SCRAMBLe : Enhancing Multimodal LLM Compositionality with Synthetic Preference Data

Samarth Mishra*, Kate Saenko, Venkatesh Saligrama

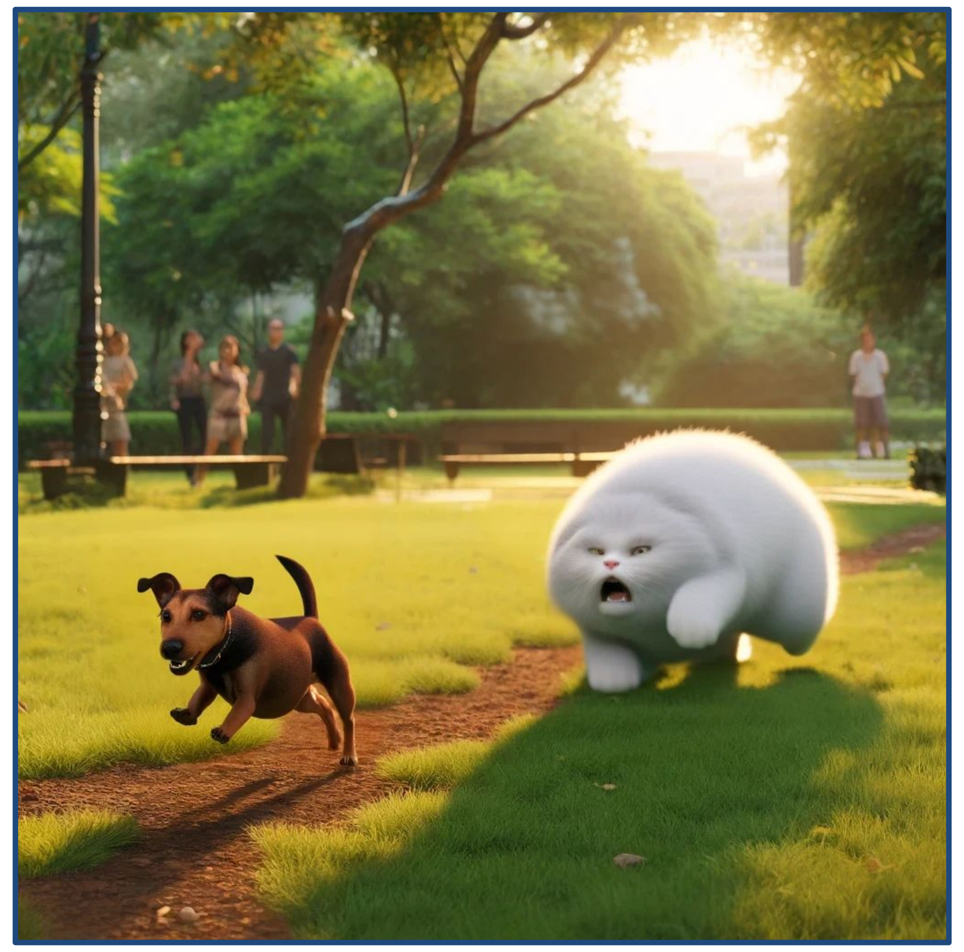
BOSTON UNIVERSITY

ICCV OCT 19-23, 2025 HONOLULU HAWAII

arXiv: 2504.04740

Can Multimodal LLMs distinguish between

a dog chasing a cat



a cat chasing a dog

On Winoground, a benchmark of such questions,
MLLM (GPT-4V) performance = 33%
Human performance = 85%

Compositionality:

Identifying parts/concepts and how they are composed to give rise to a given scene

Atomic Concepts

Apple

Banana

Infinite Compositions - Nothing New

Two apples and a banana

An apple and two bananas

A banana on an apple

An apple on a banana

Improving Compositionality with Preference Tuning

Data Source : Existing Image Caption Dataset (COCO)

Images

Preferred Captions

- A giraffe standing next to a zebra.
- A post with a clock and several birds sitting on it.
- A woman cutting into a cake with a man standing behind her.

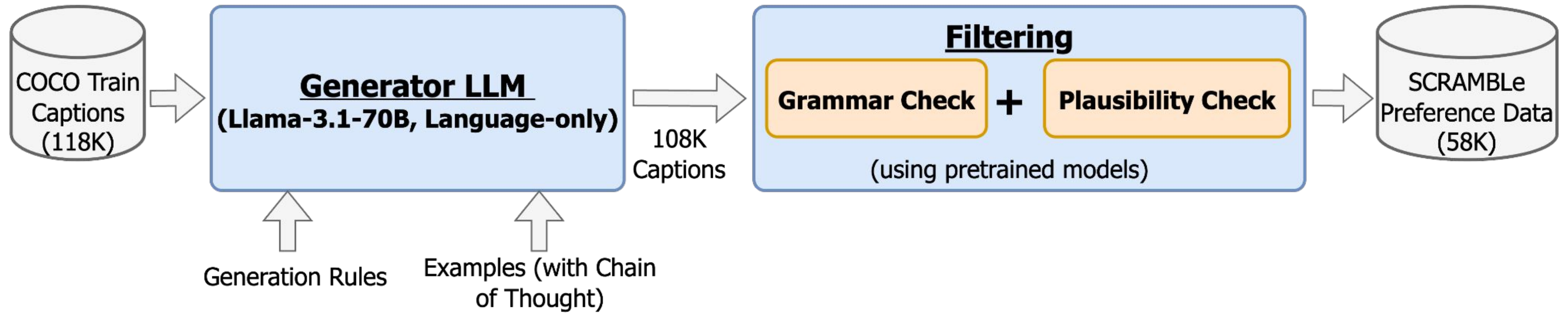
⋮

Dispreferred Captions

?

Solution : Source synthetically with an LLM

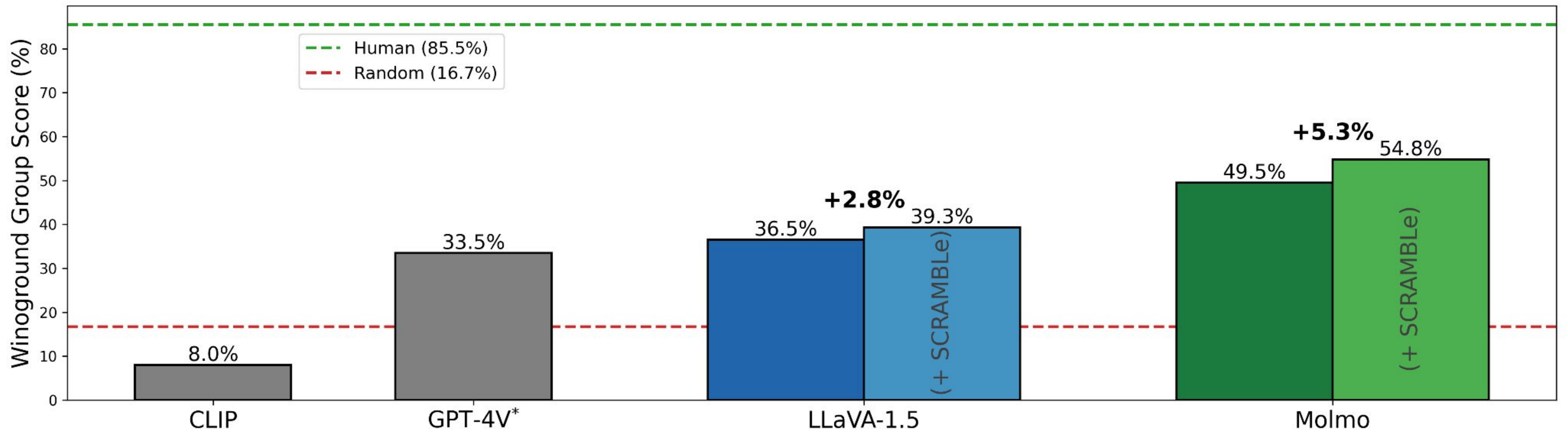
Synthetic Data Generation



References

[1] Mitra, Chancharik, et al. "Compositional chain-of-thought prompting for large multimodal models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[2] Lin, Zhiqiu, et al. "Evaluating text-to-visual generation with image-to-text generation." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

Winoground Accuracy



*GPT-4V eval based on verbose responses and LLM as a judge [1]. In contrast LLaVA and Molmo use VQAScore [2] based on model output logprobs.

Hmm, could this be overfitting?

Seemingly not. Models retain or improve general QA performance (on Control Benchmarks)

	Compositionality Benchmarks			Control Benchmarks	
Model Name	Winoground	COLA	ConMe	SEED-Bench	MM-Vet
LLaVA-1.5-13B	36.5	49.5	62.3	**68.23**	36.2 ± 0.3
+ SCRAMBLe	**39.3**	**55.7**	**64.5**	68.19	$\mathbf{38.6 \pm 0.1}$
MoLMo-7B	49.5	57.1	72.2	74.04	59.3 ± 0.2
+ SCRAMBLe	**54.8**	**60.5**	**74.6**	**74.61**	$\mathbf{60.9 \pm 0.4}$

Chats with SCRAMBLe-Molmo

* Samarth is now affiliated with Waymo. Work done while he was at Boston University