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

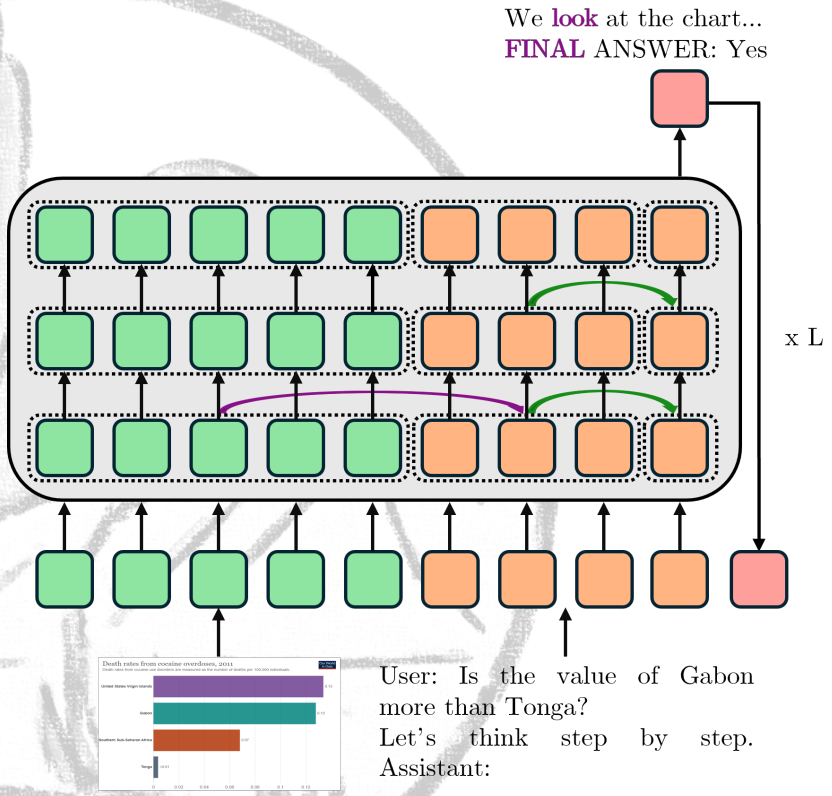
A. Saporita, V. Pipoli, L. Baraldi, E. Ficarra, A. Acquaviva, F. Bolelli

AImage Lab
University of Modena and Reggio Emilia

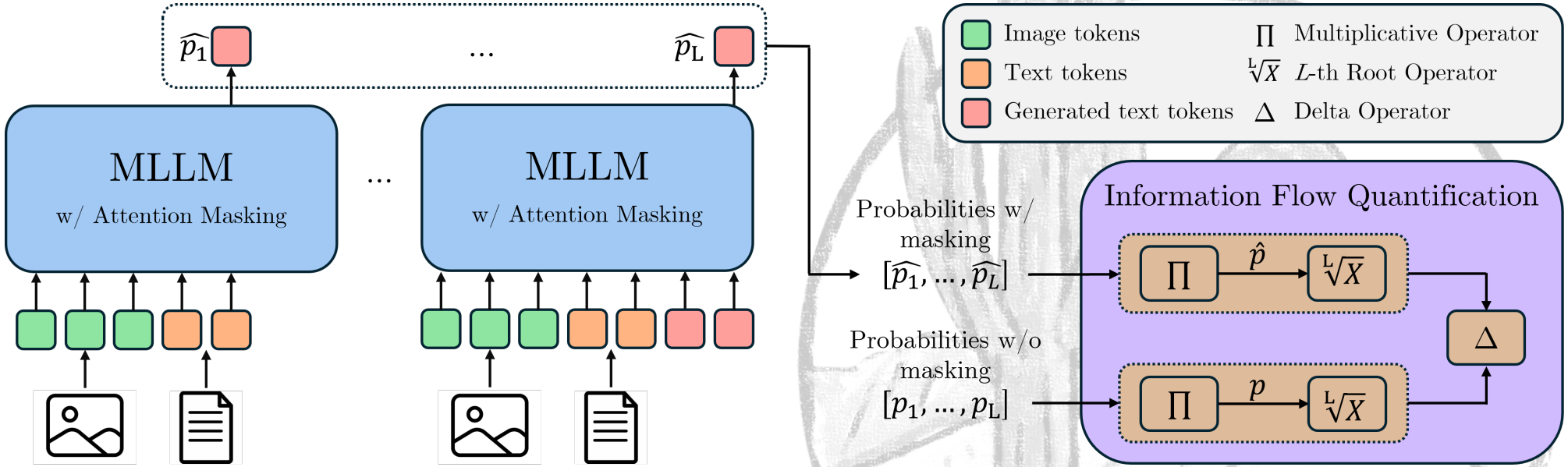
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Motivation

- 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.
- **We present the first systematic investigation of the internal mechanisms of MLLMs during free-form generation** in underexplored domains such as image captioning and chain-of-thought reasoning.



