CVPR JUNE 3-7, 2026 DENVER COLORADO

Downscaling Intelligence

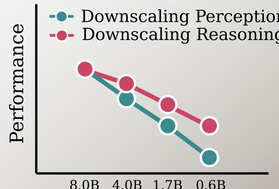
Exploring Perception and Reasoning Bottlenecks in Small Multimodal Models

Mark Endo
Serena Yeung-Levy

Scaling up multimodal models has enabled remarkable advances in visual understanding and reasoning, but practical demands call for smaller, efficient systems. However, the consequences of *downscaling intelligence* remain poorly understood. Namely, when a smaller language model serves as the backbone, which capabilities degrade the most, and *why*?

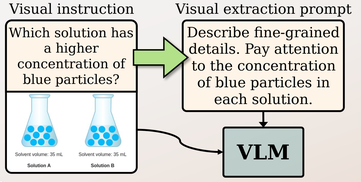
① Examining Effect of LLM Downscaling Across Tasks

② Decoupled Perception +Reasoning Analysis

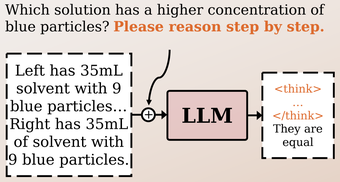


③ Alleviating Perception / Reasoning Bottlenecks

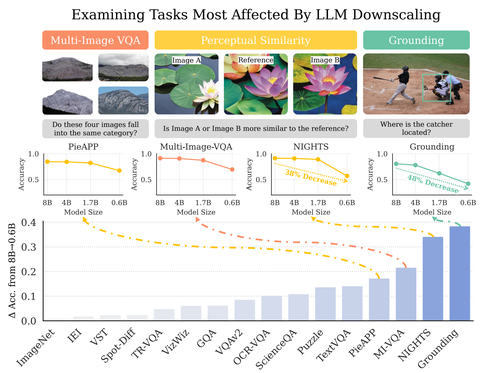
Unifying Perception Skills via *Visual Extraction Tuning*

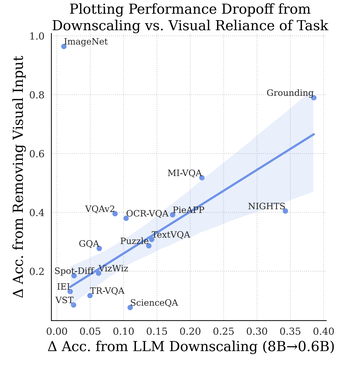


Step-by-Step Thinking About Extracted Visual Details



Examining Effect of LLM Downscaling Across Tasks

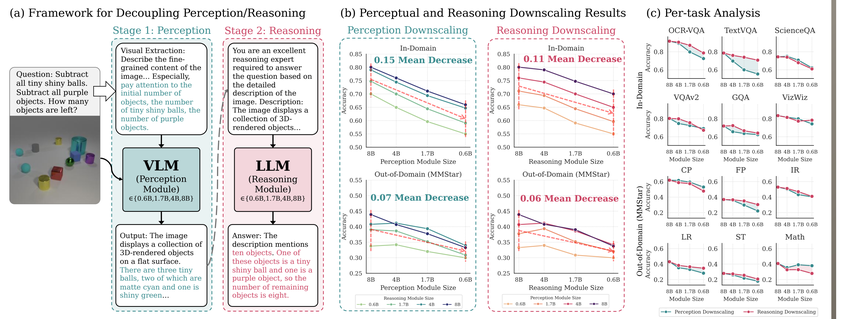




- Tasks with largest performance drops under LLM downscaling rely heavily on visual processing, not the base LLM
- The greater the impact of LLM downscaling, the more the task relies on visual information to be solved

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

Decoupled Perception / Reasoning Analysis



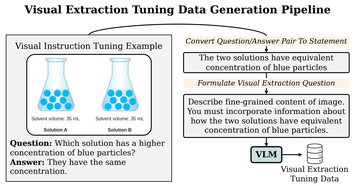
- LLM downscaling expectedly hinders visual reasoning
- LLM downscaling markedly impairs perceptual abilities, affecting a wide spectrum of tasks

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.

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

EXTRACT+THINK

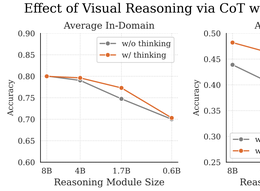
Visual Extraction Tuning: We propose visual extraction tuning as a solution to unify perceptual abilities of the perception module while enabling it to extract question-relevant information and operative effectively across the diverse domains present in visual instruction tuning.



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

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

Step-by-step Visual Reasoning: In the two-stage framework, text bridges perception and reasoning, enabling CoT reasoning over visual details without requiring explicit visual reasoning supervision.



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

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 and a reasoning module enhanced with CoT reasoning.

	LLM Size	#Vis. Data	In-Domain (Multiple-Choice)						Out-of-Domain (MMStar)						
			OCR-VQA	TextVQA	ScienceQA	VQAv2	GQA	VizWiz	Average	CP	FP	IR	LR	ST	Math	Average
End-to-End																
LLaVA-OV	0.5B	8.8M	69.5	77.2	55.7	75.7	73.6	74.7	71.1	63.2	33.1	42.1	35.8	30.0	31.4	39.0
InternVL2.5	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	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	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	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	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	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	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	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	17B	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	17B	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	17B	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**

- EXTRACT+THINK outperforms decoupled baselines (with ~12× smaller perception, ~41× smaller reasoning than PrismCaptioner) and competes with end-to-end models trained at vast scale
- Visual extraction tuning is highly data-efficient, improving over LLaVA-OV-0.5B with 95% fewer visual samples