Linear Mechanisms for Spatiotemporal Reasoning in Vision Language Models

Raphi Kang*, Hongqiao Chen*, Georgia Gkioxari, Pietro Perona
California Institute of Technology
(ICLR 2026)

Caltech

ICLR

arXiv

GitHub

Linear Mechanisms for Spatiotemporal Reasoning in Vision Language Models

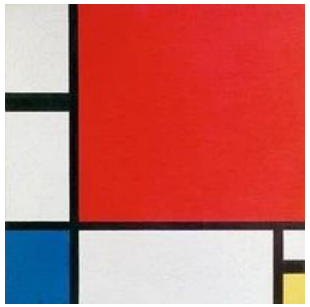
Raphi Kang*, Hongqiao Chen*, Georgia Gkioxari, Pietro Perona
California Institute of Technology
(ICLR 2026)

Caltech

ICLR

About Me:

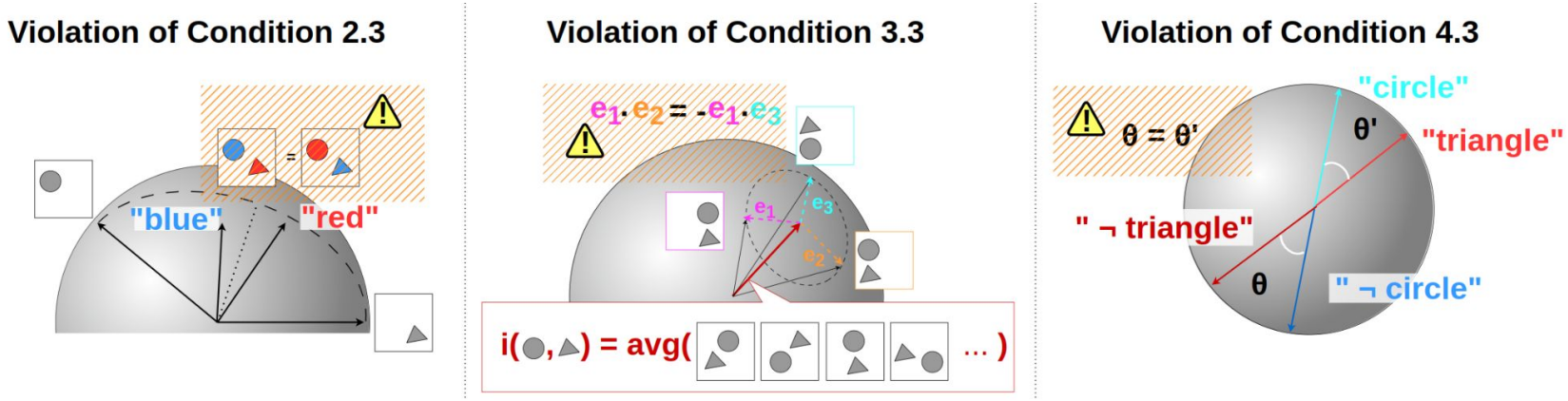
Caltech



@ Vision Lab:

How do the geometries of VLM latent spaces enable or inhibit visual reasoning?

Previous work: we showed that dual-encoder VLMs cannot perform spatial reasoning due to latent geometry limitations.



Is CLIP ideal? No. Can we fix it? Yes!

Raphi Kang Yue Song Gerogia Gkioxari Pietro Perona
California Institute of Technology
Pasadena, CA
rkang@caltech.edu

Kang, Raphi, et al. "Is CLIP ideal? No. Can we fix it? Yes!." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025.

Observation: Autoregressive VLMs exhibit compositional reasoning capacities.

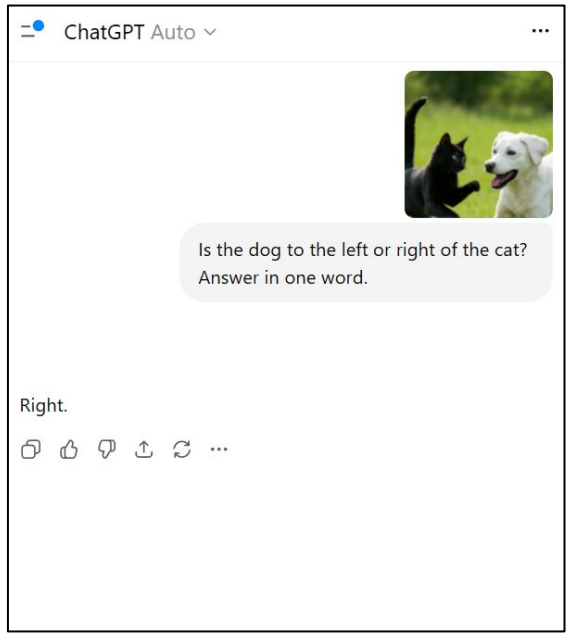
Observation: Image VLMs can perform simple *spatial reasoning*.

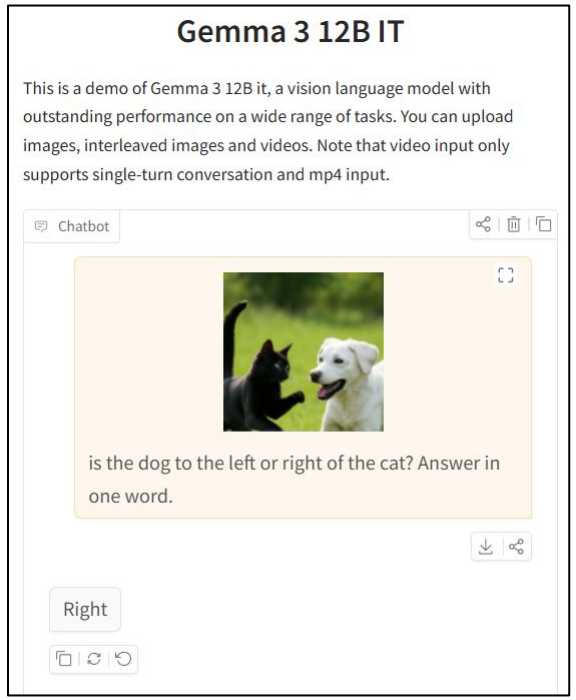
Given a prompt like:

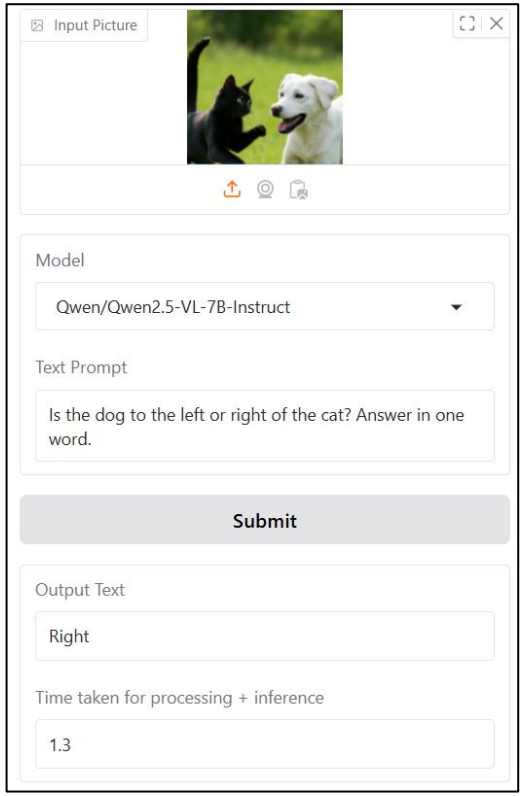
“Is the dog to the left or right of the cat?
Answer in one word.”

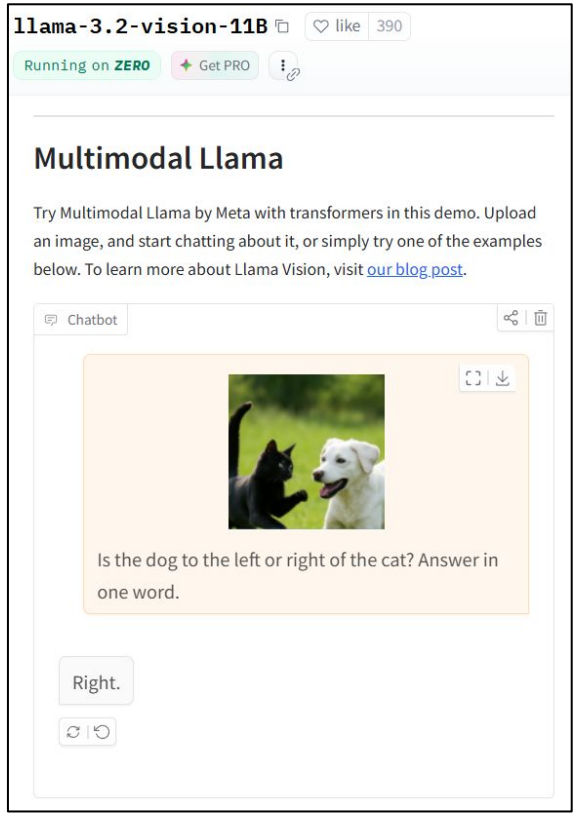
Observation: Image VLMs can perform simple *spatial reasoning*.

"Is the dog to the left or right of the cat? Answer in one word."









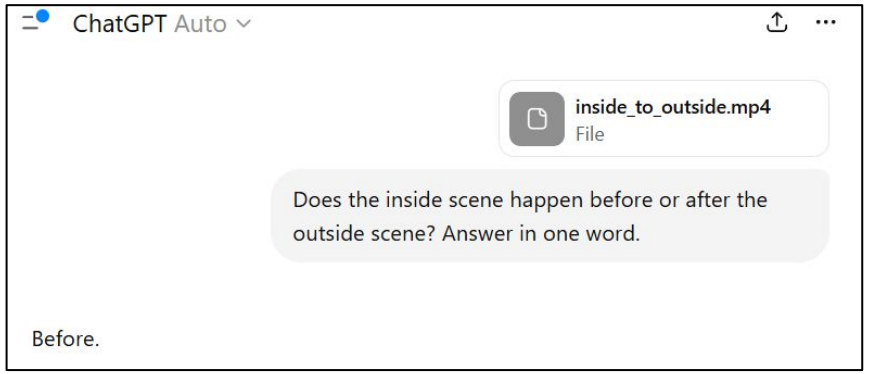
Observation: and Video VLMs can do *temporal reasoning*.

