# Cross-modal Information Flow in Multimodal Large Language Models

Zhi Zhang*, Srishti Yadav*†, Fengze Han‡, Ekaterina Shutova*

*ILLC, University of Amsterdam, Netherlands

†Dept. of Computer Science, University of Copenhagen, Denmark

‡Dept. of Computer Engineering, Technical University of Munich, Germany

# Project Objective

- **Objective:**

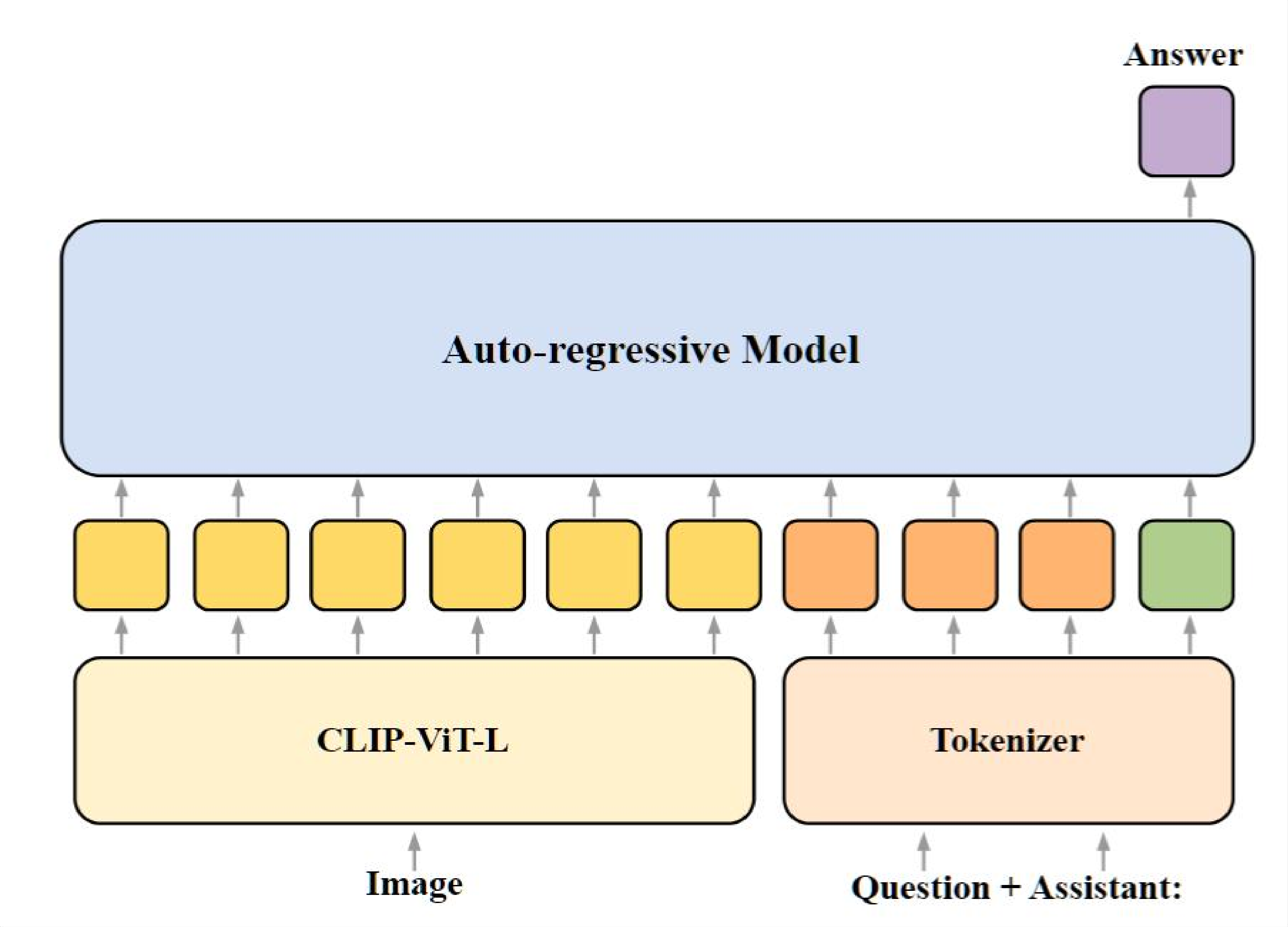
  Investigate the inner working mechanism of the multi-modal large language model (MLLM) when performing multi-modal tasks

