Visual Compositional Tuning

Xindi Wu[1*] Hee Seung Hwang[1*] Polina Kirichenko[2] Esin Tureci[1] Olga Russakovsky[1]

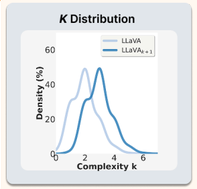
[1]Princeton University [2]Meta AI [*]Equal contribution

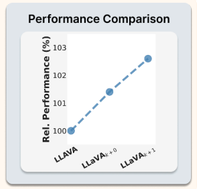
ICLR

Paper Website

Motivation

Visual instruction tuning (VIT) datasets have grown rapidly in scale, yet the **informativeness** of training samples has largely been overlooked.





K Question Samples

- k=1 "**What animal** is in the image?"
- k=2 "**What action** is the **giraffe** performing?"
- k=3 "**How many giraffes** are standing **behind** the **wire fence**?"

Why Complexity?

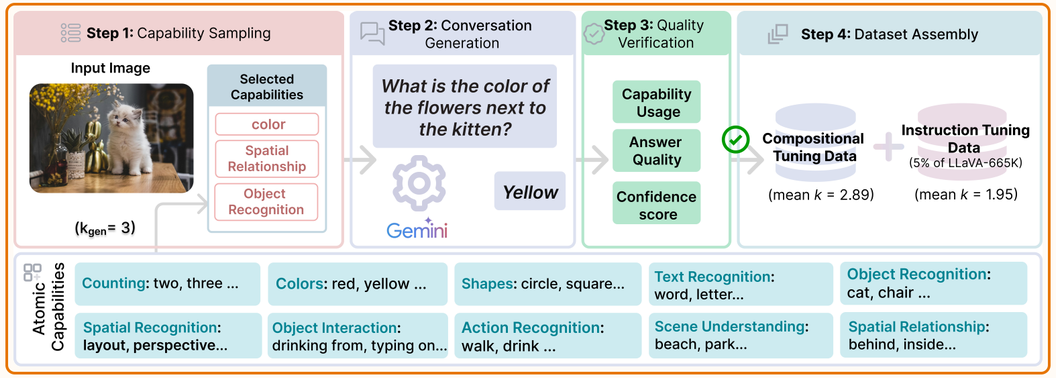
VIT datasets are dominated by simple questions that under-exploit the rich visual content in images. Composing multiple capabilities per sample yields substantially stronger visual reasoning.

10 Atomic Visual Capabilities

Group	Capability
Attribution	Color
	Shape
Recognition	Object Recognition
	Action Recognition
	Text Recognition
	Counting
	Spatial Recognition
Relation	Spatial Relationship
	Object Interaction
	Scene Understanding

Scaling data *complexity* is more effective than scaling data *volume*

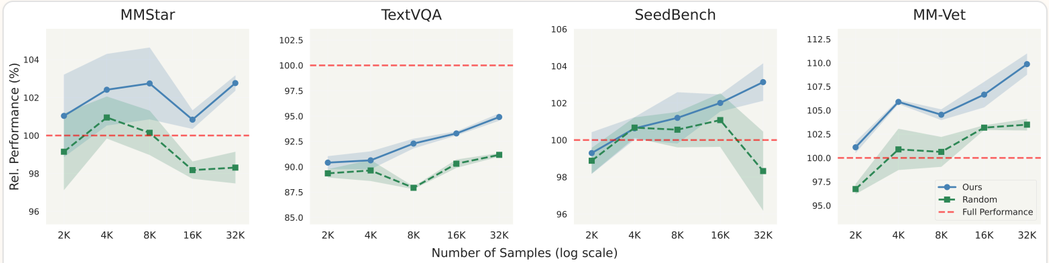
COMPACT: **COMP**ositional **A**tomic-to-complex Visual **C**ompositional **T**uning



8 Multimodal Benchmarks (LLaVA-v1.5-7B-LoRA), using only 10% of the data

Method	#Data	InfoVQA	MM-Vet	MMStar	SeedB2+	CV-Bench	TextVQA	Rel.%
LLaVA-665K (full)	665K	20.80	29.22	35.11	41.72	**60.92**	**46.99**	100.0
Random	65K	20.05	30.46	34.13	41.85	54.71	42.88	95.4
EL2N	65K	20.52	33.53	33.82	42.95	50.92	42.41	97.1
D2-Pruning	65K	20.90	31.61	**36.63**	**43.70**	48.49	41.82	97.1
ICONS	65K	21.00	31.23	35.96	42.03	55.96	43.12	97.5
COMPACT (ours)	**65K**	23.68	31.74	36.13	**43.13**	**55.28**	44.37	100.2

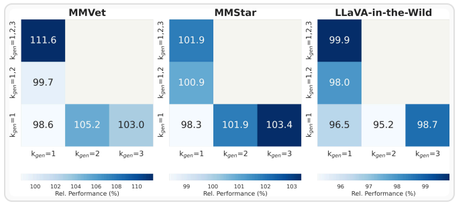
With only 10% of the data, COMPACT matches or exceeds the full 665K baseline across various benchmarks.



Compositional tuning data scales more efficiently than conventional VIT, especially on spatially complex tasks.

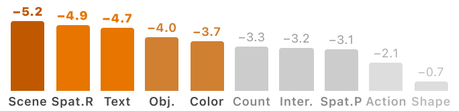
Experiments

Effect of k-value Range



$k_{gen} \in \{1,2,3\}$ outperforms k_{gen}=3 alone; **a range from simple to complex is optimal**.

Which Capabilities Drive Performance? (Leave-One-Out)



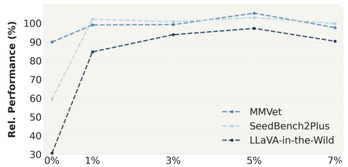
All 10 capabilities contribute non-trivially; scene understanding and spatial relationship drive the largest gains.

Complexity Distribution Shift

Statistic	LLaVA-665K	COMPACT
Mean k-value	1.95	2.89
Mode k-value	2	3
Samples with k ≤ 2	77%	35%
Zero-capability (k=0)	1.1%	0%

COMPACT shifts the distribution toward higher compositional complexity.

Instruction Tuning Ratio



Performance stabilizes at ~5% VIT mix. Instruction-following is **orthogonal** to visual reasoning.

Zero-Capability Samples in LLaVA-665K

~1.1% of LLaVA-665K requires **no visual capabilities**:

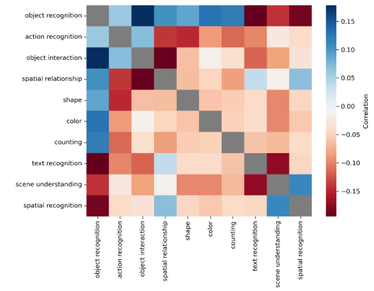
- *"Should I move to London?"*
- *"Can you explain Map Reduce to me?"*
- *"How to do coding"*
- *"I'm looking to create a podcast, can you help me?"*

Experiments

Token-Level Comparison

Metric	COMPACT	LLaVA	Diff.
Input tokens	12.83	16.85	31% shorter
Output tokens	1.70	21.74	92% shorter
Tokens per Q&A turn	14.53	38.59	62% fewer
Total tokens per entry	**104.87**	197.42	47% reduction

Capability Correlation



Spatial capabilities are locally correlated; object recognition co-occurs with most other capabilities.

Limitations

- Data generated from closed-source models (Gemini) may inherit their compositional biases.
- Focused on vision-centric compositionality; we leave exploration on knowledge-intensive tasks for future work.

Takeaways

Visual compositional tuning is a scalable, data-efficient pathway toward multimodal models that can solve **multi-capability tasks**.

Sample complexity, not volume, drives data-efficient VIT. 10% data matches the full 665K baseline.

High-k training *transfers*: k=4 test questions improve by **+33.5%**. Simple and complex samples together are optimal.

Science & Tech **+9.9%**, Instance Reasoning **+8.1%** on MMStar; broad generalization, not benchmark overfitting.