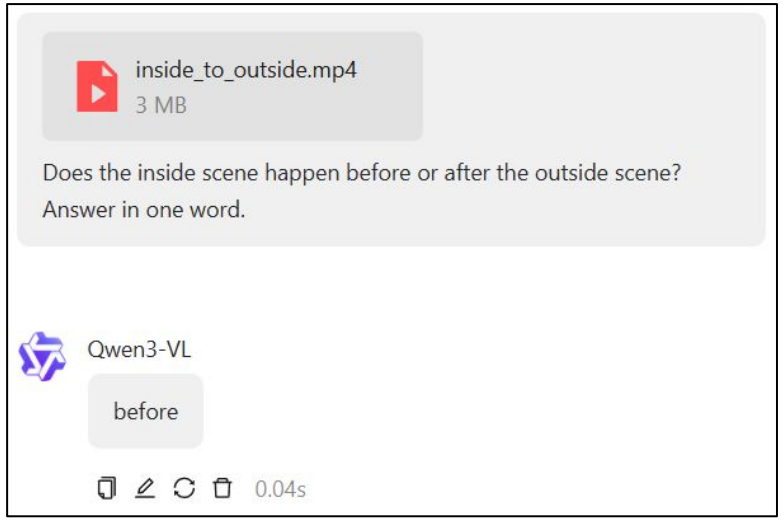
Given a prompt like:

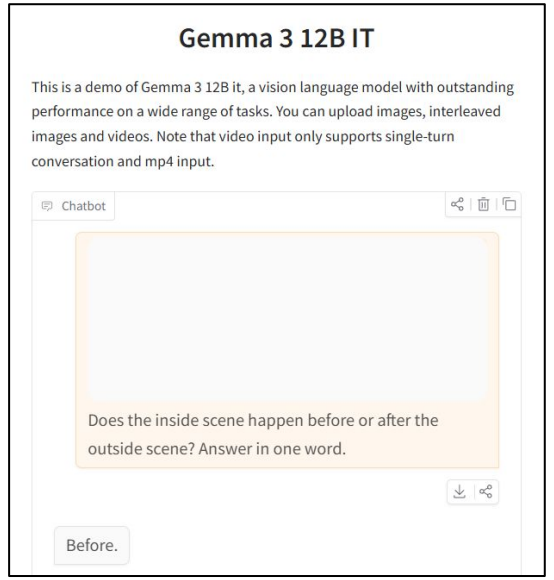
“Does the inside scene happen before or after the outside scene? Answer in one word.”

Observation: and Video VLMs can do *temporal reasoning*.

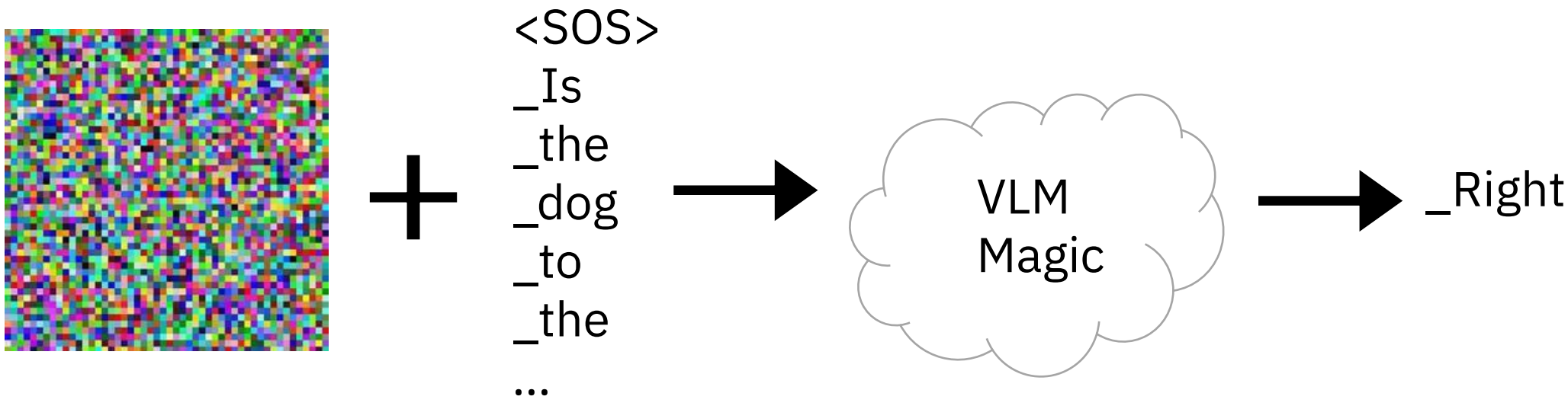
"Does the inside scene happen before or after the outside scene? Answer in one word."







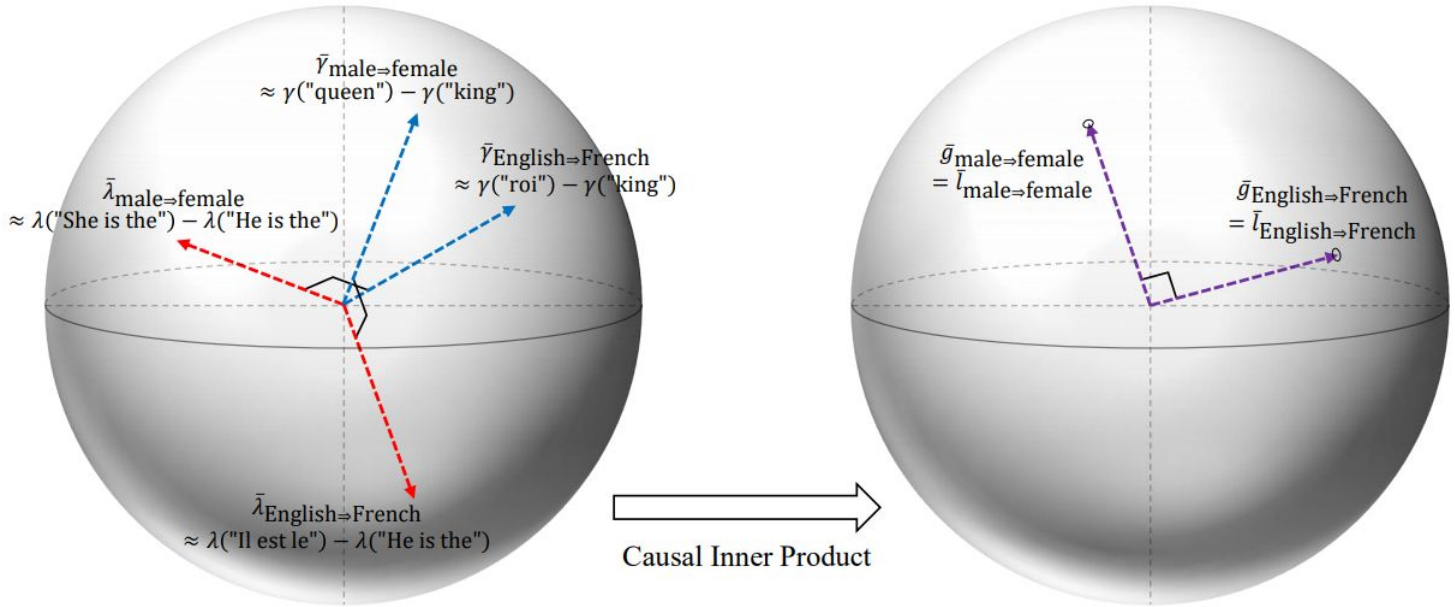
...So how do they do it?



...So how do they do it?

Hypothesis: There must be some *linear* mechanism!

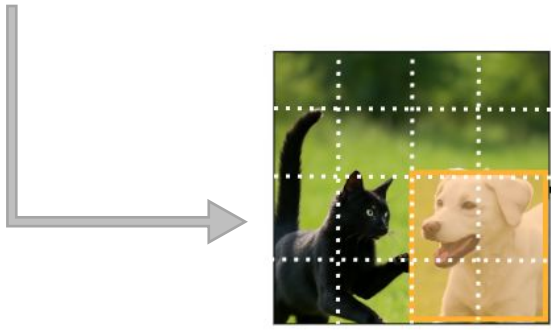
The Linear Representation Hypothesis in LLMs:



Park, Kiho, Yo Joong Choe, and Victor Veitch. "The linear representation hypothesis and the geometry of large language models." *Proceedings of the 41st International Conference on Machine Learning. 2024.*

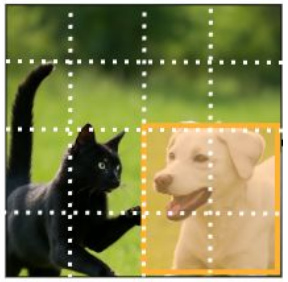
We want to characterize a linear mechanism for spatial reasoning in VLMs:

“Is the dog to the left or right of the cat? Answer in one word.”



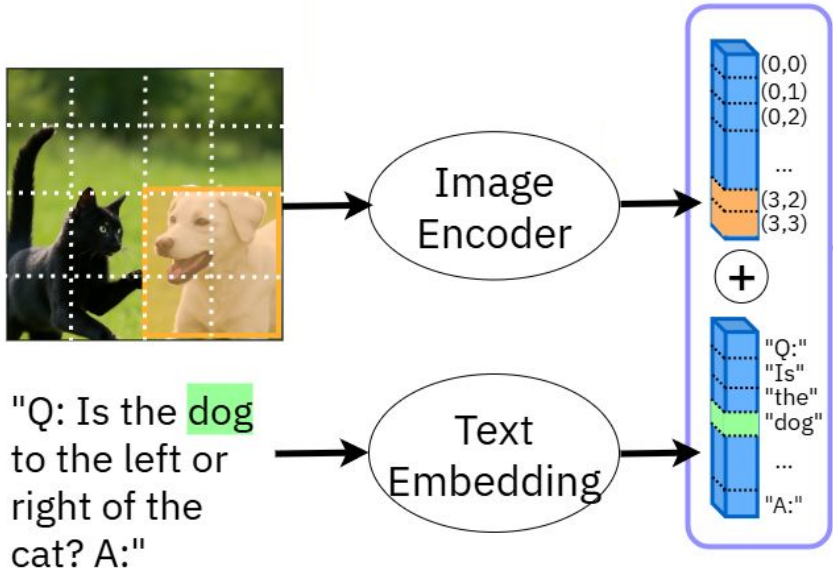
"Q: Is the dog to the left or right of the cat? A:"

We want to characterize a linear mechanism for spatial reasoning in VLMs:

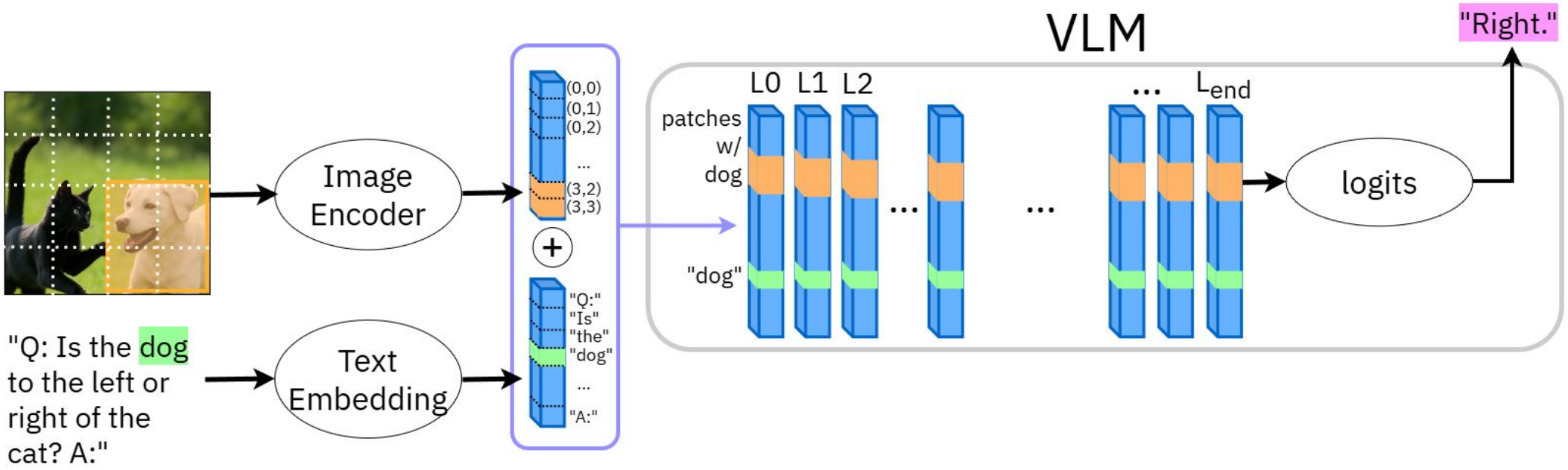


"Q: Is the dog to the left or right of the cat? A:"

