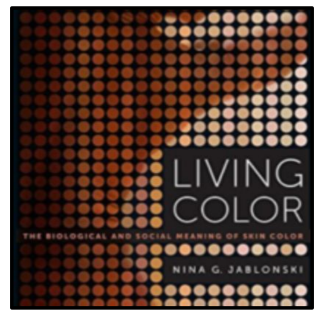
Learning Dynamics of Multitask Training Data in Vision Language Models

Tyler Zhu[1]   Koome Murungi[2]   Polina Kirichenko[3]   Olga Russakovsky[1]

1 2 3

Motivation

VLMs achieve great performance on a wide array of benchmarks. However, the source of this generalization is poorly understood.

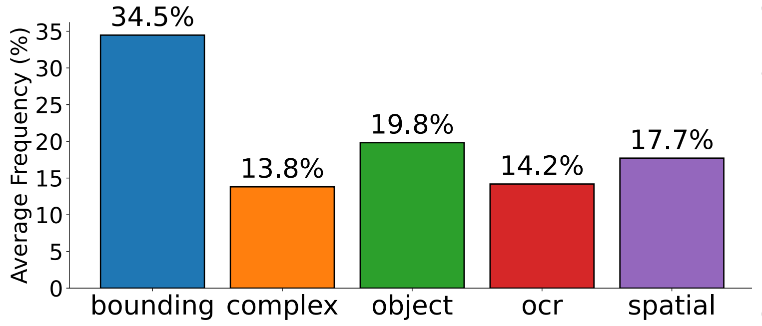
Bounding	Complex	Object	OCR	Spatial
				
Describe the area: [.52, .61, .98, .78]	What are some reasons behind the bird's behavior?	What is the unusual topping on this pizza?	What is the title of this book?	Are there any people to the left of the animals?
A white cow with a black tail.	The bird's behavior ... could indicate ... preening. Preening is the act of grooming oneself ...	Spinach.	Living Color: The Biological and Social Meaning of Skin Color	Yes.

How do we disentangle *memorization* from *generalization*?

1) Evaluate train & val accuracy in 1-epoch setting (*seen* and *unseen*)
2) Break examples into visual task-specific categories for analysis
3) Find inefficiencies (Bounding) and misleading accuracies (OCR)

Data Collection and Evaluation

We used the LLaVA 1.5 dataset w/ 665k multi-turn examples — 3.4M QA pairs [1], and labeled each QA pair into 1 of 5 categories for finer visual knowledge analysis



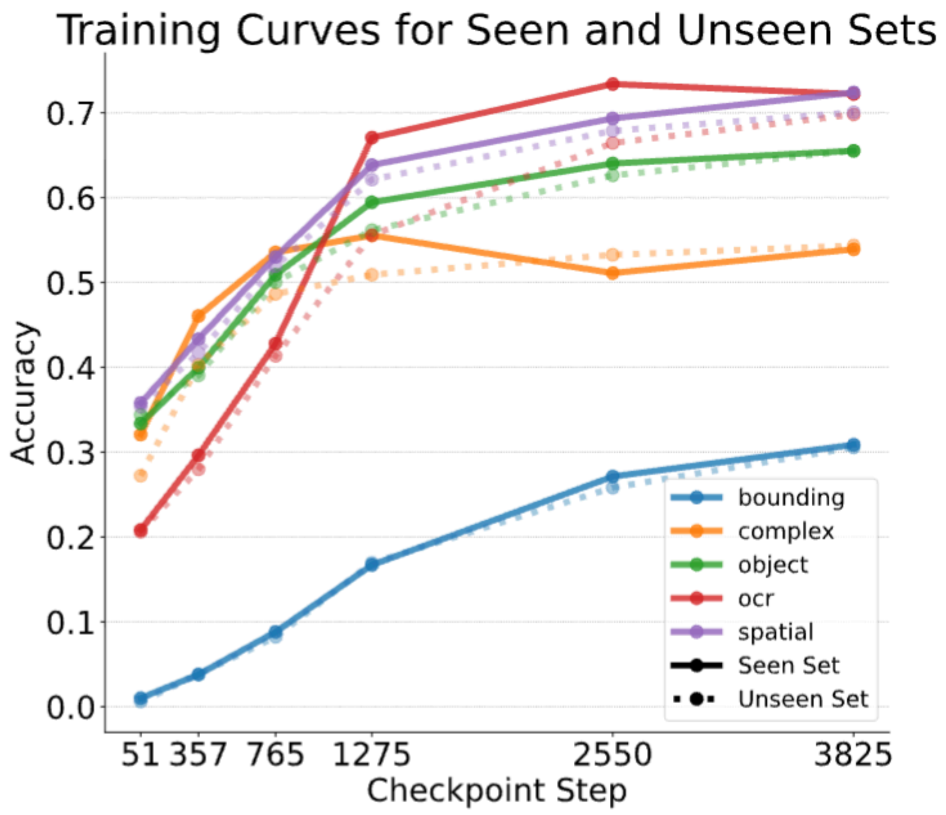
Category	Description
Bounding	Provide a $[x_1, y_1, x_2, y_2]$ bounding box or describe such a region
Complex	Multi-step reasoning and logical deduction, often image-free
Object	Relating to specific object(s), e.g., recognizing, counting
OCR	Reading printed text and semantic queries about them
Spatial	Spatial relationships of object(s) and between them