- **Steps:**

  Within the auto-regressive MLLMs, we aim to answer the following questions:

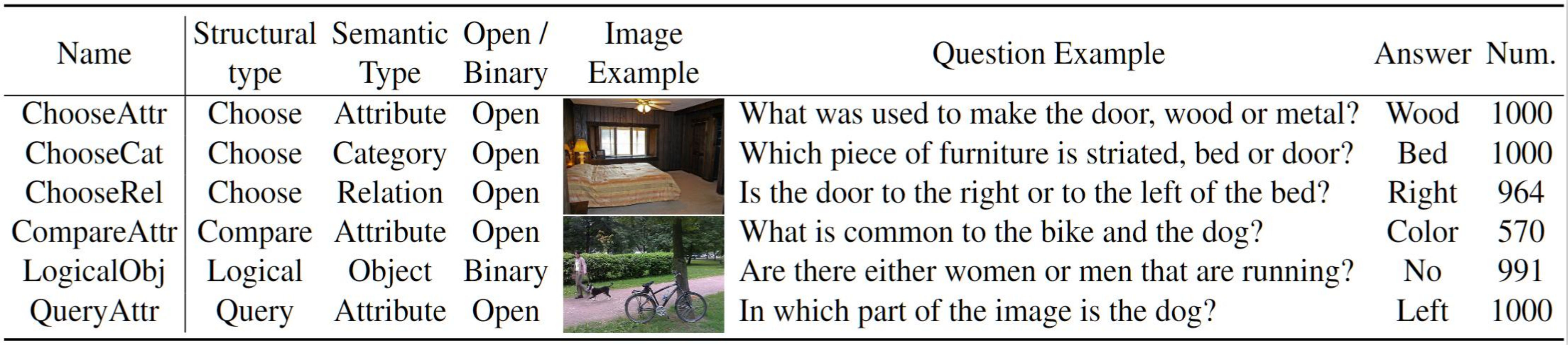
  - Where is visual and linguistic information integrated?
  - How is visual and linguistic information integrated?
  - How is the final prediction generated?

# Multi-modal large language model -- Llava



# Multi-modal task

- 6 VQA Tasks:



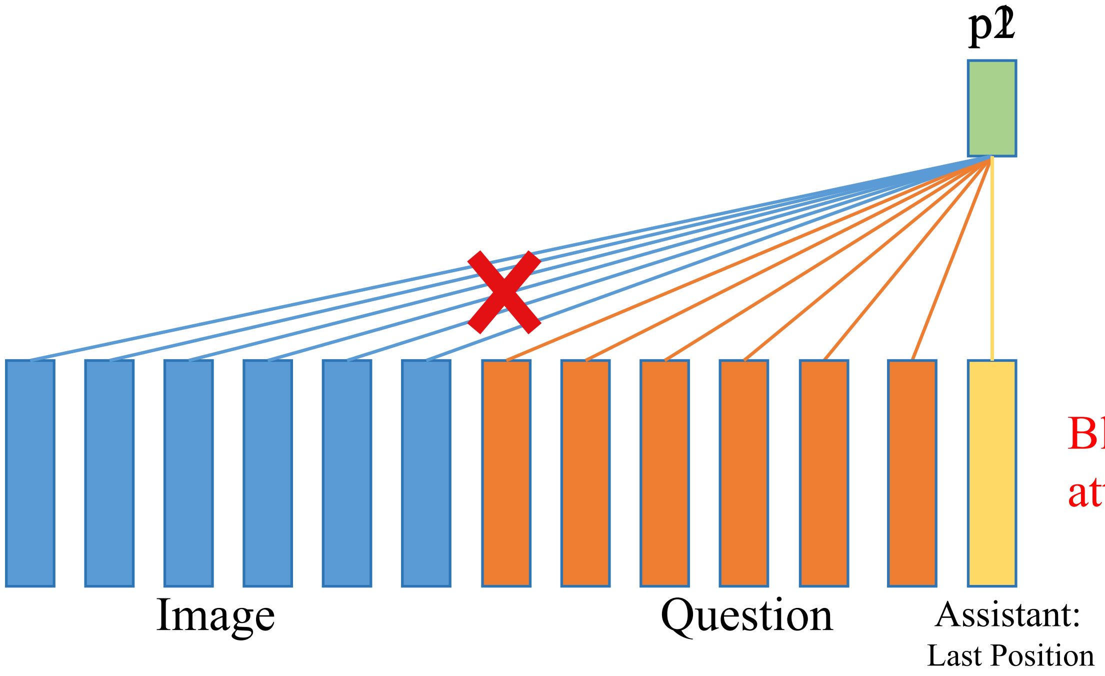
| Name | Structural type | Semantic Type | Open / Binary | Image Example | Question Example | Answer | Num. |
|---|---|---|---|---|---|---|---|
| ChooseAttr | Choose | Attribute | Open | | What was used to make the door, wood or metal? | Wood | 1000 |
| ChooseCat | Choose | Category | Open | | Which piece of furniture is striated, bed or door? | Bed | 1000 |
| ChooseRel | Choose | Relation | Open | | Is the door to the right or to the left of the bed? | Right | 964 |
| CompareAttr | Compare | Attribute | Open | | What is common to the bike and the dog? | Color | 570 |
| LogicalObj | Logical | Object | Binary | | Are there either women or men that are running? | No | 991 |
| QueryAttr | Query | Attribute | Open | | In which part of the image is the dog? | Left | 1000 |

- Input form:

  [Image, Question, Assistant:]

# Method: Attention Knockout

- We use attention knockout method to investigate the information flow between diffenerent parts of the input

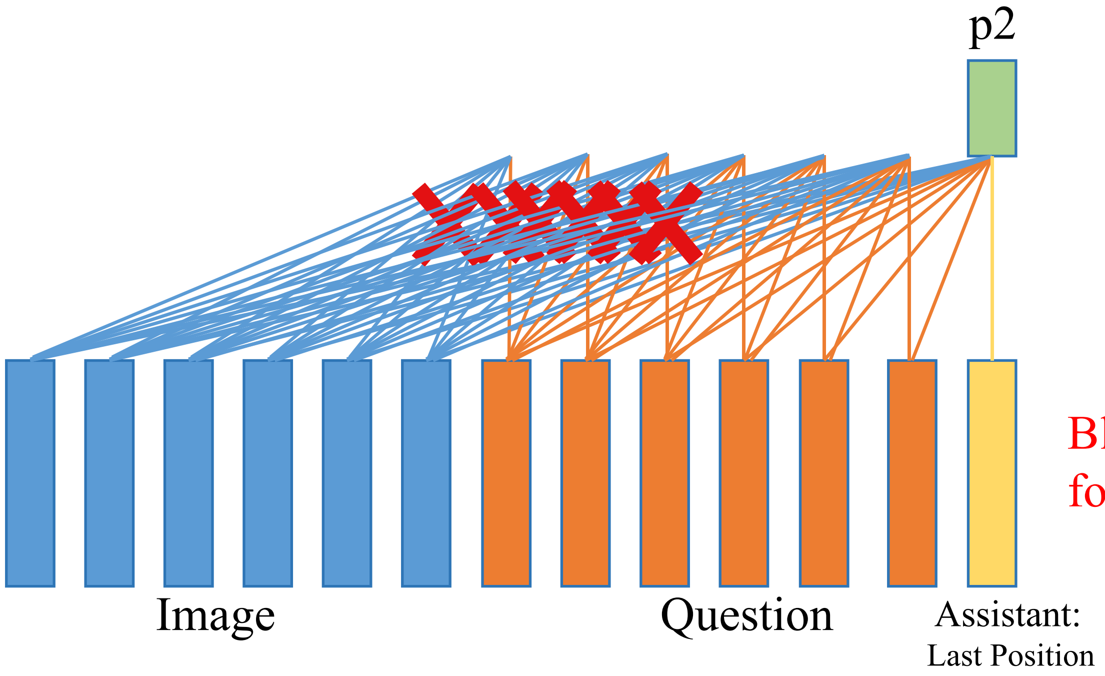


(p1- p2) /p1
relative change in probability

Block ***last position*** token from attending ***image*** for final prediction

# Method: Attention Knockout

- We use attention knockout method to investigate the information flow between diffenerent parts of the input



Block ***question*** from attending ***image*** for final prediction

# Information Flow: To *Last*

- ***Last-> Last***:

  Significant change in probability.

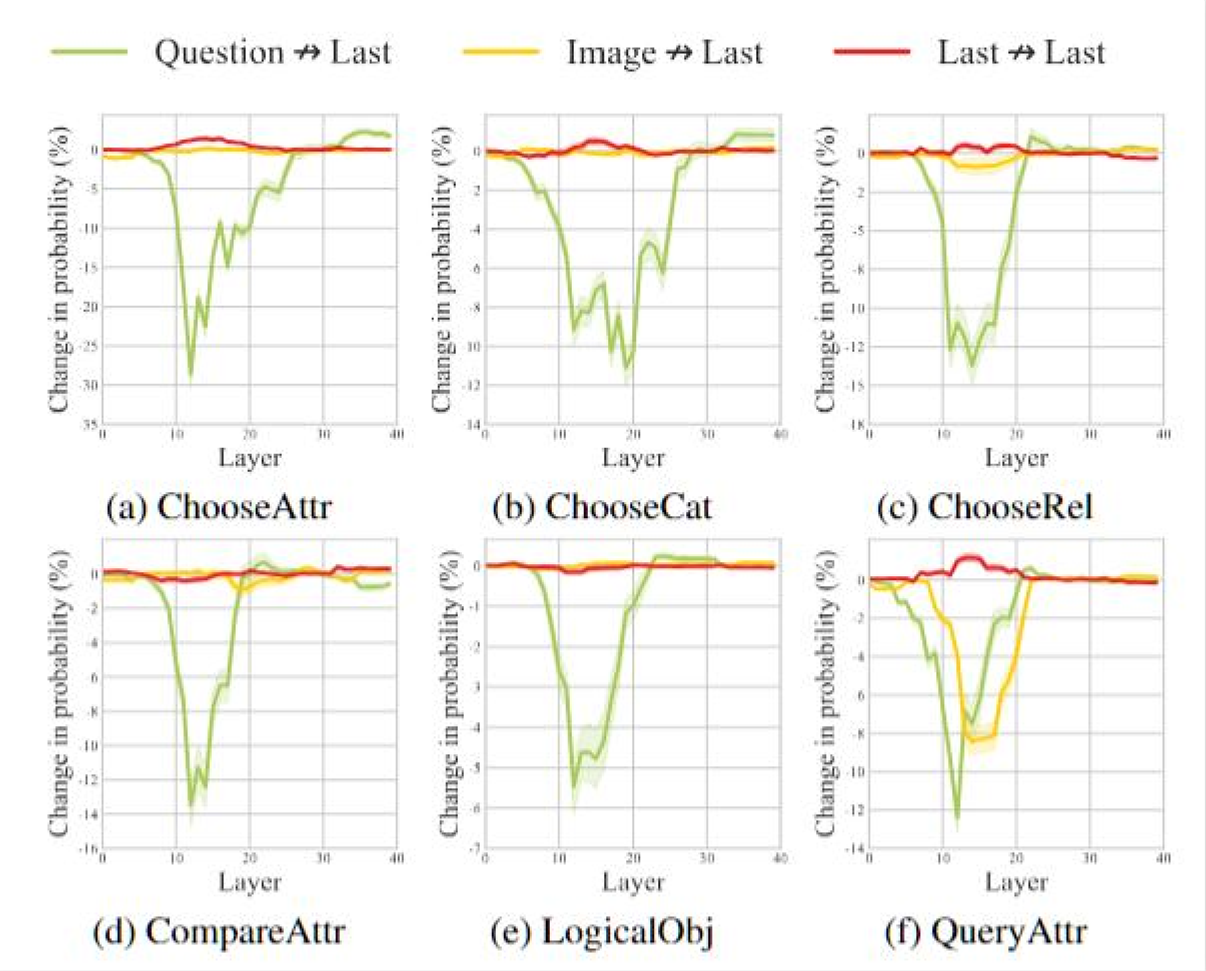
- ***Image -> Last***:

  No changes in the probability.

- ***Question->Last***:

  Significant change in probability.

Only the ***Question*** directly transfer crucial information to ***last position*** for the final prediction **in the middle layers.**



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

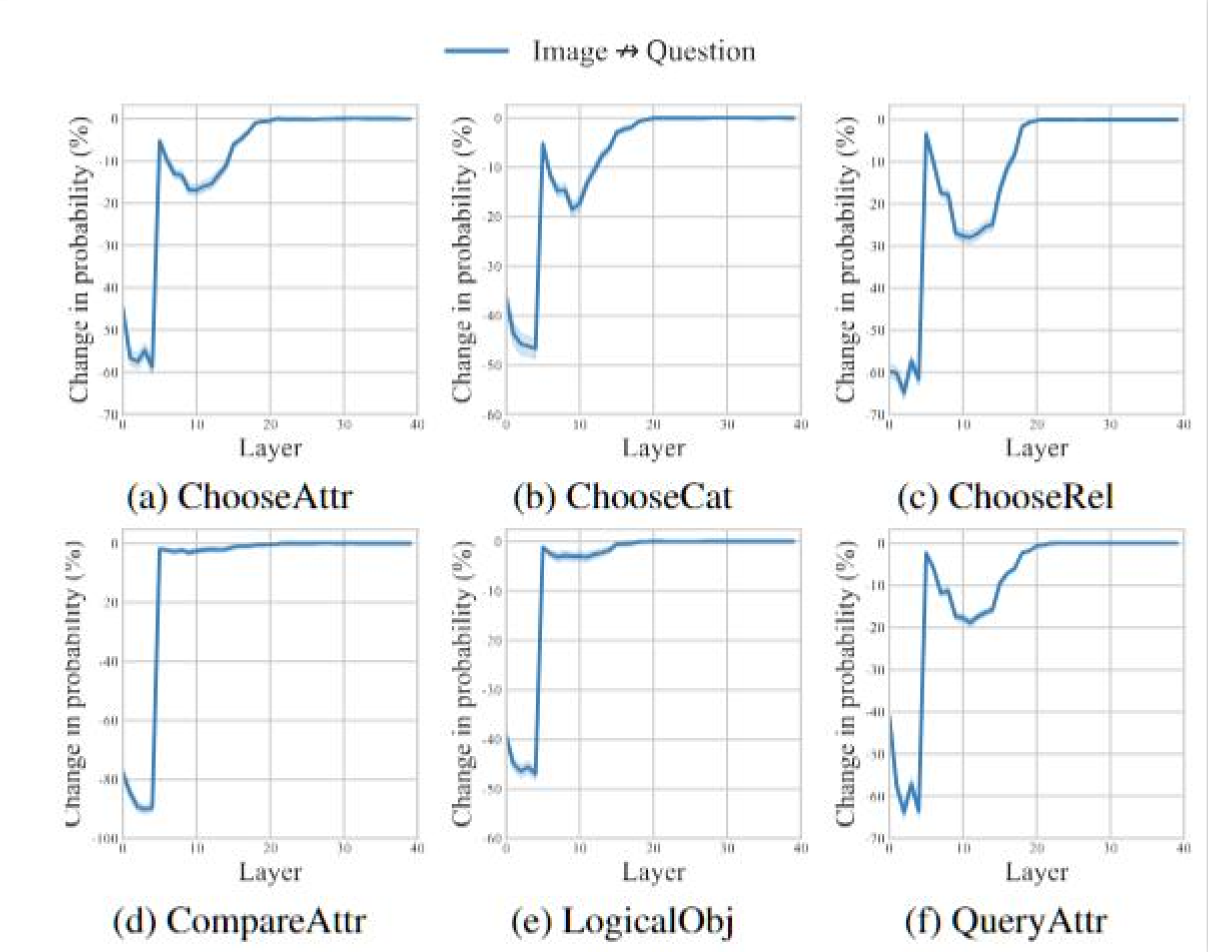
(d) CompareAttr (e) LogicalObj (f) QueryAttr

# Information Flow: *Image* To *Question*

***Image -> Question***

Image information flows twice:

- **lower layers**
- **lower-middle layers**

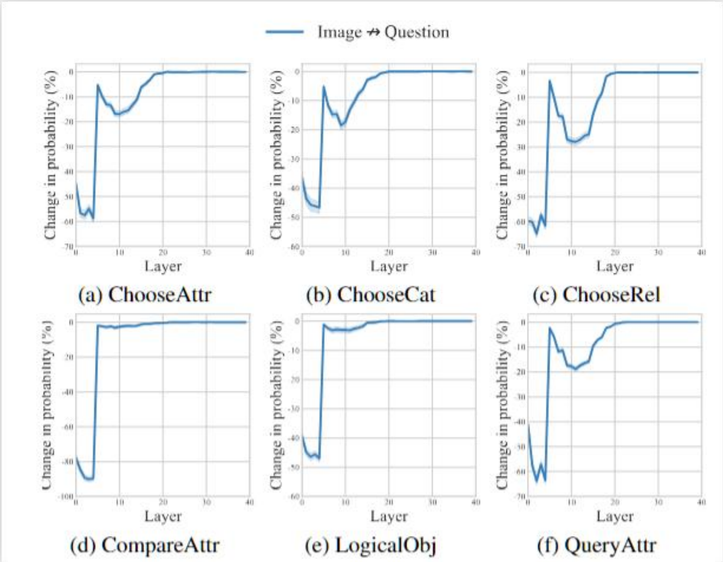


(a) ChooseAttr (b) ChooseCat (c) ChooseRel

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

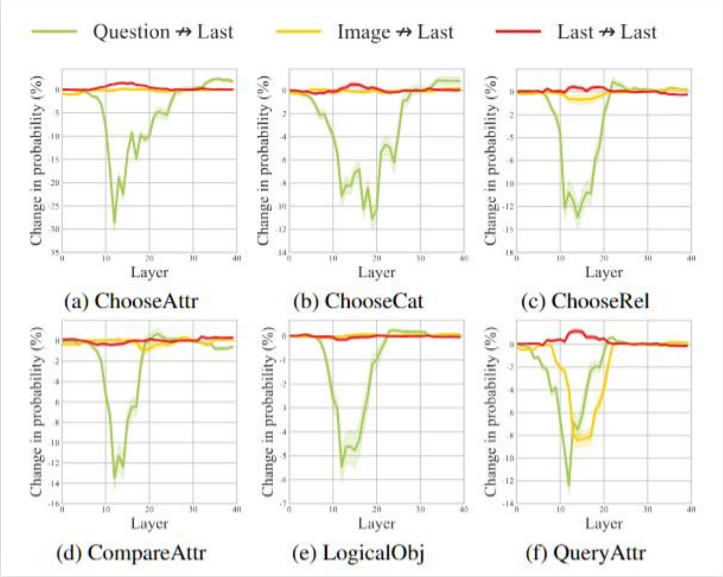
# Information Flow Conclusion

- First, in **lower** layers, ***Image*** propagate information to ***Question***;
- Then, in **lower-middle** layers, same information flow from ***Image*** to ***Question***;
- Finally, in **upper-middle** layers, information in ***Question*** is transferred to ***last position*** for final prediction.



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

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



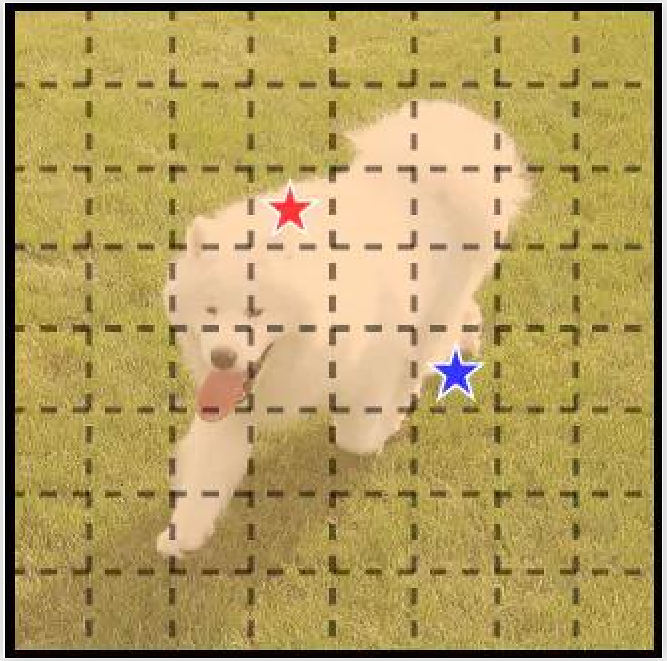
(a) ChooseAttr (b) ChooseCat (c) ChooseRel

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

# What is the difference between the two image information flow

**Image patches**

- We split image patches into two groups:
  - image patches (***Related Image Patches***)
  - image patches (***Other Image Patches***)



# Image patches To Question

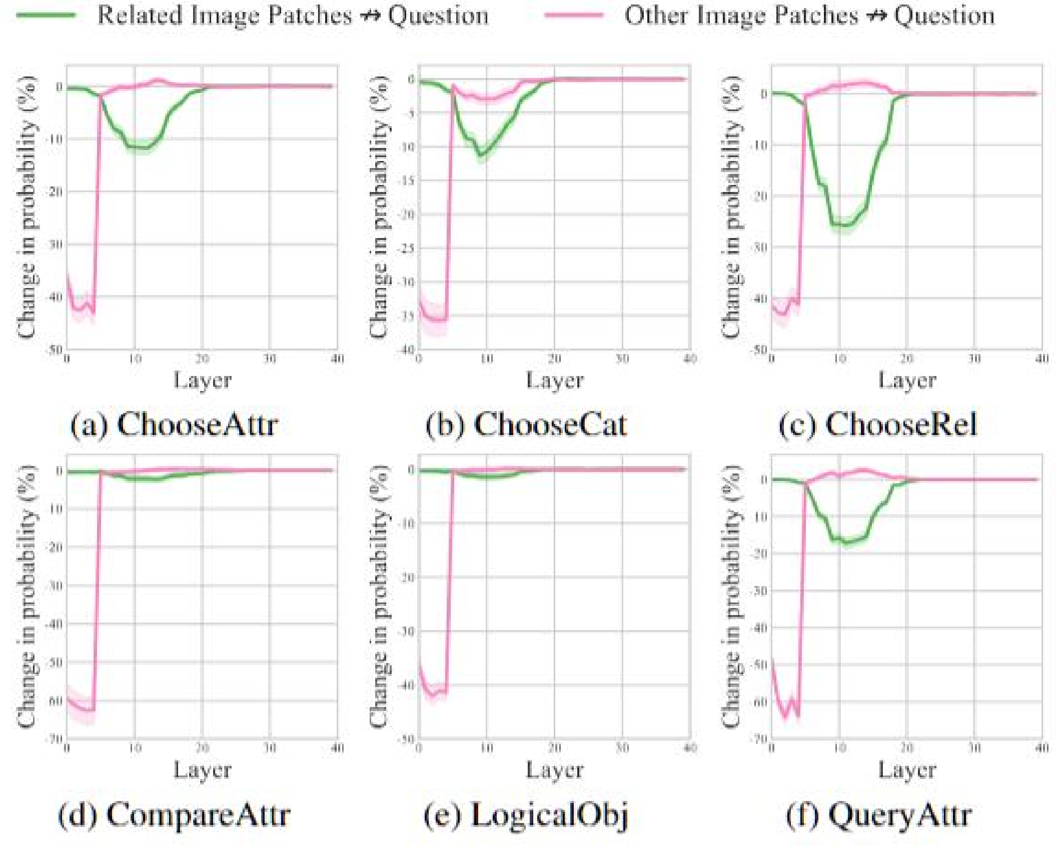
**Clear difference:**

- ***Other image patches*** propagate image information during first image information flow

- ***Related Image Patches*** propagate image information during second image information flow

**Conclusion:**

- **First**, in lower layer, the model integrates the whole image information to full question building a more generic representation,
- **Then**, in the subsequent layer (upper-middle layers), the model starts only to pay attention to the related parts to the object in the image fusing the features of the object with the full question.



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

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

# Answer word generation

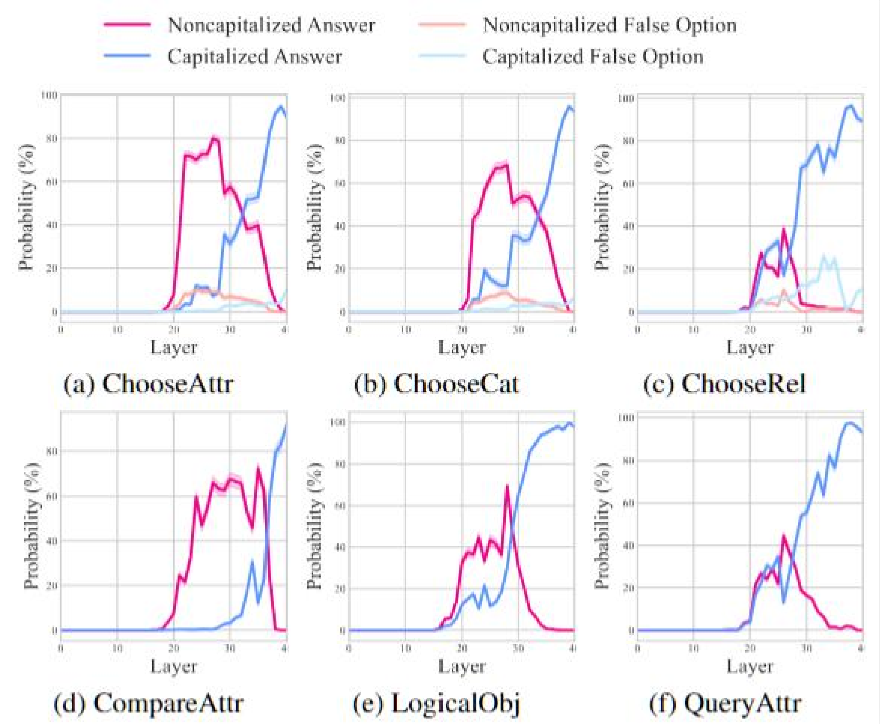
- ***Capitalized Answer***: Answer word with capital initial letter
- ***Noncapitalized Answer***: Answer word with lower-case for all letters