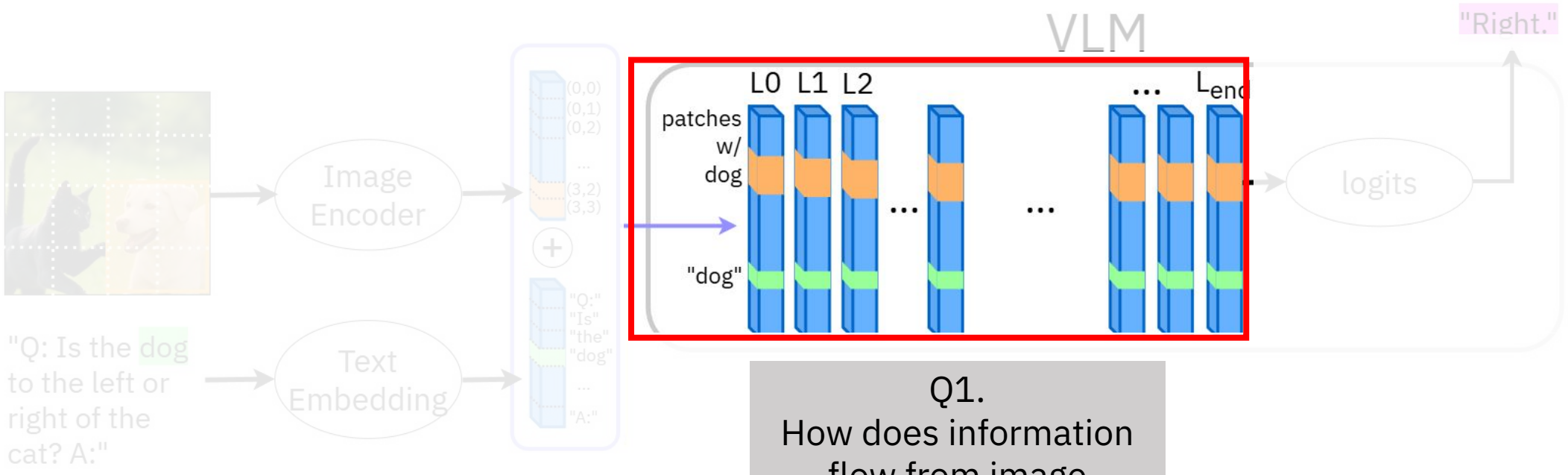
We want to characterize a linear mechanism for spatial reasoning in VLMs:



We want to characterize a linear mechanism for spatial reasoning in VLMs:

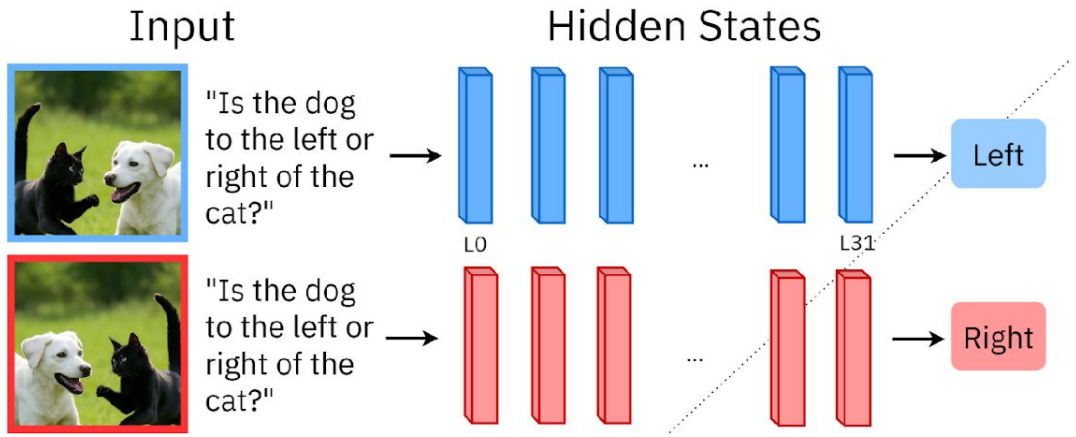


We want to characterize a linear mechanism for spatial reasoning in VLMs:

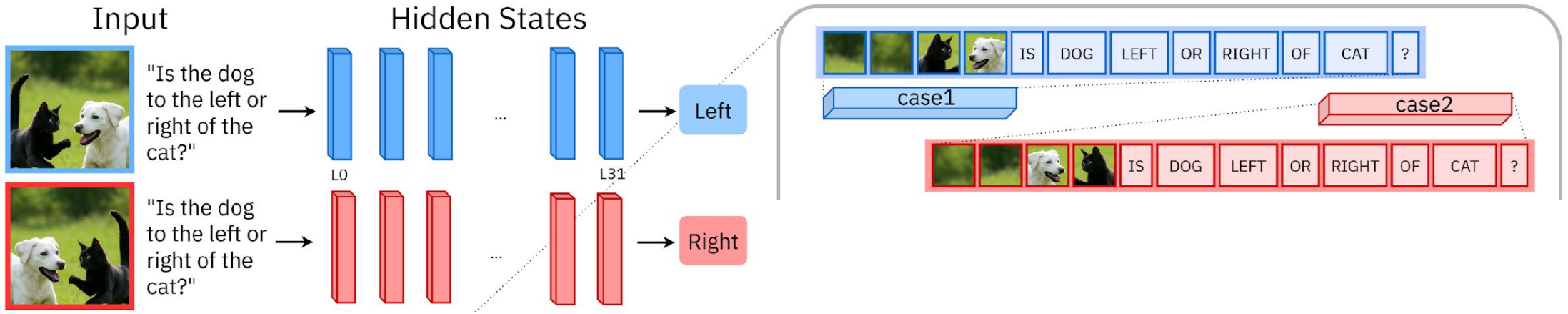


Q1.
How does information flow from image patches to text tokens?

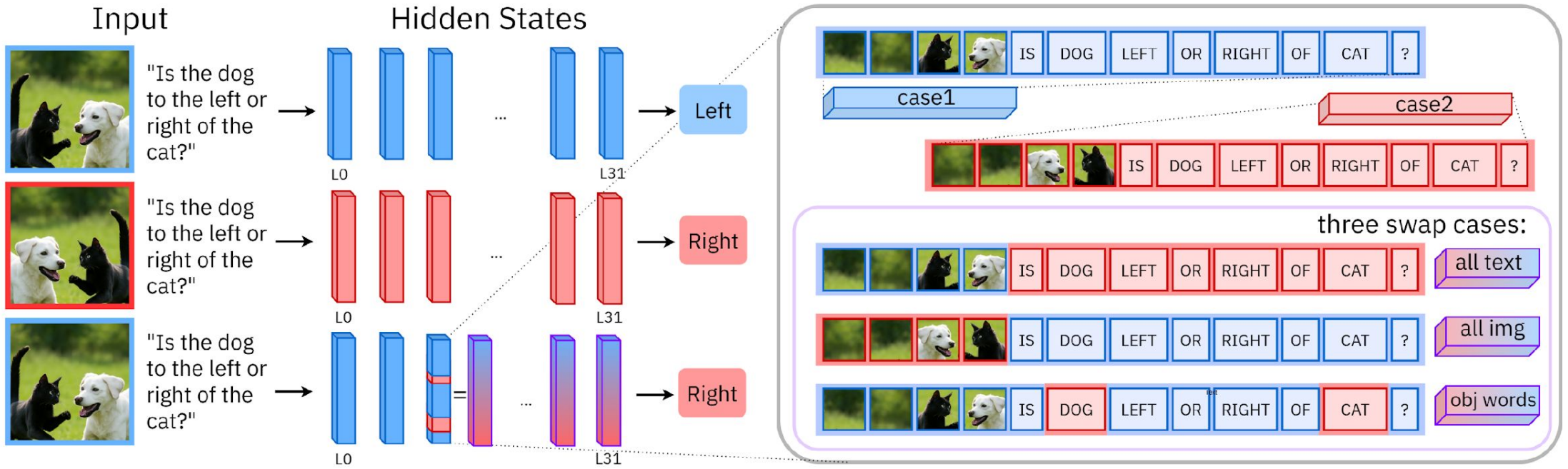
The “Mirror Swapping” Experiment:



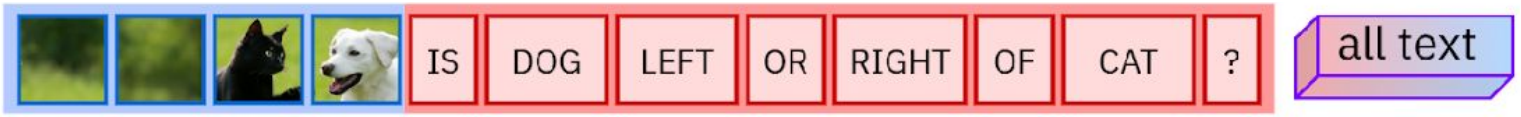
The “Mirror Swapping” Experiment:

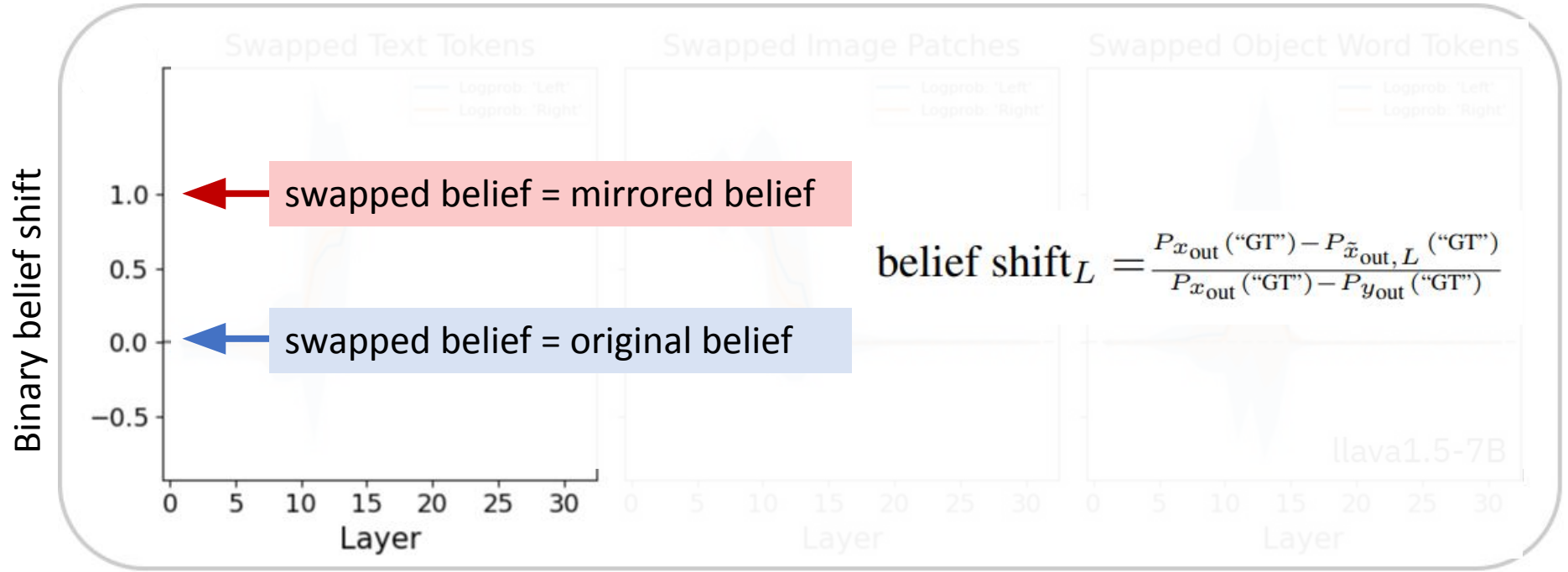


The “Mirror Swapping” Experiment:

