This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Downscaling Intelligence: Exploring Perception and Reasoning Bottlenecks in Small Multimodal Models

Mark Endo, Serena Yeung-Levy
Stanford University
https://web.stanford.edu/~markendo/projects/downscaling_intelligence

Abstract

Scaling up multimodal models has enabled remarkable advances in visual understanding and reasoning, but practical demands call for smaller, efficient systems. In this work, we conduct a principled analysis of downscaling intelligence in multimodal models, examining how reduced large language model (LLM) capacity affects multimodal capabilities. Our initial findings reveal an interesting trend: LLM downscaling disproportionately affects visual capabilities, rather than abilities inherited from the LLM. We then examine whether this drop mainly reflects the expected decline in visual reasoning or a more fundamental loss of perceptual abilities. Isolating the effect of LLM downscaling on perception, we find performance still drops sharply, often matching or exceeding the impact on reasoning. To address this bottleneck, we introduce visual extraction tuning, which explicitly trains the model to extract instruction-relevant visual details consistently across tasks. With these extracted visual details, we then apply step-by-step reasoning to generate answers. Together, these components form our EXTRACT+THINK *approach, setting a new standard for efficiency and performance in this space.*

1. Introduction

Multimodal large language models (MLLMs) have become a dominant area of research in artificial intelligence, with large-scale systems demonstrating remarkable capabilities across areas spanning visual understanding and reasoning [12, 28]. Because of their impact and wide-ranging applications, increasing work has focused on understanding the scaling laws of these methods, investigating how increasing parameters and training data enhances their capabilities [1, 54, 58]. However, there exists a widespread demand for smaller, efficient models suitable for on-device applications. While this need for compact architectures has spurred the development of many small models [34, 40, 44, 47], the consequences of *downscaling intelligence* remain poorly understood. Namely, when smaller language models serve as the backbone of a multimodal system, which capabilities degrade most, and *why*?

In this work, we systematically investigate how downscaling large language model (LLM) size impacts multimodal behavior in order to (1) understand their practical limitations, (2) uncover the mechanisms behind their failures, and (3) develop targeted solutions to improve their performance (Figure 1). Starting with a controlled exploration across diverse visual instruction tuning tasks, we observe a striking pattern: tasks with the largest performance drop rely mainly on visual capabilities rather than the base LLM's abilities. Based on this observation, we use a decoupled framework separating perception and reasoning, allowing us to assess whether loss of visual capabilities stems mainly from an expected decline in visual reasoning or also from a more fundamental ability to interpret and extract visual information. Notably, we find that isolating the impact of LLM downscaling on perception still results in severe performance drops across tasks, often matching or exceeding the drops observed when isolating reasoning.

To address the limitations of small multimodal models, we first focus on the discovered perception bottleneck. Because instruction tuning exposes the model to diverse ways of interpreting and utilizing visual information, we hypothesize that this bottleneck arises from the model needing to acquire diverse skills to extract relevant visual information. Thus, we propose *visual extraction tuning*, a training paradigm in which the model explicitly learns to extract the visual details relevant to each instruction. We then enhance reasoning by applying step-by-step thinking over the extracted visual details, substantially enhancing performance without requiring any additional supervision on visual data. Our final two-stage approach, named EXTRACT+THINK, demonstrates extreme parameter and data efficiency. For example, our smaller variant surpasses the baseline two-stage PrismCaptioner framework [52] across a wide range of tasks using a perception module roughly ***12***× smaller and a reasoning module ***41***× smaller. Even when training from scratch utilizing visual extraction tuning, our ap-

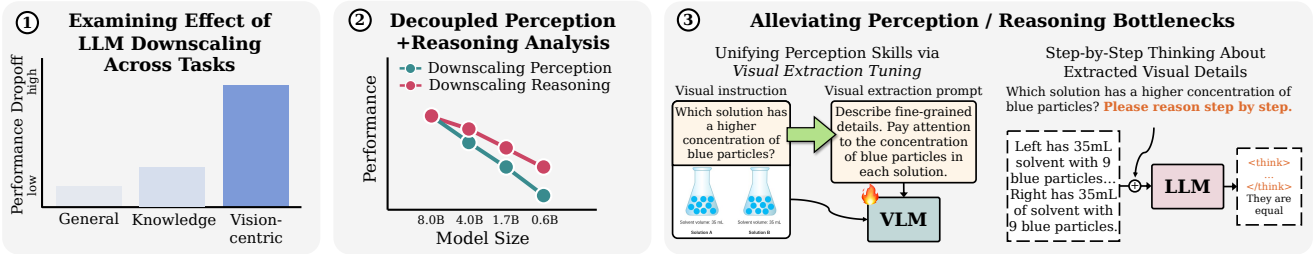


Figure 1. **Overview.** (*1*) We first analyze how downscaling language model size affects multimodal performance, finding that tasks which rely more heavily on the base LLM (e.g., general or knowledge tasks) are largely unaffected, whereas visually-demanding tasks show a disproportionate drop. (*2*) To uncover the mechanisms underlying the deteriorating visual capabilities under LLM downscaling, we perform a decoupled analysis of perception and reasoning, revealing that perception (alongside reasoning) is a critical bottleneck in small multimodal models. (*3*) To address these limitations, we present a two-stage perception–reasoning framework, featuring *visual extraction tuning*–which trains the model to extract instruction-relevant visual details consistently across tasks–coupled with step-by-step reasoning about the extracted visual details.

proach improves over LLaVA-OneVision-0.5B [35] while using ***95%*** fewer visual training samples. Together, our work offers the first systematic characterization of downscaling effects in multimodal models and introduces effective solutions to their bottlenecks, laying the groundwork for future advances in small-scale multimodal intelligence.

