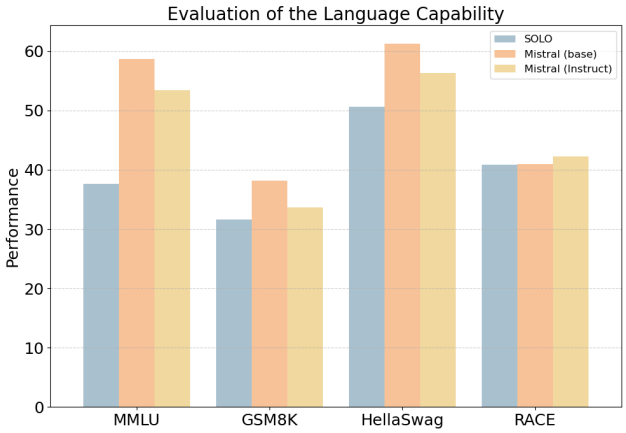


Figure 13: The evaluation of language capability of SOLO and the base Mistral LLM.