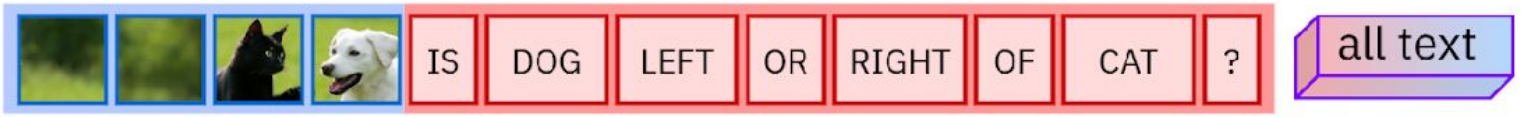


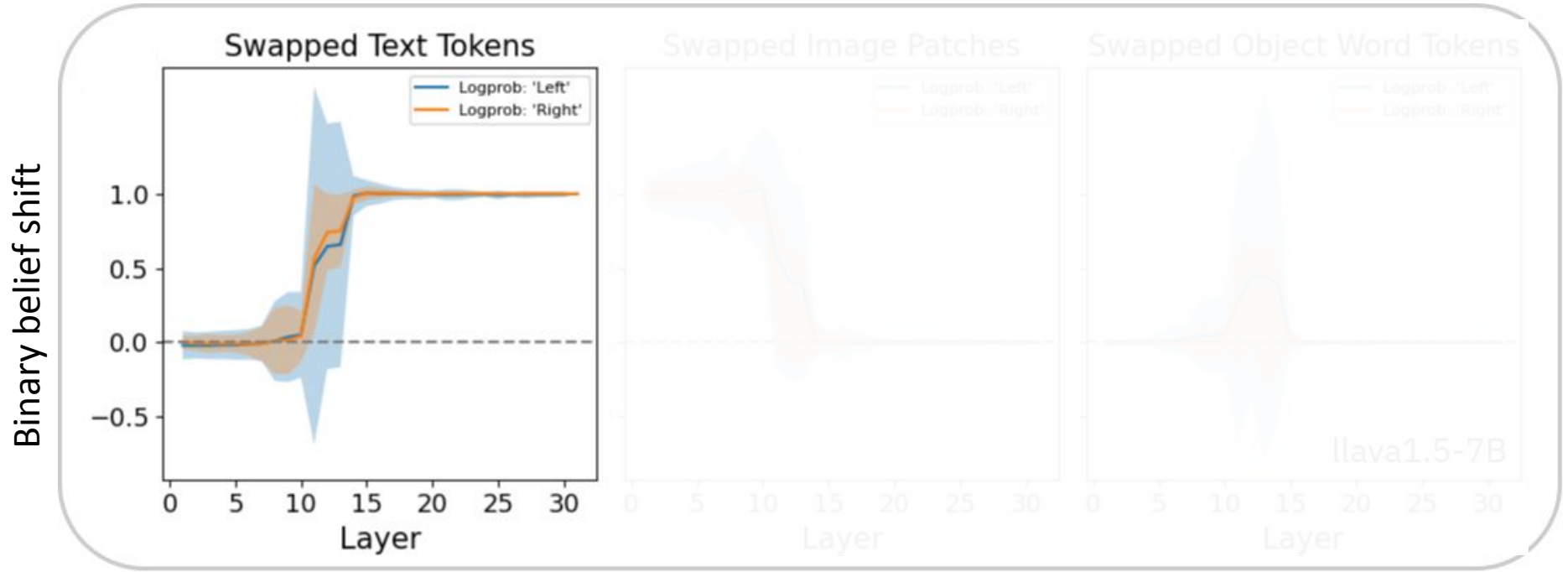
The “Mirror Swapping” Experiment:





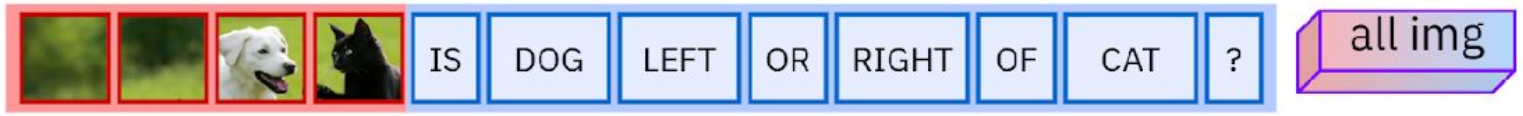
The “Mirror Swapping” Experiment:



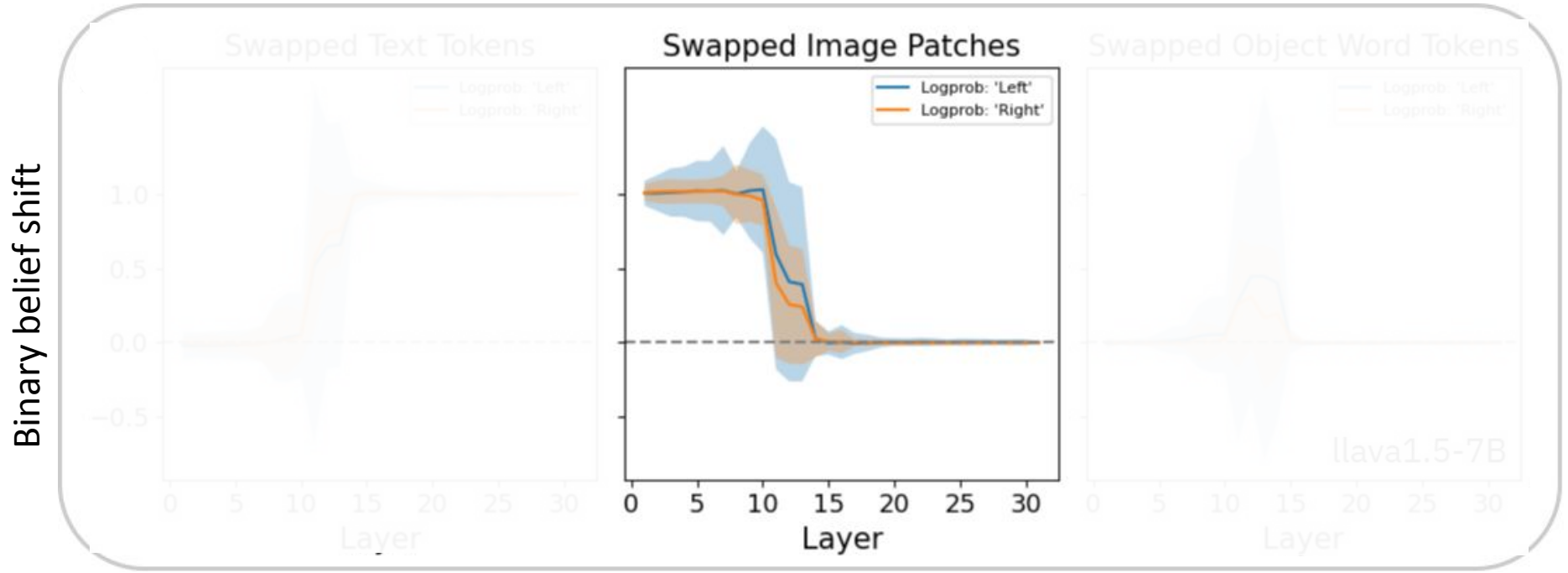


$$\text{belief shift}_L = \frac{P_{x_{\text{out}}}(\text{“GT”}) - P_{\tilde{x}_{\text{out},L}}(\text{“GT”})}{P_{x_{\text{out}}}(\text{“GT”}) - P_{y_{\text{out}}}(\text{“GT”})}$$

The “Mirror Swapping” Experiment:

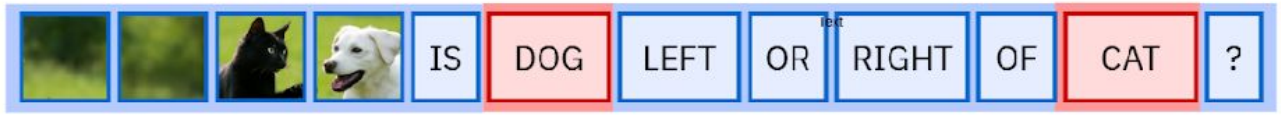


all img

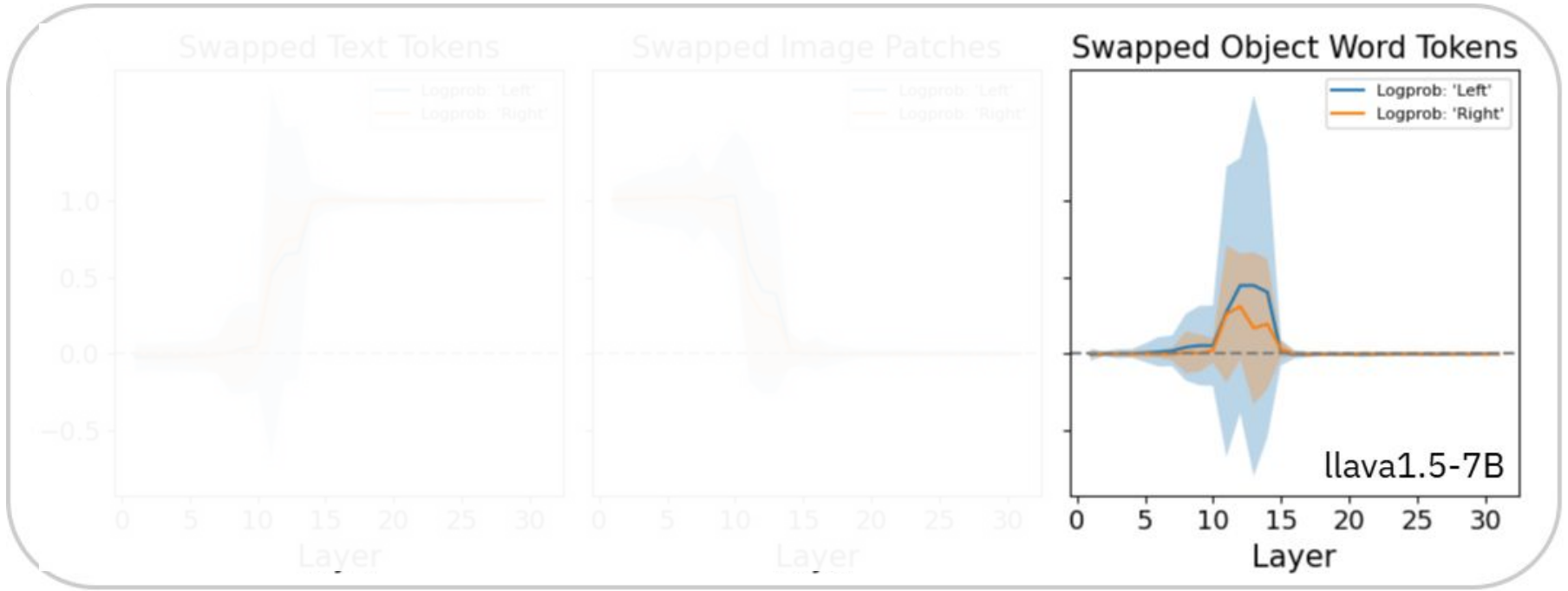


$$\text{belief shift}_L = \frac{P_{x_{\text{out}}}(\text{“GT”}) - P_{\tilde{x}_{\text{out},L}}(\text{“GT”})}{P_{x_{\text{out}}}(\text{“GT”}) - P_{y_{\text{out}}}(\text{“GT”})}$$