**Observation:**

- ***Capitalized Answer*****:** probability increase and then decrease
- ***Noncapitalized Answer***: probability increase when the probability of ***Capitalized Answer*** decrease

**Conclusion**:

- The model has already **semantically** inferred the answer in the **middle layers**
- Then, the model starts to refine the syntactic correctness of the answer in **higher layers**



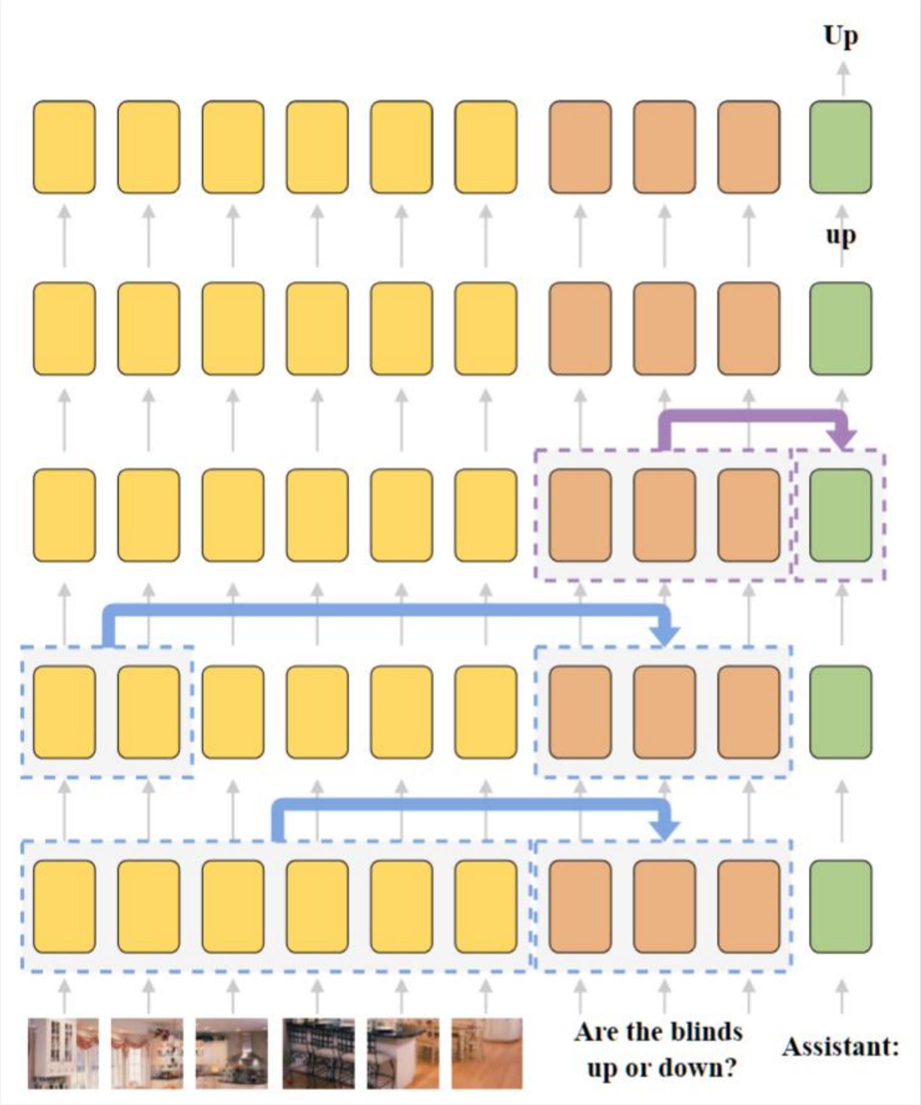
(a) ChooseAttr (b) ChooseCat (c) ChooseRel

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

# Conclusion

**From bottom to top layers:**

1. **First**, the model propagates general visual information from the whole image into the linguistic hidden representation;
2. **Next**, selected visual information relevant to answering the question is transferred to the linguistic representation;
3. **Finally**, the integrated multimodal information within the hidden representation of the question flows to last position facilitating the final prediction.
4. **In addition**, the answers are initially generated in lowercase form and then converted to uppercase for the first letter.


Up
up
Are the blinds up or down?
Assistant: