# Multimodal In-Context Learning

Anish Kachinthaya

# What Makes Multimodal In-Context Learning Work?

Folco Bertini Baldassini[1] Mustafa Shukor[1] Matthieu Cord[1,2] Laure Soulier[1]
Benjamin Piwowarski[1]
[1]Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
[2] Valeo.ai, Paris, France

# Multimodal ICL

- Similar to ICL, pass in sets of image, text (instruction/question), and response

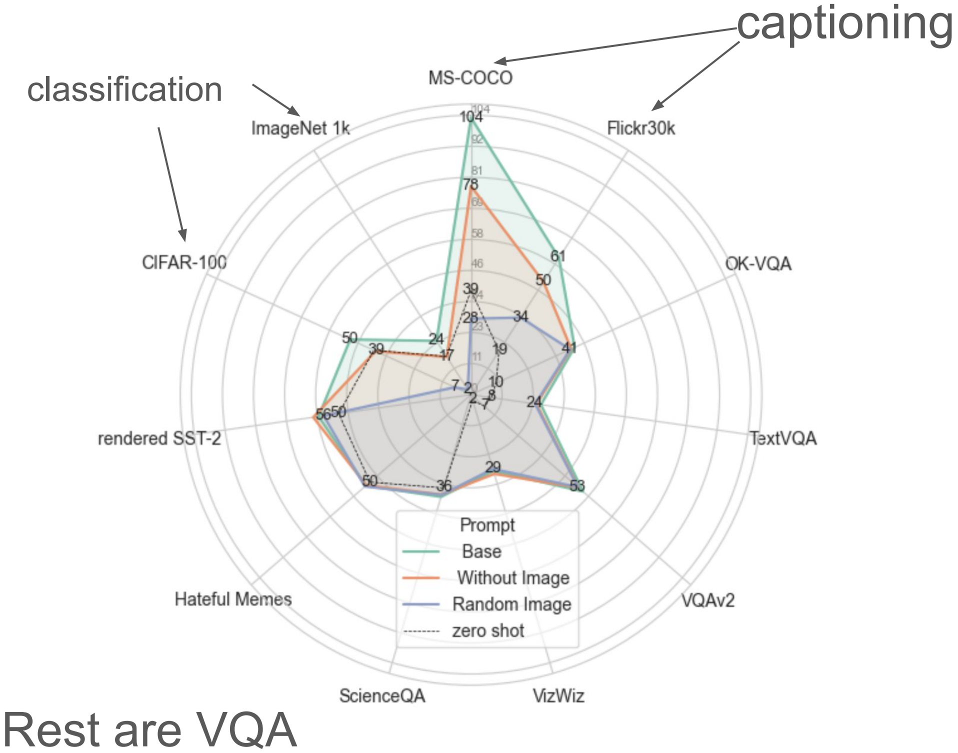
**How does each modality influence M-ICL?**

- For either the image or the text (instruction/question), either randomly replace or completely remove

**Which kind of shortcuts influence M-ICL?**

- Evaluate performance based on similarity of query <> outcomes. Is the model just copying what it sees in the demonstrations?
- Compare random sampling to retrieval based context selection (RICES)

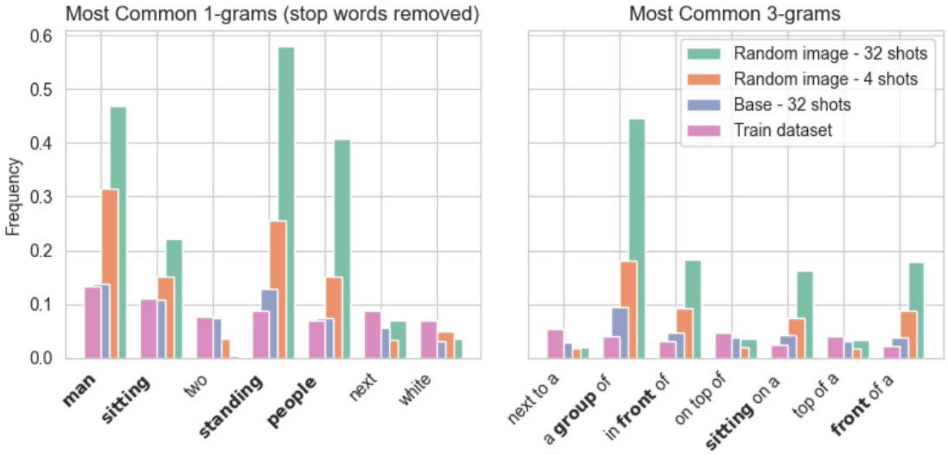
# Altering Images in M-ICL




(a) Altering image - 16 shots

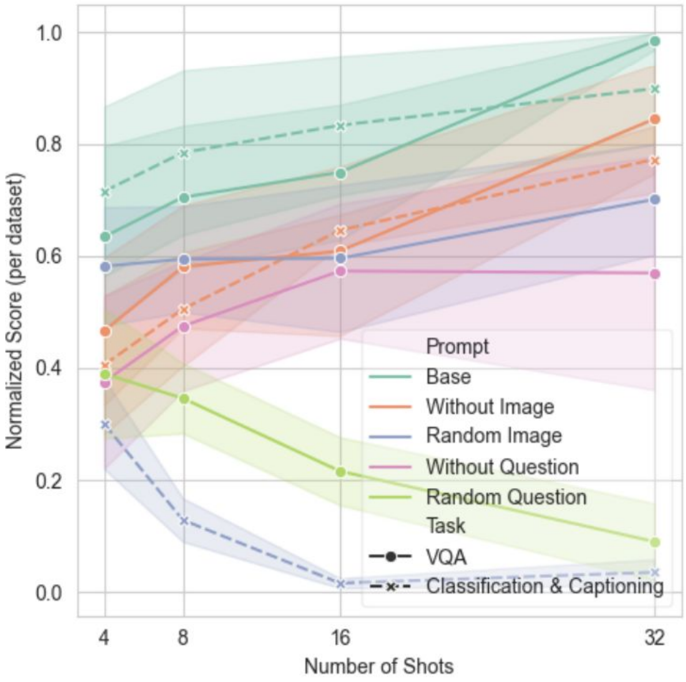
- Image-to-text tasks (captioning + classification) are heavily affected compared to VQA (less reliant on image)
- Performance is close to zero-shot/worse

# Altering Images—"Generic" text mode

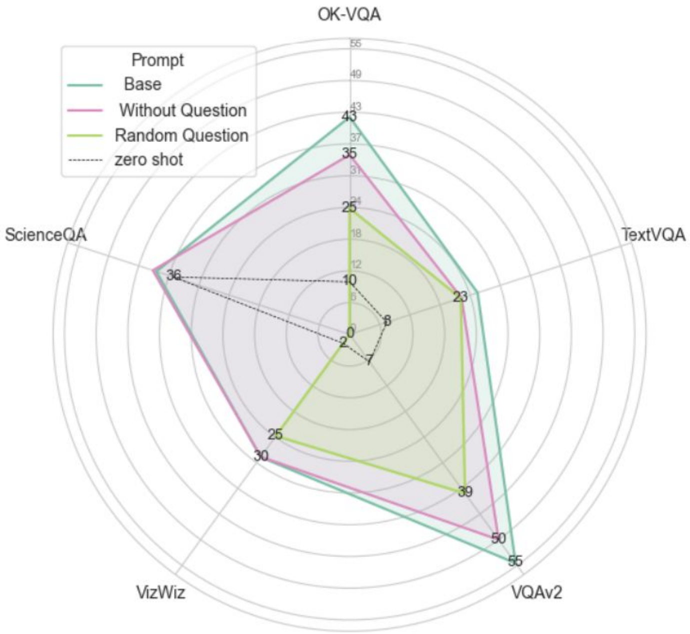



When images and text in the context demonstrations do not align, the model tends to output the most frequent words in the demonstrations.

# Altering Text in VQA




(b) Performance vs number of demonstrations.




(c) Altering question - 16 shots

# Text drives M-ICL

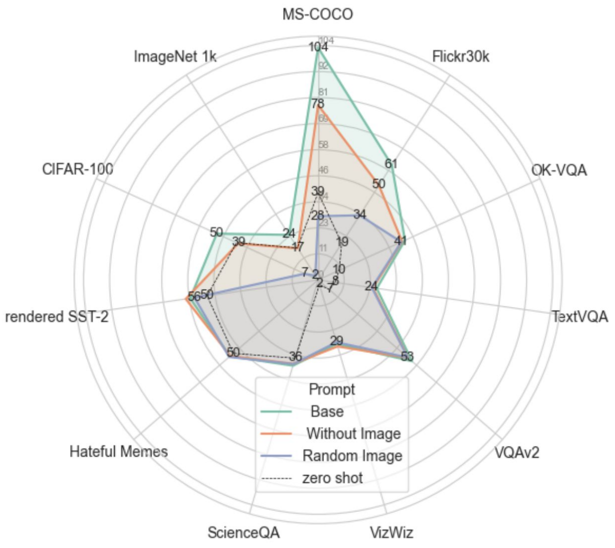
Text takes "precedence" in determining performance.

VQA: larger drop when altering text than when altering images.