FG-TRACER



We present a novel framework that selectively blocks communication between token groups to isolate their individual contributions. Information flow is defined as the change in output probability, normalized by sequence length to enable robust estimation in long, free-form outputs.

Normalization Factor

The overall sequence probability is computed as:

$$\mathrm{p} = \prod_{i=0}^{L} p_i\,, \quad \hat{p} = \prod_{i=0}^{L} \hat{p}_i$$

where *L* is the number of tokens in the generated response, p and $\hat{p}$ are the output sequence probabilities without and with masking. We then normalize the probabilities by computing the *L*-th root of the probabilities:

$$\Delta = \left(\frac{\hat{p}^{\frac{1}{L}} - p^{\frac{1}{L}}}{p^{\frac{1}{L}}} \right)$$

This normalization is crucial for long-form answers, as otherwise small changes in individual token probabilities would result in exponentially large changes in the overall answer probability when the number of tokens *L* increases.

Normalization Factor: demonstration

Consider a scenario in which the probability of each token undergoes a small relative change δ such that:

$$\hat{p}_i = (1 - \delta)p_i$$

Under this perturbation, the relative change in sequence-level probability is given by:

$$\Delta = \left(\frac{\hat{p}^{\frac{1}{L}} - p^{\frac{1}{L}}}{p^{\frac{1}{L}}} \right) \times 100 = ((1 - \delta)^L - 1) \times 100$$

This relative change decreases exponentially with increasing *L* and asymptotically approaches -100%.

