# Cross-modal Information Flow in Multimodal Large Language Models

## Supplementary Material

| | Object | Attribute | Category | Relation | Global |
|---|---|---|---|---|---|
| **Verify** | 86.21% | 83.00% | – | 87.82% | 95.56% |
| **Query** | – | 71.20% | 62.88% | 52.84% | 55.74% |
| **Choose** | – | 90.17% | 92.03% | 87.19% | 96.76% |
| **Logical** | 88.92% | 76.17% | – | – | – |
| **Compare** | – | 71.23% | – | – | – |

Table 2. The accuracy of the validation set of GQA dataset[22] on *LLaVA-1.5-13b* [28]. and represent binary (*yes/no*) and open question respectively. represents that this category contains both binary and open questions.

## A. Dataset collection

We collect our data from the validation set of the *GQA* dataset [22], which is designed for visual reasoning and compositional question-answering. Derived from the Visual Genome dataset [23], *GQA* provides real-world images enriched with detailed scene graphs. Questions in *GQA* dataset are categorized along two dimensions: structure (5 classes, defining question formats) and semantics (5 classes, specifying the main subject's semantic focus). Structural classes include: (1) *verify* (*yes/no* questions), (2) *query* (open questions), (3) *choose* (questions with two alternatives), (4) *logical* (logical inference), and (5) *compare* (object comparisons). Semantic classes are: (1) *object* (existence questions), (2) *attribute* (object properties or positions), (3) *category* (object identification within a class), (4) *relation* (questions about relational subjects/objects), and (5) *global* (overall scene properties like weather or location). Based on the combination of these two dimensions, the questions in *GQA* are categorized into 15 groups, as shown in Table 2.

We select 6 out of 15 groups according to the following steps. First, we exclude the *verify* type, as it is quite simple and involves only straightforward binary questions (*e.g.,* "Is the apple red?"). Then we focus on types with an average accuracy above 80% on *LLaVA-1.5-13b* model [28], retaining *ChooseAttr*, *ChooseCat*, *ChooseRel*, and *LogicalObj*. *ChooseGlo* is excluded due to its limited sample size in the validation set of *GQA* (only 556 instances). After that, to enhance question-type diversity, we select high-performing subtypes (accuracy>80%) in *CompareAttr* and *QueryAttr* from the GQA dataset. Specifically, we use the *positionQuery* subtype for spatial-relation questions in *QueryAttr* and the *twoCommon* subtype for comparing common attributes between two objects in *CompareAttr*. Finally, for each type of the six selected types, we sample at most 1000 data that are predicted correctly on model *LLaVA-1.5-13b* from the validation set of *GQA* resulting in our final data in this paper, as shown in Table 1.

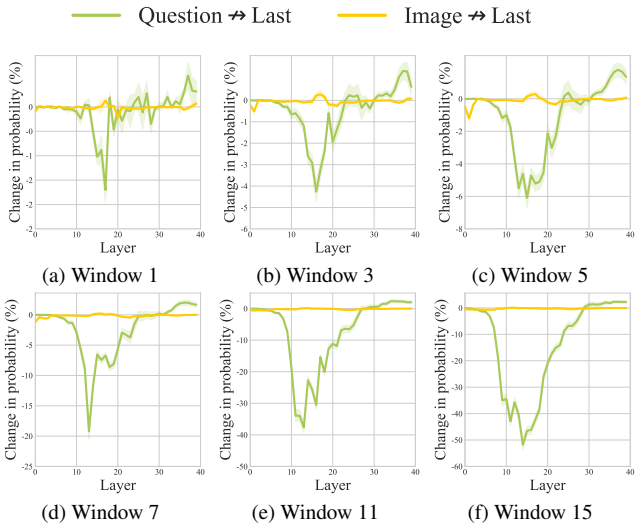



Figure 7. The relative changes in prediction probability on *LLaVA-1.5-13b* with the tasks of *ChooseAttr* for different window size. The *Question↛ Last* and *Image↛ Last* represent preventing *last position* from attending to *Question* and *Image* respectively.

## B. Informaion flow for different window size k

In the main body of the paper, we use a window size $k = 9$ for an easier analysis of the internal working mechanism of the multimodal large language models when performing multimodal tasks. We present the relative change in probability on *LLaVA-1.5-13b* and the task of *ChooseAttr* with different window sizes of $k = 1, 3, 5, 7, 9, 11, 15$. The resulting information flow between different parts of the input sequence (*image* and *question*) and *last position*, and between *image* and *question* are shown in Figure 7 and Figure 8, respectively. Overall, the observations on the information flow are consistent across different window sizes $k$. Specifically, the critical information flow from *question* to *last position* occurs in the middle layers while the critical information flow from *image* to *last position* is not observed across different window sizes, as shown in Figure 7. For critical information from *image* to *question*, the two different critical information flows are observed across different window sizes where both occur in lower layers and sequentially follow each other, as illustrated in Figure 8. In addition, we observe that as the window size increases, the two information flows gradually merge into one, which is

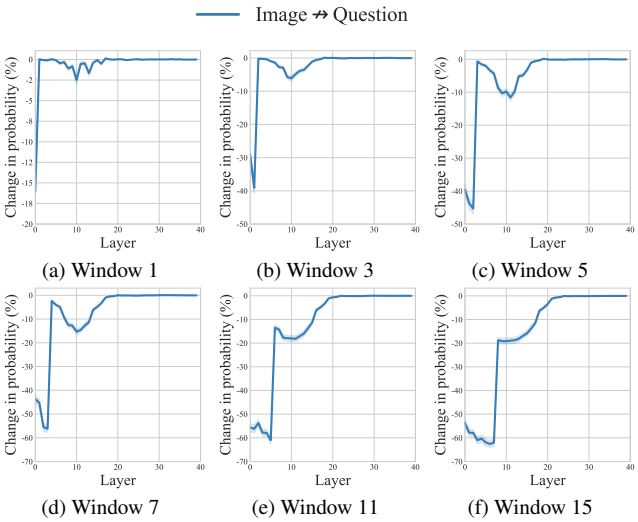



Figure 8. The relative changes in prediction probability when blocking attention edges from the *question* positions to the *image* positions on *LLaVA-1.5-13b* with the tasks of *ChooseAttr* for different window sizes.

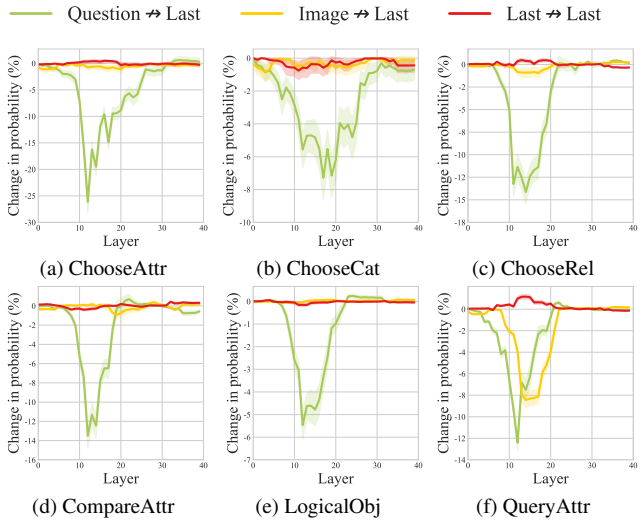



Figure 9. The relative changes in prediction probability for the final generated sub-word of the *answer* on *LLaVA-1.5-13b* with six *VQA* tasks. The *Question↛ Last*, *Image↛ Last* and *Last↛ Last* represent preventing *last position* from attending to *Question*, *Image* and itself respectively.

because the larger window encompasses layers involved in both information flows. Moreover, the decrease in the prediction probability becomes more pronounced with the increase of the value $k$. This is expected, as blocking more attention edges in the computation hinders the model's ability to properly contextualize the input.

## C. Changes in probability of the last sub-word generation

In this paper, the *answer* in our used dataset normally contains one word or one phrase, which might result in several sub-word tokens. In the main body of the paper, we present the relative change in probability of the first generated sub-word while the final generated sub-words also yield similar results. Specifically. we conduct the same experiments as in the main body of the paper: six tasks (*ChooseAttr*, *ChooseCat*, *ChooseRel*, *LogicalObj*, *QueryAttr* and *CompareAttr*) on *LLaVA-1.5-13b* model. Instead of calculating the relative change in probability for the first generated sub-word token, we calculate that for the final generated sub-word token of the correct answer word. As shown in Figure 9, Figure 10 and Figure 11, the information flow from different parts of the input sequence (*image* and *question*) to *last position*, from *image* to *question* and from different image patches (*related image patches* and *other image patches*) to *question* are consistent with the observations in Figure 3, Figure 4 and Figure 5 in the main body of the paper.

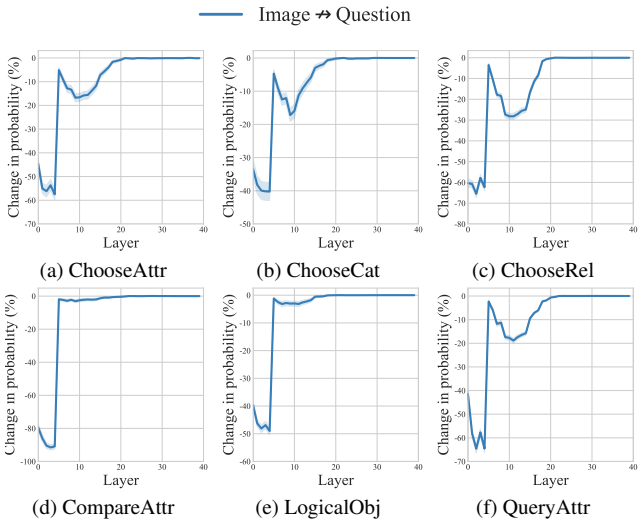



Figure 10. The relative changes in prediction probability for the final generated sub-word of the *answer* when blocking attention edges from the *question* positions to the *image* positions on *LLaVA-1.5-13b* with six *VQA* tasks.

## D. Constructing multimodal semantic representations

We have investigated how multimodal information is integrated through the MHAT module in Section 6. We now take a closer look at how the multimodal semantic representation is constructed.

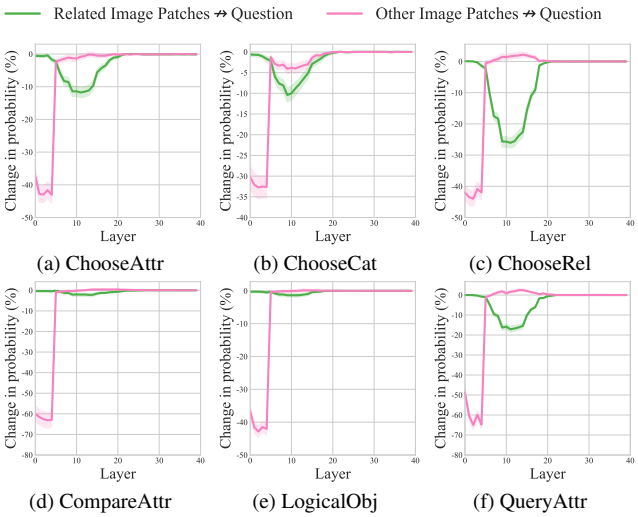



Figure 11. The relative changes in prediction probability for the final generated sub-word of the *answer* on *LLaVA-1.5-13b* with six *VQA* tasks. *Related Image Patches↛question* and *Other Image Patches↛question* represent blocking the position of *question* from attending to that of different image patches, region of interest and remainder, respectively.

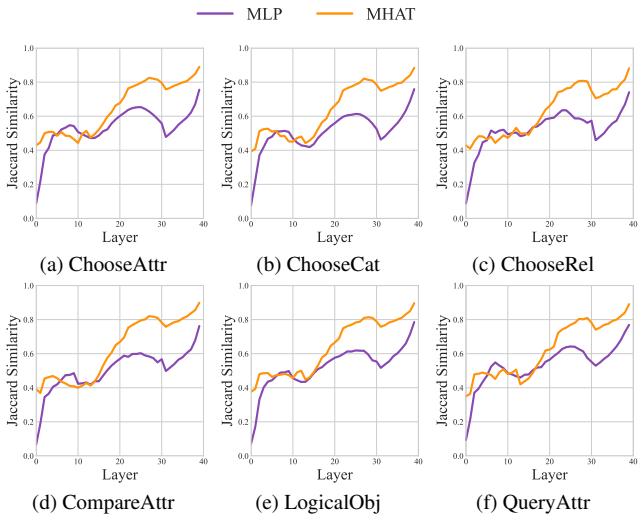



Figure 12. The Jaccard similarity between the predicted words of the original model and those of the intervened model, with the MLP and MHAT modules removed individually (*LLaVA-1.5-13b*).

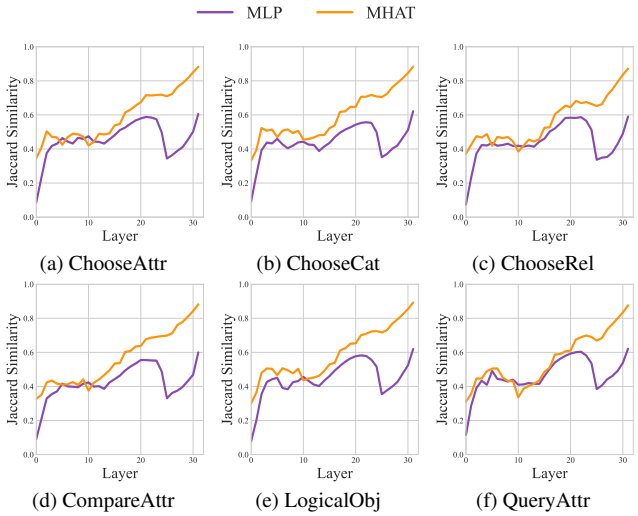



Figure 13. The Jaccard similarity between the predicted words of the original model and those of the intervened model, with the MLP and MHAT modules removed individually (*LLaVA-1.5-7b*).

**Experiment** To identify which module in the transformer contributes to the formulation of multimodal semantic information within hidden representations, we employ a *module knockout* approach to evaluate the significance of individual transformer modules. As shown in Equation (1), the hidden representation at layer $\ell$ is computed by adding $\boldsymbol{a}_i^\ell$ and $\boldsymbol{f}_i^\ell$ to $\boldsymbol{h}_i^{\ell-1}$, where $\boldsymbol{a}_i^\ell$ and $\boldsymbol{f}_i^\ell$ are derived from the MHAT (Equation (2)) and MLP (Equation (5)) modules, respectively. This allows us to determine which module contributes to constructing semantic information by selectively zeroing out the outputs of MHAT or MLP—two additive modules in the transformer layer. Specifically, for each layer $\ell$, we intervene by setting $\boldsymbol{a}_i^{\ell'}$ or $\boldsymbol{f}_i^{\ell'}$ $(i \in \mathbb{Q})$ to zero across 9 consecutive layers $\{\ell'\}_{\ell'=\ell}^{\min\{\ell+8,L\}}$. We then measure the importance of constructing multimodal semantic information by observing the semantic change of the hidden representation corresponding to *question* position $\mathbb{Q}$ at the final layer $L$. Our focus on layer $L$ is inspired by Geva et al. [20], who highlight that semantic information peaks in the final layer. We follow Wang et al. [45], who evaluate the semantic content of a hidden representation using top-k words from this representation. We estimate semantic content using the top-10 words predicted from each hidden representation in $\mathbb{Q}$, derived from Equation (6), where $\boldsymbol{h}_N^L$ is replaced by $\boldsymbol{h}_i^L$ $(i \in \mathbb{Q})$. We then quantify the change in semantic content of hidden representation resulting from our interventions using Jaccard Similarity:

$$J(\mathbb{W}_o, \mathbb{W}_i) = \frac{|\mathbb{W}_o \cap \mathbb{W}_i|}{|\mathbb{W}_o \cup \mathbb{W}_i|} \tag{9}$$

where $\mathbb{W}_o$ and $\mathbb{W}_i$ denote the sets of $10 \cdot |\mathbb{Q}|$ predicted words from the original and intervened models, respectively.

**Observation: The MLP module plays a greater role in constructing semantic representations compared to the**

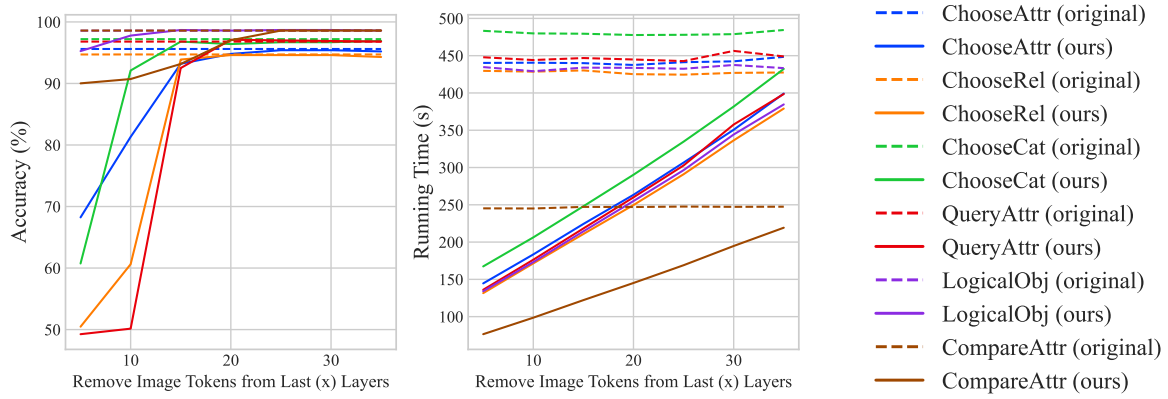



Figure 14. Accuracy and inference time of the original *LLaVA-v1.6-13B* and a variant removing image tokens in last certain layers ($X$) across six VQA datasets.

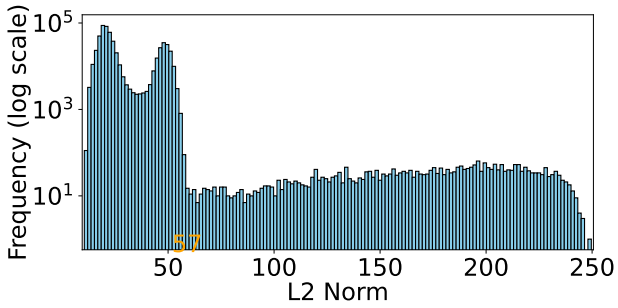



Figure 15. Distribution of image patch norms

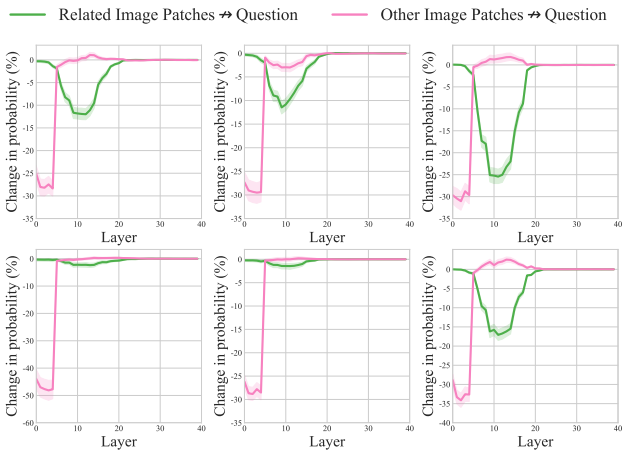



Figure 16. The relative changes in prediction probability for the first generated sub-word of the *answer* on *LLaVA-1.5-13b* with six *VQA* tasks. *Related Image Patches↛question* and *Other Image Patches↛question* represent blocking the position of *question* from attending to that of different image patches (excluding those with high norms), region of interest and remainder, respectively.

**MHAT Module** As shown in Figure 12 for model *LLaVA-1.5-13b*, removing the MLP module severely impacts semantic representation, reducing average Jaccard Similarity across six tasks by ∼90% when MLP is removed in the first layer and ∼25% in the last layer. In contrast, removing the MHAT module has a smaller effect, with reductions of ∼60% and ∼10% at the first and last layers, respectively. This highlights the MLP module's important role in generating multimodal semantic representations. These results align with the findings from [10, 19, 32], who demonstrate that factual information is primarily stored in the MLP module, emphasizing its contribution to enriching semantic information. This is also observed in the model *LLaVA-1.5-7b*, as shown in Figure 13.

## E. Attention sink in image encoder

The work in [13] shows high-norm image patch tokens hold global rather than local information, we analyze the distribution of token norms over our six datasets and identify a threshold of 57, as shown in Fig. 16. Excluding patches exceeding 57 (3.3 out of 576 patches per image on average), we split the remaining patches into two groups: *Related Image Patched* and *Other Image Patches* and replicate the experiments from our main body of paper (Section 6). As shown in Fig. 16, the results across six tasks remain consistent with our original findings in the main body of the paper.

## F. Potential application — model efficiency

Our findings in the main body of the paper indicate image information primarily propagates to other tokens in the lower layers, suggesting a potential strategy for improving the efficiency of MLLMs. Specifically, during inference, we can remove image tokens in the higher layers to reduce computational costs without significantly compromis-

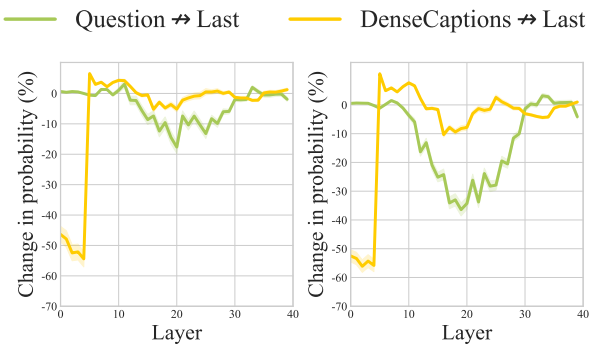



Figure 17. Information flow from dense captions or questions to last position in LLM (*vicuna1.5-13b*), the initial LLM in *LLaVA1.5-13b* (MLLM) on two tasks.

ing performance (average #image tokens 2,021.6 v.s #question tokens 11.3 in *LLaVA-v1.6-13B* over our dataset). To quantitatively analyze this, we remove image tokens in the last 5, 10, 15, 20, 25, 30, 35 layers respectively and test the performance and inference time. As shown in Fig. 14, removing image tokens from final 20 layers keeps accuracy intact while reducing inference time by ∼40%, demonstrating a promising approach for optimizing MLLMs efficiency.

## G. Difference with unimodal LLM

To compare the information flow in LLMs and MLLMs, we replace image with dense captions (obtained from VG dataset [23]) and analyze the resulting information flow. As shown in Fig. 17, we observe that information flow from the question to last position primarily occurs in middle layers which is consistent with MLLMs. However, in LLMs, information flow from dense captions to the last position in lower layers differs from MLLMs where almost no information flow is observed from the image. This suggests that captions and questions follow distinct processing stages, aligning with findings from works in [20], which identify separate stages of information flow in attribute extraction tasks within LLMs.

## H. Experiment on other factual tasks

Since the main focus of our paper is the interaction between modalities, we evaluate tasks where cross-modal alignment plays a crucial role, while excluding those requiring external knowledge for reasoning to minimize confounding factors and ensure a more precise analysis. Nevertheless, to further validate our findings, we extend our experiments to two additional factual multimodal tasks, OKVQA[31] and AOKVQA[39] involving more complex, fact-based reasoning requiring external commonsense and world knowledge. Figs. 18 and 19 shows the results align with our findings,

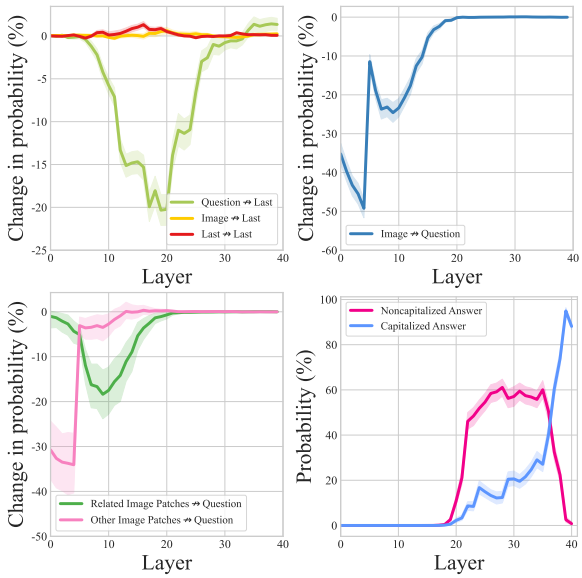



Figure 18. The information flow and probability of answer word tracking on AOKVQA dataset.

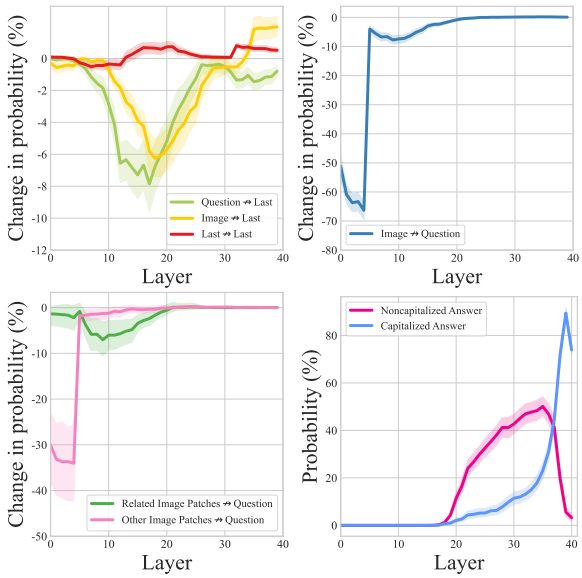



Figure 19. The information flow and probability of answer word tracking on OKVQA dataset.

i.e. from lower to higher layers: a two-stage multimodal integration, information flow from question to last positions and semantic generation and syntactic refinement.

## I. Experiments on other models

We conduct the same experiments (six *VQA* task types) as in the main body of the paper with the other three models. Six *VQA* task types include (*ChooseAttr*, *ChooseCat*, *ChooseRel*, *LogicalObj*, *QueryAttr* and *CompareAttr*).

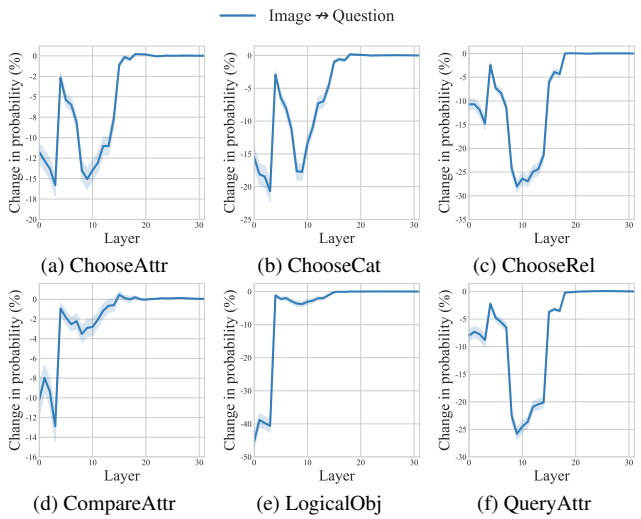



Figure 21. The relative changes in prediction probability when blocking attention edges from the *question* positions to the *image* positions on *LLaVA-1.5-7b* with six *VQA* tasks.

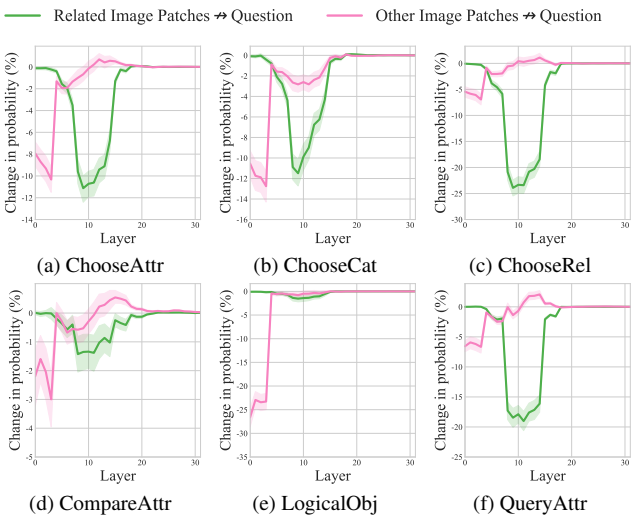



Figure 22. The relative changes in prediction probability on *LLaVA-1.5-7b* with six *VQA* tasks. *Related Image Patches↛question* and *Other Image Patches↛question* represent blocking the position of *question* from attending to that of different image patches, region of interest and remainder, respectively.

The other three models include *LLaVA-1.5-7b*, *LLaVA-v1.6-Vicuna-7b* and *Llama3-LLaVA-NEXT-8b*.


Figure 20. The relative changes in prediction probability on *LLaVA-1.5-7b* with six *VQA* tasks. The *Question↛ Last*, *Image↛ Last* and *Last↛ Last* represent preventing *last position* from attending to *Question*, *Image* and itself respectively.

## I.1. *LLaVA-1.5-7b*

*LLaVA-1.5-7b* is a small version of *LLaVA-1.5-13b* presented in the main body of the paper. It contains 32 transformer blocks (layers) instead of 40 layers in *LLaVA-1.5-13b*. The information flow from different parts of the input sequence (*image* and *question*) to *last position*, from *image* to *question* and from different image patches (*related image patches* and *other image patches*) to *question*, as shown in Figure 20, Figure 21 and Figure 22 respectively, are consistent with the observations for the *LLaVA-1.5-13b* model, as shown in Figure 3, Figure 4 and Figure 5 respectively, in the main body of the paper. Specifically, the model first propagates critical information twice from the *image* positions to the *question* positions in the lower-to-middle layers of the MLLM. For the twice multimodal information integration, the first one focuses on producing the generative representations over the whole image while the second one tends to construct question-related representation. Subsequently, in the middle layers, the critical multimodal information flows from the *question* positions to the *last position* for the final prediction. The difference between the two models is the magnitude of reduction in the probability when blocking the attention edge between *image* and *question*. In model *LLaVA-1.5-7b*, the first drop is rather smaller than that in model *LLaVA-1.5-13b*. However, this does not conflict with our conclusion that the information flows from *image* to *question* twice and one after the other in the main body of the paper. Moreover, the probability change of the answer word across all layers as shown in Figure 23 is also consistent with the result in Figure 6 in the main body of the paper. Specifically, the model first generates the answer semantically in the middle layers and then refines the syntactic correctness of the answer in the higher layers.

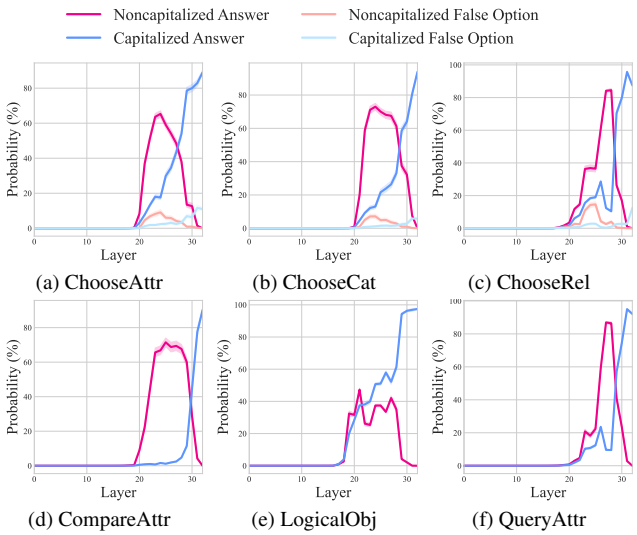



Figure 23. The probability of the answer word at the *last position* across all layers in LLaVA-1.5-7b with six VQA tasks. *Capitalized Answer* and *Noncapitalized Answer* represent the answer word with or without the uppercase of the initial letter, respectively. As the tasks of *ChooseAttr*, *ChooseCat* and *ChooseRel* contain *false option*, we also provide the probability of it.

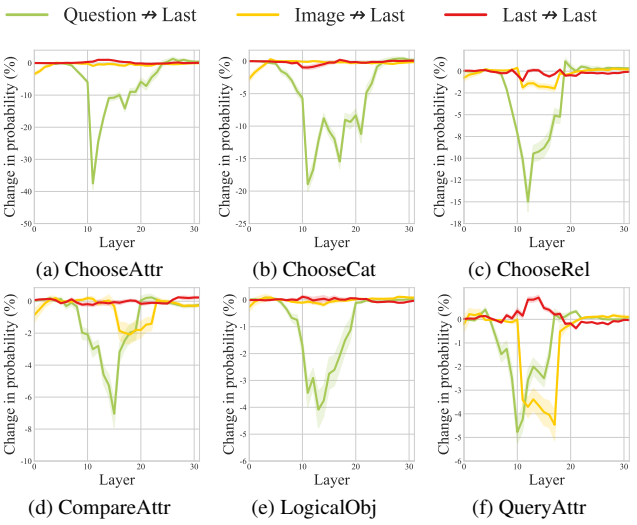



Figure 24. The relative changes in prediction probability on *LLaVA-v1.6-Vicuna-7b* with six *VQA* tasks. The *Question↛ Last*, *Image↛ Last* and *Last↛ Last* represent preventing *last position* from attending to *Question*, *Image* and itself respectively.

## I.2. *LLaVA-v1.6-Vicuna-7b*

*LLaVA-v1.6-Vicuna-7b* has the similar architecture with *LLaVA-1.5-13b* in the main body of the paper. The difference between them includes the layer number and the the way processing image patch features. The *LLaVA-v1.6-Vicuna-7b* has 32 layers versus 40 layers in *LLaVA-1.5-13b*. *LLaVA-1.5-13b* directly feeds the original fixed-length image patch features from the image encoder into the LLM as input tokens. In contrast, *LLaVA-v1.6-Vicuna-7b* employs a dynamic high-resolution technique, which dynamically adjusts image resolution, resulting in variable-length image patch features with higher resolution. Specifically, the higher resolution is implemented by splitting the image into grids and encoding them independently.

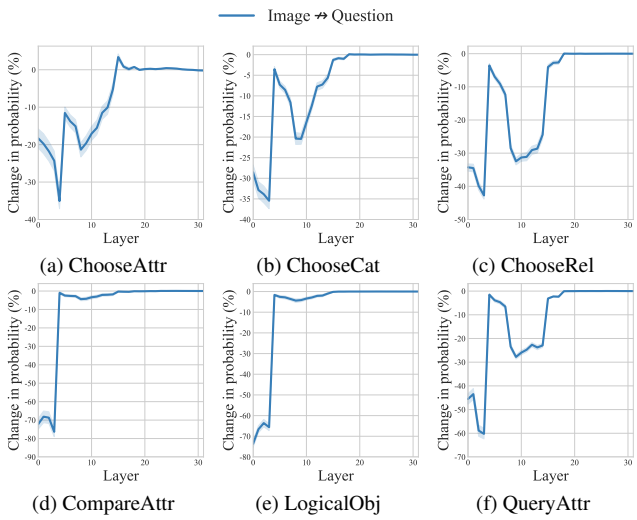



Figure 25. The relative changes in prediction probability when blocking attention edges from the *question* positions to the *image* positions on *LLaVA-v1.6-Vicuna-7b* with six *VQA* tasks.

The information flow from different parts of the input sequence (*image* and *question*) to *last position*, from *image* to *question* and from different image patches (*related image patches* and *other image patches*) to *question*, as shown in Figure 24, Figure 25 and Figure 26 respectively, are consistent with the observations for the *LLaVA-1.5-13b* model, as shown in Figure 3, Figure 4 and Figure 5 respectively, in the main body of the paper. Specifically, the model first propagates critical information twice from the *image* positions to the *question* positions in the lower-to-middle layers of the MLLM. For the dual-stage multimodal information integration, the first stage emphasizes generating holistic representations of the entire image, while the second stage focuses on constructing representations that are specifically aligned with the given question. Subsequently, in the middle layers, the critical multimodal information flows from the *question* positions to the *last position* for the final prediction. Moreover, the probability change of the answer word across all layers as shown in Figure 27 is also consistent with the result in Figure 6 in the main body of the paper. Specifically, the model first generates the answer semantically in the middle layers and then refines the syntactic correctness

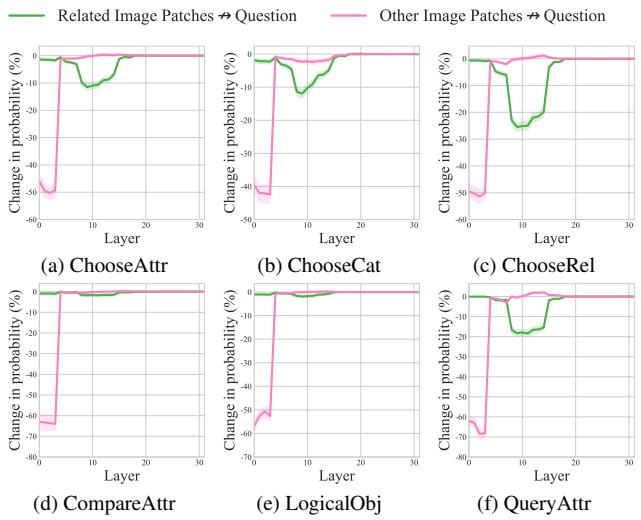



Figure 26. The relative changes in prediction probability on *LLaVA-v1.6-Vicuna-7b* with six *VQA* tasks. *Related Image Patches↛question* and *Other Image Patches↛question* represent blocking the position of *question* from attending to that of different image patches, region of interest and remainder, respectively.

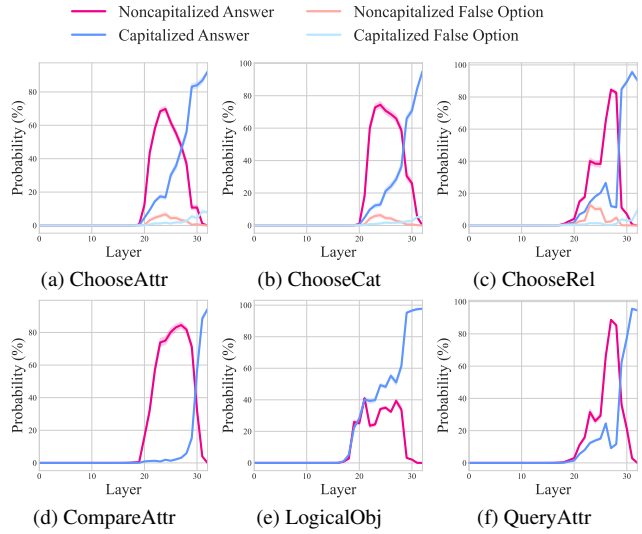



Figure 27. The probability of the answer word at the *last position* across all layers in LLaVA-v1.6-Vicuna-7b with six VQA tasks. *Capitalized Answer* and *Noncapitalized Answer* represent the answer word with or without the uppercase of the initial letter, respectively. As the tasks of *ChooseAttr*, *ChooseCat* and *ChooseRel* contain *false option*, we also provide the probability of it.

of the answer in the higher layers.

### I.3. *Llama3-LLaVA-NEXT-8b*

*Llama3-LLaVA-NEXT-8b* has quiet different architecture with *LLaVA-1.5-13b* in the main body of the paper. The

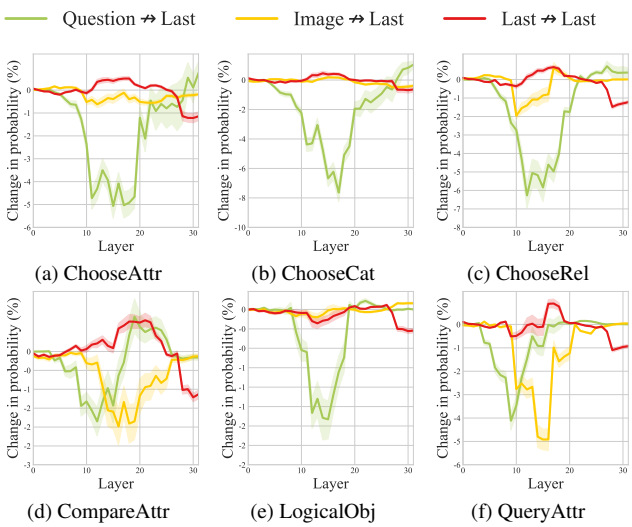



Figure 28. The relative changes in prediction probability on *llama3-llava-next-8b* with six *VQA* tasks. The *Question↛ Last*, *Image↛ Last* and *Last↛ Last* represent preventing *last position* from attending to *Question*, *Image* and itself respectively.

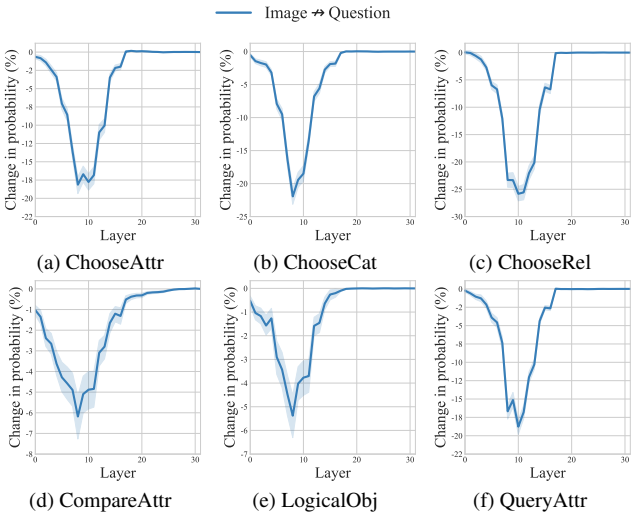



Figure 29. The relative changes in prediction probability when blocking attention edges from the *question* positions to the *image* positions on *llama3-llava-next-8b* with six *VQA* tasks.

difference between them includes the layer number, the way of processing image patch features, and the attention mechanism. The *Llama3-LLaVA-NEXT-8b* has 32 layers verse 40 layers in *LLaVA-1.5-13b*. *LLaVA-1.5-13b* directly feeds the original fixed-length image patch features from the image encoder into the LLM as input tokens. In contrast, *Llama3-LLaVA-NEXT-8b* employs a dynamic high-resolution technique, which dynamically adjusts image resolution, resulting in variable-length image patch features with higher res-

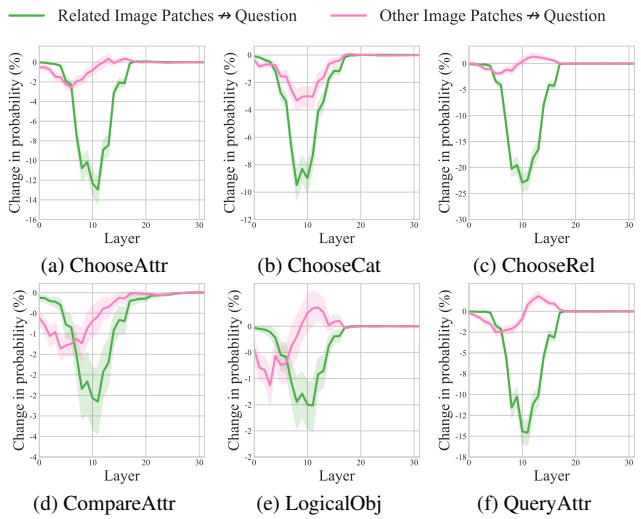



Figure 30. The relative changes in prediction probability on *llama3-llava-next-8b* with six *VQA* tasks. *Related Image Patches↛question* and *Other Image Patches↛question* represent blocking the position of *question* from attending to that of different image patches, region of interest and remainder, respectively.

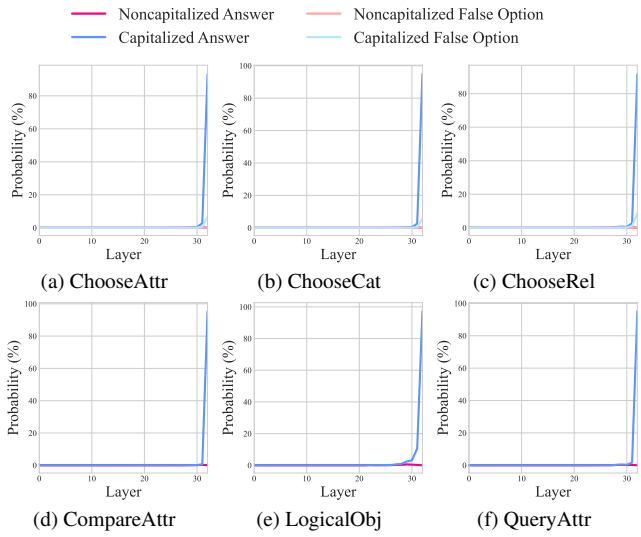



Figure 31. The probability of the answer word at the *last position* across all layers in llama3-llava-next-8b with six VQA tasks. *Capitalized Answer* and *Noncapitalized Answer* represent the answer word with or without the uppercase of the initial letter, respectively. As the tasks of *ChooseAttr*, *ChooseCat* and *ChooseRel* contain *false option*, we also provide the probability of it.

olution. Specifically, the higher resolution is implemented by splitting the image into grids and encoding them independently. As for the attention mechanism, *LLaVA-1.5-13b* use a standard and dense transformer architecture [44] while *Llama3-LLaVA-NEXT-8b* adopts grouped query attention [4] where the queries are grouped and the queries in the same group has shared key and value.

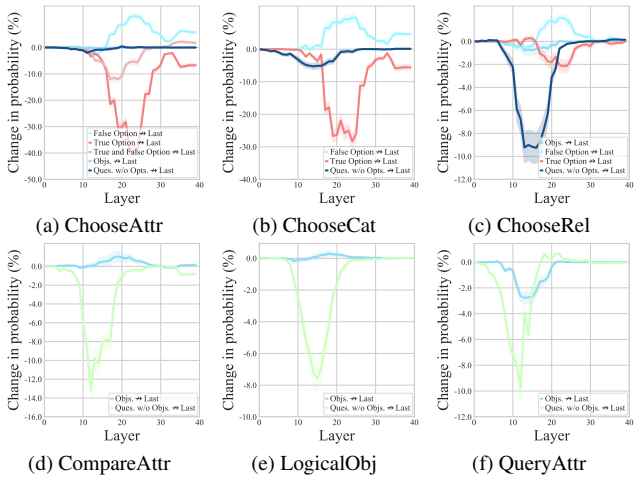
The information flow from different parts of the input sequence (*image* and *question*) to *last position*, from *image* to *question* and from different image patches (*related image patches* and *other image patches*) to *question*, as shown in Figure 28, Figure 29 and Figure 30 respectively, are consistent with the observations for the *LLaVA-1.5-13b* model, as shown in Figure 3, Figure 4 and Figure 5 respectively, in the main body of the paper. Although the information flow from *image* to *question* in Figure 29 appears to exhibit only a single drop, the Figure 30 reveals that, in lower layers, the information flow from *Other Image Patches* to the *question* play a dominant role compared to that from *Related Image Patches* to *question* and in following layers, information flow from *Related Image Patches* to *question* are more notable than that form *Other Image Patches* to *question*. This observation indicates that the model still has a two-stage multimodal information integration process. Specifically, in the first stage, the model focuses on generating holistic representations of the entire image. In the second stage, it refines these representations to align them more closely with the specific given question. Subsequently, in the middle layers, the critical multimodal information flows from the *question* positions to the *last position* for the final prediction. Moreover, the probability changes for the *Capitalized Answer* across all layers, as illustrated in Figure 31, align closely with the results in the main body of the paper while no such pattern is observed for the *Noncapitalized Answer*. This suggests that the model generates the syntactically correct answer directly, without a distinct intermediate step of semantic generation followed by syntactic correction. A potential explanation for this behavior is that when *Llama3* generates an answer to a given question, it first outputs a "\n" token, which may act as a cue to produce an answer word starting with an uppercase letter.

## J. The fine-grain analysis for information flow

In the main body of the paper, we primarily focus on analyzing the information flow between one specific combination of (*image*, *question*, and *last position*) for analyzing the multimodal information integration. In this section, we will further investigate the information flow between fine-grain parts of the input sequence, including the *question without options*, *true option*, *false option*, *objects* in the question, *question without objects*, *related image patches* and *other image patches*. We also use the same *attention knockout* method to block the attention edge between them to investigate the information flow between them.

### J.1. Different parts of the *question* to the *last position*

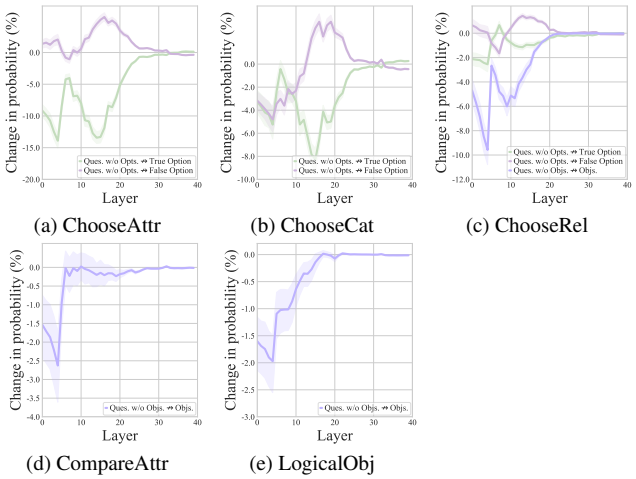
In the tasks of *ChooseAttr*, *ChooseCat* and *ChooseRel*, for each layer $\ell$, we block *last position* from attending to dif-




(a) ChooseAttr (b) ChooseCat (c) ChooseRel

(d) CompareAttr (e) LogicalObj (f) QueryAttr

Figure 32. The relative changes in prediction probability on *LLaVA-v1.5-13b* with six *VQA* tasks. Preventing *Last Position* from attending to different parts of *Question*, such as *True Option*, *False Option*, *Objects* in question, *Question without Options*, *Question without Objects*, both *N*True Option and *False Option* together.




(a) ChooseAttr (b) ChooseCat (c) ChooseRel

(d) CompareAttr (e) LogicalObj

Figure 33. The relative changes in prediction probability on *LLaVA-v1.5-13b* with six *VQA* tasks. Preventing information flow from *Question without Option* to *Options* and from *Question without Objects* to *Objects*.

ferent parts of *question*, including *question without options*, *true option*, *false option*, with the same window size ($k = 9$) around the $\ell$-th layer and observe the change in the probability of the answer word at the *last position*. In the tasks of *CompareAttr LogicalObj* and *QueryAttr*, we conduct the same operations with the above tasks except for blocking *last position* from attending to *objects* or *question without objects* as these tasks do not contain *options* in the question.

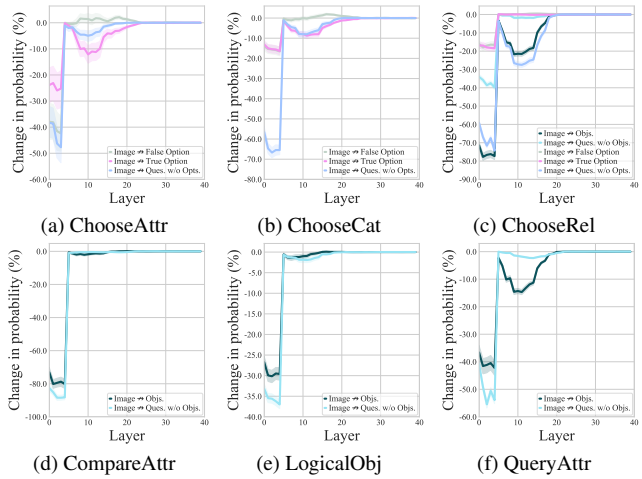
As shown in Figure 32 (a), (b) and (c), for the tasks of *ChooseAttr*, *ChooseCat* and *ChooseRel*, the *true option* and *false option* flowing the information to the *last position* occur in similar layers (higher layers) in the model. When blocking *last position* from attending *true option*, the probability obtain a reduction, while blocking *last position* from attending *false option* increases the probability of the correct answer word. The increase is reasonable because the question without the false option becomes easy for the modal. For the tasks of *ChooseAttr* and *ChooseCat*, in the information flowing to *last position*, the *options* play a dominant role while *question without options* only results in a small reduction for the probability fo the correct answer word. In contrast, for the *ChooseRel* task, the *true option* does not significantly reduce the probability of the correct answer word. This may stem from the format of the *ChooseRel* questions, where the options are positioned in the middle of the question, rather than at the end as in the *ChooseAttr* and *ChooseCat* tasks. As a result, the options in *ChooseRel* are less effective at aggregating the complete contextual information of the question within an autoregressive transformer decoder. Consequently, the flow of information from the *option* to the *final position* becomes less critical in determining the correct answer.

As the questions in our dataset target one or more specific objects in the image, we also conduct experiments on blocking *last position* from attending to *objects* or *question without objects*. As shown in Figure 32 (d), (e) and (f), the critical information from the *objects* does not directly transfer into the *last position* compared to that form *question without objects* to *last position*. This implies that the *objects* might affect the final prediction in an indirect way.

## J.2. Different parts of the *question* to different parts of the *question*

In the tasks of *ChooseAttr*, *ChooseCat* and *ChooseRel*, for each layer $\ell$, we block *options* from attending to *question without options* with the same window size ($k = 9$) around the $\ell$-th layer and observe the change in the probability of the answer word. In the tasks of *CompareAttr* and *LogicalObj*, we conduct the same operations with the above tasks except for blocking *objects* from attending to *question without objects*.

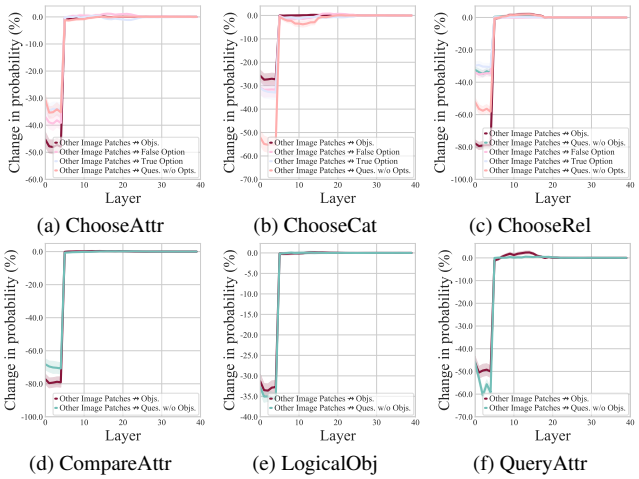
As shown in Figure 33 (a), (b) and (c), for the tasks of *ChooseAttr*, *ChooseCat* and *ChooseRel*, the information flow from *question without options* to *true option* occurs in similar transformer layers with that from *question without options* to *false option*. We also observe that these indirect information flows from *question without options* to *false option* occur before the information flow from *options* to *last position* as shown in Figure 32. This indicates that the information of the question is aggregated into the *options* in lower layers and then the information in *options* is transferred to the *last position* for the prediction of the fi-




(a) ChooseAttr (b) ChooseCat (c) ChooseRel

(d) CompareAttr (e) LogicalObj (f) QueryAttr

Figure 34. The relative changes in prediction probability on *LLaVA-v1.5-13b* with six *VQA* tasks. Blocking the information flow from *Image* to different parts of the *question*, including *True Option*, *False Option*, *Objects* in question, *Question without Objects*, *Question without Options*.




(a) ChooseAttr (b) ChooseCat (c) ChooseRel

(d) CompareAttr (e) LogicalObj (f) QueryAttr

Figure 35. The relative changes in prediction probability on *LLaVA-v1.5-13b* with six *VQA* tasks. Blocking the information flow from *Other Image Patches* to different parts of the question, including *True Option*, *False Option*, *Objects* in question, *Question without Objects*, *Question without Options*.

nal answer in higher layers. For the tasks of *CompareAttr* and *LogicalObj*, we observe that the information flow from *question without objects* to *objects* occurs in lower layers.

### J.3. *Image* to different parts of *question*

In the tasks of *ChooseAttr*, *ChooseCat* and *ChooseRel*, for each layer $\ell$, we block attention edge between *image* and different parts of *question*, including *question without options*, *true option* and *false option*, with the same window size ($k = 9$) around the $\ell$-th layer and observe the change in the probability of the answer word. In the tasks of *CompareAttr*, *LogicalObj* and *QueryAttr*, we carry out the same operations as in the above tasks except for blocking the edge of the attention between *image* and *question without objects* or *objects* respectively.

As illustrated in Figure 34, the overall information flow from *image* to different parts of the *question* aligns consistently with the information flow from the *image* to the entire *question*, as depicted in Figure 4 in the main body of the paper. Specifically, two distinct flows are from the *image* to the *question*. Notably, however, different parts of the *question* exhibit varying magnitudes of probability change, especially in the second-time drop in probability, which may be because different kinds of questions have different attention patterns to the image. For example, during the second-time drop in probability, in the tasks of *ChooseAttr* and *ChooseCat*, the *image* information does not transfer to *false option* while it transfers much more information to *true option*. However, this pattern isn't observed in the task of *ChooseRel*, where most *image* information is transferred into *question without options* and *objects*.

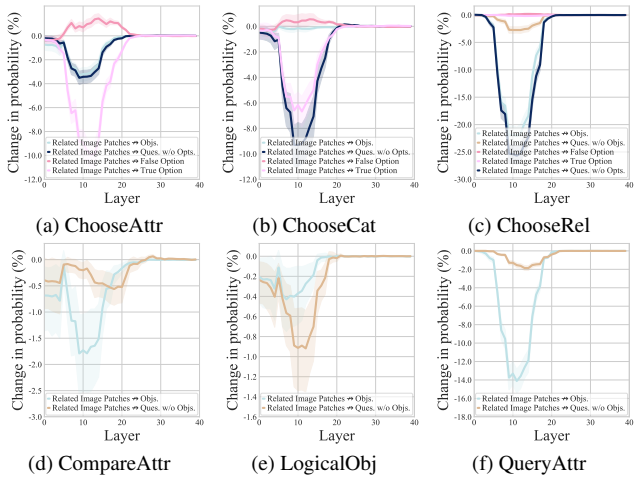
### J.4. *Other image patches* to different parts of *question*

In the tasks of *ChooseAttr*, *ChooseCat* and *ChooseRel*, for each layer $\ell$, we block attention edge between *other image patches* and different parts of *question*, including *question without options*, *true option*, *false option*, *objects* and *question without objects*, with the same window size ($k = 9$) around the $\ell$-th layer and observe the change in the probability of the answer word. In the tasks of *CompareAttr*, *LogicalObj* and *QueryAttr*, we conduct the same operations with the above tasks except for blocking attention edge between *other image patches* and *question without objects* or *objects* respectively.

As shown in Figure 35, the information flow from *other image patches* to different parts of the *question* for all six tasks consistently aligns the flow observed from *other image patches* to the entire *question*, as illustrated in Figure 5 in the main body of the paper. Specifically, the information flow dominantly occurs in the first-time drop in the probability in the lower layers, regardless of which part of the *question* is being blocked.

### J.5. *Related image patches* to different parts of *question*

In the tasks of *ChooseAttr*, *ChooseCat* and *ChooseRel*, for each layer $\ell$, we block the attention edge between *Related image patches* and different parts of *question*, including *question without options*, *true option*, *false option*, *objects* and *question without objects*, with the same window size ($k = 9$) around the $\ell$-th layer and observe the change in the probability of the answer word. In the tasks of *CompareAttr*, *LogicalObj* and *QueryAttr*, we conduct the same operations with the above tasks except for blocking the at-




(a) ChooseAttr (b) ChooseCat (c) ChooseRel

(d) CompareAttr (e) LogicalObj (f) QueryAttr

Figure 36. The relative changes in prediction probability on *LLaVA-v1.5-13b* with six *VQA* tasks. Blocking the information flow from *Related Image Patches* to different parts of the question, including *True Option*, *False Option*, *Objects* in question, *Question without Objects*, *Question without Options*.

tention edge between *Related image patches* and *question without objects* or *objects* respectively.

The observations of the overall information flow from *related image patches* to different parts of the *question* for all six tasks shown in Figure 36 consistently align the flow observed from *related image patches* to the entire *question*, as illustrated in Figure 5 in the main body of the paper. Specifically, the information flow dominantly occurs in the second-time drop in the probability in the lower-to-middle layers (around 10th). However, there are some parts of *question* that don't obtain the information followed from the *related image patches*. For example, the *objects* in the task of *ChooseCat*, or *false option* and *true option* in the task of *ChooseRel*.

## K. The influence of *images* on the semantics of *Questions*

We already know that the *image* information is integrated into the representation corresponding to the position of *question*. To investigate whether the *image* affects the final semantics of the *question*, for each layer $\ell$, we prevent the *question* from attending to the *question*, with the same window size ($k = 9$) around the $\ell$-th layer and observe the change of semantics of the *questtion* in the final layer. The semantics of the *question* is evaluated by the *Jaccard Similarity* as in Appendix D.

As illustrated in Figure 37, the *Jaccard Similarity* demonstrates a significant decline in the lower layers, resembling the behaviour observed in layers where information flows from the *image* to the *question*. This pattern highlights the critical role of *image* information in constructing the final multimodal semantic representation.

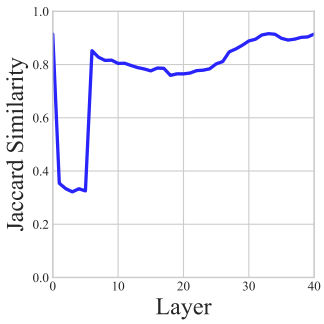



Figure 37. The Jaccard similarity between the predicted words of the original model *LLaVA-1.5-13b* and those of the intervened model blocking *question* from attending to *image* on the task of *ChooseAttr*.