FG-TRACER: Tracing Information Flow in Multimodal Large Language Models in Free-Form Generation

Alessia Saporita[1,2], Vittorio Pipoli[1,3], Federico Bolelli[1], Lorenzo Baraldi[1], Andrea Acquaviva[3], Elisa Ficarra[1]
[1]University of Modena and Reggio Emilia, [2]University of Bologna, [3]University of Pisa
{name.surname@unimore.it}

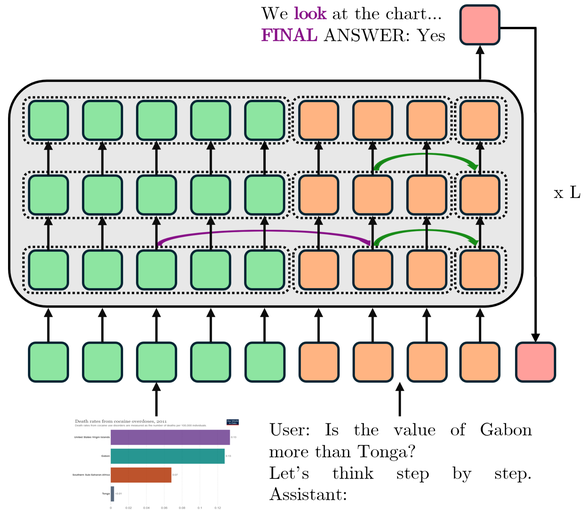
WACV TUCSON, AZ 2026 3/6 - 3/10

How Do MLLMs integrate VL during Free-Form Generation?

- Despite their impressive performance, how MLLMs integrate vision and language remains largely unexplored.
- Previous works investigate cross-modal information flow in constrained settings, such as single-token VQA answers.

In this work, we present the first systematic investigation of the internal mechanisms of MLLMs during free-form generation.

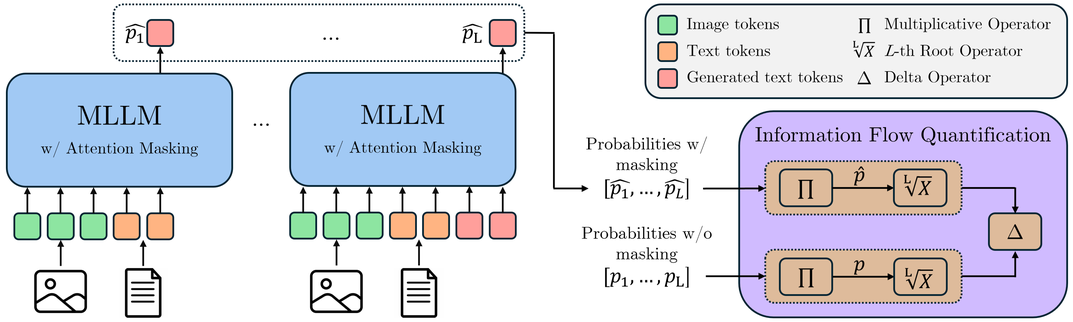
We introduce a novel framework to analyze multimodal information flow in underexplored domains such as image captioning and chain-of-thought reasoning.



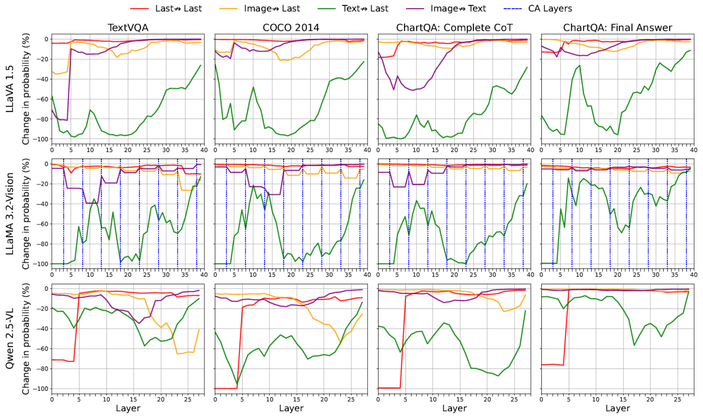
Our approach reveals how multimodal fusion emerges across models, tasks, and generated words.

Proposed Method: FG-TRACER

- We **selectively block communication between token groups** to isolate their individual contributions and measure the information flow as the change in output probability.
- We introduce a **normalization method** based on the answer sequence length **to ensure robust information flow estimation in long, free-form outputs.**



Insights from Information Flow Dynamics



- **Multimodal fusion** mainly occurs in early-to-mid layers, but the fusion of visual and textual information is model- and task-dependent.
- **In OCR-based VQA,** visual information directly influences generation, as fine-grained visual features are required to interpret embedded text.
- **In image captioning,** MLLMs rely on visual input after multimodal fusion to generate accurate descriptions.
- **In CoT reasoning,** visual information contributes throughout the step-by-step reasoning process, while the final answer is derived primarily from the generated linguistic chain.

Word-Level Analysis

- We investigate **which words rely most heavily on visual input** and which can be inferred from linguistic context alone
- The words requiring more visual grounding differ between the MLLMs. Nonetheless, these words exhibit shared characteristics.
- Words conveying spatial, temporal, contextual, and procedural information exhibit significant visual grounding, reflecting their dependence on the visual input.
- Structural and template-based words show limited visual dependence and are predominantly driven by linguistic priors.

	LLaVA 1.5		LLaMA 3.2-Vision		Qwen 2.5-VL	
	Word	Drop	Word	Drop	Word	Drop
ChartQA	first	-66.20	examine	-76.91	examine	-62.22
	corresponding	-62.30	look	-69.93	following	-38.59
	shown	-60.43	arrange	-64.44	find	-37.14
	since	-53.73	subtract	-57.71	section	-34.20
	identify	-50.09	final	-57.04	associated	-34.03
	need	-2.03	need	-2.61	the	-3.13
	a	-2.95	of	-3.15	of	-3.75
	we	-3.85	at	-5.21	to	-5.73
	to	-4.85	we	-6.68	need	-5.80
	of	-5.57	the	-7.12	in	-8.96
COCO 2014	that	-81.77	showcasing	-79.46	while	-65.50
	another	-70.15	setting	-79.21	setting	-64.82
	using	-59.67	features	-65.01	enjoying	-56.08
	near	-55.39	situated	-63.75	another	-54.52
	while	-50.14	featuring	-50.80	next	-53.39
	food	-1.25	image	-0.34	image	-0.60
	of	-2.48	of	-3.53	by	-3.48
	to	-4.84	depicts	-4.65	of	-4.90
	his	-6.42	to	-9.07	are	-7.41
	a	-9.98	a	-9.72	to	-7.53

Ablation: Normalization Factor

- We quantify information flow with and without normalization to validate our formulation.
- Our proposed normalization factor removes length-induced distortions, enabling clear and fair comparisons across tasks with varying output lengths.