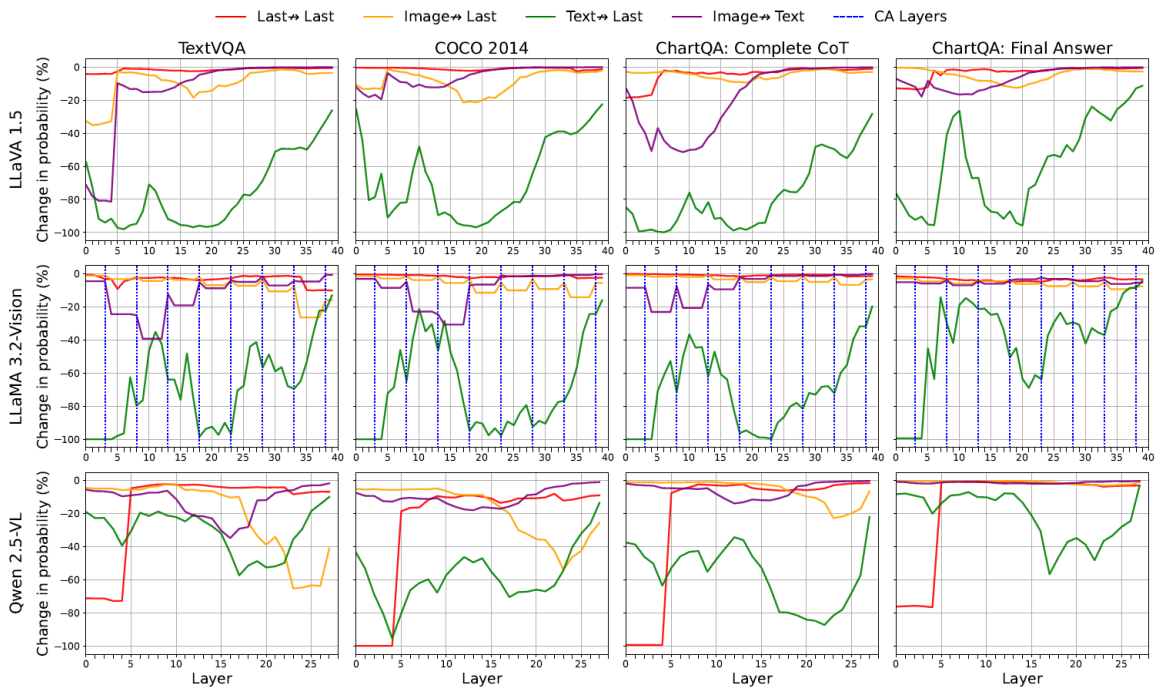
Multimodal Backbones and Datasets

- In our analysis, we study **three state-of-the-art MLLM backbones**: **LLaVA 1.5–13B**, **LLaMA 3.2-11B-Vision**, and **Qwen 2.5-VL-7B** which employ different strategies for multimodal integration. .
- We evaluate our framework on **three distinct vision–language tasks**:
  - **TextVQA** is a benchmark for text-based VQA, which requires Optical Character Recognition capabilities.
  - **COCO 2014** is a widely used dataset for captioning.
  - **ChartQA** is a benchmark of chart understanding, where the questions typically require multi-step, chain-of-thought reasoning to infer information from visual data.

Insights from Information Flow Dynamics

Our analysis yields several key findings:

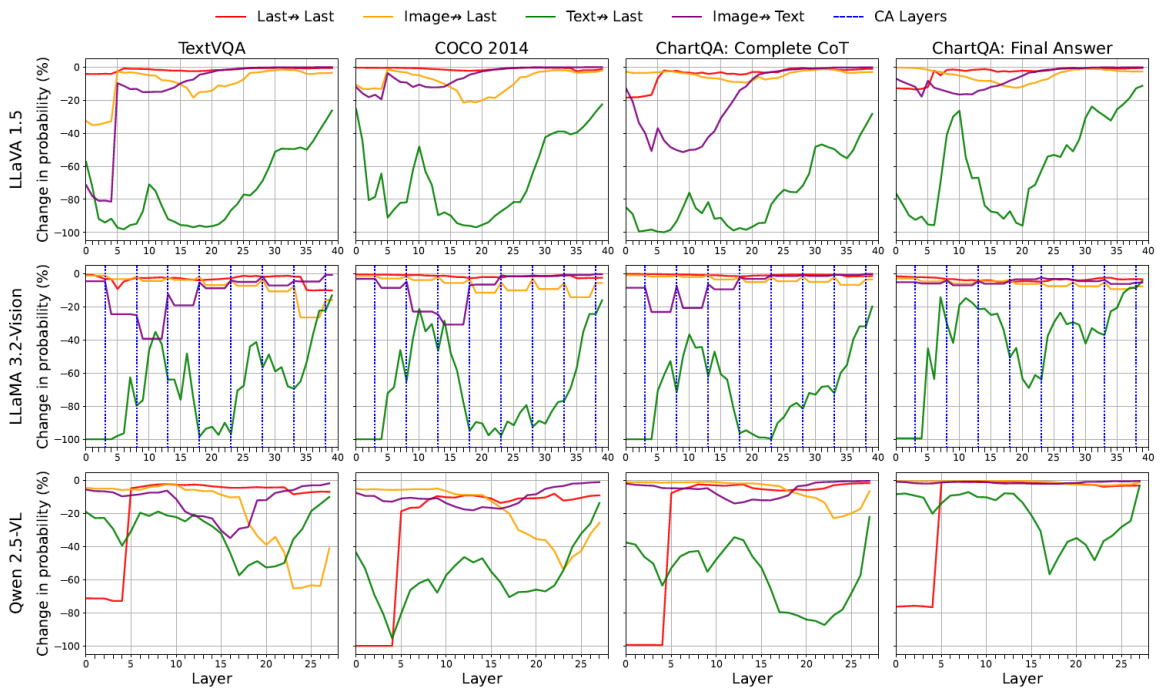
1. **Multimodal fusion** occurs early in the MLLMs. However, distinct patterns of multimodal fusion emerge across models and datasets, indicating that visual–linguistic integration is both task- and model-dependent.