2. Related Work

Small MLLMs. The development of small yet powerful vision-language models (VLMs) has been a significant focus of recent research, aiming to provide strong multimodal capabilities in resource-constrained environments. This includes models like Moondream [34], Phi-3-Vision [47], SmolVLM [44], and MiniCPM [24], as well as compact variants of Gemma 3 [57], DeepSeek-VL [40], Qwen-VL series [3, 11, 60], LLaVA-OneVision [35], and InternVL [9]. While these models demonstrate impressive general capabilities, their failure modes–especially those concerning visual capabilities, and in particular perception–remain poorly understood. Findings across prior works are inconsistent: some studies suggest that scaling LLM size has little effect on perception [35, 52], while others find that perception-heavy tasks such as OCR and Chart VQA are highly sensitive to model size [22]. These discrepancies highlight the need for an in-depth analysis, which we undertake to examine how downscaling LLMs affects visual capabilities and the mechanisms behind their failures.

Failures of MLLMs. A number of works have revealed shortcomings of state-of-the-art MLLMs on perceptual and visual reasoning tasks. For example, [17] discovers that even the best-performing multimodal models perform near-randomly on perceptual tasks that humans can solve quickly. For visual spatial planning, [62] identifies fundamental deficiencies in the models' visual perception and reasoning abilities. Many works demonstrate that VLMs often struggle with visual reasoning puzzles that require strong pattern recognition and abstract reasoning [10, 42, 63]. Examining why VLMs fail on visual tasks, studies often find that visual information from the vision encoder is inadequately utilized by the language model [16, 39, 69], attributing failures to limited exposure to relevant visual data and mitigating this with more representative training data. However, these works often focus on much bigger and more powerful models, leaving the failures from LLM downscaling largely unexplored.

3. LLM Downscaling Exploration

In the first part of this work, we conduct a controlled study to examine how reducing language model size impacts multimodal task performance, aiming to understand the limitations of small models as general visual assistants and the causes of their failures. After covering model and setup preliminaries (§3.1), we present our results on how downscaling model size impacts performance across various tasks (§3.2). We find that the tasks most affected by downscaling language model size are **not** those that heavily rely on the base LLM, but rather those emphasizing visual processing. Next, we investigate whether the decline in performance under LLM downscaling stems primarily from weakened visual reasoning or if it also reflects a more fundamental impairment to perception (§3.3). When isolating the effect of LLM downscaling on perception, we find that performance still drops sharply–often matching or exceeding the decline observed when isolating its effect on reasoning–indicating that a central limitation of small multimodal models arises from a degradation in their ability to recognize and understand visual information.

3.1. Preliminaries

In this work, we focus on the popular multimodal LLM approach of taking a language model trained on a broad corpora of text as the foundation, integrating a pre-trained vi-

sion encoder with a simple projector to connect the visual representations to the LLM token space, and training the combined system with visual instruction tuning data. We go over details about each component of our setup below.

Architecture. Our architectural decisions are guided by the principle of using well-established and widely validated design choices, ensuring that the findings are broadly applicable and impactful for future work. Hence, we use Qwen3 series (8B, 4B, 1.7B, and 0.6B sizes) [65] for the LLM, SigLIP [66] as the vision encoder, and a 2-layer MLP as the connector. We use the Higher AnyRes with Bilinear Interpolation scheme from [35] for visual processing.

Data. We use a broad range of visual instruction tuning datasets for our exploration. To enable a more controlled setting for analyzing task performance, we focus on data that includes both training sets and benchmark evaluations. Specifically, for single-image tasks we leverage [6], and for multi-image tasks we utilize the subset of M4-Instruct data that includes evaluation benchmarks [36]. We additionally include PieAPP [51] to ensure sufficient data for the Perceptual Similarity task. All datasets are listed in Table 1. For the connector pretraining stage, we utilize BLIP558K.

Training Recipe. Based on [35], after pre-training the connector for language-image alignment, we perform visual instruction tuning, fine-tuning all parameters on single-image data (574K) and then on a combination of the multi-image data (309K) and 150K randomly sampled single-image examples. We use the same batch size, learning rates for model parameters, and image resolutions as [35].

Visual Instruction Tuning Data	
Single-Image (574K)	OCR-VQA [48] (165K)
VQAv2 [18] (82.8K)	ImageNet [13] (130K)
VizWiz [21] (20.5K)	Grounding (55.9K)
ScienceQA [41] (12.7K)	└ RefCOCO [31]
TextVQA [55] (34.6K)	└ RefCOCO+ [43]
GQA [26] (72.1K)	└ RefCOCOg [43]
Multi-Image (309K)	Text-Rich VQA (21.3K)
Spot the Difference (28.9K)	└ WebQA [5]
└ Spot-the-Diff [30]	└ TQA [32]
└ Birds-to-Words [14]	└ OCR-VQA [48]
└ CLEVR-Change [23, 50]	└ DocVQA [45]
Image Edit Instruction (67.7K)	Multi-Image-VQA (22.4K)
└ HQ-Edit [27]	└ MIT-StateCoherence [29]
└ MagicBrush [68]	└ MIT-PropertyCoherence [29]
└ IEdit [56]	└ RecipeQA-ImageCoherence [64]
Visual Story Telling (67.5K)	└ VISION [2]
└ AESOP [53]	Puzzle (Raven) [67] (35K)
└ FlintstonesSV [20]	Perceptual Similarity (66.4K)
└ PororoSV [37]	└ NIGHTS [15]
└ VIST [25]	└ PieAPP [51]

Table 1. List of used visual instruction tuning data. Task colors indicate their relative proportion in the data mixture.