Classification: "without image" (only text) performs as poorly as zero-shot (but better than random image)

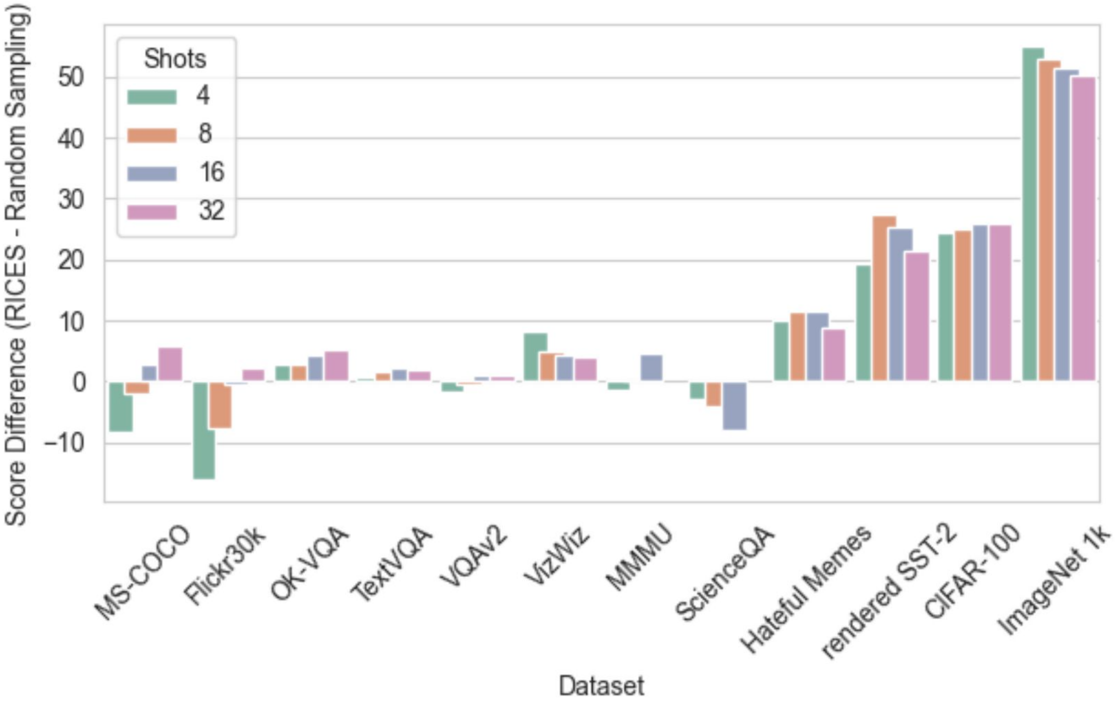
Captioning: only text captures the "style" of the captions or the distribution of words, improving over zero-shot by 31% (image provides additional 20%).

Both text and image modalities are important, but adding text provides a bigger boost.

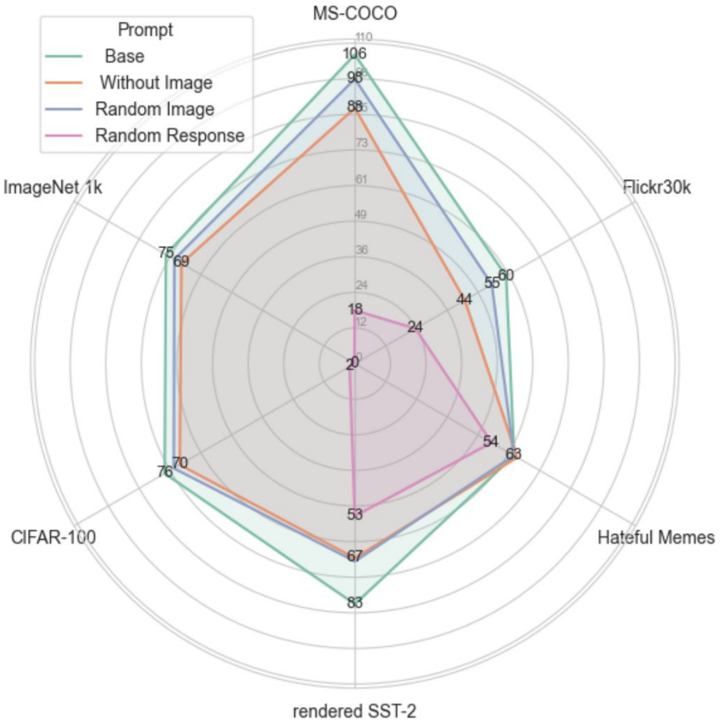



(a) Altering image - 16 shots

# Retrieving Similar Demonstrations



# Retrieving Similar Demonstrations



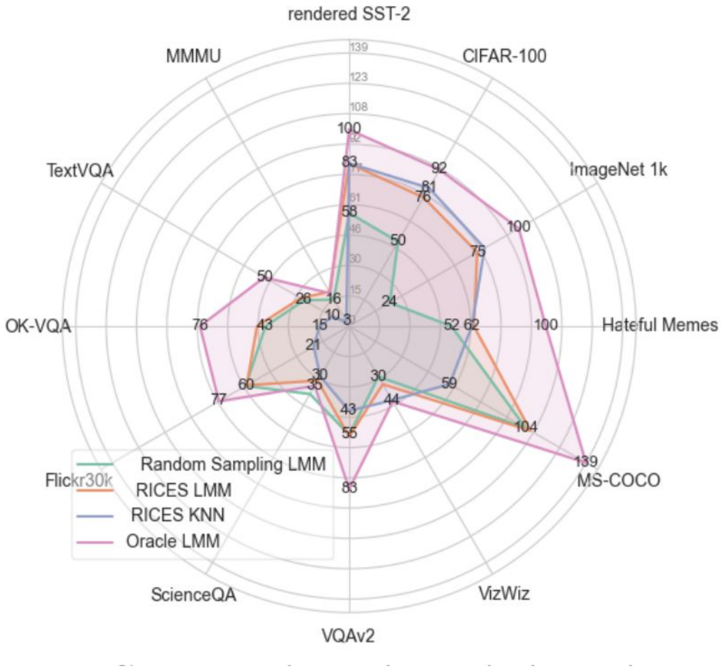

Random images do not degrade performance as much as random responses

# Shortcuts

Retrieving similar demonstrations for context achieves higher performance: but this is because the model is actually *learning* from the demonstrations or just “copying” them (using them as a shortcut)?

So, they compare RICES KNN (majority vote on similar examples) with RICES M-ICL.

# “Shortcut” Eval



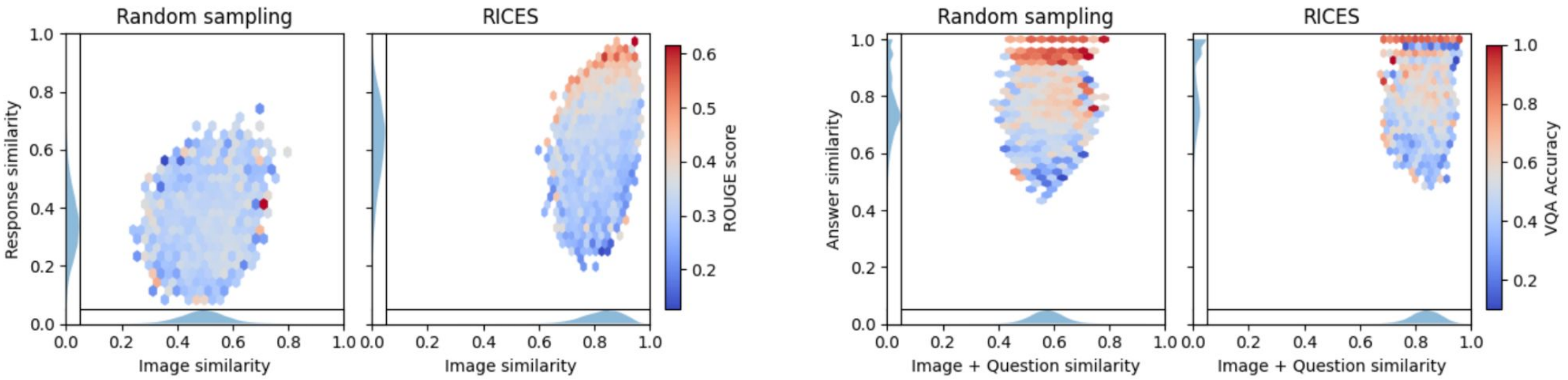

KNN achieves similar performance to M-ICL with similarity retrieval—suggests that M-ICL is using the distribution of the context responses rather than actually learning.

In open-ended generation, though, KNN is insufficient.

Oracle LMM is RICES based on ground truth response—the ideal case if retrieval was perfect, shows

- m-ICL can do intelligent soft copy when provided close responses
- just RICES similarity does not select good enough demonstrations to be *ideal*

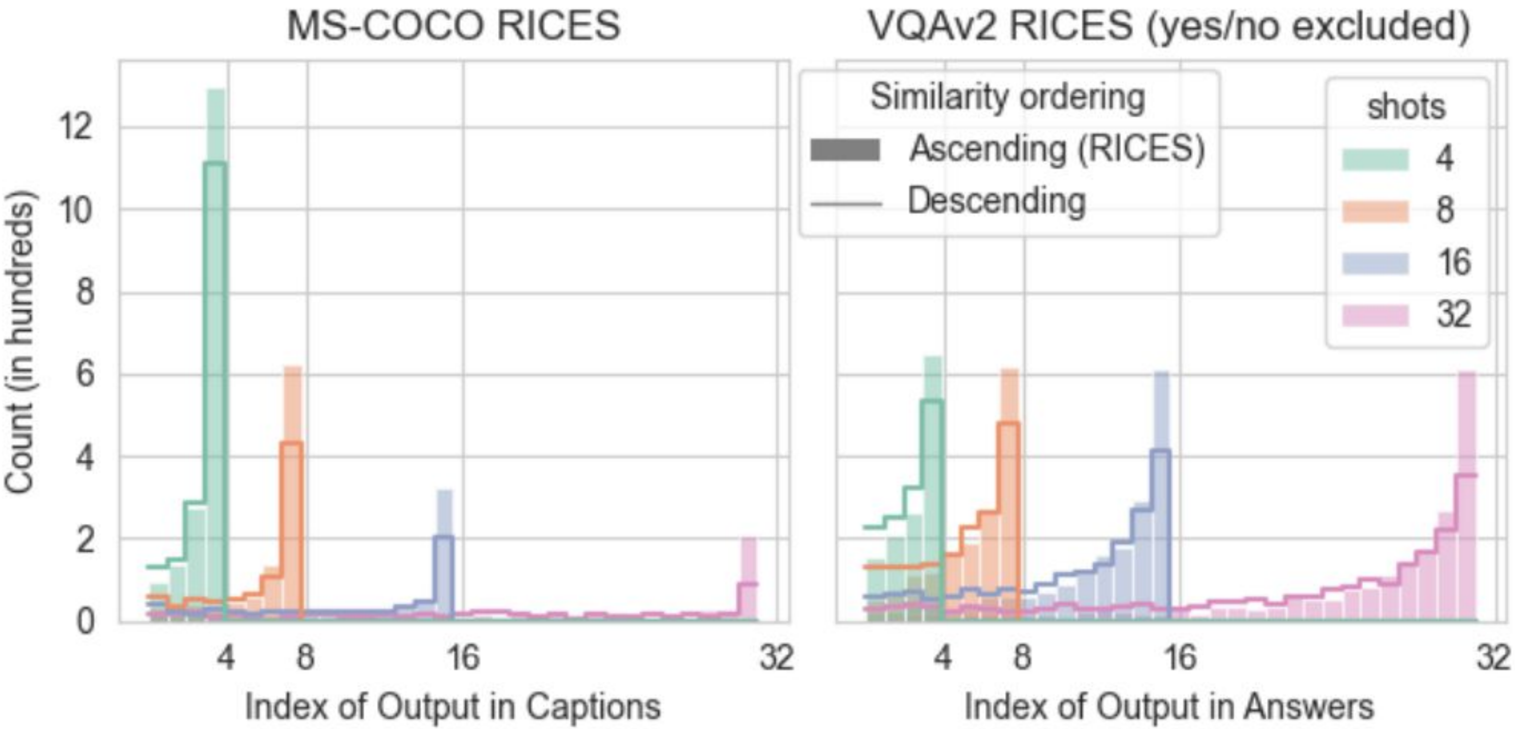
# Higher Response Similarity <> Performance

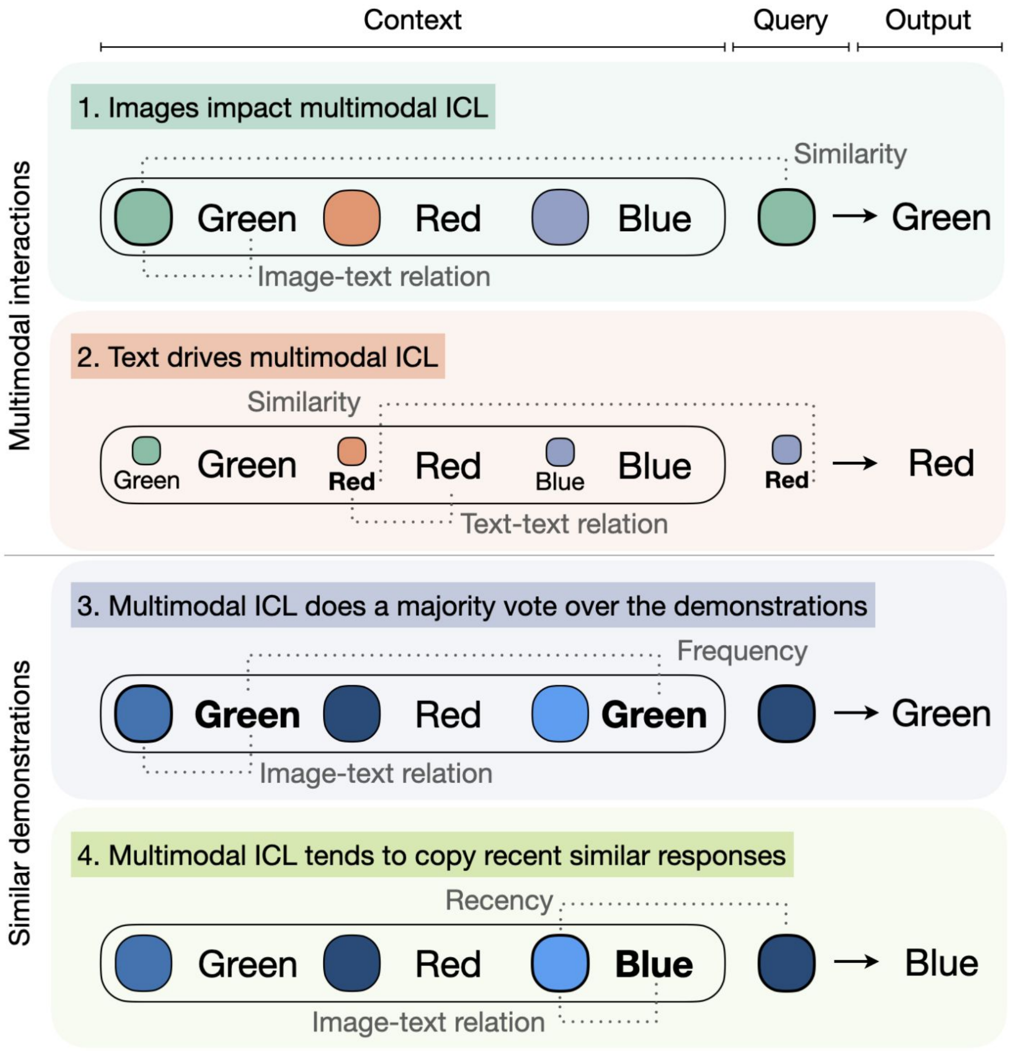



(a) COCO dataset

(b) VQA dataset

# Copying Later Demonstrations





Context
Query
Output
Multimodal interactions
1. Images impact multimodal ICL
Similarity
Green
Red
Blue
→ Green
Image-text relation
2. Text drives multimodal ICL
Similarity
Green
Green
Red
Red
Blue
Blue
Red
→ Red
Text-text relation
Similar demonstrations
3. Multimodal ICL does a majority vote over the demonstrations
Frequency
Green
Red
Green
→ Green
Image-text relation
4. Multimodal ICL tends to copy recent similar responses
Recency
Green
Red
Blue
→ Blue
Image-text relation

# Interpreting Visual Information Processing in VLMs

# Towards Interpreting Visual Information Processing in Vision-Language Models

**Clement Neo**[†*]

**Luke Ong**[†], **Philip Torr**[‡], **Mor Geva**[◇], **David Krueger**[♡]

**Fazl Barez**[‡,§]

[†]Nanyang Technological University [‡]University of Oxford [◇]Tel Aviv University

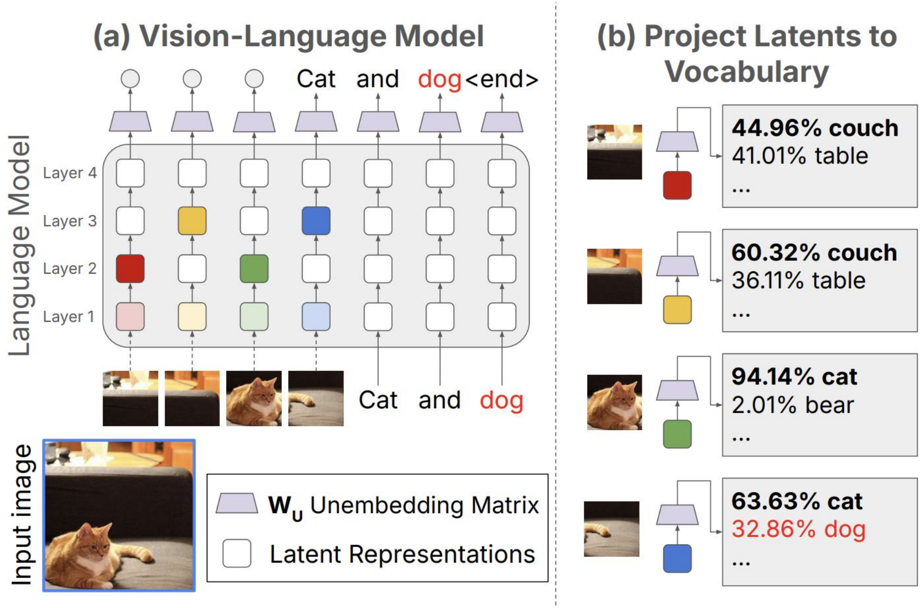
[♡]MILA [§]Tangentic

# Interpreting and Editing Vision-Language Representations to Mitigate Hallucinations

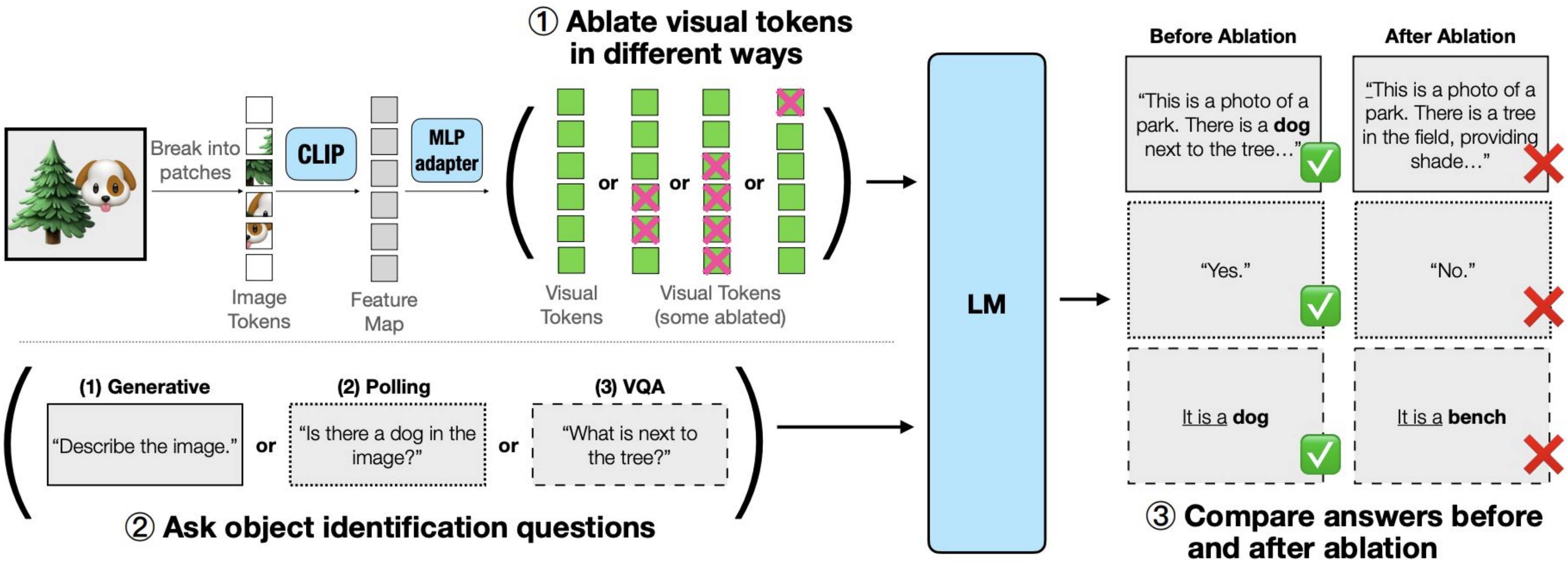
**Nick Jiang**[*], **Anish Kachinthaya**[*], **Suzie Petyrk**[†] ,**Yossi Gandelsman**[†]
University of California, Berkeley
{nickj,anishk,spetryk,yossi_gandelsman}@berkeley.edu

# Logit Lens

Unembed intermediate hidden states to retrieve probability distribution over the vocabulary at an intermediate layer



# Do visual tokens contain specific object information?




**Results:** object token ablation consistently results in larger performance decreases across all settings as compared to the gradient-based and random baselines. This suggests that the information about that object is localised to the region of the object token.

# At which layers is object information processed?

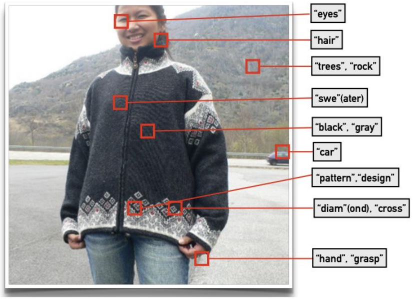
Blocking attention from the object tokens (and their buffers) to the final token in mid-late layers leads to noticeable performance degradation.

The model directly extracts object-specific information in these later stages.

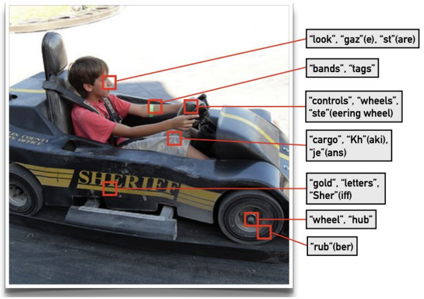
| Attention Blocking | | Layer | | | | | |
|---|---|---|---|---|---|---|---|
| From* | To† | Early | Early-Mid | Mid | Mid-Late | Late | All |
| O | LTP | 1.00 | 1.00 | 0.96 | 0.88 | 0.93 | 0.82 |
| O+1 | LTP | 1.00 | 0.99 | 0.90 | 0.82 | 0.89 | 0.67 |
| O+2 | LTP | 1.00 | 1.00 | 0.91 | 0.80 | 0.91 | 0.68 |
| I-(O+1) | LTP | 0.88 | 1.00 | 0.97 | 0.96 | 0.98 | 0.82 |
| O+1 | LVR | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| I-LVR | LVR | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

*__From:__ O = Object Tokens, O+$n$ = O + $n$ Buffer,
I-(O+1) = All visual tokens except O+1, I-LVR = All visual tokens except last row
†__To:__ LTP = Last Token Position, LVR = Last Visual Token Row

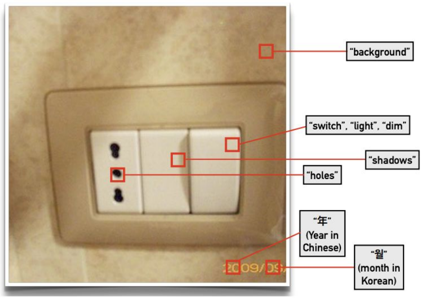
# Localization



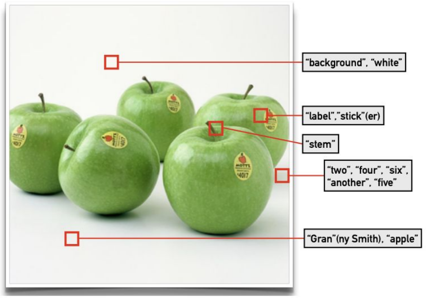

(a) An image of a lady in the sweater. The logit lens identifies tokens that correspond to specific detail of the sweater, such as "pattern" and "diam"(ond).




(b) An image of a child in a go-kart. The representations sometimes encode specific details, such as "look" and "gaz"(e) instead of just "face".

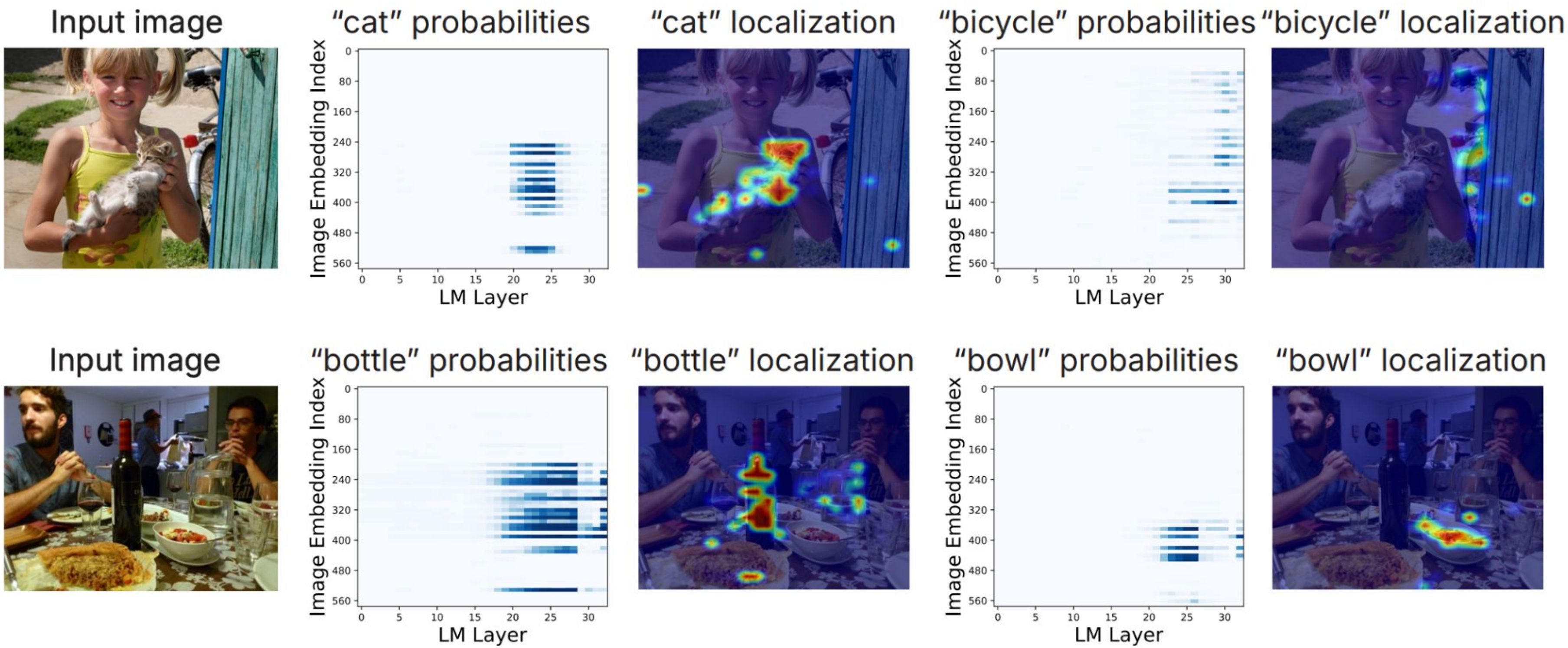



(c) An image of a switch. In the intermediate layers, the year and month tokens are encoded in non-English characters.

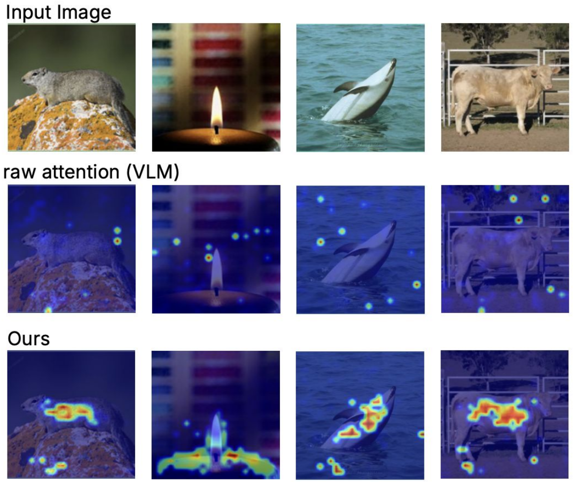



(d) An image of a bunch of apples. Global features, like count, show up in background tokens, though this may be an artifact of the LM processing.

# More Localization

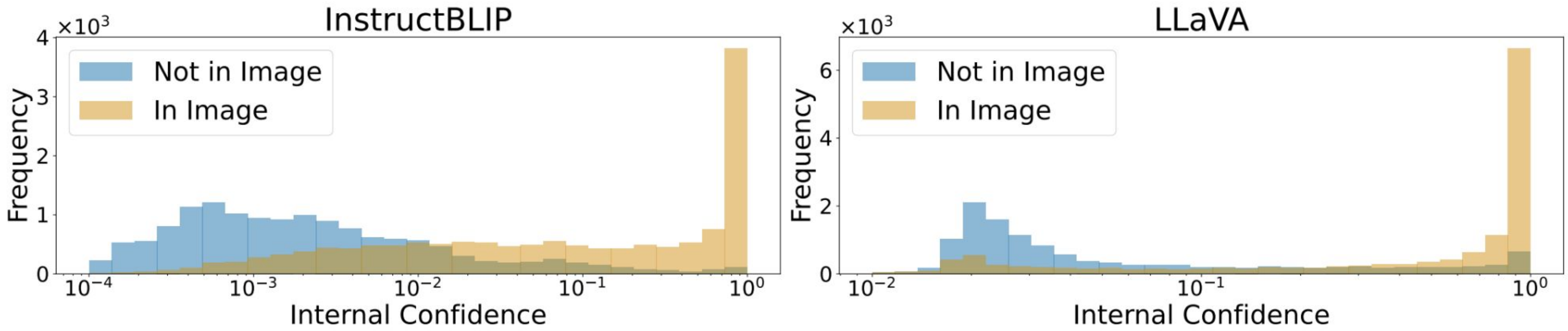


# Localization Performance



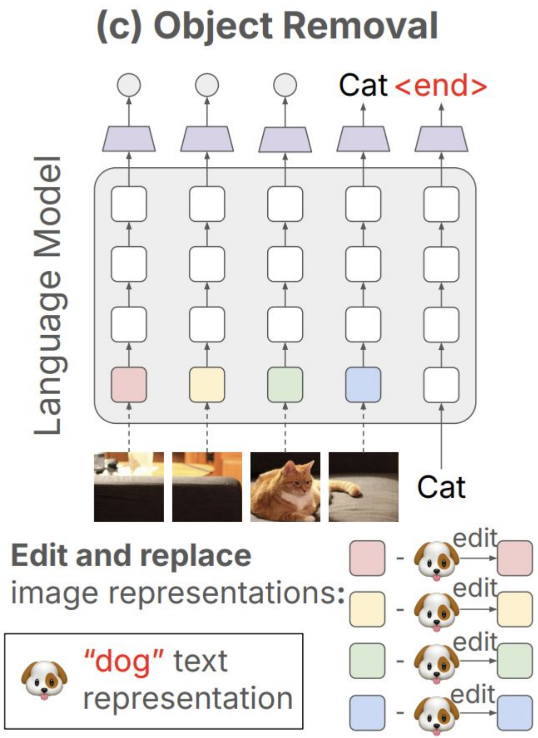

| Model | Method | Pixel Acc. ↑ | mIoU ↑ | mAP ↑ |
|---|---|---|---|---|
| raw attention (CLIP) | Image Encoder | 69.81 | 45.19 | 77.30 |
| TextSpan (Gandelsman et al., 2024b) | Image Encoder | 75.57 | 53.60 | **80.22** |
| raw attention (VLM) | VLM | 67.28 | 39.27 | 73.96 |
| Ours | VLM | **76.16** | **54.26** | 79.90 |

# Hallucination Detection




| Method | InstructBLIP mAP ↑ | InstructBLIP ROC AUC ↑ | LLaVA mAP ↑ | LLaVA ROC AUC ↑ |
|---|---|---|---|---|
| Baseline | 0.53 | 0.55 | 0.49 | 0.47 |
| Ours | 0.78 | 0.83 | 0.60 | 0.68 |

# Linearly Editing Visual Tokens (Hallucination Removal)




| Model | Method | CHAIR$_i$ ↓ | CHAIR$_s$ ↓ | Hallucinated Objects ↓ |
|---|---|---|---|---|
| InstructBLIP | Greedy | 57.0 | 23.3 | 512 |
| | Nucleus | 58.0 | 24.0 | 508 |
| | Beam Search | 53.4 | 14.6 | 564 |
| | OPERA | 45.6 | 13.9 | 472 |
| | Ours | **43.8** | **12.5** | **419** |
| LLaVA | Greedy | 49.2 | 14.2 | 532 |
| | Nucleus | 55.8 | 17.1 | 618 |
| | Beam Search | 52.4 | 15.0 | 583 |
| | OPERA | 44.8 | 12.8 | 462 |
| | Ours | **42.0** | **12.2** | **444** |

# Qualitative Examples


**Before:** In the image, a group of four young **men** are gathered around a picnic table, posing for a photo. They are holding **frisbees** in their hands and seem to be having a good time together. The picnic **table** is located in the middle of the scene, providing a central point for the group's gathering. There are several **chairs** placed around the table, indicating that the group may have been enjoying a meal or socializing while playing with the frisbees. A **car** can be seen parked in the background, possibly belonging to one of the group members.

**After:** A group of four young **men** are gathered around a picnic **table**, posing for a photo while holding frisbees. They seem to be enjoying their time together, possibly playing a friendly game of **frisbee**. The picnic table is located in the middle of the scene, providing a convenient spot for the group to gather and play. In addition to the frisbees, there are several **backpacks** scattered around the area, suggesting that the group may have come prepared for an outdoor adventure.

**Before:** The image features a delicious slice of **pizza** sitting on a white paper plate. The pizza is topped with various ingredients, including pepperoni, olives, and mushrooms, making it a mouth-watering dish. The plate is placed on a **table**, adding to the overall presentation of the pizza. In addition to the pizza slice, there are several other items on the table, including two **cups**, a **fork**, a **knife**, and a **spoon**. These items add to the overall atmosphere of the scene, creating a cozy and inviting setting for enjoying the pizza.

**After:** The image features a delicious slice of **pizza** on a paper plate, placed on a **dining table**. The pizza is topped with various ingredients, including pepperoni, olives, and mushrooms, making it a mouth-watering meal. The pizza slice dominates the scene, taking up most of the space on the paper plate. The plate is placed on top of a dining table, which can be seen in the background.

**Before:** The image features a black and white **cat** sitting in the middle of a cluttered living room floor. The cat is surrounded by various items such as clothes, bags, **books**, and a **backpack**. There are two **chairs** placed near the cat, one on the left side and the other on the right side of the room. In addition to the cat, there are two people visible in the scene. One **person** is standing on the left side of the room, while the other person is located on the right side of the room. Both individuals seem to be engrossed in their own activities, possibly unaware of the cat's presence.

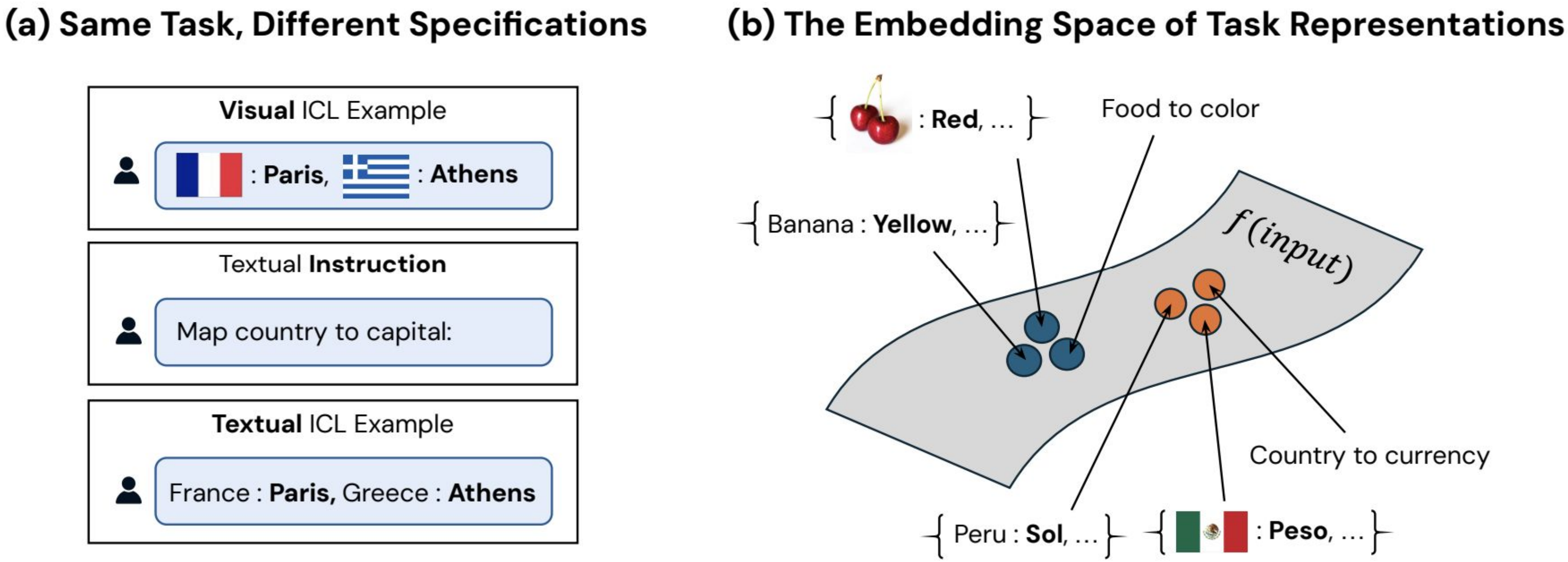
**After**: The image depicts a black and white **cat** sitting in the middle of a cluttered room. The cat is surrounded by a variety of items, including **suitcases**, **backpacks**, clothes, and shoes. There are at least three suitcases scattered around the room, with one located closer to the cat and the other two further away. A backpack can be seen on the left side of the room, and a pair of shoes can be spotted on the right side. In addition to these items, there are several clothes spread out on the floor, including a shirt, a jacket, and a pair of pants. The cluttered environment suggests that the room may have been recently used for packing or preparing for a trip.

# Task Vectors are Cross-Modal

# Task Vectors are Cross-Modal

**Anonymous authors**
Paper under double-blind review

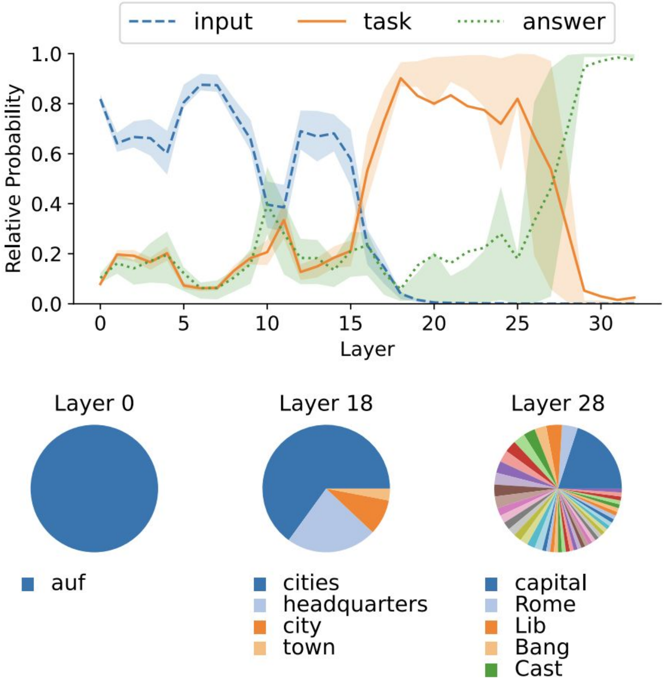
# Task Representations



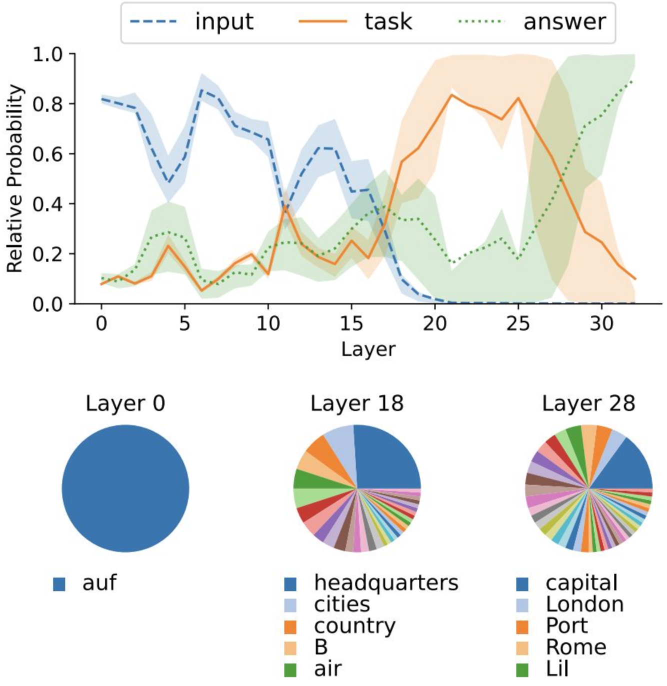
# Cross-Modal Tasks

| Task | Instruction | Text ICL Example | Image ICL Example |
|---|---|---|---|
| Country-Capital | *The capital city of the country:* | {Greece : **Athens**} | {  : **Athens**} |
| Country-Currency | *The last word of the official currency of the country:* | {Italy : **Euro**} | {  : **Euro**} |
| Animal-Latin | *The scientific name of the animal's species in latin:* | {Gray Wolf : **Canis lupus**} | {  : **Canis lupus**} |
| Animal-Young | *The term for the baby of the animal:* | {Common Dolphin : **calf**} | {  : **calf**} |
| Food-Color | *The color of the food:* | {Persimmon : **orange**} | {  : **orange**} |
| Food-Flavor | *The flavor descriptor of the food:* | {Strawberry : **sweet**} | {  : **sweet**} |

# Evolution of Layer Outputs




(a) Text ICL



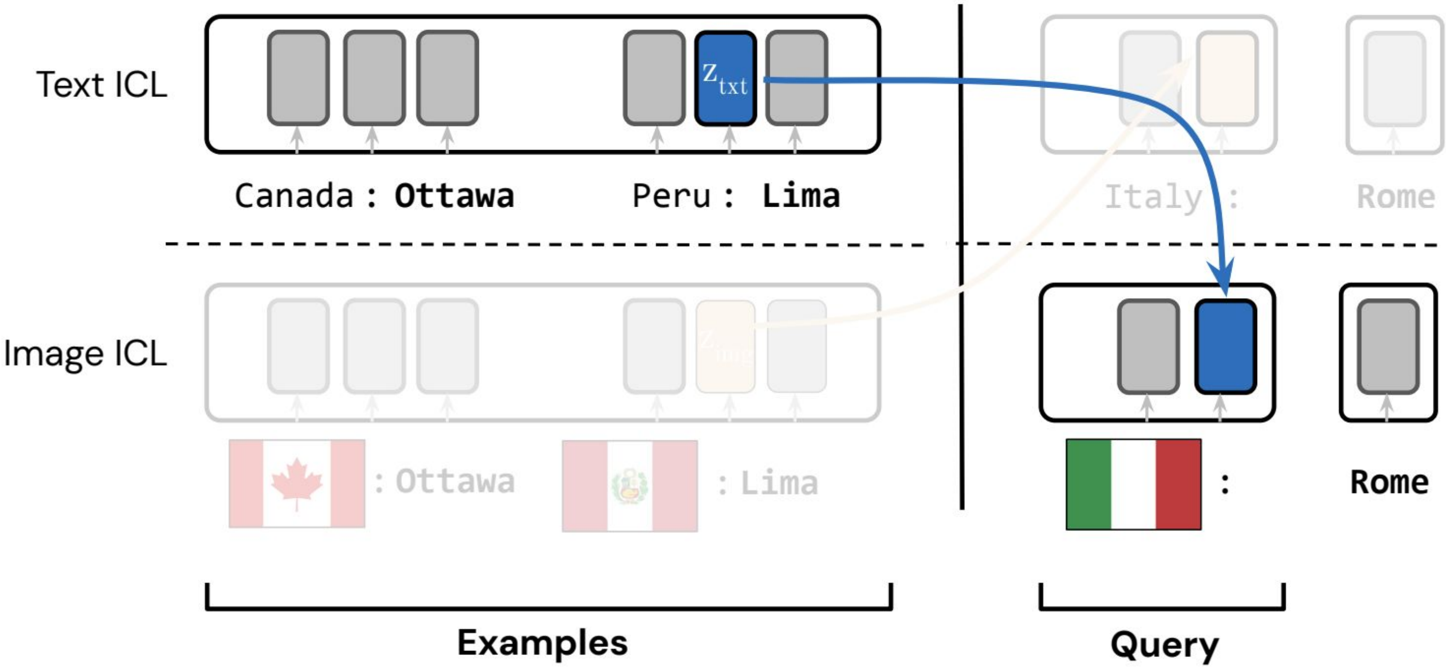

(b) Image ICL

# Logit Lens on Task Representation

| Task | Text ICL | Image ICL |
|---|---|---|
| Country-Capital | *headquarters*, *cities*, *city*, *cidade*, *centro* | *headquarters*, *administr*, *cities*, *city*, ◊ |
| Country-Currency | *currency*, *currency*, *dollar*, *dollars*, *Currency* | *currency*, ◊, *currency*, *undefined*, *dollars* |
| Animal-Latin | *species*, *genus*, *habitat*, *mamm*, *american* | *species*, *genus*, *mamm*, *spec*, *creature* |
| Animal-Young | *pup*, *babies*, *baby*, *called*, *young* | *young*, *species*, *scriptstyle*, *animal*, *teenager* |
| Food-Color | *yellow*, *pink*, *green*, *purple*, *orange* | *green*, *yes*, *yellow*, *verd*, *yes* |
| Food-Flavor | *flavor*, *taste*, *mild*, *flav*, *tastes* | *yes*, *none*, *anger*, *cerca*, *vegetables* |

Decodes task summaries!

# Cross-Modal Transfer



# Approaches Compared

Using an image query:

Image ICL Base: Provide image examples in context.

Image ICL Patch: Provide the image task vector (derived from images in context).

Text ICL xBase: Provide text examples in context.

Text ICL xPatch: Provide text task vector (derived from text in context).

# Cross-Modal Transfer Results

| Model | Country-Capital | Country-Currency | Animal-Latin | Animal-Young | Food-Color | Food-Flavor | Avg. |
|---|---|---|---|---|---|---|---|
| Random | 0.00 | 0.12 | 0.00 | 0.18 | 0.24 | 0.31 | 0.14 |
| **LLaVA-v1.5** | | | | | | | |
| No Context | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Image ICL Base | - | - | - | - | - | - | - |
| Image ICL Patch | - | - | - | - | - | - | - |
| Text ICL xBase | 0.02 | 0.18 | 0.03 | <u>0.23</u> | 0.28 | <u>0.37</u> | 0.18 |
| Text ICL xPatch | <u>0.31</u> | <u>0.30</u> | <u>0.26</u> | 0.18 | <u>0.53</u> | 0.31 | **0.32** |
| **Mantis-Fuyu** | | | | | | | |
| No Context | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Image ICL Base | 0.11 | 0.13 | 0.24 | 0.05 | 0.34 | 0.23 | 0.18 |
| Image ICL Patch | 0.17 | 0.03 | 0.16 | 0.05 | 0.50 | 0.31 | 0.20 |
| Text ICL xBase | 0.09 | 0.06 | 0.08 | 0.02 | 0.23 | 0.04 | 0.09 |
| Text ICL xPatch | <u>0.32</u> | <u>0.23</u> | <u>0.36</u> | <u>0.09</u> | <u>0.51</u> | <u>0.36</u> | **0.31** |
| **Idefics2** | | | | | | | |
| No Context | 0.03 | 0.00 | 0.03 | 0.00 | 0.01 | 0.01 | 0.01 |
| Image ICL Base | <u>0.71</u> | <u>0.57</u> | 0.43 | 0.12 | 0.41 | 0.35 | 0.43 |
| Image ICL Patch | 0.58 | 0.32 | 0.40 | 0.03 | 0.39 | 0.17 | 0.31 |
| Text ICL xBase | 0.11 | 0.03 | 0.41 | 0.13 | 0.21 | 0.18 | 0.18 |
| Text ICL xPatch | 0.61 | 0.40 | <u>0.48</u> | <u>0.62</u> | <u>0.53</u> | <u>0.39</u> | **0.51** |

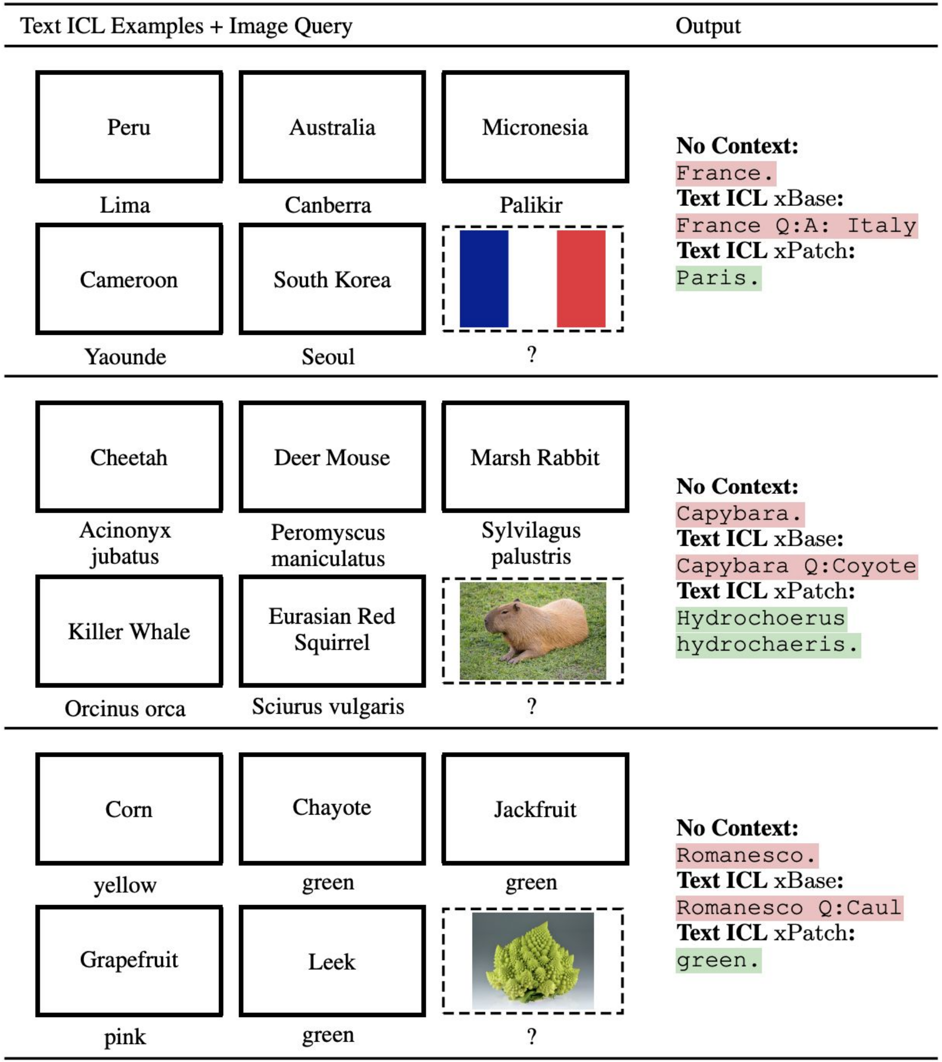
For image queries, patching cross-modal task vectors (Text ICL xPatch) outperforms text ICL in the same context window (Text ICL xBase) and the strong unimodal image ICL baseline (Image ICL Base, Patch).

# Qualitative Examples

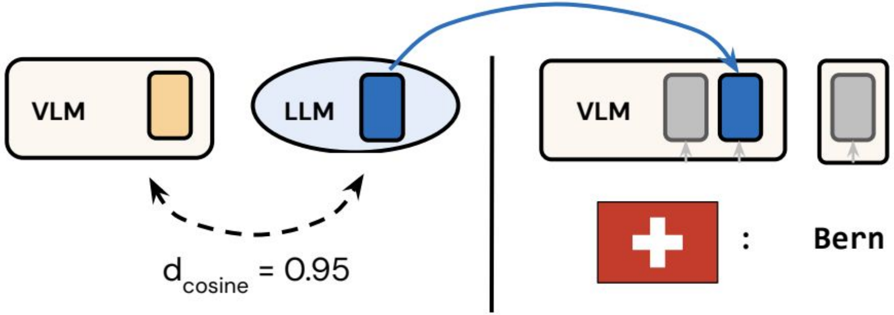
Authors hypothesize that image ICL requires an additional visual recognition step to understand the task compared with text ICL, which may lead to noisier task representations.

We know from M-ICL paper that text is more important than images for multimodal ICL.

Could this contribute to why Text xPatch outperforms Image ICL?



# Inter-Model Transfer (LLM to VLM)




Patching text task vectors from the LLM: even more improvement!

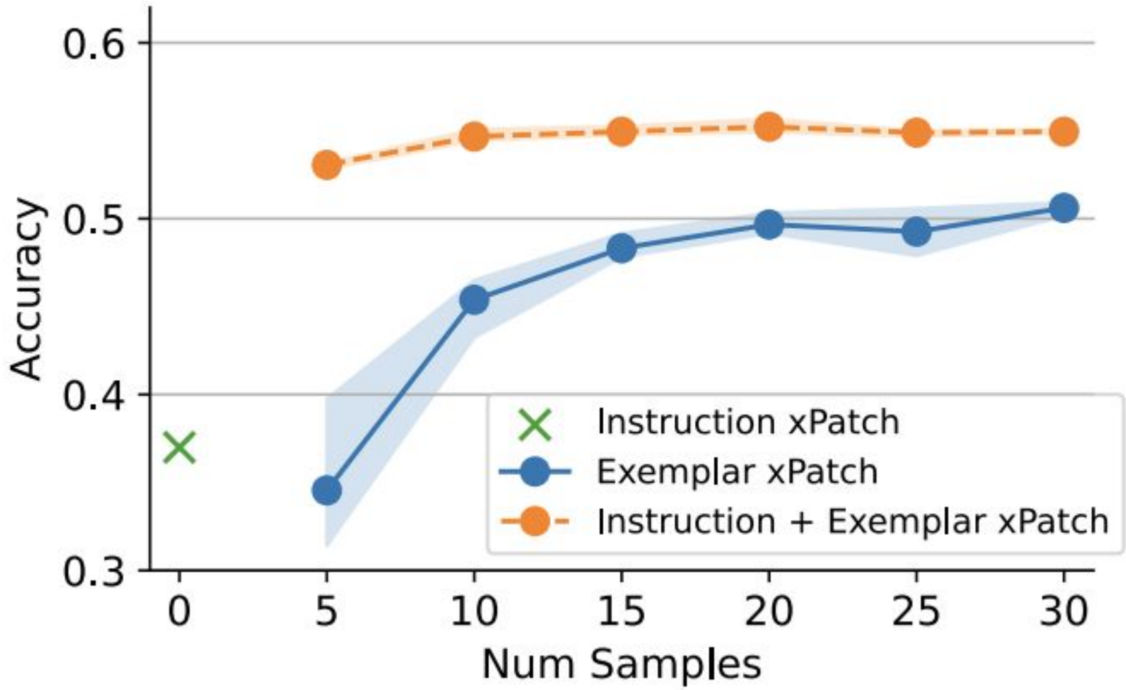
| Model | Cosine Sim. | Avg. |
|---|---|---|
| Random | 0.58 | 0.14 |
| **LLaVA-v1.5** | | |
| VLM-VLM xPatch | - | 0.32 |
| LLM-VLM xPatch | 0.95 | **0.37** |
| **Idefics2** | | |
| VLM-VLM xPatch | - | 0.51 |
| LLM-VLM xPatch | 0.89 | **0.52** |

# Instruction Vectors

| Instruction | Image Query | Output |
|---|---|---|
| The term for the baby of the animal: |  | **No Context:** `A kangaroo.` **Instruction xPatch:** `joey.` |
| The scientific name of the animal's species in latin: |  | **No Context:** `Elephant.` **Instruction xPatch:** `Elephas maximus.` |

Figure 7: **Instruction Vectors.** Task vectors can also be defined via brief instructions and patched onto image queries (Instruction xPatch).

# Instruction + Example Vectors (Averaged)



# Task **Conflict**

This mirrors a practical challenge where the user may prompt for a task that conflicts with the global system instruction

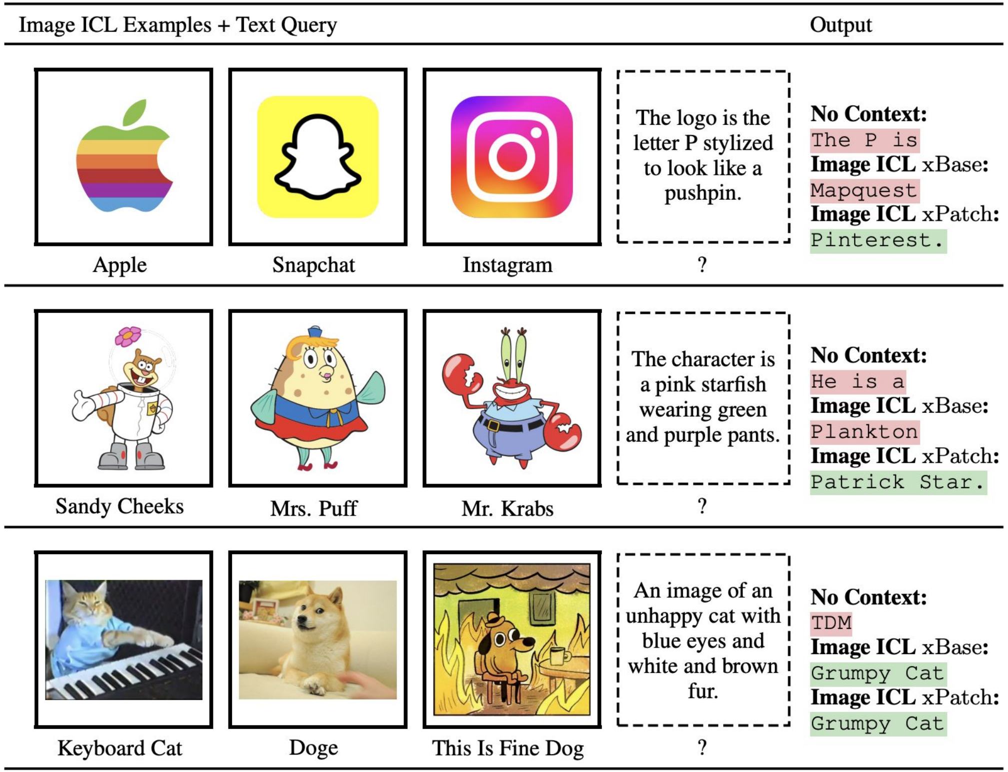
Global vector patching is able to **override** local prompting in many cases, fails sometimes when task vector is more complicated than local prompt.

| Instruct. xBase | | Instruct. xPatch | Image Query | Output |
|---|---|---|---|---|
| What is on top of the meat | vs. | What is the green vegetable |  | **Instruction** xBase:<br>`Sauce.`<br>**+ Instruction** xPatch:<br>`broccoli` |
| What color are the letters | vs. | What does the sign say |   | **Instruction** xBase:<br>`Black. What`<br>**+ Instruction** xPatch:<br>`Street car crossing be alert` |
| What color is the van | vs. | Who is the manufacturer of this van |  | **Instruction** xBase:<br>`It is blue.`<br>**+ Instruction** xPatch:<br>`blue and white.` |
| Write something very mean | vs. | Write something nice |  | **Instruction** xBase:<br>`Get off the leaves you little b******.`<br>**+ Instruction** xPatch:<br>`A dog is in a pile of leaves and it is adorable.` |

# Patching Text Queries with Image Task Vectors

Mixing modalities in context does not perform well, but patching with Image task vector does.

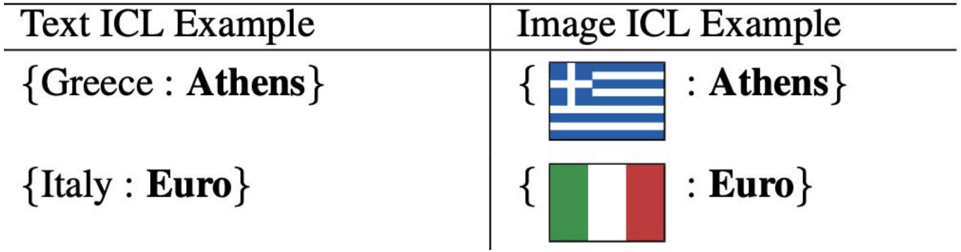
However, image ICL generally doesn't outperform the strong text ICL.



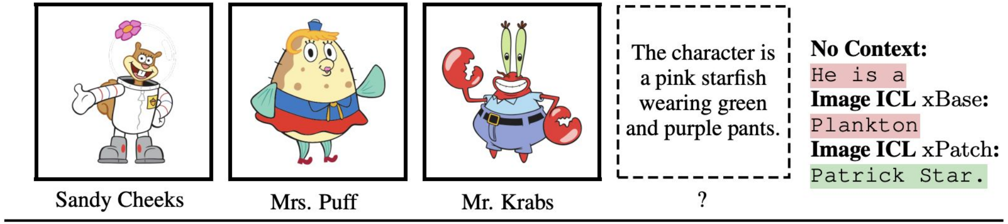
# Image Task Vectors vs. Text Task Vectors

Image Task Vectors < Text Task vectors in most cases.

Why? “Image ICL also has to complete an implicit recognition task mapping the image to the underlying textual concept. For example, if the model cannot match the flag to the correct country name, it will not be able to predict the correct currency.”




“However, if recognition is instead required in text space, image ICL may better encode the task.” Because describing “Patrick Star” is a very visual process.



# Aside: Platonic Representation Hypothesis

**The Platonic Representation Hypothesis**

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.

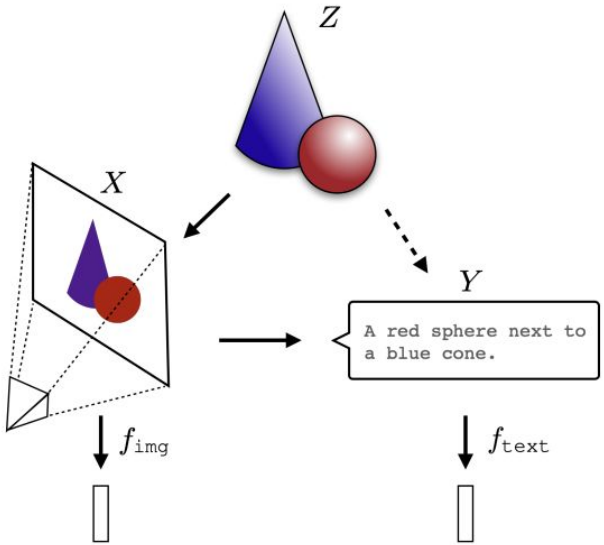



*Figure 1.* **The Platonic Representation Hypothesis:** Images ($X$) and text ($Y$) are projections of a common underlying reality ($Z$). We conjecture that representation learning algorithms will converge on a shared representation of $Z$, and scaling model size, as well as data and task diversity, drives this convergence.

# Discussion

(Patrick) Interpreting VLMs: The paper only focuses on LLaVA-based models, which directly concatenates visual and text tokens for LLM. However, there are several models like Flamingo that use cross-attention in the downstream model. I'm personally not convinced by the result as LLaVA is only a branch of methods.

(Zeeshan) This paper mostly focuses on object identification tasks. How might the results change for more complex visual reasoning tasks that require understanding relationships between multiple objects or abstract concepts in images?