To measure accuracy during training [2], we use GPT-4o-mini to judge QA + model responses using a human-aligned rubric.

Compared to humans, LLMs graded QA+Rs similarly, esp. w/ rubric!

	Acc	Prec	Recall	F_1 Score
LLM (w/ rubric)	**89.4%**	93.0%	**89.8%**	**0.91**
LLM (w/o rubric)	85.1%	**97.9%**	78.0%	0.87

Results: Dynamics emerge alongside training & validation frontiers

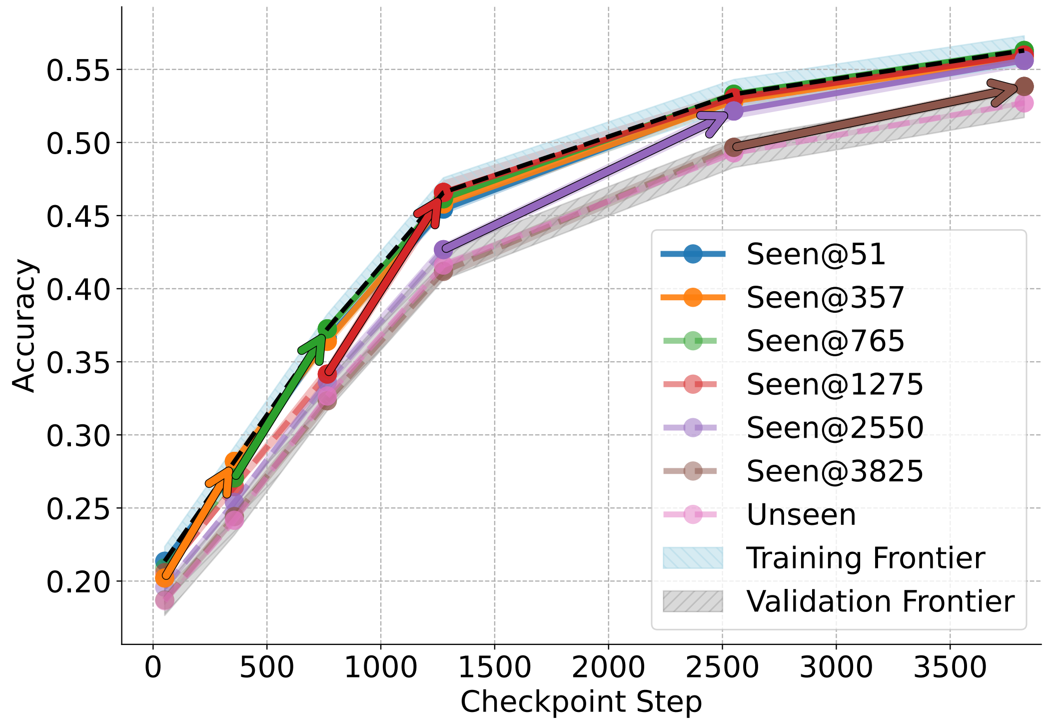


Object and **Spatial** progress normally, **OCR** learns sharply, **Bounding** struggles, and **Complex** has an early plateau.

Additionally, there is a clear *seen* vs. *unseen* performance gap highlighting how much the model *actually* generalizes.

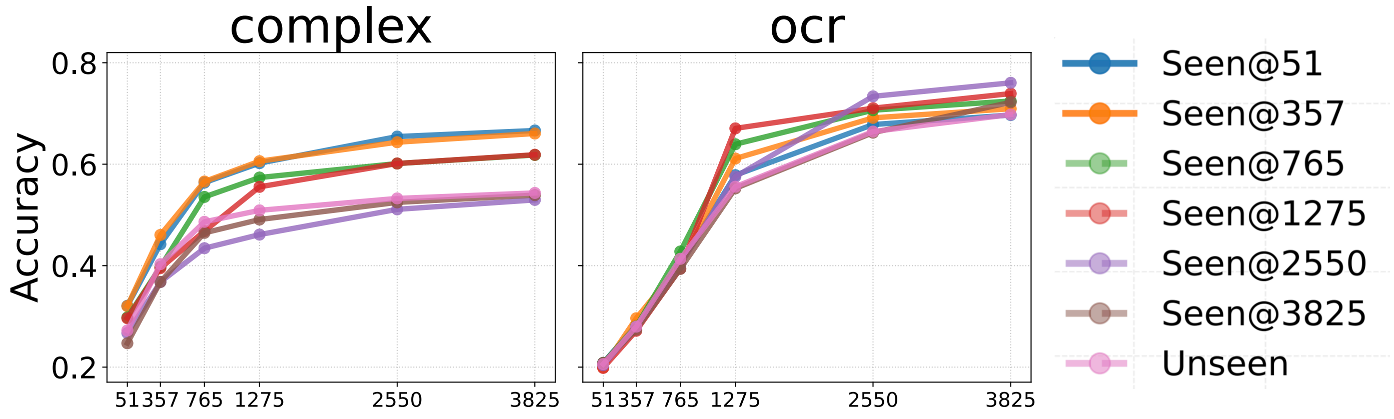
Evaluate at six checkpoint steps for reduced cost.

Divide examples into *seen* vs *unseen* for train & val.

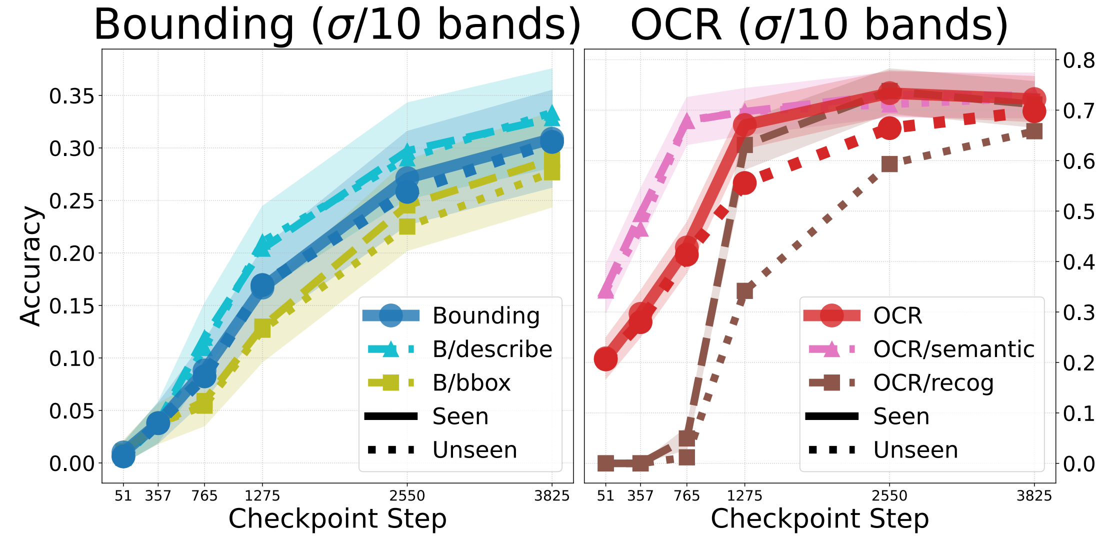


We identify distinct training and validation frontiers with sharp jumps (arrows) when examples become seen. Filtering duplicate images is crucial!

Additional Investigations: Time dynamics, Bounding and OCR subdynamics



- Later **Complex** questions perform worse, likely due to a lack of language SFT data.
- Best performing set for **OCR** is the most recent seen set, pointing to memorization.



VLMs learn semantics before recognition and likely rely on other visual cues to solve semantics rather than reading the text.

[1] Haotian Liu, Chunyuan Li, Yuheng Li, Yong Jae Lee. "Improved Baselines with Visual Instruction Tuning." CVPR 2024.
[2] Siddharth Karamcheti et al.,. "Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models." ICML 2024.