3.2. Analyzing impact across tasks

Tasks with largest performance drops under LLM downscaling rely heavily on visual processing, not the base

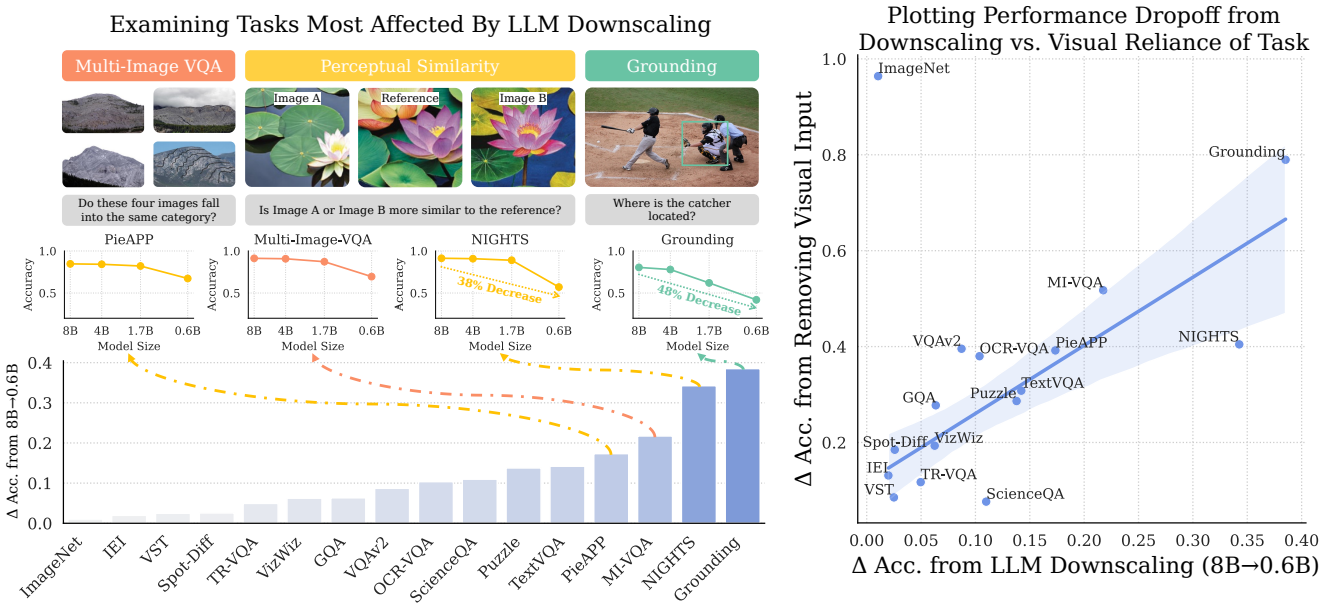


Figure 2. **LLM downscaling exploration.** (*Left*) **Performance dropoff from LLM downscaling most notable for visually demanding tasks.** Tasks like Grounding and Perceptual Similarity (e.g., NIGHTS and PieAPP) which primarily focus on visual processing are most affected by LLM downscaling, rather than tasks which rely heavily on the base LLM (such as ScienceQA evaluating knowledge or GQA assessing general abilities). (*Right*) **The more a task's performance declines under LLM downscaling, the greater it depends on visual information**. As the impact of LLM downscaling increases (8B → 0.6B), so does the task's reliance on visual information (measured by performance difference with and without visual input). IEI=Image Edit Instruction, VST=Visual Story Telling, Spot-Diff=Spot the Difference, TR-VQA=Text-Rich VQA, MI-VQA=Multi-Image-VQA. Full plots for all datasets are provided in the supplemental material.

LLM. As shown in Figure 2(Left), most tasks exhibit modest performance decline when downscaling the language model size from 8B to 0.6B, except for a few tasks which exhibit much larger deterioration. Interestingly, rather than these tasks depending heavily on the base LLM (such as ScienceQA, which assesses knowledge, or GQA, which evaluates general abilities), they instead rely primarily on visual processing. For example, Grounding drops 48% and NIGHTS (Perceptual Similarity) declines 38% when downscaling from 8B to 0.6B.

The greater the impact of LLM downscaling, the more the task relies on visual information to be solved. While our analysis so far has focused on the few tasks most affected by model downscaling, here we extend the analysis to the full set of datasets. To better understand how a dataset's sensitivity to LLM downscaling relates to how vision-centric the task is, we plot the performance difference between the 8B and 0.6B LLMs against the difference in performance with and without visual input (using the 8B LLM). As shown in Figure 2(Right), most datasets exhibit an approximately linear trend: as the impact of LLM downscaling increases, so does the task's reliance on visual information. The exception is ImageNet, where the small model achieves very strong performance but blind performance is near zero. This likely occurs because the perception required is notably simple and the task comprises a large portion of the visual instruction tuning data.

Takeaway 1: LLM downscaling is most detrimental to vision-centric capabilities rather than base LLM abilities.

Discussion: While previous studies connect poor utilization of visual representations in multimodal models to limited training data [16, 39, 69], we observe a distinct behavior: ***even when the training mixture ensures coverage across all evaluated tasks, visually-intensive tasks deteriorate most as LLM size decreases***. Overall, our findings suggest that in multimodal models trained using visual instruction tuning, processes related to understanding and/or reasoning about visual information are significantly impaired by downscaling the language model.

3.3. Decoupled perception / reasoning analysis

Our findings in the previous section are intriguing, but ***the reason behind the observed trend remains unclear***. Namely, vision-centric tasks generally require two essential capabilities: *perception*, the foundational ability to recognize, extract, and understand visual details, and *reasoning*, the downstream ability to operate on extracted visual information to formulate answers. While our analysis showed that the visual capabilities of multimodal models degrade significantly under LLM downscaling, it did not reveal the mechanisms underlying these failures. Given that reasoning depends on model scale for textual tasks [38], we expect visual reasoning to decline under downscaling; how-

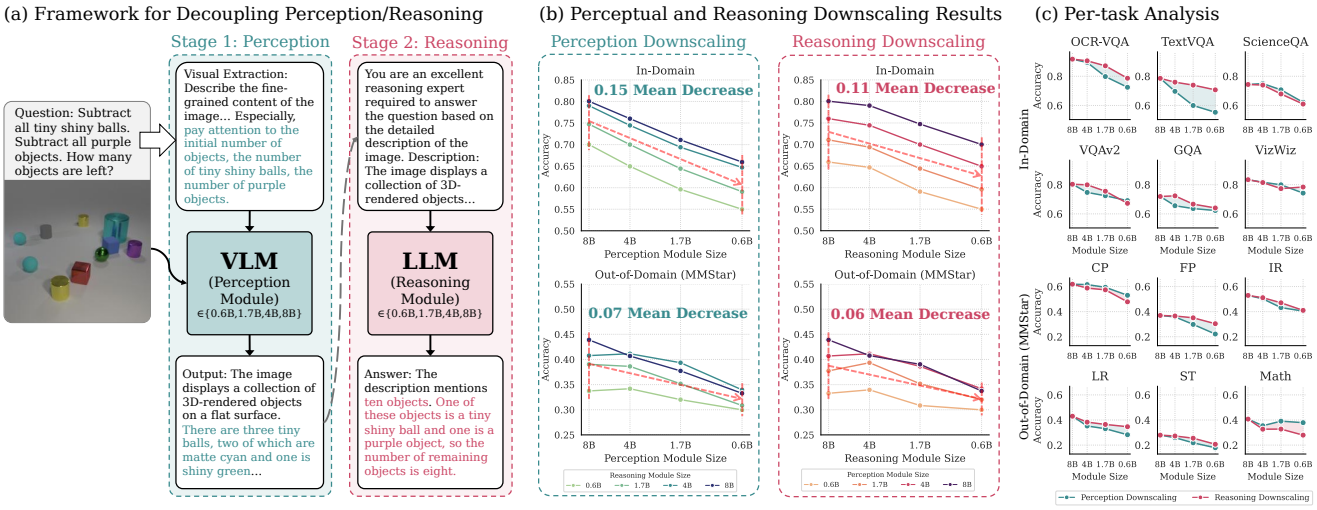


Figure 3. **Decoupled perception and reasoning downscaling analysis.** *(a)* **Decoupled Setup.** We disentangle perceptual and reasoning abilities using a two-stage framework: the perception module (VLM) first extracts visually relevant information, then the reasoning module (LLM) generates answers based on the extracted visual information. *(b)* **Perception and reasoning emerge as key bottlenecks under LLM downscaling.** We see that LLM downscaling of either the perception module or reasoning module largely degrades in-domain and out-of-domain task performance. *(c)* **Perceptual degradation limits performance across tasks.** Even for tasks targeting visual reasoning (e.g., IR and LR), downscaling perception has an impact comparable to–or even exceeding–that of downscaling reasoning. In this per-task analysis, the non-downscaled module is set at 8B. CP=Coarse Perception, FP=Fine-grained Perception, IR=Instance Reasoning, LR=Logical Reasoning, ST=Science & Technology.

ever, the effect on the more foundational process of perception is highly uncertain and warrants further study. Thus, in this section, we perform a rigorous analysis separating the effects of LLM downscaling on perception and reasoning to better understand the causes of the observed behavior. We detail our decoupled setup for this analysis and present our results below.

Setup. To study perception and reasoning independently under downscaling, we apply the Prism framework [52], which separates these two processes. As shown in Figure 3(a), each question is answered in two stages. In the first stage, the question is converted into a prompt to extract all visually relevant information, and both this question-specific instruction and the image are fed into a multimodal model (perception module) to obtain the important visual information. In the second stage, an LLM (reasoning module) uses the extracted visual information to reason and generate the final answer. Using this setup, we independently downscale the LLM in each module to measure how the two abilities are affected by LLM size.

In our analysis, we utilize the same multimodal models from §3.2 as the perception module and their corresponding Qwen3 series models [65] as the reasoning module. Differing from [52], we convert the prompts offline using one model type (Qwen3-8B), so that the questions remain consistent across setups and the model's ability to generate question-specific instructions does not influence our analysis of perception and reasoning. For evaluation datasets, as the reasoning module is not trained on the output distribution of the visual tasks from §3, we utilize the converted multiple-choice format of these datasets from AutoConverter [70], which has proven to enable objective evaluation under variability in natural language responses. We exclude Grounding and ImageNet from this analysis as these are purely perceptual tasks. We additionally evaluate on the carefully curated, out-of-domain benchmark MMStar [8], which assesses both perceptual and reasoning abilities.

Results. ***LLM downscaling expectedly hinders visual reasoning.*** As shown in Figure 3(b), we find that downscaling the reasoning module size has a considerable impact on performance across tasks, confirming that visual reasoning is a critical bottleneck for small multimodal models.

LLM downscaling markedly impairs perceptual abilities, affecting a wide spectrum of tasks. More notably, in Figure 3(b) we also observe that LLM downscaling of the perception module has as substantial an effect on performance, where downscaling from 8B to 0.6B causes an average accuracy drop of 0.15 for in-domain data and 0.07 for out-of-domain data. As shown in Figure 3(c), even for tasks that target visual reasoning (such as Instance Reasoning and Logical Reasoning), downscaling the perception module has an impact on performance comparable to, or even exceeding, that of downscaling the reasoning module. This likely occurs because the foundational ability to understand visual information is a prerequisite for successfully performing downstream reasoning.

Takeaway 2: While LLM downscaling expectedly impairs visual reasoning, isolating its impact solely on perception still reveals severe performance degradation across a wide range of tasks, often matching or exceeding its effect on reasoning.

Discussion. This section highlights an important and previously undiscovered phenomenon. The original Prism work [52] used a relatively small LLM for the perception module (e.g., InternLM2-1.8B [4]) and a much larger LLM for the reasoning module (LLaMA-3-70B [19] and ChatGPT)), based on the assumption that perception is far less sensitive to LLM scale than reasoning. Although reasoning is naturally expected to degrade more than perception, we find that its impact on performance is surprisingly similar to that of perceptual abilities. Thus, ***perception*** *(alongside reasoning)* ***emerges as a central bottleneck in small multimodal models***. Given that the visual representations are fixed across model setups, *what drives this perceptual decline*?

We hypothesize that this perception bottleneck arises from a fundamental limitation of the visual instruction tuning paradigm under LLM downscaling. Namely, visual instruction tuning exposes the model to various ways of recognizing, understanding, and extracting visual information. We posit that this variability requires the model to acquire diverse skills for interpreting instructions and extracting the relevant visual information. The *Quantization Model* of neural scaling laws [46] offers a theoretical lens: model skills can be "quantized" into discrete chunks (quanta), and scaling laws limit the total number a model can effectively learn from the training data. Because visual instruction tuning requires the model to learn many skills to process visual information across diverse tasks, smaller models have weaker perceptual capabilities.

Hypothesis: LLM downscaling's effect on perception arises from the heterogeneity of how perception is learned under visual instruction tuning.

As part of the following section, we leverage this hypothesis to guide method advancements aimed at improving the perceptual abilities of small multimodal models.

4. EXTRACT+THINK

Having shown that LLM downscaling weakens both foundational perception and downstream reasoning, we conclude by proposing solutions that address these limitations and move toward a high-performing generalist small multimodal model. We focus our efforts on the two-stage frame-

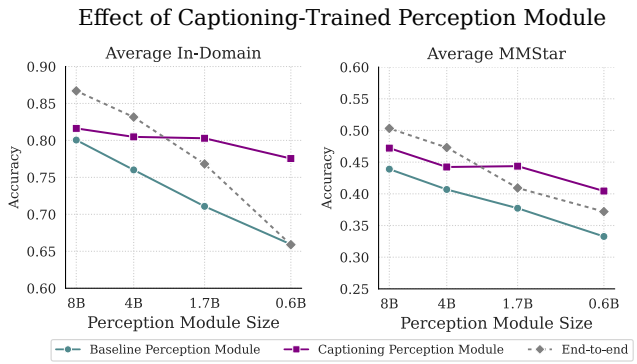
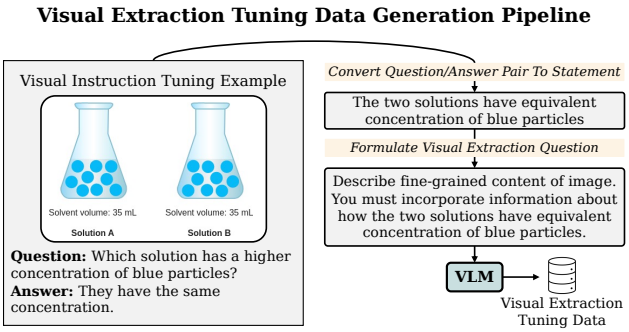


Figure 4. **Captioning alleviates perception bottleneck.** Decoupled frameworks use an 8B reasoning module.



Perception Module	Size	In-domain	MMStar
Captioning	0.6B	77.6	40.4
+ Visual Extraction	0.6B	82.8	44.0
Δ		**+5.2**	**+3.6**
Captioning	1.7B	80.3	44.4
+ Visual Extraction	1.7B	84.4	49.0
Δ		**+4.1**	**+4.6**

Figure 5. **Visual extraction tuning.** *(Top)* **Simple pipeline for generating visual extraction tuning data.** Given a visual instruction tuning example, it is converted to a *visual extraction* task by prompting a VLM to describe fine-grained visual details relevant to the original question. *(Bottom)* **Visual extraction tuning enhances perception.** Post-training on visual extraction data improves both in-domain and out-of-domain (MMStar) performance. Size indicates the number of parameters of the perception module's LLM. All setups use an 8B reasoning module.

work, as it provides a modular approach to study and improve each bottleneck (though we note it adds the challenge of capturing all relevant visual information in text; see supplementary material for more details).

First, we look at improving perception of small models by streamlining the learning of perceptual skills through a new *visual extraction tuning* paradigm (§4.1). Next, we look at how to better utilize the extracted visual information by allowing the reasoning module to reason step-by-step (§4.2). Together, these two components comprise our final approach EXTRACT+THINK, which sets a new standard in parameter- and data-efficient multimodal modeling (§4.3), offering an effective path toward generalist small models.

4.1. Visual extraction tuning

We first aim to alleviate the foundational perception bottleneck for small multimodal models. As discussed in §3.3, we hypothesize that the perception bottleneck on small multimodal models arises from the model needing to acquire a diverse set of skills to extract relevant visual information across a wide range of tasks. Thus, a natural approach to improve performance in downscaled LLMs is to increase the homogeneity of how visual information is extracted. In this section, we first assess captioning as a baseline method to achieve this, and then propose a new training paradigm, *visual extraction tuning*, which demonstrates strong abilities in enhancing perception in small multimodal models.

Captioning baseline. A simple way to unify perceptual skills for visual question answering is to train the perception module as a captioner. We therefore post-train the perception module on ALLaVA-4V [7], a 950K caption dataset. As shown in Figure 4, this approach mitigates the effect of LLM downscaling and even outperforms end-to-end baselines at smaller scales (0.6B, 1.7B). However, ***captioning introduces two key limitations***. First, the two-stage framework is not merely captioning plus reasoning; the first stage should extract question-relevant visual details, which captioning does not teach. Second, visual instruction tuning often involves specialized, domain-specific data. Training solely on general captioning datasets limits domain-specific understanding, as the model is not taught to interpret specialized visual concepts present in those domains. Thus, an alternative approach is required to address these limitations.

Visual extraction tuning. Here, we propose *visual extraction tuning* as a solution to unify the perceptual abilities of the perception module while enabling it to extract question-relevant information and operate effectively across the diverse domains present in visual instruction tuning. Provided visual instruction data, we design a simple pipeline that transfers this data to the task of *visual extraction*, where the goal is to generate all visual information relevant to answering the instruction, aligning precisely with the role of the perception module in the two-stage framework.

As shown in Figure 5(Top), given a visual instruction tuning example, we first convert the question–answer pair into a declarative statement by prompting a model. We then integrate this declarative statement into a prompt that asks the model to describe fine-grained visual details, with explicit emphasis on information relevant to the declarative statement. Finally, this instruction, together with the image, is provided to a model to generate the visual extraction re-

sponse. For simplicity, we use Qwen3VL-8B [11] throughout the entire process; although the first step is text-only, this model has shown strong performance on purely textual tasks. We apply this pipeline to 382K training samples corresponding to the assessed in-domain tasks, and post-train our captioning perception module with this data. Additional details on the generation process, including prompt templates and data examples, are provided in the supplemental material.

Results. As shown in Figure 5(Bottom), we find that additionally post-training under the visual extraction tuning paradigm offers large performance improvements over the captioning baseline on both in-domain data and the out-of-domain MMStar benchmark. Specifically, in-domain performance increases by 5.2 when the perception module uses a 0.6B LLM and by 4.1 when it uses a 1.7B LLM. On the MMStar benchmark, performance improves by 3.6 for the 0.6B LLM and by 4.6 for the 1.7B LLM.

Takeaway 3: Visual extraction tuning proves an effective and efficient solution for alleviating the perception bottleneck of small multimodal models.

4.2. Step-by-step visual reasoning

Chain-of-Thought (CoT) reasoning is a widely studied method for improving LLM reasoning [33, 61]. In our two-stage framework, although the reasoning module is not trained on visual data, text serves as an interface connecting perception and reasoning. Therefore, we expect that encouraging step-by-step in the reasoning module will directly enhance visual reasoning without requiring training.

Approach. The Qwen3 model [65], which we utilize for the reasoning module, is capable of complex, multi-step reasoning by enabling thinking mode. Thus, we activate thinking mode and modify the prompt: instead of directly requesting the answer like before, we instruct the model to reason step-by-step. Since Qwen3 produces long reasoning chains, to improve efficiency we limit self-reflection with NoWait [59] and limit the thinking budget to 4096 tokens using [49]. Additional information is available in the supplementary material.

Results. As shown in Figure 6, incorporating CoT reasoning substantially improves out-of-domain performance across all LLM sizes. For in-domain tasks, we observe a more nuanced behavior where the performance degradation under LLM downscaling becomes more concave when reasoning is enabled: the 8B and 0.6B models perform similarly with or without CoT, but at intermediate scales (4B and 1.7B), CoT yields notable gains. This suggests that ***while CoT does not fully resolve the reasoning bottleneck in smaller multimodal models, it meaningfully enhances performance***–particularly at mid-range LLM sizes, where it brings results closer to those of larger models.

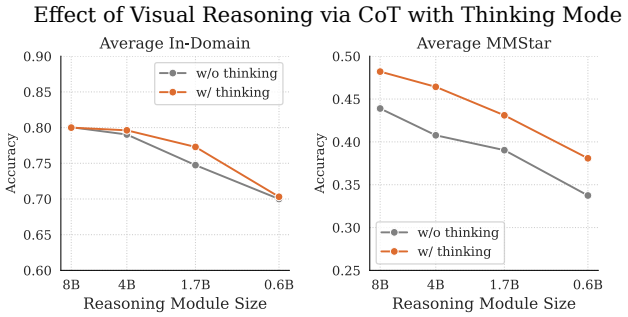


Figure 6. **CoT reasoning enhances in-domain and out-of-domain performance.** Performance gains exhibited at intermediate model scales (4B and 1.7B) for in-domain tasks, while out-of-domain performance improves across all LLM sizes. Both setups use 8B baseline perception module. Per-task plots provided in supplemental material.

Takeaway 4: Utilizing CoT boosts visual reasoning capabilities without requiring any supervision on visual data.

4.3. Distilling insights

Guided by these insights, we now present our final approach, Extract+Think. Specifically, we employ the perception module trained under our proposed visual extraction paradigm from §4.1 and a reasoning module enhanced with CoT reasoning from §4.2. Based on our finding that CoT does not fully resolve the reasoning bottleneck in smaller multimodal models, we adopt a larger LLM for the reasoning module than for the perception module (while keeping both models within a lightweight regime). We present two configurations: one where the perception module's LLM size is 0.6B and the reasoning module's is 1.7B, and a larger setup with 1.7B and 4B, respectively.

For the perception module, we test two configurations–one post-trained under the visual extraction paradigm starting from a captioning model (as is done in §4.1), and another trained from scratch without prior instruction tuning or captioning. We compare Extract+Think against both end-to-end baselines and other decoupled methods, including PrismCaptioner [52], the original decoupled setup from §3.3, and the captioning baseline in §4.1 (denoted Caption+Think).

Results. Extract+Think ***substantially outperforms decoupled baselines and even competes with end-to-end models trained at vast scale***. As shown in Table 2, even our smaller variant surpasses the largest PrismCaptioner model on both in-domain and out-of-domain tasks, with a perception module LLM roughly ***12× smaller*** and a reasoning module ***41× smaller***. It also outperforms LLaVA-OneVision-0.5B by 12.9% on in-domain data and 19.5% on the out-of-domain MMStar benchmark, while using ***73%***

	LLM Size		#Vis. Data	In-Domain (Multiple-Choice [70])							Out-of-Domain (MMStar)						
				OCR-VQA	TextVQA	ScienceQA	VQAv2	GQA	VizWiz	Average	CP	FP	IR	LR	ST	Math	Average
End-to-End																	
LLaVA-OneVision [35]	0.5B		8.8M	69.5	77.2	55.7	75.7	73.6	74.7	71.1	63.2	31.1	42.1	35.8	30.0	31.4	39.0
InternVL2.5 [9]	0.5B		64M	79.8	89.1	**89.8**	82.0	75.4	83.0	83.2	**69.9**	38.8	53.9	37.7	**39.3**	49.7	48.2
SmolVLM [44]	1.7B		unk.	72.9	81.4	79.7	75.5	70.6	75.1	75.9	69.2	30.6	45.9	37.9	29.8	34.2	41.3
Baseline (from §3)	0.6B		1.0M	41.1	71.3	67.9	71.2	69.5	74.5	65.9	58.1	30.4	39.3	35.1	27.4	32.9	37.2
Baseline (from §3)	1.7B		1.0M	73.4	83.4	76.2	77.8	74.3	75.8	76.8	63.9	35.1	45.6	38.5	27.5	34.9	40.9
Decoupled Models	P	R															
PrismCaptioner [52]	1.8B	70B	1.9M	89.2	72.7	64.6	77.8	66.0	82.3	75.4	64.0	38.8	55.8	36.7	23.0	33.1	41.9
PrismCaptioner [52]	7.0B	70B	1.9M	91.5	77.0	68.1	79.9	67.5	85.8	78.3	66.7	38.5	**61.5**	39.8	26.7	40.4	45.7
Baseline (from §3.3)	0.6B	4.0B	1.0M	71.8	50.7	63.0	67.6	62.3	72.3	64.6	58.2	25.4	38.7	26.5	20.7	34.2	34.0
Baseline (from §3.3)	1.7B	4.0B	1.0M	79.4	59.4	65.0	71.6	64.5	76.4	69.4	62.2	30.4	46.3	32.0	29.2	35.9	39.4
CAPTION+THINK	0.6B	1.7B	2.0M	84.9	80.6	60.6	74.7	66.2	83.0	75.0	60.7	37.2	51.9	38.9	27.0	42.4	43.0
CAPTION+THINK	1.7B	4.0B	2.0M	89.2	84.8	68.9	80.5	72.1	84.3	80.0	64.6	37.6	53.4	48.6	33.9	**56.2**	49.0
EXTRACT+THINK†	0.6B	1.7B	0.4M	86.9	79.8	69.9	76.6	72.5	82.1	78.0	65.2	41.7	49.7	37.5	21.9	39.8	42.6
EXTRACT+THINK†	1.7B	4.0B	0.4M	91.5	84.0	71.3	**84.6**	**77.8**	86.9	82.7	64.4	40.7	58.4	46.3	35.5	43.1	48.1
EXTRACT+THINK	0.6B	1.7B	2.4M	89.4	81.8	72.2	78.0	74.7	85.6	80.3	64.5	41.7	54.9	43.0	28.3	47.3	46.6
EXTRACT+THINK	1.7B	4.0B	2.4M	**92.9**	**90.1**	75.2	84.4	**77.8**	**91.3**	**85.3**	68.5	**47.8**	59.2	**53.3**	33.0	53.8	**52.6**

Table 2. **EXTRACT+THINK demonstrates extreme effectiveness as a generalist small multimodal model.** Even the smaller EXTRACT+THINK variant surpasses LLaVA-OneVision-0.5B by up to 19.5% while using 73% fewer visual samples, and outperforms the larger PrismCaptioner model on both in-domain and out-of-domain tasks with a perception module roughly 12× smaller and a reasoning module 41× smaller. The EXTRACT+THINK† configuration, trained from scratch under the visual extraction tuning paradigm, demonstrates robust performance using very minimal data. #Vis. Data denotes the amount of visual data used for training (excluding the connector pre-training stage). P=Perception Module, R=Reasoning Module. For MMStar, CP=Coarse Perception, FP=Fine-grained Perception, IR=Instance Reasoning, LR=Logical Reasoning, ST=Science & Technology. The best results are bolded and the second best are underlined. Inference latency analysis included in supplementary material.

fewer visual samples.

Visual extraction tuning offers a data-efficient solution for generalist small multimodal models. Looking at our configuration trained from scratch without prior visual training (denoted as EXTRACT+THINK† in Table 2), the smaller variant improves over LLaVA-OneVision-0.5B by 9.7% on in-domain data while using ***95%*** fewer visual samples. This setup also outperforms the 1.7B baseline trained directly on the in-domain instruction tuning data, and even exceeds the in-domain performance of the comparable CAPTION+THINK configuration, which was trained on both the in-domain instruction tuning data and 950K additional captioning examples. Overall, these results demonstrate that visual extraction tuning is an extremely effective and efficient paradigm for training small multimodal models.

5. Conclusion

In this work, we provide a systematic study of how language model downscaling affects multimodal task performance, revealing that visually demanding tasks are disproportionately impacted. Through a decoupled analysis, we identify that both foundational perception and downstream reasoning abilities are central bottlenecks when downscaling LLMs. To address these limitations, we introduce a two-stage perception–reasoning framework that employs visual extraction tuning to enhance the model's ability to extract relevant visual details across tasks and applies step-by-step reasoning over the extracted data without requiring additional visual training. Our final approach establishes a highly parameter- and data-efficient paradigm for training small multimodal models, setting a new standard for efficiency and performance in this space.

This work lays the groundwork for future research on downscaling of multimodal models. On the analysis side, future studies can explore downscaling across a broader range of model sizes, assess how the downscaling of visual representations compares to that of language models, and incorporate data size as a variable to examine how downscaling behavior varies across different scales. On the methodological side, future research can further investigate the visual extraction tuning paradigm in comparison to visual instruction tuning and evaluate its effectiveness with larger language models.

Acknowledgments. This work is supported in part by the National Science Foundation (NSF) under Grant No. 2026498 and the NSF Graduate Research Fellowship Program under Grant No. DGE-2146755 (for M.E.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any other entity.

References

[1] Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pages 265–279. PMLR, 2023. 1

[2] Haoping Bai, Shancong Mou, Tatiana Likhomanenko, Ramazan Gokberk Cinbis, Oncel Tuzel, Ping Huang, Jiulong Shan, Jianjun Shi, and Meng Cao. Vision datasets: A benchmark for vision-based industrial inspection. *arXiv preprint arXiv:2306.07890*, 2023. 3

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[4] 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. 5

[5] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504, 2022. 3

[6] Cheng Chen, Junchen Zhu, Xu Luo, Heng T Shen, Jingkuan Song, and Lianli Gao. Coin: A benchmark of continual instruction tuning for multimodel large language models. *Advances in Neural Information Processing Systems*, 37: 57817–57840, 2024. 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 lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024. 6

[8] 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? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 5

[9] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 2, 8

[10] Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16259–16273, 2024. 2

[11] Alibaba Cloud. Qwen3-vl. https://huggingface.co/collections/Qwen/qwen3-vl, 2025. Online. 2, 7

[12] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[14] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge J. Belongie. Neural naturalist: Generating fine-grained image comparisons. In *Conference on Empirical Methods in Natural Language Processing*, 2019. 3

[15] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, pages 50742–50768, 2023. 3

[16] Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025. 2, 4

[17] 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. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024. 2

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3

[19] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5

[20] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 598–613, 2018. 3

[21] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 3

[22] Junlin Han, Shengbang Tong, David Fan, Yufan Ren, Koustuv Sinha, Philip Torr, and Filippos Kokkinos. Learning to see before seeing: Demystifying llm visual priors from lan-

guage pre-training. *arXiv preprint arXiv:2509.26625*, 2025. 2

[23] Mehrdad Hosseinzadeh and Yang Wang. Image change captioning by learning from an auxiliary task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2725–2734, 2021. 3

[24] 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. 2

[25] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239, 2016. 3

[26] 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*, pages 6700–6709, 2019. 3

[27] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024. 3

[28] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[29] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015. 3

[30] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034, 2018. 3

[31] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 3

[32] 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. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384, 2017. 3

[33] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 7

[34] Vik Korrapati. Moondream. https://moondream.ai/, 2025. Online. 1, 2

[35] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 3, 8

[36] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 3

[37] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6329–6338, 2019. 3

[38] Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*, 2025. 4

[39] Junteng Liu, Weihao Zeng, Xiwen Zhang, Yijun Wang, Zifei Shan, and Junxian He. On the perception bottleneck of vlms for chart understanding. *arXiv preprint arXiv:2503.18435*, 2025. 2, 4

[40] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 1, 2

[41] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 3

[42] Mikołaj Małkiński, Szymon Pawlonka, and Jacek Mańdziuk. Reasoning limitations of multimodal large language models. a case study of bongard problems. *arXiv preprint arXiv:2411.01173*, 2024. 2

[43] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 3

[44] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 1, 2, 8

[45] 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. 3

[46] Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36:28699–28722, 2023. 5

[47] Microsoft. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 1, 2

[48] 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)*, pages 947–952. IEEE, 2019. 3

[49] Zach Mueller. Limiting qwen 3's thinking: How to make qwen3 think less. https://muellerzr.github.io/til/end_thinking.html, 2025. Published April 30, 2025. 7

[50] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633, 2019. 3

[51] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 3

[52] Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of vlms. *Advances in Neural Information Processing Systems*, 37:111863–111898, 2024. 1, 2, 5, 7, 8

[53] Hareesh Ravi, Kushal Kafle, Scott Cohen, Jonathan Brandt, and Mubbasir Kapadia. Aesop: Abstract encoding of stories, objects, and pictures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2052–2063, 2021. 3

[54] Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. *arXiv preprint arXiv:2504.07951*, 2025. 1

[55] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 3

[56] Hao Tan, Franck Dernoncourt, Zhe L. Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. In *Annual Meeting of the Association for Computational Linguistics*, 2019. 3

[57] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 2

[58] Changyao Tian, Hao Li, Gen Luo, Xizhou Zhu, Weijie Su, Hanming Deng, Jinguo Zhu, Jie Shao, Ziran Zhu, Yunpeng Liu, et al. Navil: Rethinking scaling properties of native multimodal large language models under data constraints. *arXiv preprint arXiv:2510.08565*, 2025. 1

[59] Chenlong Wang, Yuanning Feng, Dongping Chen, Zhaoyang Chu, Ranjay Krishna, and Tianyi Zhou. Wait, we don't need to" wait"! removing thinking tokens improves reasoning efficiency. *arXiv preprint arXiv:2506.08343*, 2025. 7

[60] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[61] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 7

[62] Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. Vsp: Assessing the dual challenges of perception and reasoning in spatial planning tasks for vlms. *arXiv preprint arXiv:2407.01863*, 2024. 2

[63] Antonia Wüst, Tim Woydt, Lukas Helff, Inga Ibs, Wolfgang Stammer, Devendra S Dhami, Constantin A Rothkopf, and Kristian Kersting. Bongard in wonderland: Visual puzzles that still make ai go mad? *arXiv preprint arXiv:2410.19546*, 2024. 2

[64] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, 2018. 3

[65] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3, 5, 7

[66] 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. 3

[67] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327, 2019. 3

[68] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023. 3

[69] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *Advances in Neural Information Processing Systems*, 37:51727–51753, 2024. 2, 4

[70] Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, et al. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29580–29590, 2025. 5, 8