Insights from Information Flow Dynamics

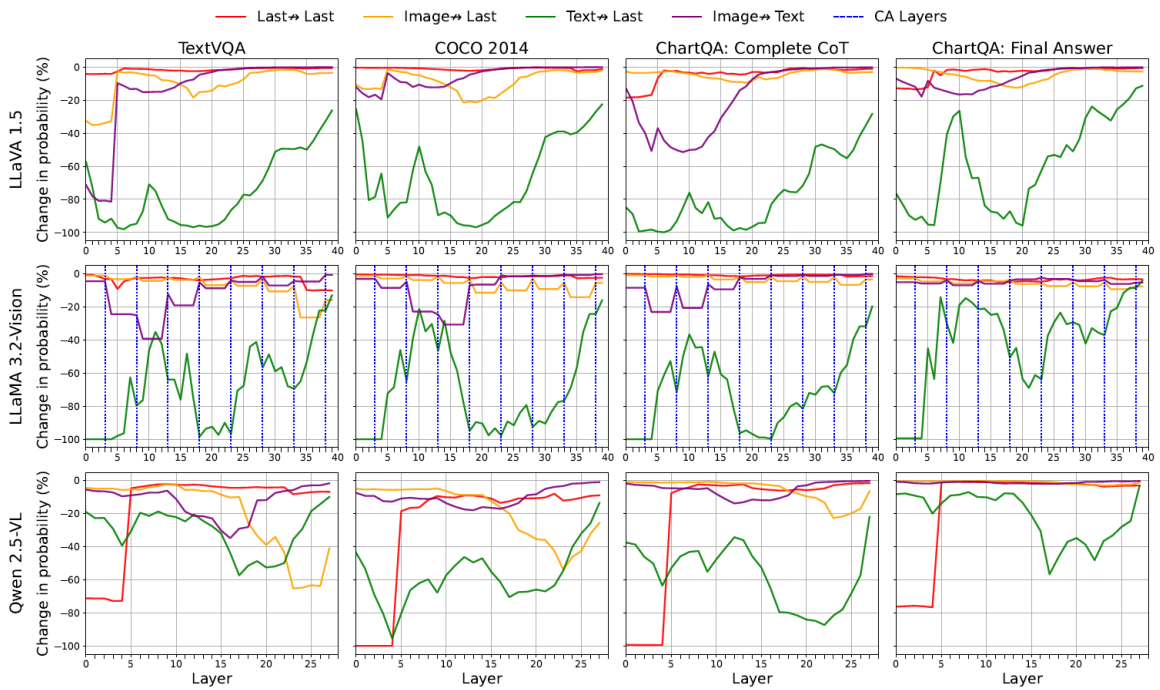
- **In OCR-based VQA tasks**, visual information directly influences generation, as fine-grained visual features are required to interpret embedded text.

- **In image captioning,** MLLMs rely on visual input after multimodal fusion to generate accurate descriptions.



Insights from Information Flow Dynamics

- **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, demonstrating that MLLMs utilize intermediate reasoning to encode the necessary information for answer generation.



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

Word-Level Anlysis

- Within the contexts of image captioning and CoT reasoning, **we investigate which words rely most heavily on visual input** and which can instead be inferred from linguistic context alone, without depending on multimodal integration.
- The table lists at the top the most frequent words with the highest image-to-text drop value and at the bottom the most frequently generated words with the lowest drop value.

	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

Word-Level Anlysis

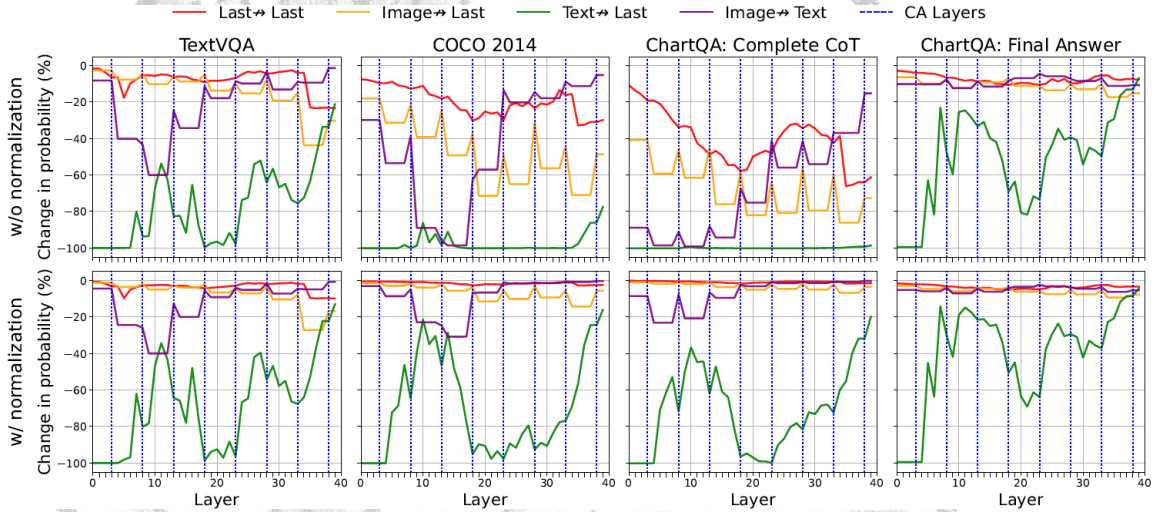
Our findings suggest that:

1. **Content words** that convey spatial, temporal, procedural, or contextual relationships are typically **visually grounded words**.
2. **Structural words** such as pronouns, articles, and prepositions, as well as **terms commonly used in template-based expressions, exhibit minimal visual dependence**, indicating that they are primarily generated from linguistic priors rather than visual content.

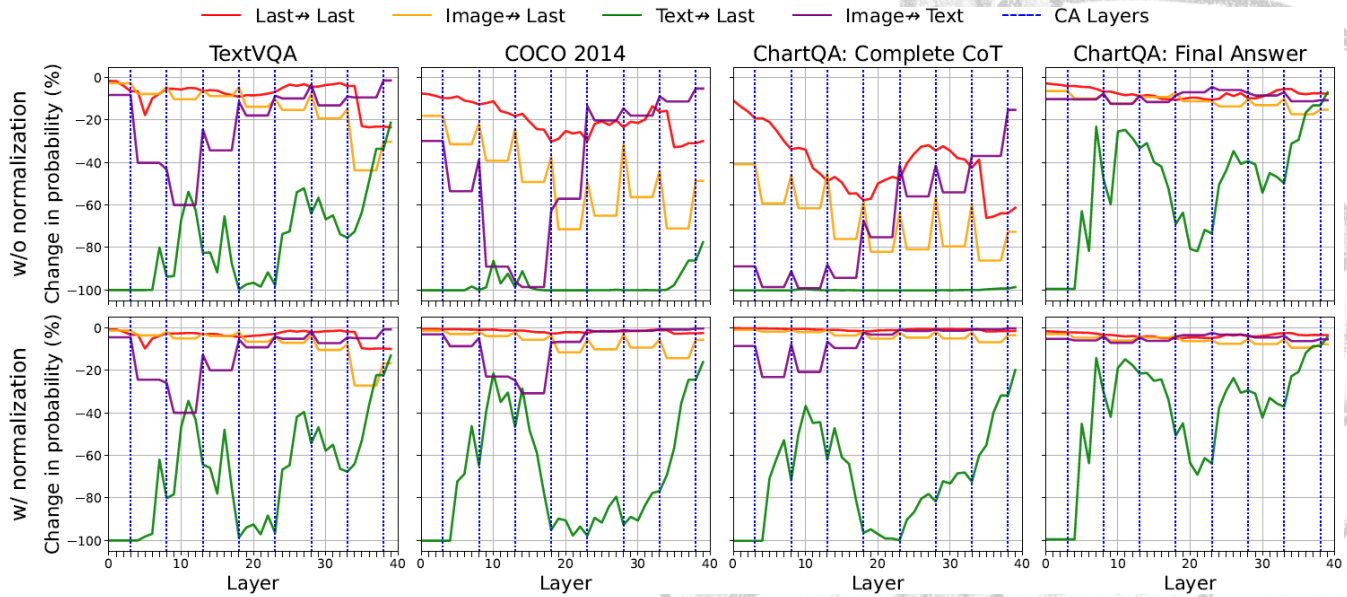
	LLaVA 1.5		LLaMA 3.2-Vision		Qwen 2.5-VL	
	Word	**Drop**	**Word**	**Drop**	**Word**	**Drop**
ChartQA	first	-66.20	examine	-76.91	examine	-62.22
	corresponding	-62.30	look	-69.93	following	-38.59
	shown	-60.43	arrange	-64.44	find	-37.14
	since	-53.73	subtract	-57.71	section	-34.20
	identify	-50.09	final	-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

Insights from Information Flow Dynamics

- **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, demonstrating that MLLMs utilize intermediate reasoning to encode the necessary information for answer generation.



Ablation: Effectiveness of Normalization



The information flow patterns become obscured without normalization as the unnormalized sequence probabilities induce length distortion in the information flows by diminishing exponentially with increasing output length.

Scan Me!

WACV 2026

TUCSON, AZ

3/6 - 3/10

A. Saporita

V. Pipoli

L. Baraldi

E. Ficarra

A. Acquaviva

F. Bolelli

AImage Lab
University of Modena and Reggio Emilia

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA