

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

1. ChartQA: Word Information Flows

We analyze the information flows—last-to-last, image-to-last, text-to-last, and image-to-text—related to the content words (*examine, look, arrange, subtract, final*) and structural words (*need, of, at, we, the*) of ChartQA [1] dataset, using LLaMA 3.2-Vision [4]. The results are presented in Fig. 5. We observe that content words exhibit a strong image-to-text information flow, with multimodal fusion occurring primarily in the early to mid layers. These results demonstrate the strong coupling between these words and visual input. In contrast, for structural words, the image-to-text flow is negligible, while the text-to-last flow remains strong, indicating that their activation is driven more by linguistic priors rather than visual context.

2. ChartQA: Responses with CoT Reasoning

Additionally, we provide four examples of question–answer pairs from the ChartQA dataset, generated using chain-of-thought (CoT) prompting with LLaMA3.2-Vision, to illustrate the structure of the model’s reasoning process and contextualize the content words (highlighted in bold).

Question 1: *What’s the value of the largest yellow bar?*

Response 1: To determine the value of the largest yellow bar, we must **examine** the chart and locate the bar with the highest percentage. The chart displays the share of representatives among each group who have served 1-3 terms in the House (114th Congress).

The yellow bars represent the percentage of representatives in each group.

Key information from the chart: the largest yellow bar corresponds to “Freedom Republicans” with a value of 72.

FINAL ANSWER: 72.

Question 2: *What is the number of deaths per 100,000 in 2002 in Colombia?*

Response 2: To find the number of deaths per 100,000 in 2002 in Colombia, we can **look** at the bar graph provided. The x-axis lists the countries, and the y-axis represents the number of deaths per 100,000 persons.

Looking at the bar for Colombia, we can see that it is the second bar from the left.

The y-axis starts at 0 and increases in increments of 50%,

with labels at 50, 100, 150, 200, and 250.

Given the scale, the bar for Colombia appears to be slightly below the 175 mark. Therefore, based on the bar graph, the number of deaths per 100,000 in 2002 in Colombia is approximately 175.

FINAL ANSWER: 175.

Question 3: *Is the median value of the light green bars greater than 45?*

Response 3: To determine the median value of the light green bars, we need to identify the values represented by these bars in the chart. The light green bars correspond to the percentage of people who trust information from the Chinese government regarding the coronavirus outbreak.

To find the median, we **arrange** these values in ascending order: 13%, 40%, 51%. Since there are an odd number of values (three), the middle value is the median. In this case, the median value is 40%.

Comparing this median value to 45%, we can see that the median value of the light green bars (40%) is less than 45%.

FINAL ANSWER: No.

Question 4: *What is the difference in the value of High blood sugar and High blood pressure?*

Response 4: To determine the difference between High blood sugar and High Blood pressure, we **subtract** the value of High Blood pressure from the value of High blood sugar. Values:

High blood sugar: 879

High Blood pressure: 676

Calculation: Difference = High blood sugar - High Blood pressure = 879 - 676 = 203

FINAL ANSWER: 203.

3. COCO 2014: Word Information Flows

We also analyze the information flows—last-to-last, image-to-last, text-to-last, and image-to-text—associated with the content words (*showcasing, setting, features, situated, featuring*) and the structural words (*image, of, depicts, to, a*) of COCO 2014 [6] dataset (Tab. 1 of the main paper), using LLaMA 3.2-Vision. The results are presented in Fig. 6. For the content words, a consistent image-to-text information flow is observed across all terms, suggesting a close association between these words and visual input. Conversely,

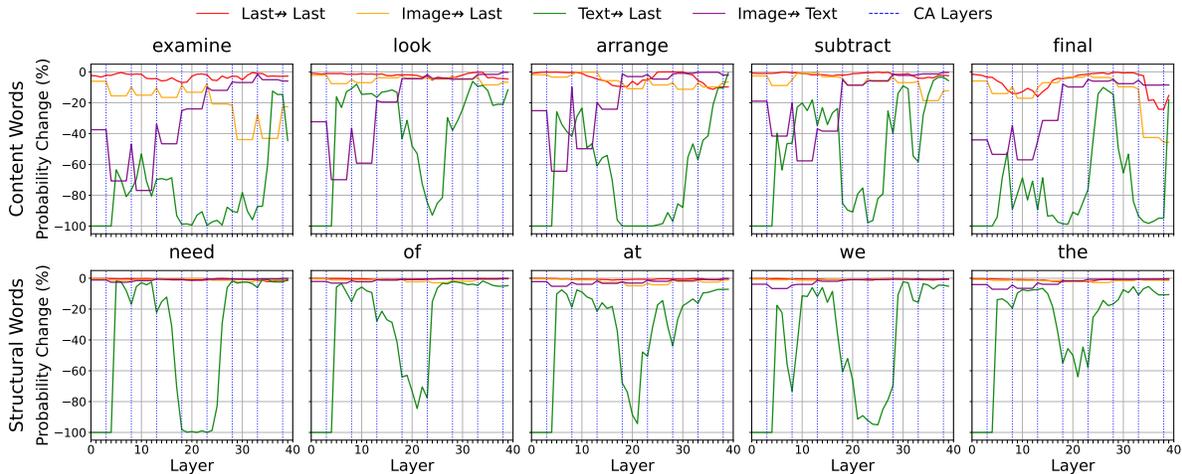


Figure 5. Information flow patterns—last-to-last, image-to-last, text-to-last, and image-to-text—for LLaMA 3.2-Vision on the ChartQA dataset, analyzed for content and structural words. Dotted blue lines indicate cross-attention layers.

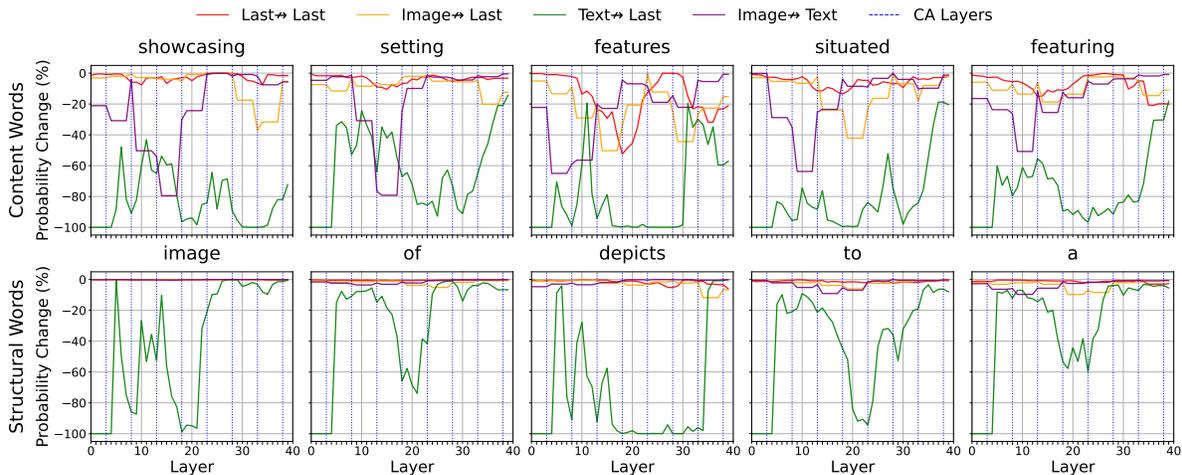


Figure 6. Information flow patterns—last-to-last, image-to-last, text-to-last, and image-to-text—for LLaMA 3.2-Vision on the COCO 2014 dataset, analyzed for content and structural words. Dotted blue lines indicate cross-attention layers.

structural words exhibit negligible image-to-text information flow, but show a strong text-to-last information flow, indicating that their activation is largely influenced by linguistic structure rather than visual cues.

4. COCO 2014: Image Captions

Furthermore, we present four representative COCO captions processed by LLaMA 3.2-Vision to illustrate the contextual usage of the content words (highlighted in bold) reported in Tab. 1 of the main paper.

Caption 1. This image is a close-up photograph of a toucan in a cage, **showcasing** its vibrant colors and unique beak.

Caption 2. The image **features** a plate of food, a bottle of wine, and a plant on a table in an outdoor **setting**.

Caption 3. The image depicts a traffic light with a green smiley face on it, **situated** in a suburban area.

Caption 4. The image depicts a fire hydrant spraying water into the air on a city street, with a brick building in the background **featuring** graffiti and a sign indicating a parking restriction.

Caption 5. The image depicts a Starbucks coffee cup, a muffin, and a banana on a table, showcasing a casual breakfast or snack **setting**.

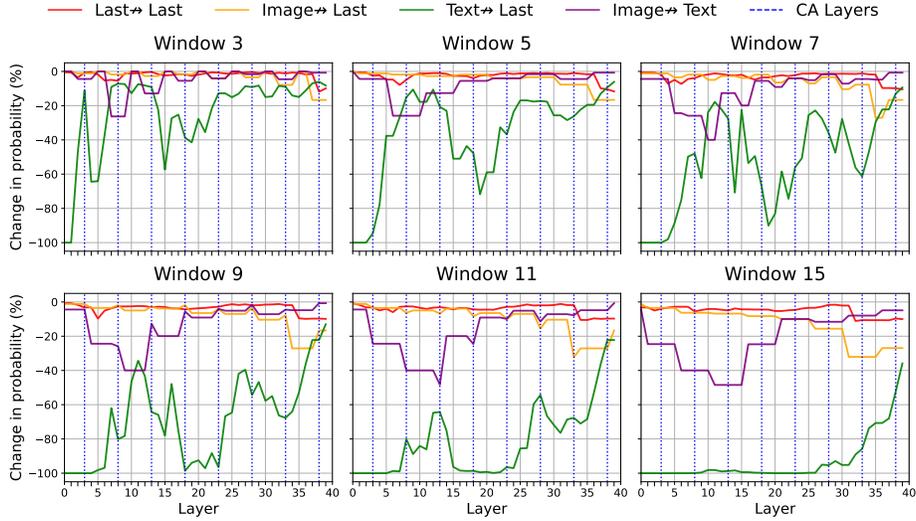


Figure 7. Relative changes in prediction probabilities for LLaMA 3.2-Vision on the TextVQA dataset across different window sizes.

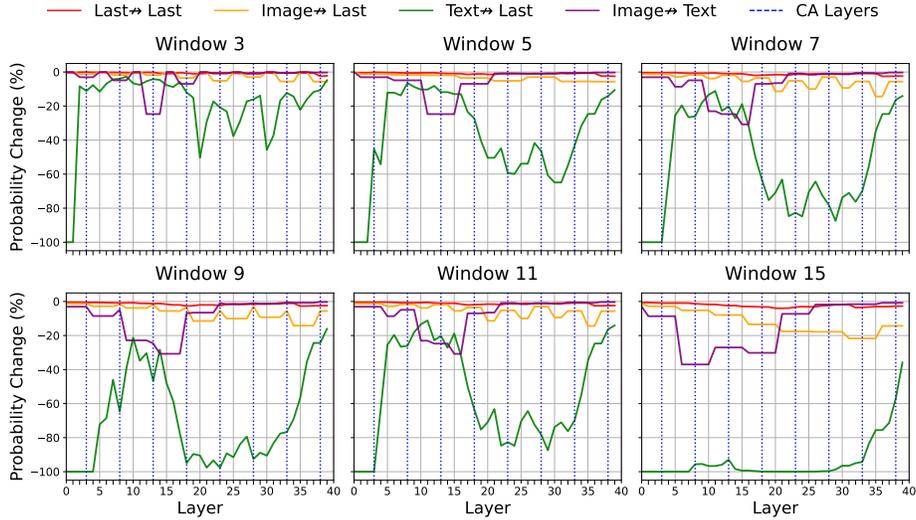


Figure 8. Relative changes in prediction probabilities for LLaMA 3.2-Vision on the COCO 2014 dataset across different window sizes.

5. Information Flows for Different Window Sizes

In the experiments reported in the main paper, we employed a window size of $k = 9$. To evaluate the effect of window size on model behavior, we analyze the relative change in answer probabilities in LLaMA 3.2-Vision on the TextVQA [2] and COCO 2014 [6] datasets across a range of window sizes: $k = 3, 5, 7, 9, 11, 15$. The results are presented in Fig. 7 and Fig. 8. Overall, the patterns of the information flows—last-to-last, image-to-last, text-to-last, and image-to-text—remain consistent across different values of k in both datasets, with multimodal integration occurring primarily in the early to mid layers. However, the mag-

nitude of change in output probability increases with larger window sizes. This trend is expected, as more restrictive attention windows reduce the number of accessible attention edges, thereby limiting the model’s capacity to contextualize the input effectively. Notably, with small k (e.g., 3 or 5), masking applies only to a few layers, and we mainly observe local information flow patterns, which limits our ability to study broader multimodal dynamics. Instead, with large k (e.g., 11, 15), masking spans many layers, overly suppressing cross-modal interactions. The choice of using $k=9$ provides a balanced setting: it is large enough to reveal meaningful effects beyond local patterns, yet not so large that it disrupts information flow excessively.

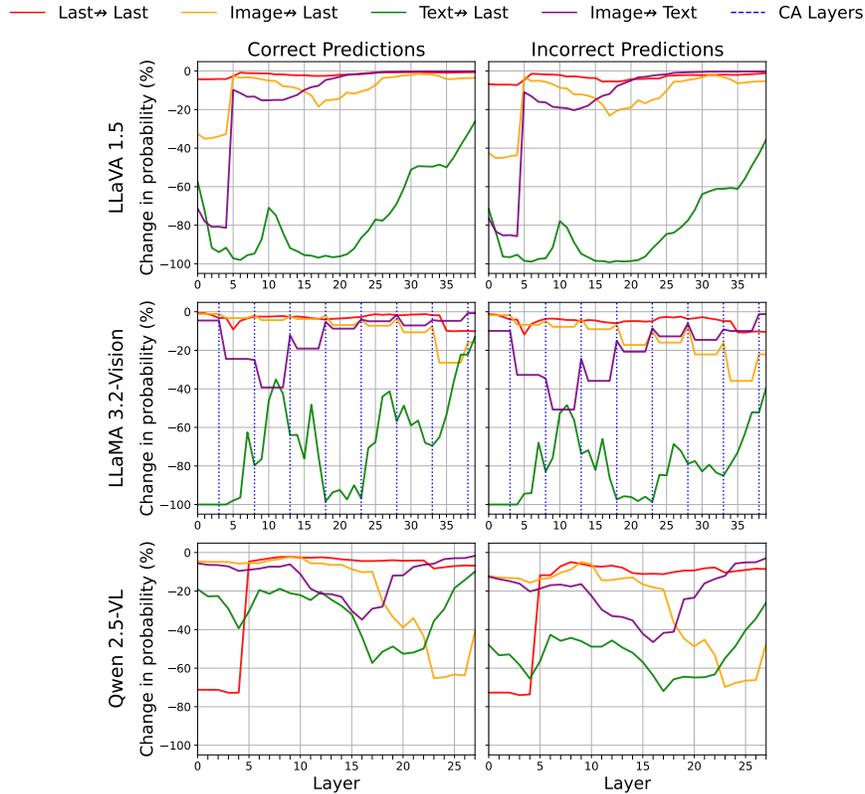


Figure 9. Relative changes in prediction probabilities for LLaVA 1.5, LLaMA 3.2-Vision, and Qwen 2.5-VL on the TextVQA dataset using only correct or wrong samples.

6. Information Flows for Wrong Samples

In the main paper, we reported results based exclusively on correctly answered samples for each model and dataset, in order to minimize potential noise introduced by incorrect predictions. To examine whether errors might reveal distinct patterns of multimodal fusion, we conducted an additional analysis on the subset of incorrect samples from the TextVQA dataset—*i.e.*, cases where the models produced wrong predictions—on LLaVA 1.5 [5], LLaMA 3.2-Vision, and Qwen 2.5-VL [3]. The experimental methodology was identical to that employed in the main study on correct samples. Specifically, we computed the same information flows—last-to-last, image-to-last, text-to-last, and image-to-text—thereby enabling a direct comparison between correct and incorrect subsets. The corresponding results are shown in Fig. 9. Notably, across all three MLLMs, the results obtained for incorrect predictions remain consistent with those observed for the correct subset: the same patterns of information flow emerge. These findings indicate that the underlying reasoning dynamics are a general property of model behavior, and that failures are not attributable to disruptions or breakdowns in multimodal fusion. These results highlight the robustness of information flow patterns

across both successful and unsuccessful predictions.

References

- [1] Masry Ahmed et al. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning, 2022. 1
- [2] Singh Amanpreet et al. Towards VQA Models That Can Read, 2019. 3
- [3] Shuai Bai et al. Qwen2.5-vl technical report, 2025. 4
- [4] Aaron Grattafiori et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. 1
- [5] Liu Haotian et al. Improved Baselines with Visual Instruction Tuning, 2024. 4
- [6] Lin Tsung-Yi et al. Microsoft COCO: Common Objects in Context, 2015. 1, 3