

VLsI: Verbalized Layers-to-Interactions from Large to Small Vision Language Models

Byung-Kwan Lee^{1,2†}

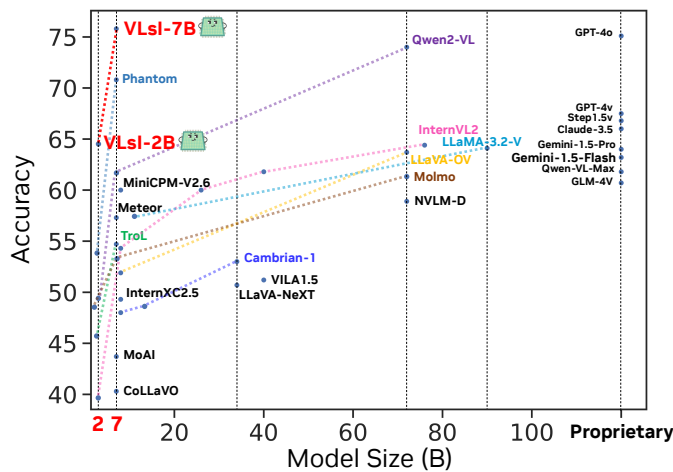
Ryo Hachiuma¹

Yu-Chiang Frank Wang^{1,3}

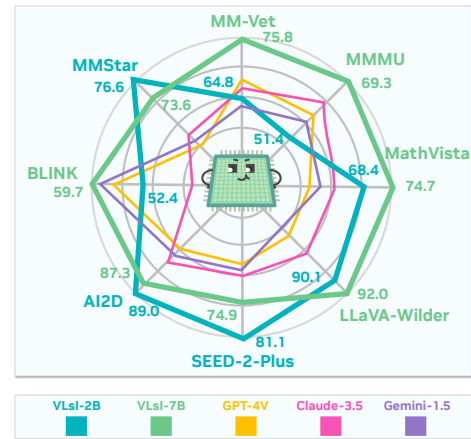
Yong Man Ro^{2‡}

Yueh-Hua Wu^{1‡}

¹NVIDIA, ²KAIST, ³National Taiwan University



(a) MM-Vet across various model sizes



(b) VLsI v.s. Closed-source VLMs

Figure 1. Performance overview of VLsI on vision-language benchmarks. (a) Accuracy on MM-Vet [100] for various model sizes, showing that VLsI (2B and 7B) achieves competitive performance compared to proprietary closed-source VLMs. (b) Comparative evaluation on multiple challenging benchmarks, where VLsI (green and blue) outperforms leading closed-source VLMs, including GPT-4V [79], Claude-3.5-Sonnet [1], and Gemini-1.5-Pro [87], highlighting its efficiency and effectiveness across diverse tasks.

Abstract

The recent surge in high-quality visual instruction tuning samples from closed-source vision-language models (VLMs) such as GPT-4V has accelerated the release of open-source VLMs across various model sizes. However, scaling VLMs to improve performance using larger models brings significant computational challenges, especially for deployment on resource-constrained devices like mobile platforms and robots. To address this, we propose **VLsI: Verbalized Layers-to-Interactions**, a new VLM family in 2B and 7B model sizes, which prioritizes efficiency without compromising accuracy. VLsI leverages a unique, layer-wise distillation process, introducing intermediate “verbalizers” that map features from each layer to natural language space, allowing smaller VLMs to flexibly

align with the reasoning processes of larger VLMs. This approach mitigates the training instability often encountered in output imitation and goes beyond typical final-layer tuning by aligning the small VLMs’ layer-wise progression with that of the large ones. We validate VLsI across ten challenging vision-language benchmarks, achieving notable performance gains (11.0% for 2B and 17.4% for 7B) over GPT-4V without the need for model scaling, merging, or architectural changes. [Project Page](#).

1. Introduction

The integration of large language models (LLMs) with vision-language models (VLMs) has significantly enhanced the interpretive and processing capabilities of visual systems [5, 23, 78, 96]. By leveraging architectures such as CLIP-aligned vision encoders [11, 105], these VLMs have achieved unprecedented performance in understanding and responding to visual inputs. The core of these advance-

† Work Done during Internship.

‡ Corresponding Author.

ments is visual instruction tuning [18, 65], which pairs images with question-answer texts to provide VLMs with rich, instruction-based training. Closed-source VLMs like GPT-4V [78] and Gemini-Pro [87] have led the way, generating high-quality instruction tuning samples and setting new performance standards for visual language understanding.

In response, open-source VLMs of various sizes, including LLaVA-OneVision(OV) [56] and Qwen2-VL [92], have rapidly emerged. While these models demonstrate the advantage of scaling for performance gains in vision-language tasks, the computational cost of larger models presents a critical barrier to deployment in real-world, resource-limited settings, such as mobile devices and autonomous robots. Consequently, the challenge lies in designing high-performing, efficient VLMs capable of handling complex visual tasks without requiring extensive hardware resources.

Traditional approaches to address these constraints often involve adding specialized modules [53] or modifying model architectures [49]. However, these methods introduce significant engineering complexity and can lead to compatibility issues during deployment, particularly for on-device applications where low-latency and memory efficiency are paramount. Furthermore, recent evaluations using benchmarks like MM-Vet [100] and MMMU [102] reveal that these structural modifications still struggle with advanced visual reasoning tasks. This raises the question: *Can we achieve a similar or superior level of performance without scaling, merging, or architectural changes?*

To address this, we introduce  **VLsI: Verbalized Layers-to-Interactions**, a new VLM family that leverages an innovative, natural language-based distillation process to efficiently transfer knowledge from large to small VLMs. Unlike traditional distillation methods, which often directly imitate outputs from a larger model,  VLsI introduces a layer-wise approach where each intermediate layer generates verbal responses in natural language space, enhancing interpretability and alignment with larger models. This is achieved through a three-step process: (1) the *verbalization step*, which uses “verbalizers” to project intermediate features into the language space, making them interpretable as text-based responses; (2) the *interaction step*, which performs adaptive layer matching to align the reasoning progression between large and small VLMs; and (3) the *SFT step*, which finetunes the distilled VLMs for task-specific instruction-following responsiveness.

Our experiments validate  VLsI’s effectiveness across ten challenging benchmarks, demonstrating significant performance gains of 11.0% (2B model) and 17.4% (7B model) over GPT-4V. Notably, these improvements are achieved without increasing model size, merging modules, or modifying the architecture. Our contributions are as:

- We introduce  VLsI, a new VLM family that applies natural language-based, layer-wise distillation, offering

a scalable solution to high-performing yet efficient-scale VLMs without requiring scaling or structural changes.

- We present sampling distribution-based dynamic layer matching before distillation, in order to address the layer number gap between large and small VLMs.
-  VLsI is easy to implement and adaptable across various model architectures, making it practical and deployable for real-world applications.

2. Related Works

Evolution of Vision-Language Models. The emergence of visual instruction tuning: LLaVA [65] and Instruct-BLIP [18] initially brings in not only introducing slight variations of VLMs [3, 8, 17, 48, 54, 107, 112], but also curating high-quality visual instruction samples [7, 9, 27, 32, 91, 99, 109]. Since that time, there has been a growing interest in enhancing visual understanding; thus the simple visual input technique of enlarging image resolution or dividing images into smaller sections with fixed or dynamic rules has got standardized [12, 33, 60, 66, 74, 94]. Furthermore, merging additional visual encoders [26, 46, 80, 105] or multiple computer vision models [13, 22, 76, 97] into LLMs have also become a major focus [34, 52, 53, 70, 82, 113]. Besides, Meteor [51] employs an additional rational projector that embeds multifaceted reasoning information to cover diverse capabilities, including chart, diagram, document, and math reasoning. More recently, Cambrian-1 [89], LLaVA-OneVision [56] InternVL2 [12], Molmo [20], and Qwen2-VL [92] have released large scale models in order to follow or surpass the performances of closed-source VLMs. While these advancements are both rapid and impactful, relatively few studies focus on achieving high-performing yet efficient-scale VLMs within limited architectures. Furthermore, many current approaches rely heavily on GPT-based instruction datasets, and only the final layer learns the target responses in visual instruction tuning. In contrast,  VLsI presents how to effectively harness the capabilities of larger VLMs—already outperforming the closed-source ones—to transfer internally embedded knowledge from large to small VLMs.

Efficient Modeling Strategy. In the field of lightweight LLMs, MobiLlama [88], OpenELM [75], MobileLLM [68] have leveraged various engineering techniques such as shared feed-forward network (FFN) design, layer-wise scaling, and embedding-language head sharing to efficiently reduce model parameters. Their primary objective is not to close the gap with closed-source LLMs but rather to reduce parameters impacting less performance degradation. In the field of VLMs, there are variations about how to utilize pretrained lightweight LLMs [2, 35, 108] in order to make efficient-scale VLMs [14, 15, 62, 81, 110, 111],

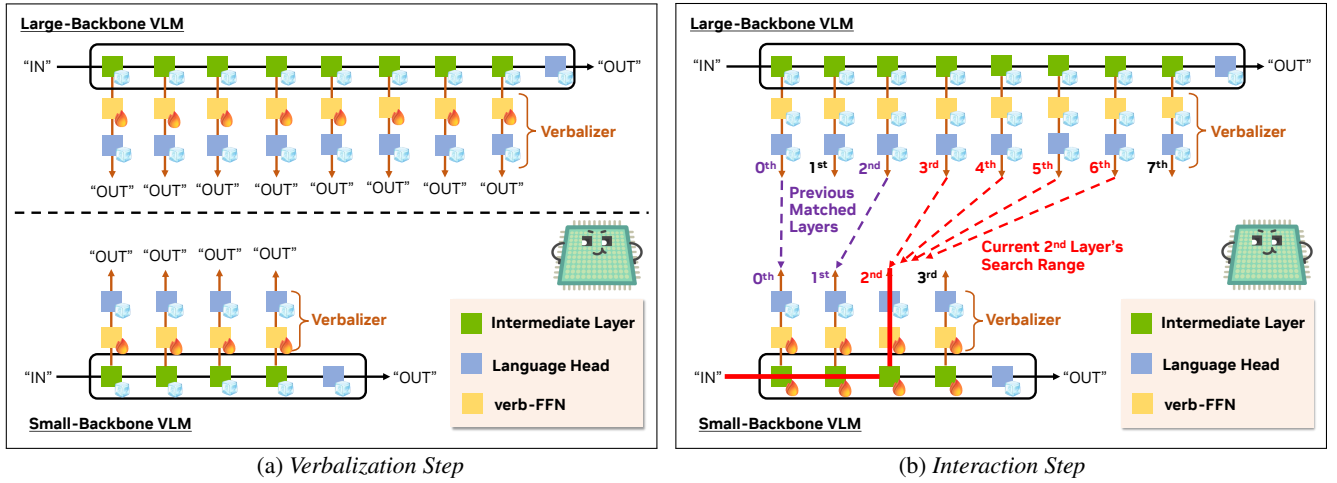


Figure 2. Illustration of the training process in VLsI, showing (a) the verbalization step and (b) the interaction step. (a) In the verbalization step, intermediate layers in both the large- and small-backbone VLMs are equipped with a “verbalizer”, allowing their outputs to be projected into natural language space. Autoregressive loss is applied to align these verbalized outputs with the target responses. (b) In the interaction step, each intermediate layer in the small-backbone VLM searches for a matching layer in the large backbone VLM within a specified range. For example, once the 2nd layer of the small VLM is matched with the 4th layer in the large VLM, the next matching search for the 3rd layer in the small VLM will proceed from the 5th to the 7th layers of the large VLM, ensuring progressive alignment. Note that, “fireball” icon represents parameters that are actively trained or updated, while the “ice cube” icon indicates untrained parameters.

but these works are also not a fundamental solution to embed more vision-language knowledge within limited structures. Notably, two VLM approaches, TroL [50] and Phantom [49], aim to expand learning own capabilities within limited structures by doubling forward propagation steps and enlarging the latent dimension without physically increasing model sizes, thereby showing large improvements. Unfortunately, these approaches face limitations such as key-value cache storage constraints and extensive architectural modifications, which may hinder direct application to real-world scenarios. Besides, a few works in distilling language models [61, 77, 85] have emerged, and TED [61] utilizes language head but it is not a version of LLM covering general tasks, and it does not handle layer number gap between large and small models. The other studies simply use the final layer distillation. For VLMs, LLaVA-MoD [83] and LLaVA-KD [4] have also been recently proposed, but they use the same way. Align-KD [28] distills both vision features and logits from the language head, while MoVE-KD [6] employs multiple vision encoders for large VLMs and a mixture of LoRA [40] for small VLMs. Regarding the distillation metric, DistiLLM [47] and MiniLLM [36] emphasize the effectiveness of reverse KL divergence, whereas LLAVADI [95] and Minitron [77] demonstrate that normal KL divergence is more effective. On the other hand, VLsI makes layer-wise distillation process where we leverage natural language in order to make small VLMs mimic the reasoning progression of large VLMs across layers. We hope that incorporating natural language will facilitate smoother communication between large and small VLMs, alleviating the complexities of feature alignment.

3. VLsI: Verbalized Layers-to-Interactions

Overview of Model Architecture. As illustrated in Fig. 2, VLsI comprises two main components: the backbone VLM and verbalizers. For the backbone VLM, we use Qwen2-VL [92], selected for its high performance on the OpenVLM-Leaderboard [16]. The verbalizer consists of a simple FFN [90] and the language head from the backbone VLM. To distinguish between FFN components, we refer to the FFN within the verbalizer as “verb-FFN” to avoid confusion with the FFN in the LLM transformer decoder block. This design is inspired by the recent speculative decoding paradigm [59], which demonstrates the effectiveness of using a smaller LLM constructed with a frozen word embedding and the language head of larger LLMs to emulate the performance of those larger models. Building on this insight, we incorporate the language head of the backbone VLM into its intermediate layers. Specifically, the verb-FFN is placed between each intermediate layer and the language head, introducing trainable parameters that allow effective projection into the language space. In other words, the verb-FFN enables outputs from the intermediate layers to project to the language space via the language head. In the following sections, we detail the three critical training stages: *verbalization*, *interaction*, and *SFT*.

3.1. Verbalization Step

In this step, we introduce a verbalizer for each target intermediate layer (see Fig. 2(a)), allowing the outputs of these layers to be mapped to the natural language space. Each verbalizer comprises a verb-FFN (the yellow block in Fig. 2(a)) and a language head (the blue block in Fig. 2(a)), which is the same language head used in the corresponding backbone VLM. It is important to note that the outputs from in-



Figure 3. Example of verbalized outputs from each intermediate target layer in an alternative small-backbone VLM (without VLsI enhancements) and the VLsI. The visual question prompts VLM to predict the missing image in a sequence pattern. The outputs illustrate how each layer progressively interprets the visual cues, with VLsI accurately identifying the answer as ‘a star with a dot’ in the final layer, while the alternative small-backbone VLM incorrectly predicts ‘a diamond with a dot’. This demonstrates the improved interpretative capability of VLsI through layer-wise, language-based distillation.

intermediate layers are not directly translatable to the natural language space. To address this, we apply an additional verbalizer to process these intermediate layer outputs. To optimize the mapping from embeddings to natural language, we leverage instance pairs from visual instruction tuning datasets and apply an autoregressive loss to ensure the verbalized output aligns with the target response. Since the weights in both backbone VLMs are fixed, the gradient updates for each verbalizer at each layer remain independent. Our goal here is not for the intermediate layers to generate the correct response but rather to showcase their capacity to express verbal information given specific visual and text inputs. Fig. 3 presents the verbalization results, displaying the output of each layer along with its corresponding verbalizer. The results show a gradual improvement in reasoning and response accuracy as layers progress deeper. This verbalization enables our VLsI to track the evolution of verbal responses in the natural language space, facilitating a clearer understanding of the ‘key developments’ required to generate desired responses and thereby offering a more efficient distillation approach.

3.2. Interaction Step

After the *verbalization step*, we proceed with a distillation process that leverages the natural language output from each

intermediate layer. The main objective of this *interaction step* is to establish an effective mapping between the layers of the large-backbone and small-backbone VLMs. This approach ensures that the small-backbone VLM mirrors the reasoning progression in the large one as layers deepen. Furthermore, because the computational requirements for each ‘key development’ in generating the desired response vary, we propose an adaptive layer-matching strategy that dynamically aligns corresponding layers.

Extracting Verbal Information for Distillation. To extract verbal information from intermediate layers, we use the vocabulary probabilities from the language head in the verbalizer based on input from the visual instruction training dataset. This method avoids the high computational cost of text generation (*e.g.*, greedy or beam search [29]), which would be prohibitively resource-intensive for each layer across both the large and small VLMs.

Layer Matching for Distilling Reasoning Progression. To achieve effective layer matching, we develop a strategy to pair layers between the large- and small-backbone VLM, allowing the small model to learn the reasoning progression encoded in the large model. We define the number of target layers in the large and small VLMs as t_l and

VLMs	QBench	AI2D	ChartQA	POPE	HallB	MME	MathVista	MMB	MMB ^{CN}	MM-Vet	MMMU
LLaVA-NeXT-7B [66]	-	-	-	86.5	-	1851	34.6	69.6	63.3	43.9	35.1
LLaVA-NeXT-8B [66]	-	71.6	69.5	-	-	1972	37.5	72.1	-	-	41.7
LLaVA-NeXT-13B [66]	-	70.0	62.2	86.7	-	1892	35.1	70.0	68.5	47.3	35.9
MM1-7B [74]	-	-	-	86.6	-	1858	35.9	72.3	-	42.1	37.0
MM1-MoE-7B×32 [74]	-	-	-	87.8	-	1992	40.9	72.7	-	45.2	40.9
MiniGemini-HD-7B [60]	-	-	-	-	-	1865	32.2	65.8	-	41.3	36.8
MiniGemini-HD-13B [60]	-	-	-	-	-	1917	37.0	68.6	-	50.5	37.3
Cambrian-1-8B [89]	-	73.0	73.3	-	-	-	49.0	75.9	-	-	42.7
Cambrian-1-13B [89]	-	73.6	73.8	-	-	-	48.0	75.7	-	-	40.0
Eagle-8B [82]	-	76.1	80.1	-	-	-	52.7	75.9	-	-	43.8
Eagle-13B [82]	-	74.0	77.6	-	-	-	54.4	75.7	-	-	41.6
VILA1.5-8B [63]	-	-	-	85.6	-	-	-	75.3	69.9	43.2	38.6
VILA1.5-13B [63]	-	-	-	86.3	-	-	-	74.9	66.3	44.3	37.9
VILA ² -8B [27]	-	-	-	86.7	-	-	-	76.6	71.7	50.0	38.3
CogVLM2-8B [38]	-	73.4	81.0	-	-	1870	-	80.5	-	60.4	44.3
LLaVA-OneVision-7B [56]	-	81.4	80.0	-	-	1998	63.2	80.8	-	57.5	48.8
InternVL2-8B [11]	-	<u>83.8</u>	83.3	-	-	2210	58.3	81.7	<u>81.2</u>	54.2	49.3
MiniCPM-V2.5-8B [98]	-	-	-	-	-	2025	54.3	77.2	74.2	-	45.8
MiniCPM-V2.6-8B [98]	-	-	-	-	-	2348	60.6	-	-	60.0	49.8
TroL-7B [50]	73.6	78.5	71.2	87.8	65.3	2308	51.8	<u>83.5</u>	<u>81.2</u>	54.7	49.9
Phantom-7B [49]	<u>73.8</u>	79.5	87.8	87.7	65.4	2126	70.9	84.8	84.7	<u>70.8</u>	51.2
Qwen2-VL-7B [92]	77.5	77.5	83.0	88.9	<u>65.7</u>	2327	58.2	83.0	80.5	<u>62.0</u>	<u>54.1</u>
 VLsI-7B	77.5	87.3	<u>86.1</u>	<u>88.6</u>	74.2	<u>2338</u>	74.7	86.3	85.5	75.2	69.3

Table 1. Evaluation of existing open-source VLMs and  VLsI on various vision-language benchmarks: QBench [93], AI2D [44], ChartQA [72], POPE [58], HallB [64], MME [30], MathVista [71], MMB [67], MMB^{CN} [67], MM-Vet [100], and MMMU [102]. **Bold** and underline indicate the top and second-best results, respectively.

t_s , respectively, with ranges $i_l = 0^{\text{th}}, 1^{\text{st}}, \dots, (t_l - 1)^{\text{th}}$ and $i_s = 0^{\text{th}}, 1^{\text{st}}, \dots, (t_s - 1)^{\text{th}}$, where $t_s \leq t_l$. Our matching strategy iterates over the t_s layers in the small VLM, ensuring that each layer has a corresponding target layer in the large VLM. Note that, layers in the large-backbone VLM may not have corresponding layers in the small one. To maintain the reasoning progression as layers deepen, we employ a strategy that respects two key criteria: (i) Order Preservation—the matched layer j (large-backbone VLM) of layer i (small-backbone VLM) should be deeper than the matched layer k of layer $i - 1$, ensuring $j > k$; and (ii) Layer-wise Exploration (Multinomial Sampling)—to encourage novel and effective configurations for layer matching, we sample layers based on a distribution that is inversely proportional to the KL divergence between the verbal distributions of the matched layers in the large and small VLMs. Specifically, we compute this sampling distribution using a softmax of the scaled, negative KL divergence values, as summarized in Algorithm 1.

3.3. Supervised Finetuning (SFT) Step

In the final stage of our distillation framework, we finetune the entire small-backbone VLM, including word embeddings, multi-head attention, FFN, and the language head on the visual instruction dataset in a supervised-learning manner. This *SFT step* is inspired by pruning-based distillation methods [77, 85], which require additional training after pruning to counteract potential performance drops from structural changes. While our *interaction step* does not alter the model’s structure, fully absorbing the rich information from the large VLM remains challenging for the small VLM

Algorithm 1 Pseudo-Code for Interaction Loss

```

1: Input:  $t_s, t_l$ 
2: Initialize: loss: 0,  $i_l^*$ : 0,  $\epsilon$ : 1e-6, scale: 2
3: for  $i_s$  in  $0 \leq i_s < t_s$  do
4:    $kld\text{-}list = []$ 
5:   for  $i_l$  in  $i_l^* \leq i_l \leq t_l - t_s + i_s$  (Search Range) do
6:      $kld\text{-}list.append(\text{compute-}kld(i_s, i_l))$ 
7:   end for
8:    $T \leftarrow \frac{\text{scale}}{kld\text{-}list.max - kld\text{-}list.min + \epsilon}$  {Temperature}
9:    $p\text{-}list \leftarrow \text{Softmax}(-kld\text{-}list/T)$ 
10:   $i_l^* \leftarrow 1 + \text{Multinomial}(p\text{-}list)$  {Sampling Index}
11:  loss  $\leftarrow$  loss +  $kld\text{-}list[i_l^*]$ 
12: end for
13: Return: loss

```

within a single *interaction step*. The SFT step allows the small VLM to better align and integrate the acquired knowledge across layers, enhancing response quality through autoregressive loss. We avoid optimizing autoregressive loss during the *interaction step* (where KL divergence loss is optimized instead) as this can degrade performance, as demonstrated in our ablation study (Sec. 4). Therefore, to improve task-specific instruction-following responsiveness and accuracy, we incorporate the SFT step with autoregressive loss to further finetune the small-backbone VLM.

4. Experiments

4.1. Implementation Details

We present four key technical components of  VLsI to ensure reproducibility: (a) the configuration of the backbone vision-language model, (b) the architecture of the ver-

VLMs	QBench	AI2D	ChartQA	POPE	HallB	MME	MathVista	MMB	MMB ^{CN}	MM-Vet	MMMU
MiniCPM-2.4B [41]	-	56.3	-	-	-	1650	28.9	64.1	62.6	31.1	-
MiniCPM-V2-2.8B [41]	-	62.9	-	-	-	1809	38.7	69.1	66.5	41.0	-
MM1-3B [74]	-	-	-	87.4	-	1762	32.0	67.8	-	43.7	33.9
MM1-MoE-3B×64 [74]	-	-	-	87.6	-	1773	32.6	70.8	-	42.2	38.6
ALLaVA-3B [7]	-	-	-	-	-	1623	-	64.0	-	32.2	35.3
VILA1.5-3B [63]	-	-	-	85.3	-	-	-	62.8	52.2	38.6	33.3
InternVL2-4B [11]	-	78.9	81.5	-	-	2064	58.6	78.6	73.9	51.0	34.3
TroL-3.8B [50]	70.0	73.6	73.8	86.5	<u>62.2</u>	1980	55.1	79.2	<u>77.1</u>	51.1	37.5
Phantom-3.8B [49]	70.3	71.7	87.3	87.1	60.8	<u>2046</u>	60.6	<u>80.4</u>	<u>77.1</u>	<u>54.4</u>	39.2
DeepSeek-VL-1.3B [70]	-	-	-	87.6	-	-	31.1	64.6	62.9	34.8	32.2
MobileVLM-1.7B [14]	-	-	-	84.5	-	-	-	53.2	-	-	-
MobileVLM-V2-1.7B [15]	-	-	-	84.3	-	-	-	57.7	-	-	-
MoE-LLaVA-1.8B×4 [62]	-	-	-	87.0	-	-	-	59.7	-	25.3	-
Mini-Gemini-2B [60]	-	-	-	-	-	1653	29.4	59.8	-	-	31.7
InternVL2-2B [11]	-	<u>74.1</u>	76.2	-	-	1877	46.3	73.2	70.9	39.5	34.3
TroL-1.8B [50]	68.2	68.9	64.0	<u>88.6</u>	60.1	2038	45.4	76.1	74.1	45.1	35.2
Phantom-1.8B [49]	69.1	62.3	<u>87.0</u>	89.6	<u>62.2</u>	1885	<u>60.9</u>	76.6	75.1	54.1	40.6
Qwen2-VL-2B [92]	<u>70.8</u>	60.2	73.5	87.8	61.2	1872	43.0	74.9	73.5	49.5	<u>41.1</u>
 VLsI-2B	72.3	89.0	85.8	87.9	70.0	2022	68.4	81.7	78.8	64.8	51.4

Table 2. Comparison of smaller open-source VLMs and  VLsI on the same evaluation benchmarks as in Table 1.

balizer, (c) the training and inference configuration, and (d) the structure of the visual-instruction dataset, which includes a variety of capabilities crucial for effectively building  VLsI.

(a) Configuration of the Backbone VLM. We select Qwen2-VL [92] as our backbone VLM due to its flexible model scaling options, which include Qwen2-1.5B, Qwen2-7B, and Qwen2-72B [96]. Importantly, the tokenizer’s vocabulary indices remain consistent across these model sizes, allowing for seamless integration without reordering the vocabulary. This structure enables us to focus on optimizing the LLM component, where Qwen2-1.5B and Qwen2-7B each contain 28 layers, and Qwen2-72B consists of 80 layers. For the vision encoder and projector, we adopt the same modules as those in Qwen2-VL: the vision encoder is a ViT model [21] adapted from DFN [25] and enhanced with visually-adapted rotary positional embeddings [86]. The vision projector comprises an MLP with two fully-connected layers interleaved with GELU activations [37].

(b) Architecture of the Verbalizer. We design the verbalizer as a sequential feed-forward network (FFN), similar to the FFN typically used in transformer blocks [90], and as a language head of the backbone VLM. Conventionally, an FFN consists of three MLPs responsible for dimensional expansion, gating, and reduction. This configuration first expands the hidden dimension, then applies importance weighting to emphasize relevant features, and finally reduces the features back to the original hidden dimension. To enhance computational efficiency in verbalization and interaction, we opt to maintain a consistent hidden dimension throughout the process, foregoing the typical expansion and reduction steps. This streamlined FFN (verb-FFN) design reduces computational complexity while preserving

overall performance. For intermediate target layers, we select i_s : 2nd, 6th, 10th, ..., and 26th layers, and i_t : 2nd, 6th, 10th, ..., and 78th layers.

(c) Training and Inference Configuration. Training and inference for  VLsI are conducted on 8 NVIDIA A100 80GB GPUs. To enable efficient training, we apply LoRA [40] to the LLM component, setting the rank and alpha parameters to 64. We use the AdamW optimizer [69] with a cosine annealing schedule, adjusting the learning rate from $1e^{-4}$ to $1e^{-6}$ over each training step. To handle large batch sizes effectively, we employ gradient accumulation with 16 steps and gradient checkpointing [84] to optimize memory usage. Specifically, batch configurations are four batches each for the 2B and 7B models and two batches for the 72B model, resulting in total gradient update counts of 512 ($8 \times 16 \times 4$) and 256 ($8 \times 16 \times 2$), respectively. For inference, we maintain the setup used in Qwen2-VL with a greedy search for text generation.

(d) Visual-instruction dataset. Following the methodology in [49], we compiled a diverse dataset spanning a broad range of vision-language capabilities, totaling 2.9 million visual instruction tuning samples from various sources. This dataset includes foundational image understanding samples sourced from datasets such as ShareGPT4o-Images (57K) [24], ShareGPT4V (755K) [9], ALLaVA-VFLAN/Text (548K) [7], and MiniGemini (27K) [60], which are focused on tasks like DocVQA [73], ChartQA [72], DVQA [42], and AI2D [44]. To support scientific and mathematical reasoning, we incorporated samples from LLaVA-HD (116K) [109], enhancing datasets like ArXivQA [57] and TextbookQA [45]. Additionally, we included document understanding samples from mPLUG-DocOwl1.5-Downstream/Reasoning (599K) [39]. For more

Benchmarks	OmniFusion-7B	DeepSeek-VL-7B	MoVA-7B	Eagle-8B	CoLLaVO-7B	MoAI-7B	Meteor-7B	 VLsI-2B	 VLsI-7B
MMB [67]	69.0	73.2	81.3	75.9	83.0	79.3	82.9	81.7	86.3
MathVista [71]	-	-	44.3	52.7	57.6	56.2	53.4	68.4	74.7
MM-Vet [100]	39.4	41.5	-	-	40.3	43.7	57.3	64.8	70.8
MMM [102]	36.6	36.6	-	43.8	42.2	45.6	48.3	51.4	69.3

(a) Validation of open-source VLMs with additional modules and projectors compared to  VLsI: OmniFusion [34], DeepSeek-VL [70], MoVA [43], Eagle [82], CoLLaVO [52], MoAI [53], and Meteor [51].

VLMs	MM-Vet	MM-Vet-v2	MMM	MMStar	AI2D	SEED-2-Plus	MathVista	BLINK	CV-Bench	LLaVA-Wilder
LLaVA-NeXT-34B [66]	50.7	50.9	48.8	51.6	78.9	65.9	40.4	-	-	-
VILA1.5-40B [63]	51.2	-	55.1	55.2	77.8	-	49.5	-	-	-
Cambrian-34B [89]	53.2	-	50.4	54.2	79.5	65.1	50.3	-	76.9	-
Molmo-72B [20]	61.1	-	52.8	63.3	83.4	-	55.8	-	-	-
LLaVA-OV-72B [56]	60.6	-	56.6	65.8	86.2	-	68.4	-	-	72.0
LLaMA-3.2-Vision	64.1	-	60.3	55.3	69.5	68.2	58.3	48.0	-	-
Claude3.5-Sonnet [1]	66.0	71.8	65.9	62.2	80.2	71.7	61.6	28.2	-	83.1
NVLM-D-72B [19]	58.9	-	60.8	63.7	80.1	68.4	63.9	48.0	-	-
GPT-4V (0409) [9]	67.5	66.3	61.7	56.0	78.6	69.3	54.7	58.3	69.1	71.5
Gemini-1.5-Pro	64.0	66.9	60.6	59.1	79.1	70.8	57.7	59.1	-	-
InternVL2-76B [11]	64.4	68.4	58.3	67.1	87.6	70.0	65.6	57.5	-	-
GPT-4o (0806)	75.1	71.0	69.9	64.7	84.7	70.8	62.7	64.7	-	85.9
Qwen2-VL-72B [92]	73.9	68.7	64.3	68.6	<u>88.3</u>	72.3	<u>69.7</u>	<u>60.5</u>	74.3	84.1
TroL-1.8B [50]	45.1	-	35.2	45.5	68.9	-	45.4	-	-	-
TroL-7B [50]	54.7	-	49.9	51.3	78.5	-	51.8	-	-	-
Phantom-1.8B [49]	54.1	46.3	40.6	45.5	62.3	57.1	60.9	44.2	63.1	78.5
Phantom-7B [49]	70.8	60.6	51.2	57.3	79.5	65.5	70.9	58.9	74.9	82.9
 VLsI-2B	64.8	60.8	51.4	76.6	89.0	81.1	68.4	52.4	90.1	<u>90.1</u>
 VLsI-7B	75.8	70.0	<u>69.3</u>	<u>73.6</u>	87.3	<u>74.9</u>	74.7	59.7	<u>89.1</u>	92.0

(b) Comparison of  VLsI with other open-source and closed-source VLMs on challenging benchmarks: MM-Vet [100], MM-Vet-v2 [101], MMM [102], MMStar [10], AI2D [44], SEED-2-Plus [55], MathVista [71], BLINK [31], CV-Bench [89], and LLaVA-Wilder [56]. This comparison includes models embedding additional knowledge [49, 50] and larger open/closed-source VLMs.

Table 3. Detailed comparison of  VLsI with various open and closed-source VLMs on challenging evaluation benchmarks. Appendix A provides detailed descriptions of the evaluation benchmarks listed in Tables 1 and 2.

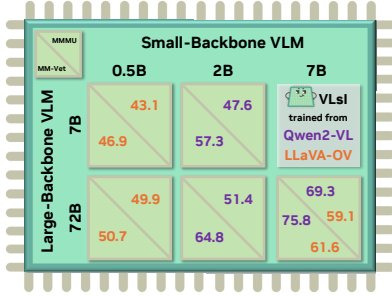


Figure 4. Comparison of performance on MM-Vet [100] and MMM [102] across different model size combinations in large and small backbone VLMs. Each cell shows the evaluation results for various interaction configurations between 0.5B, 2B, and 7B small backbone VLMs trained with either Qwen2-VL [92] or LLaVA-OV [56] as the large-backbone VLM.

general mathematical tasks, our dataset features samples from GLLaVA (177K) [32], MathVision (3K) [91], Math-Instruct (262K) [103], and MathPlus (304K) [104].

4.2. Verbalization on VLsI

Fig. 3 illustrates the verbal responses generated at each intermediate layer in small-backbone VLM and  VLsI. Using the verbalized outputs to trace each layer’s interpretive progression, this comparison highlights how both mod-

els gradually enhance understanding across layers. At the shallower layers, both models generate basic descriptions, focusing on large, simple shapes and colors. However, as  VLsI progresses to mid-level layers, it begins to recognize and articulate more complex visual structures, such as labeled shapes and their relative positions. In contrast, the small-backbone VLM’s verbal responses remain relatively vague or repetitive, often lacking in specific relational details. By the deeper layers,  VLsI demonstrates a clear advantage: its verbalizations shift towards identifying the correct pattern, explicitly referring to shapes and colors in alignment with the target response: “star with a dot”. Meanwhile, the small-backbone VLM incorrectly predicts the missing image as a “diamond with a dot”, failing to capture the specific pattern. This example underscores the effectiveness of  VLsI’s layer-wise verbalization, where each stage of verbal responses helps the small-backbone VLM align with the larger one. Note that, Appendix B shows the loss changes during verbalization. Additional examples of  VLsI’s verbalization are available in Appendix C, highlighting its capacity to interpret layer-wise verbal responses.

4.3. Comparison on Evaluation Benchmarks

As shown in Tab. 1 and Tab. 2,  VLsI achieves higher performance while maintaining an efficient model scale.

VLMs	MMB	MM-Vet	MMMU
LLaVA-OV-0.5B	52.1	29.1	31.4
VLSI-0.5B (LLaVA-OV-72B)	72.5	50.7	49.9
LLaVA-OV-7B	80.8	57.5	48.8
VLSI-7B (Qwen2-VL-72B)	86.3	75.8	69.3
VLSI-7B (LLaVA-OV-72B)	86.1	61.6	59.1
LLaVA-OV-72B	85.9	63.7	56.8

(a) Backbone VLMs

VLMs	SFT	MMB	BLINK	MM-Vet	MMMU
Qwen2-VL-2B	-	74.9	41.4	49.5	41.1
VLSI-2B	✗	73.2	40.1	47.9	39.8
VLSI-2B (Beam)	✗	77.6	48.3	58.1	45.9
VLSI-2B	✓	81.7	52.4	64.8	51.4
Qwen2-VL-7B	-	83.0	50.8	62.0	54.1
VLSI-7B	✗	82.1	49.6	60.5	52.9
VLSI-7B (Beam)	✗	85.2	56.3	70.9	63.5
VLSI-7B	✓	86.3	59.7	75.8	69.3
Qwen2-VL-72B	-	86.5	60.5	73.9	64.3

(b) Use of SFT Step

RS Training Percent.	MMB	BLINK	MM-Vet	MMMU
Qwen2-VL-7B	80.5	50.8	62.0	54.1
+50%	81.0	51.5	62.8	54.6
+100%	81.8	52.3	63.7	55.2
VLSI-7B	82.1	49.6	60.5	52.9
+50%	85.4	56.0	70.0	62.1
+100%	86.3	59.7	75.8	69.3

(c) Percentage of Training Iterations in RS

IL-Ops	LL-Ops	MMB	BLINK	MM-Vet	MMMU
CE	✗	79.2	51.3	64.5	56.5
CE	CE	77.8	50.2	63.2	55.2
CE	KLD	81.0	53.5	67.2	59.0
L2	KLD	81.5	53.2	66.8	58.0
KLD	✗	83.0	55.0	69.5	61.0
KLD	CE	81.5	54.3	68.5	59.8
KLD	KLD	86.3	59.7	75.8	69.3
L2	KLD	81.7	53.5	67.0	58.3

(d) Operations for Intermediate/Last Layers

VLMs	MMB	BLINK	MM-Vet	MMMU
Random Index	77.0	50.0	62.0	52.0
Uniform Index	79.5	51.9	66.5	55.3
Bottom-1 Index	81.2	53.8	67.8	57.7
Bottom-3 Index	81.5	54.0	68.0	58.0
Multinomial Sampling	82.0	54.5	68.5	58.5
+Search Range	83.5	55.5	69.8	60.0
+Order Preservation	86.0	59.2	75.2	68.5
+Adaptive Temperature	86.3	59.7	75.8	69.3

(e) Components in Matching Strategy

Structure	Params	MMB	BLINK	MM-Vet	MMMU
Decoder×2	3.3B	85.5	59.1	76.2	68.6
Decoder	1.6B	85.7	59.0	76.0	68.4
FFN×2	2.9B	86.3	59.2	75.9	69.3
FFN	1.4B	86.4	59.4	75.7	69.2
verb-FFN×2	539M	85.8	59.9	75.8	69.3
verb-FFN	269M	86.3	59.7	75.8	69.3
MLP×2	180M	84.2	57.5	74.1	67.0
MLP	90M	83.8	57.0	73.5	66.7

(f) Verbalizer Architecture

Table 4. Ablation studies examining the six main factors influencing the effectiveness of VLSI.

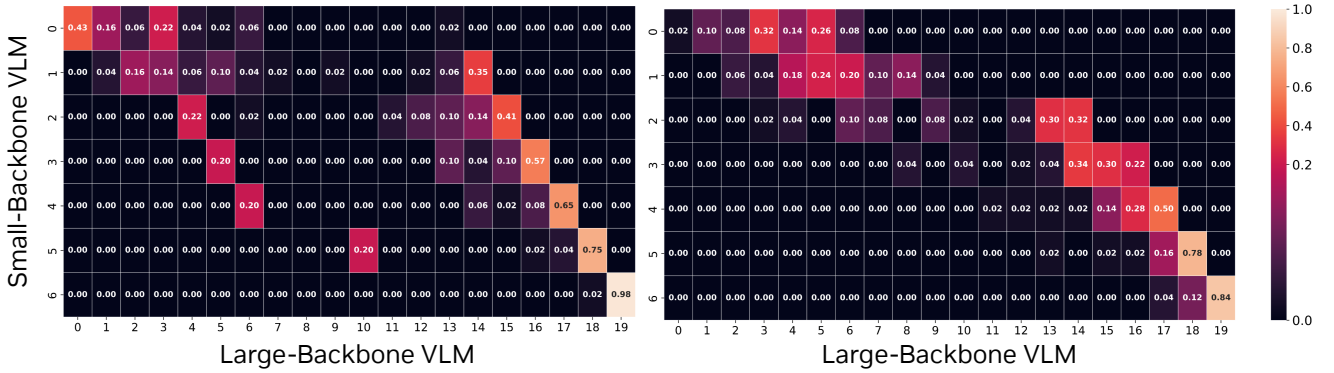


Figure 5. Distribution changes of the matched indices between small-backbone and large-backbone VLMs at the interaction step. The left figure shows the distribution at the beginning of training, while the right figure shows it at the end.

Furthermore, Tab. 3 compares VLSI’s performance with (a) open-source VLMs that incorporate multiple vision encoders, computer vision models, and additional rationale projector; (b) VLMs with modified architectures [49, 50] and various larger open- and proprietary closed-source VLMs. To evaluate where the effectiveness comes from, we analyze six key factors detailed in Fig. 4 and Tab. 4. Appendix D explains the detailed settings for ablation studies. These results indicate that (1) utilizing a more capable large-backbone VLM provides substantial performance benefits, suggesting that the choice of backbone significantly impacts the transfer of knowledge; (2) using larger-backbone VLM gets benefits; (3) KL divergence is more effective during interaction step than cross-entropy, and simultaneously using last-layer distillation boosts performances; and (4) additional SFT step is crucial for further performance gains, consistent with findings in [77, 85]. Appendix E represents VLSI’s text generation quality.

4.4. Discussion and Limitations

For Fig. 3, VLSI’s verbalizer is not re-trained but just utilized by the trained verbalizer of the small-backbone VLM. Interestingly, this verbalizer also works well at VLSI, demonstrating flexibility and indicating that it may

serve as a practical medium of interpretability. Additionally, Fig. 5 illustrates that, as the interaction step progresses, the small-backbone VLM gradually tries to learn about deeper layers’ responses of the large-backbone VLM, which can be considered accelerating the process of reaching an answer. While VLSI is highly effective, the large and small-backbone VLMs must share the same tokenizer and token index order when constructing VLSI. We will explore more general ways that accommodate different tokenizers and token index orders, potentially expanding VLSI’s applicability and scalability.

5. Conclusion

We present a new VLM family with 2B and 7B model sizes, VLSI: Verbalized Layers-to-Interactions, designed to build high-performing yet efficient-scale VLMs. This is accomplished by leveraging natural language-based distillation to transfer knowledge from large to small VLMs. We show VLSI achieves strong vision-language performances, suggesting that natural language is an important key in transferring knowledge not only for humans but also for AI. We hope to keep progress in further utilization of natural language and larger VLMs.

References

- [1] Anthropic. The claude 3 model family: Opus, sonnet, haiku. <https://www.anthropic.com>, 2024. 1, 7
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2
- [4] Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, and Xiang Bai. Llava-kd: A framework of distilling multimodal large language models. *arXiv preprint arXiv:2410.16236*, 2024. 3
- [5] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 1
- [6] Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning Ma, and Shanghang Zhang. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. *arXiv preprint arXiv:2501.01709*, 2025. 3
- [7] 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*, 2024. 2, 6
- [8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2
- [9] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 2, 6, 7
- [10] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 7, 2
- [11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 1, 5, 6, 7
- [12] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2
- [13] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2
- [14] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023. 2, 6
- [15] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024. 2, 6
- [16] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>, 2023. 3
- [17] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. <https://github.com/InternLM/xtuner>, 2023. 2
- [18] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2
- [19] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint*, 2024. 7
- [20] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 2, 7
- [21] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6
- [22] Yuning Du, Chenxia Li, Ruoyu Guo, Cheng Cui, Weiwei Liu, Jun Zhou, Bin Lu, Yehua Yang, Qiwen Liu, Xiaoguang Hu, et al. Pp-ocrv2: Bag of tricks for ultra lightweight ocr system. *arXiv preprint arXiv:2109.03144*, 2021. 2
- [23] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1
- [24] et al. Erfei Cui. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o, 2024. <https://sharegpt4o.github.io/>. 6
- [25] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data

- filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 6
- [26] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 2
 - [27] Yunhao Fang, Ligeng Zhu, Yao Lu, Yan Wang, Pavlo Molchanov, Jang Hyun Cho, Marco Pavone, Song Han, and Hongxu Yin. Vila²: Vila augmented vila. *arXiv preprint arXiv:2407.17453*, 2024. 2, 5
 - [28] Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language model. *arXiv preprint arXiv:2412.01282*, 2024. 3
 - [29] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver, 2017. Association for Computational Linguistics. 4
 - [30] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiaowu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 5, 1
 - [31] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 7, 2
 - [32] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wan-jun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023. 2, 7
 - [33] Chunjiang Ge, Sijie Cheng, Ziming Wang, Jiale Yuan, Yuan Gao, Jun Song, Shiji Song, Gao Huang, and Bo Zheng. Convlava: Hierarchical backbones as visual encoder for large multimodal models. *arXiv preprint arXiv:2405.15738*, 2024. 2
 - [34] Elizaveta Goncharova, Anton Razzhigaev, Matvey Mikhalechuk, Maxim Kurkin, Irina Abdullaeva, Matvey Skripkin, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. Omnifusion technical report. *arXiv preprint arXiv:2404.06212*, 2024. 2, 7
 - [35] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2
 - [36] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023. 3
 - [37] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 6
 - [38] Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 5
 - [39] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 6
 - [40] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 6, 7
 - [41] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. 6
 - [42] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018. 6
 - [43] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklucar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204*, 2024. 7
 - [44] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 5, 6, 7, 1
 - [45] 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*, pages 4999–5007, 2017. 6
 - [46] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2
 - [47] Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024. 3
 - [48] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*, 2023. 2
 - [49] Byung-Kwan Lee, Sangyun Chung, Chae Won Kim, Beomchan Park, and Yong Man Ro. Phantom of latent for large language and vision models. *arXiv preprint arXiv:2409.14713*, 2024. 2, 3, 5, 6, 7, 8

- [50] Byung-Kwan Lee, Sangyun Chung, Chae Won Kim, Beomchan Park, and Yong Man Ro. Trol: Traversal of layers for large language and vision models. *arXiv preprint arXiv:2406.12246*, 2024. 3, 5, 6, 7, 8
- [51] Byung-Kwan Lee, Chae Won Kim, Beomchan Park, and Yong Man Ro. Meteor: Mamba-based traversal of rationale for large language and vision models. *arXiv preprint arXiv:2405.15574*, 2024. 2, 7
- [52] Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. Collavo: Crayon large language and vision model. *arXiv preprint arXiv:2402.11248*, 2024. 2, 7
- [53] Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. Moai: Mixture of all intelligence for large language and vision models. *arXiv preprint arXiv:2403.07508*, 2024. 2, 7
- [54] 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*, 2023. 2
- [55] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multi-modal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 7, 1
- [56] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 5, 7
- [57] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024. 6
- [58] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 5, 1
- [59] Yuhui Li, Fangyu Wei, Chao Zhang, and Hongyang Zhang. Eagle: Speculative sampling requires rethinking feature uncertainty. *arXiv preprint arXiv:2401.15077*, 2024. 3
- [60] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 2, 5, 6
- [61] Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning*, pages 20852–20867. PMLR, 2023. 3
- [62] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. 2, 6
- [63] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoyebi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023. 5, 6, 7
- [64] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hal-lusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023. 5, 1
- [65] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2
- [66] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 5, 7
- [67] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 5, 7, 1
- [68] Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunsang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, et al. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. *arXiv preprint arXiv:2402.14905*, 2024. 2
- [69] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6, 7
- [70] 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. 2, 6, 7
- [71] 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. 5, 7, 1
- [72] 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. 5, 6, 1
- [73] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 6
- [74] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mml: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 2, 5, 6
- [75] Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, et al. Openelm: An efficient language model family with open-source training and inference framework. *arXiv preprint arXiv:2404.14619*, 2024. 2
- [76] Matthias Minderer, Alexey A. Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Thirty-seventh*

Conference on Neural Information Processing Systems, 2023. 2

- [77] Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. *arXiv preprint arXiv:2407.14679*, 2024. 3, 5, 8
- [78] OpenAI. Gpt-4v(ision) system card, 2023. <https://openai.com/research/gpt-4v-system-card>, Last accessed on 2024-02-13. 1, 2
- [79] OpenAI. Gpt-4v(ision) technical work and authors, 2023. <https://openai.com/contributions/gpt-4v>, Last accessed on 2024-02-13. 1
- [80] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2
- [81] Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. V1-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*, 2024. 2
- [82] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024. 2, 5, 7
- [83] Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, Siming Fu, et al. Llava-mod: Making llava tiny via moe knowledge distillation. *arXiv preprint arXiv:2408.15881*, 2024. 3
- [84] Nimit S Sohoni, Christopher R Aberger, Megan Leszczynski, Jian Zhang, and Christopher Ré. Low-memory neural network training: A technical report. *arXiv preprint arXiv:1904.10631*, 2019. 6, 7
- [85] Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Llm pruning and distillation in practice: The minitron approach. *arXiv preprint arXiv:2408.11796*, 2024. 3, 5, 8
- [86] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024. 6
- [87] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2
- [88] Omkar Thawakar, Ashmal Vayani, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Michael Felsberg, Tim Baldwin, Eric P. Xing, and Fahad Shahbaz Khan. Mobillama: Towards accurate and lightweight fully transparent gpt, 2024. 2
- [89] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2, 5, 7
- [90] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 6
- [91] 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*, 2024. 2, 7
- [92] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 2, 3, 5, 6, 7
- [93] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023. 5, 1
- [94] 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 llm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024. 2
- [95] Shilin Xu, Xiangtai Li, Haobo Yuan, Lu Qi, Yunhai Tong, and Ming-Hsuan Yang. Llavadi: What matters for multimodal large language models distillation. *arXiv preprint arXiv:2407.19409*, 2024. 3
- [96] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1, 6, 7
- [97] Jingkan Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 2
- [98] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 5
- [99] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2
- [100] 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. 1, 2, 5, 7
- [101] Weihao Yu, Zhengyuan Yang, Linfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Li-

- juan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024. [7](#), [2](#)
- [102] 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. [2](#), [5](#), [7](#)
 - [103] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023. [7](#)
 - [104] Xiang Yue, Tuncy Zheng, Ge Zhang, and Wenhui Chen. Mammoth2: Scaling instructions from the web. 2024. [7](#)
 - [105] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. [1](#), [2](#)
 - [106] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkan Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024. [2](#)
 - [107] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. [2](#)
 - [108] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024. [2](#)
 - [109] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024. [2](#), [6](#)
 - [110] Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference. *arXiv preprint arXiv:2403.14520*, 2024. [2](#)
 - [111] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024. [2](#)
 - [112] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [2](#)
 - [113] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024. [2](#)