The “Mirror Swapping” Experiment:

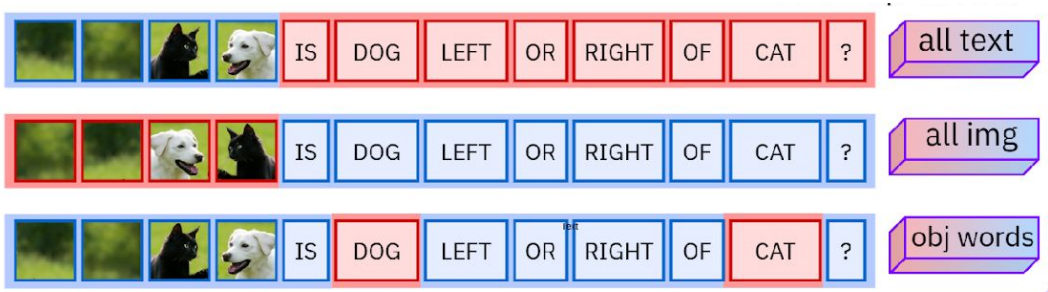


obj words

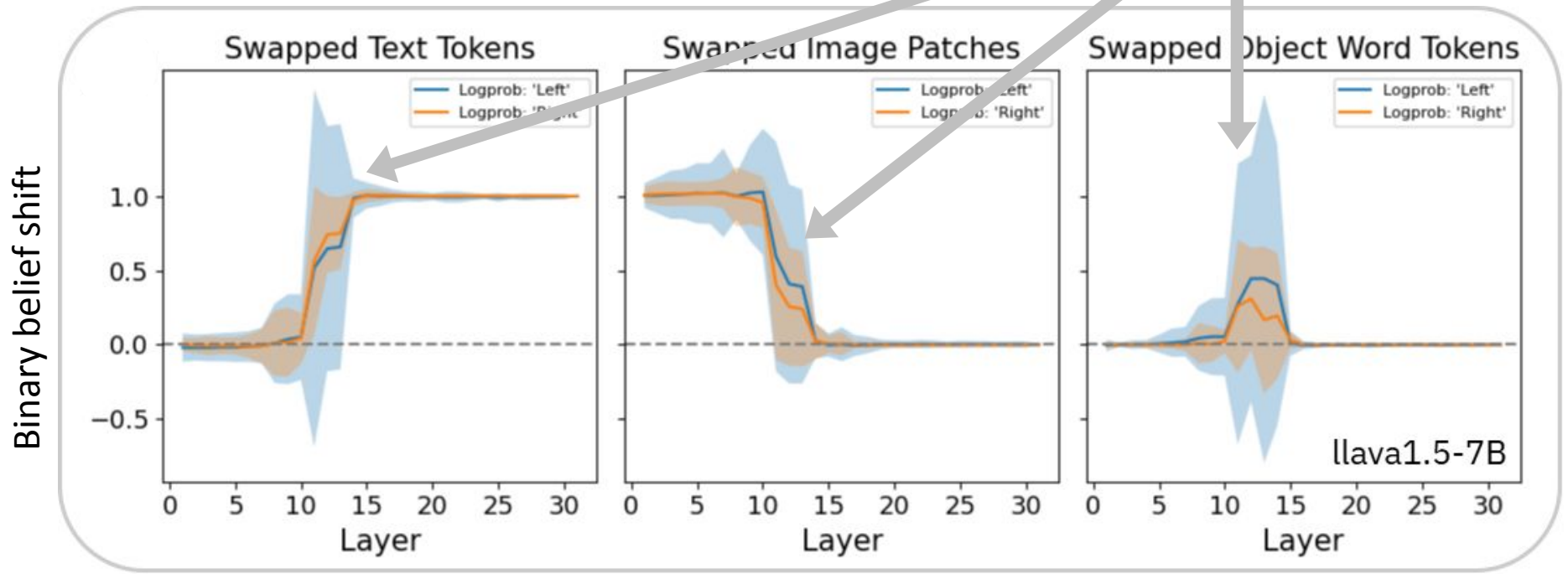


$$\text{belief shift}_L = \frac{P_{x_{\text{out}}}(\text{“GT”}) - P_{\tilde{x}_{\text{out},L}}(\text{“GT”})}{P_{x_{\text{out}}}(\text{“GT”}) - P_{y_{\text{out}}}(\text{“GT”})}$$

The “Mirror Swapping” Experiment:

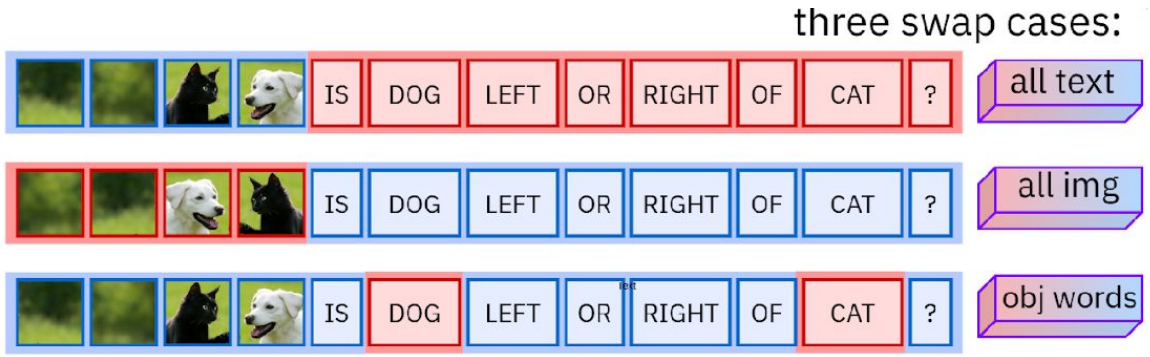


Modality Merging Point!

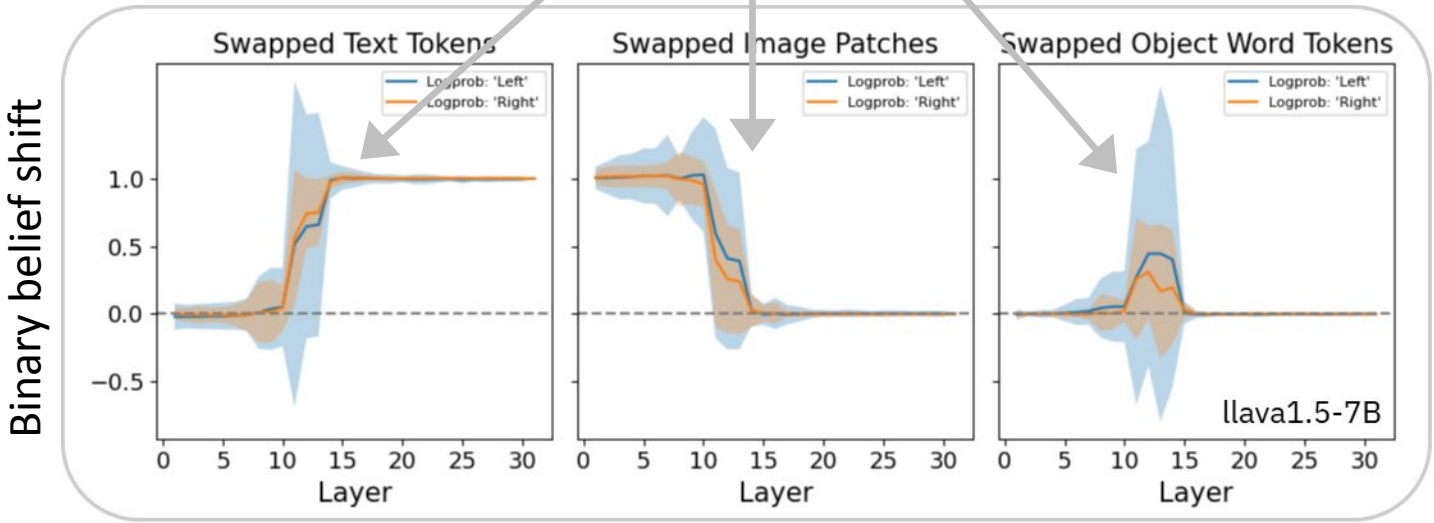


$$\text{belief shift}_L = \frac{P_{x_{\text{out}}}(\text{“GT”}) - P_{\tilde{x}_{\text{out},L}}(\text{“GT”})}{P_{x_{\text{out}}}(\text{“GT”}) - P_{y_{\text{out}}}(\text{“GT”})}$$

The “Mirror Swapping” Experiment:

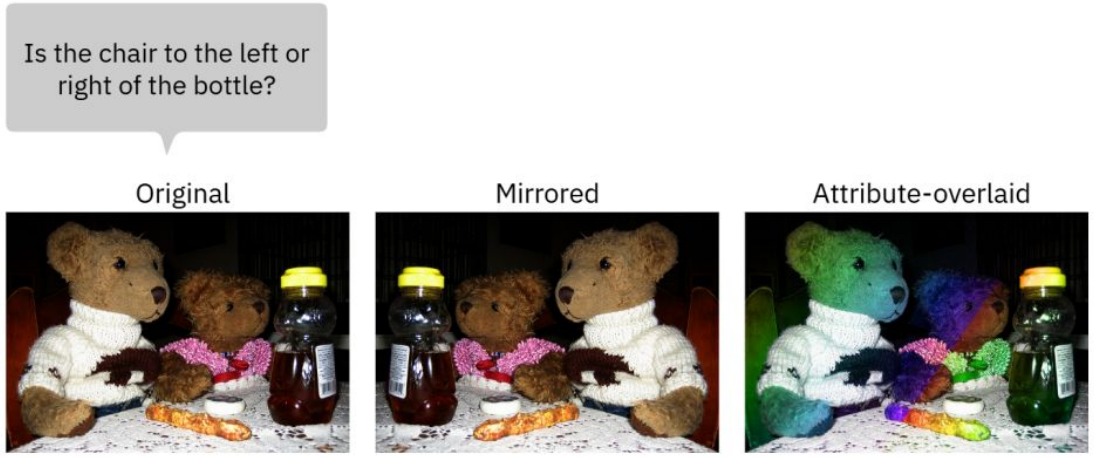


Modality Merging Point!

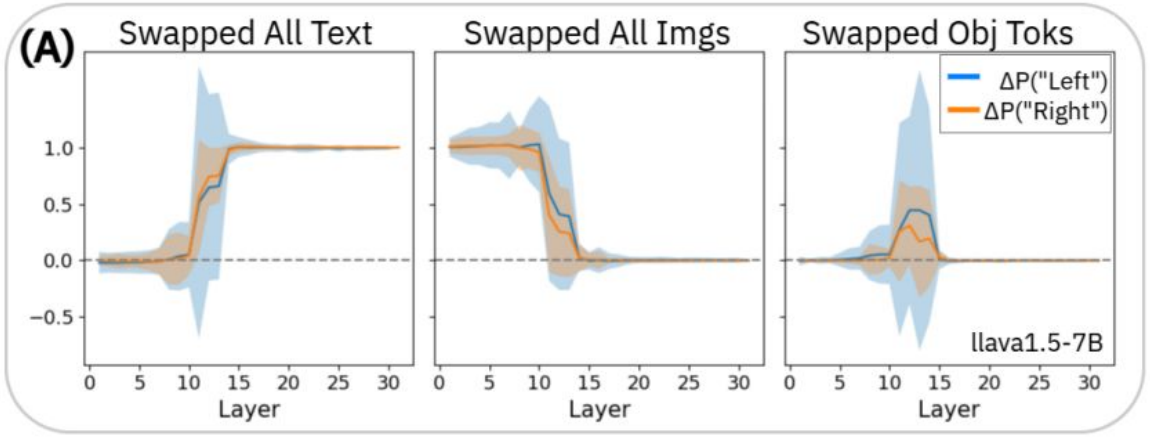


$$\text{belief shift}_L = \frac{P_{x_{\text{out}}}(\text{“GT”}) - P_{\tilde{x}_{\text{out},L}}(\text{“GT”})}{P_{x_{\text{out}}}(\text{“GT”}) - P_{y_{\text{out}}}(\text{“GT”})}$$

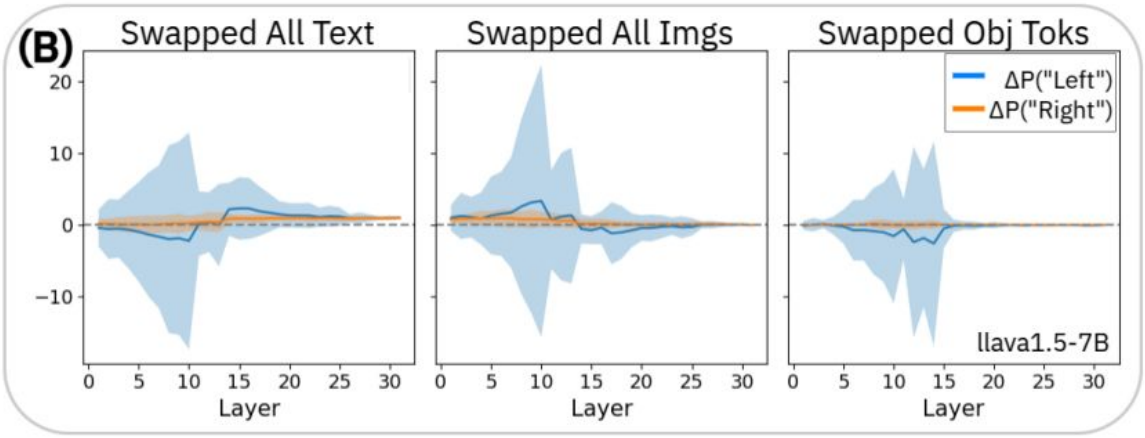
Is this effect really a result of spatial mirroring?



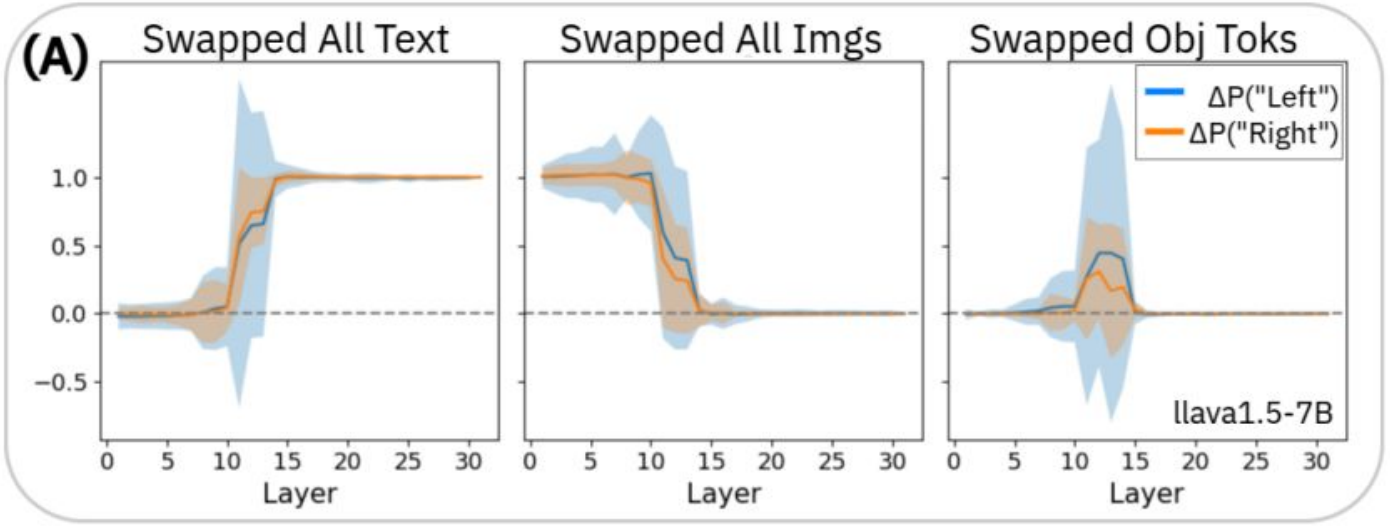
Mirror Swapping

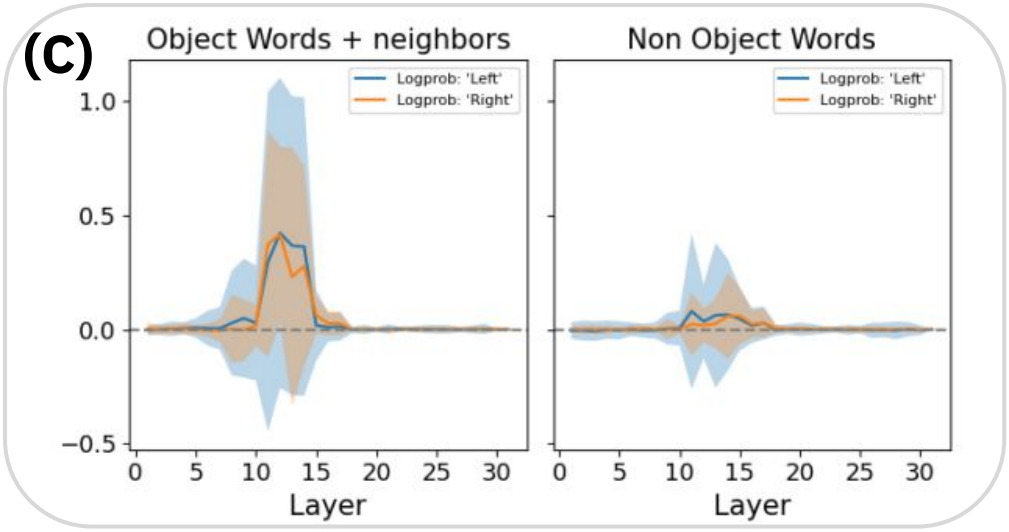


Attribute Swapping

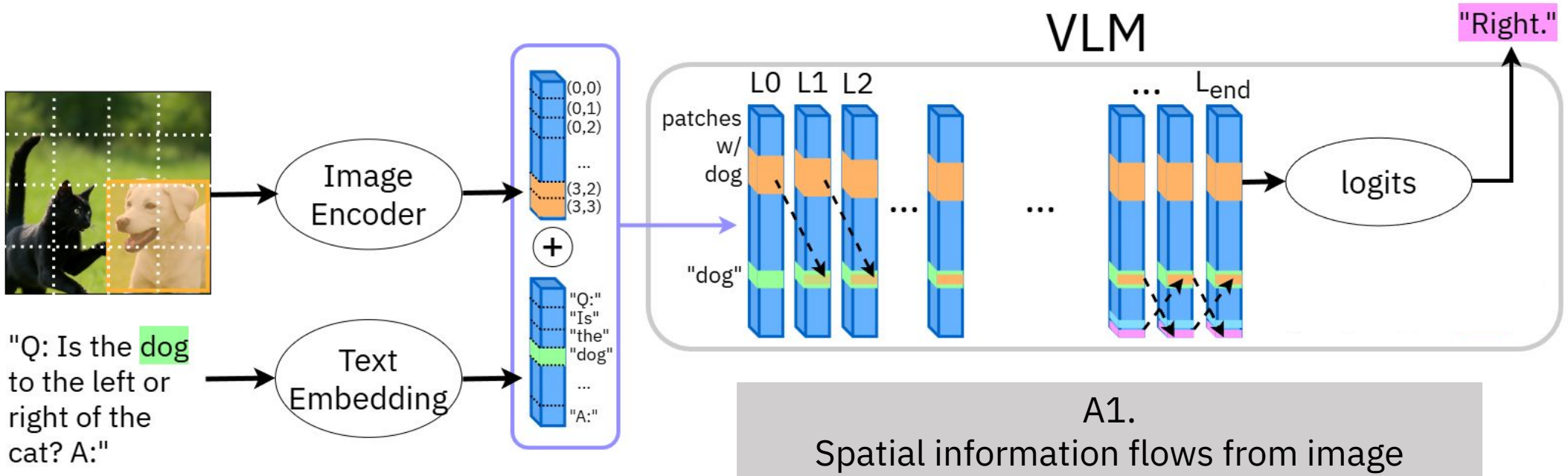


Are object word tokens actually special?



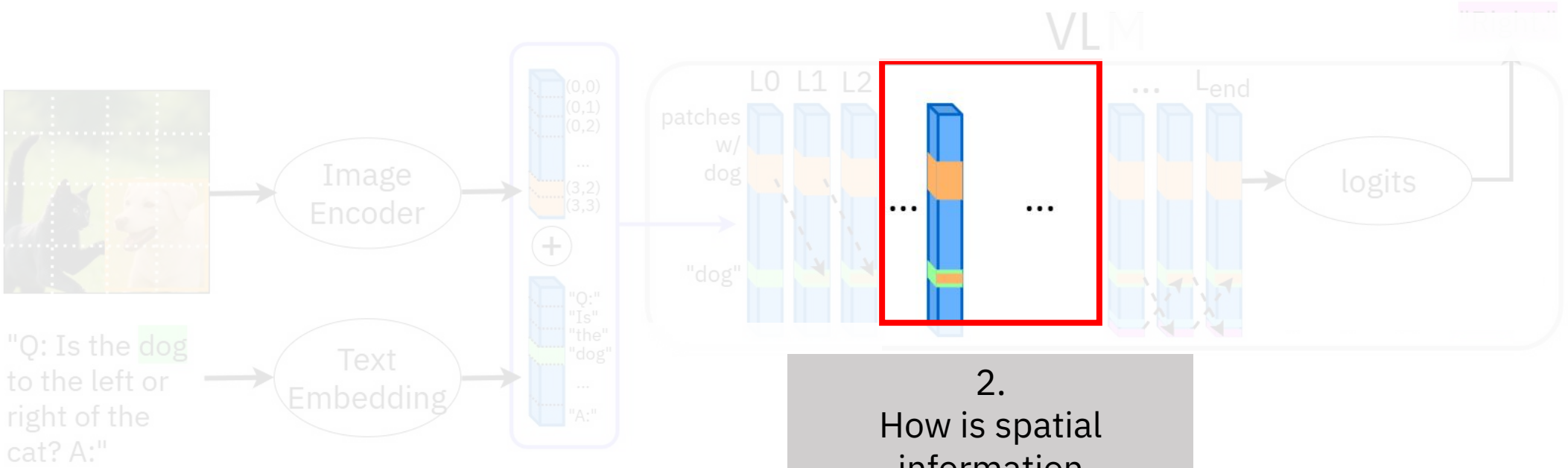


So now we know the VLM is doing something like this:



A1.
Spatial information flows from image patches to text tokens in modality merging layers, specifically onto object word tokens.

New question:



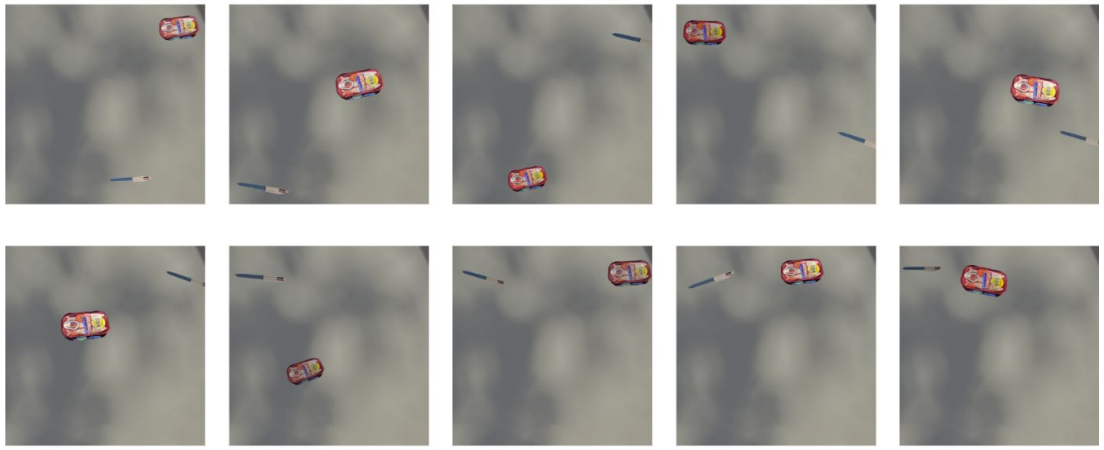
2.
How is spatial information represented in text?

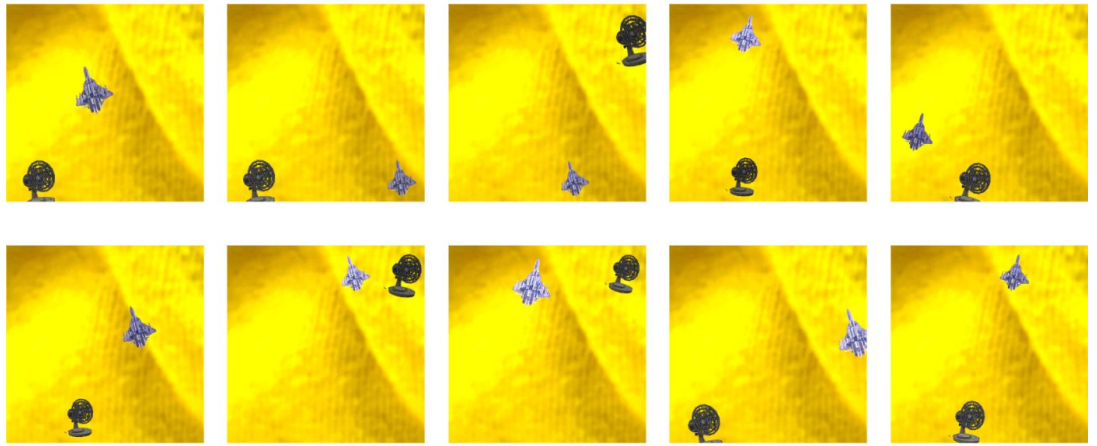
2. How is spatial information represented in text?

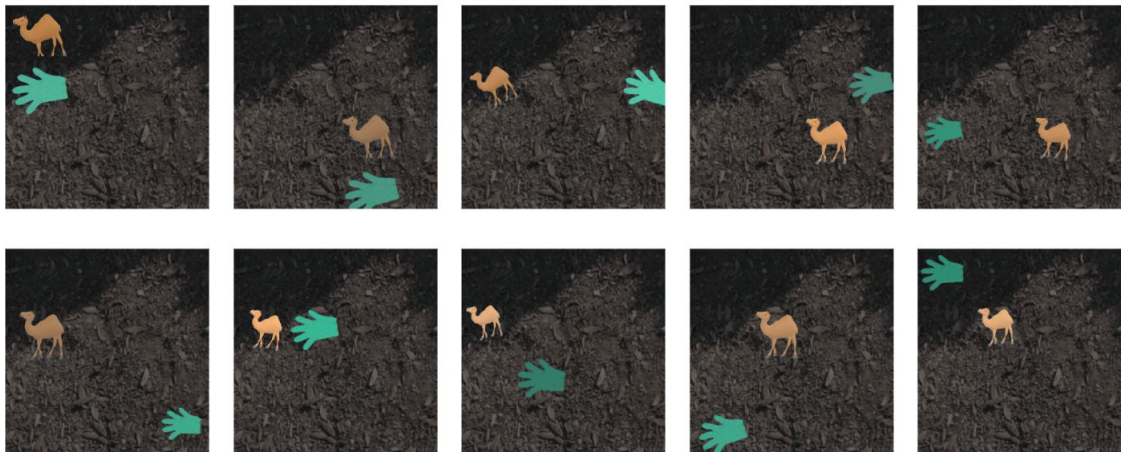
- It looks like spatial information is getting stored in object words
- But what is the format?

Lowest effort starting point: maybe it is bound linearly like this?

Simple linear derivation from synthetic data

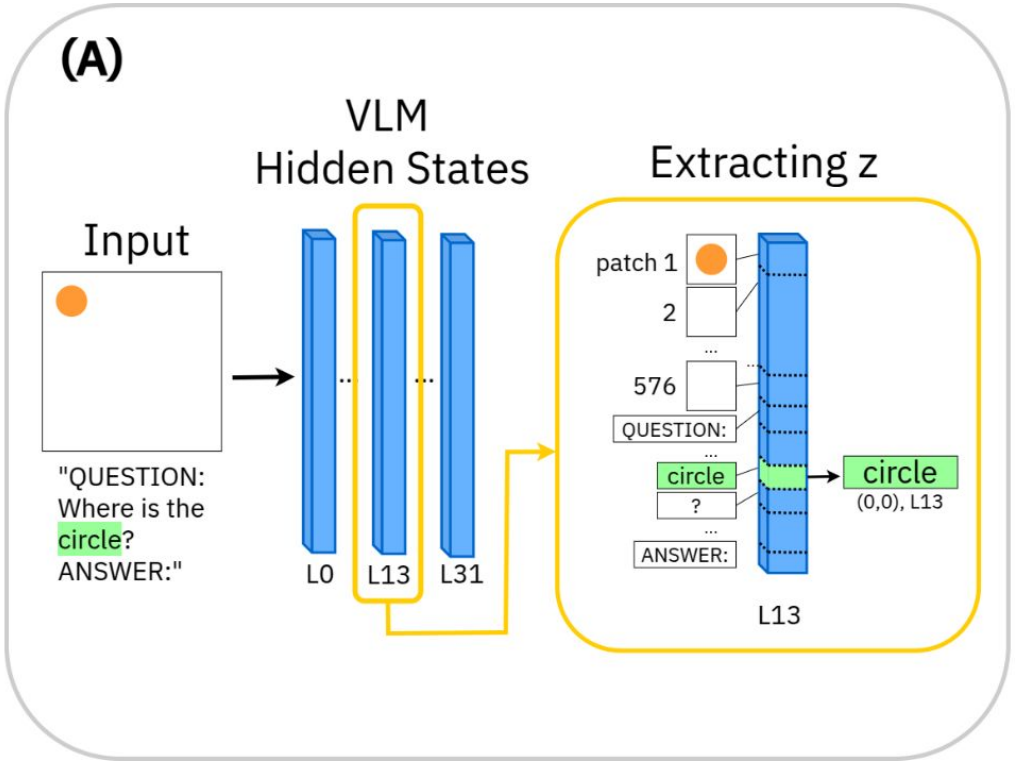




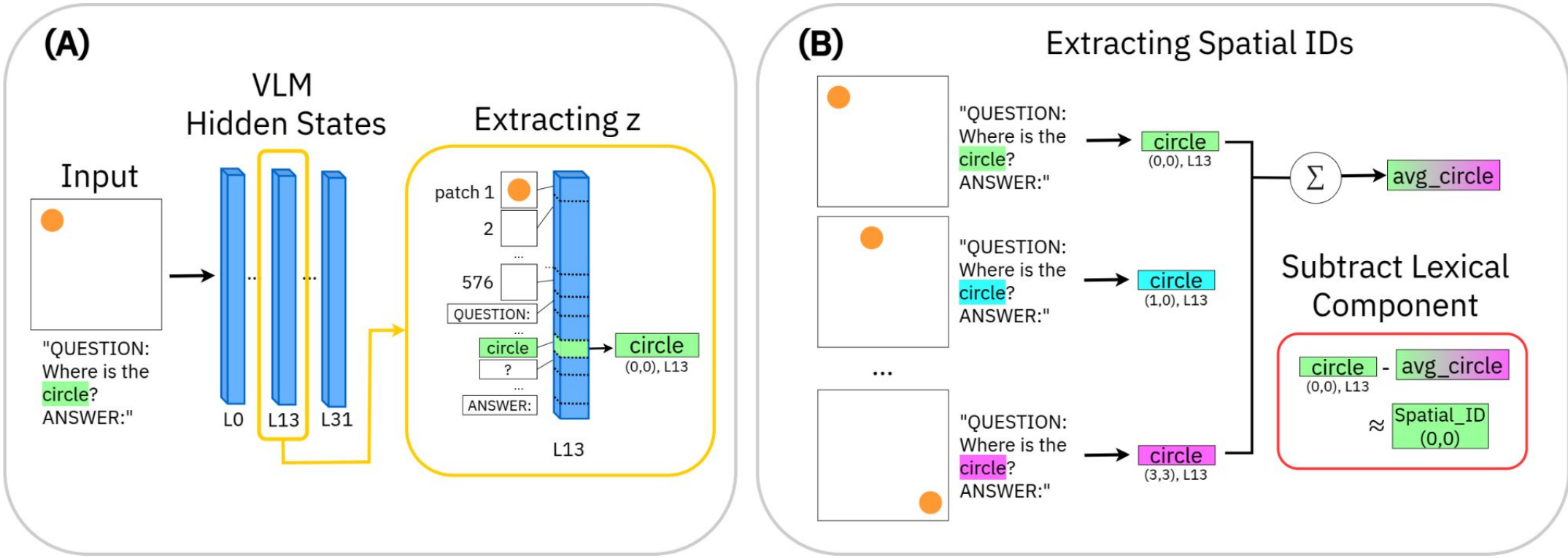


Deitke, Matt, et al. "Objaverse: A universe of annotated 3d objects." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.

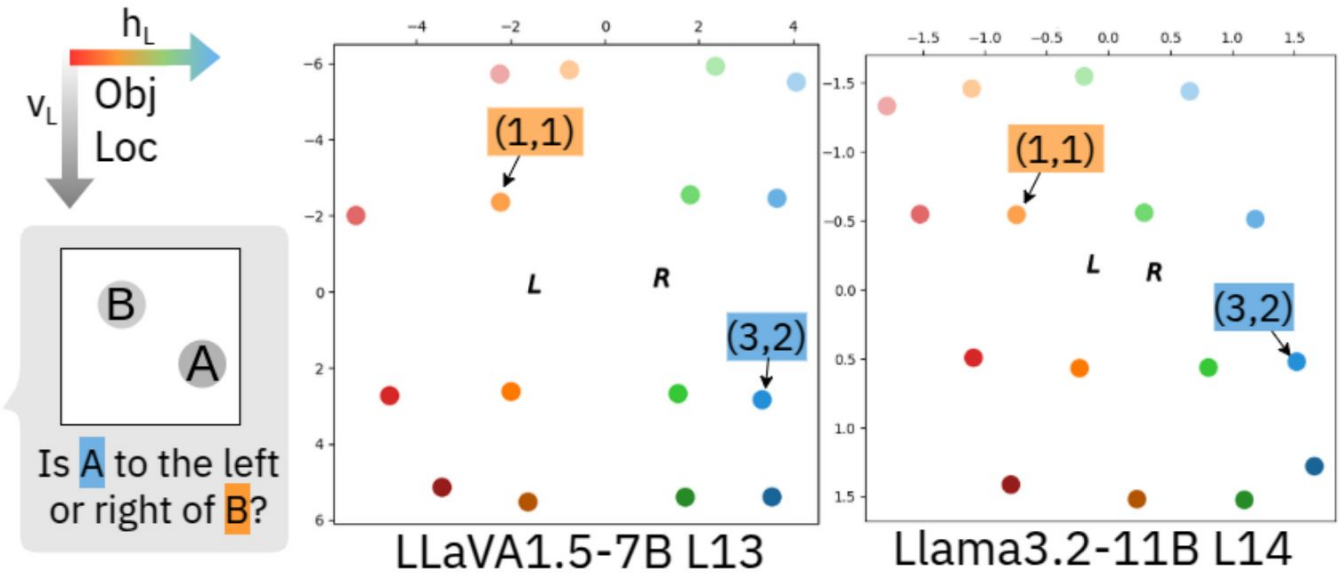
Simple linear derivation from synthetic data



Simple linear derivation from synthetic data

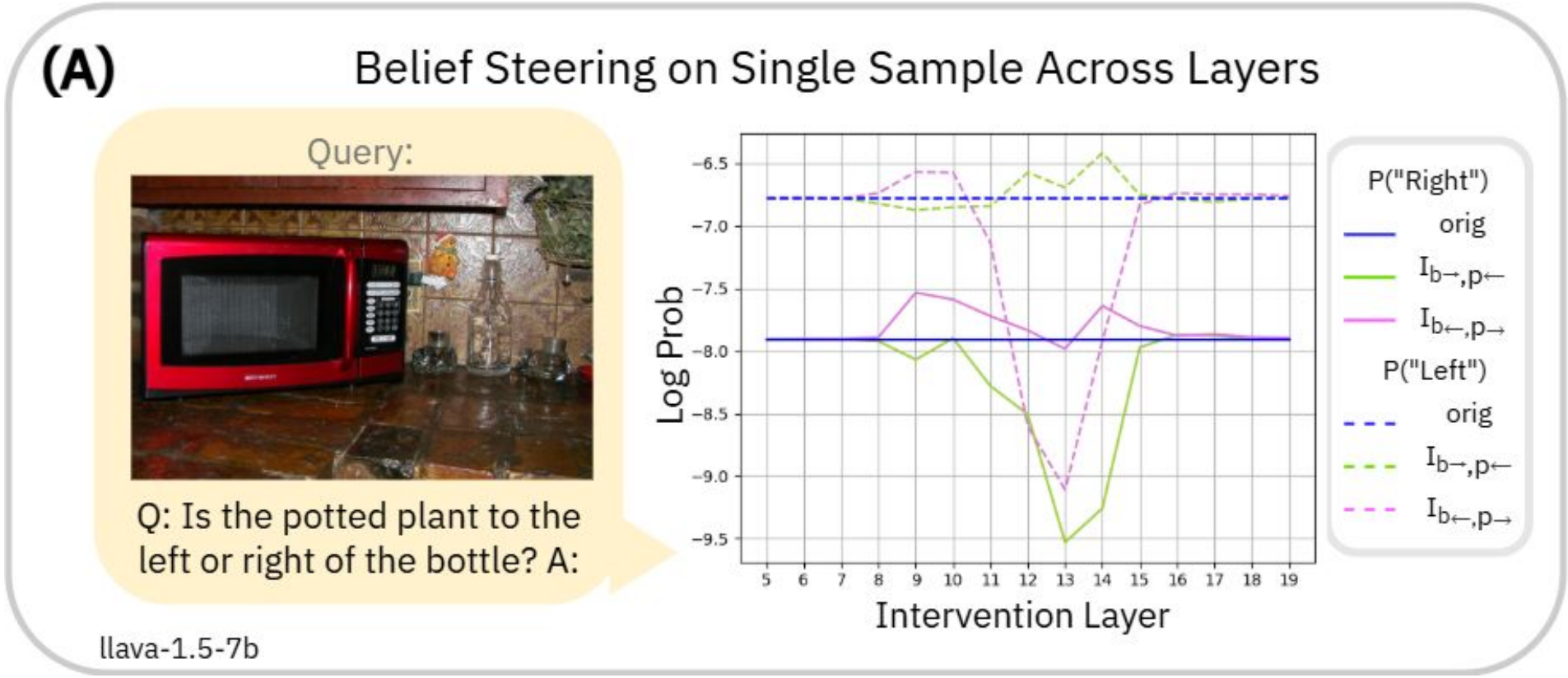


Extracted spatial IDs projected onto vertical and horizontal axis vectors preserve linear structure:



$$v_L = \frac{1}{m \cdot \binom{m}{2}} \sum_{i=0}^{m-1} \sum_{j_1 > j_2} [\Delta_L(i, j_1) - \Delta_L(i, j_2)]$$

Spatial IDs can linearly mediate model belief on *real images*

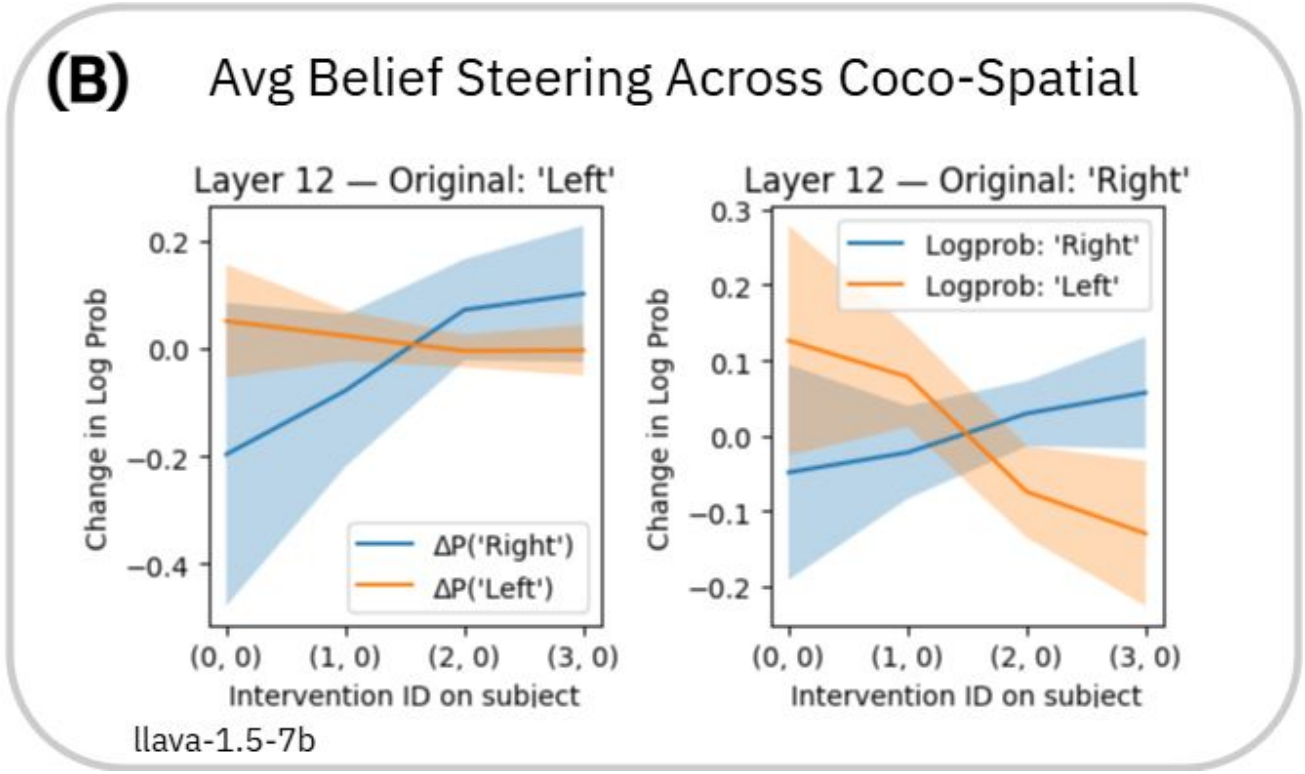


Spatial IDs can linearly mediate model belief on *real images*

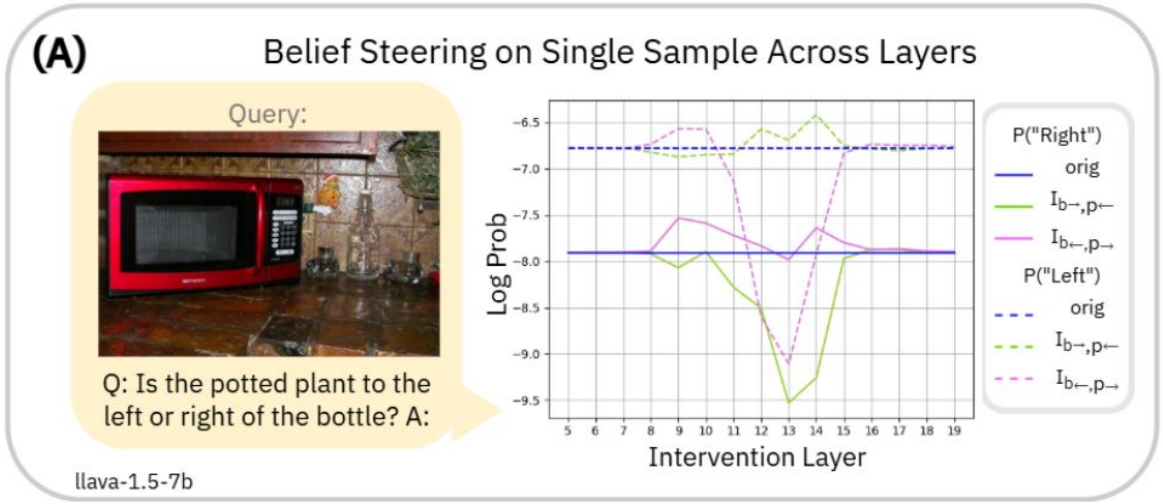
Across 100 image-query pairs in COCO, with prompt:

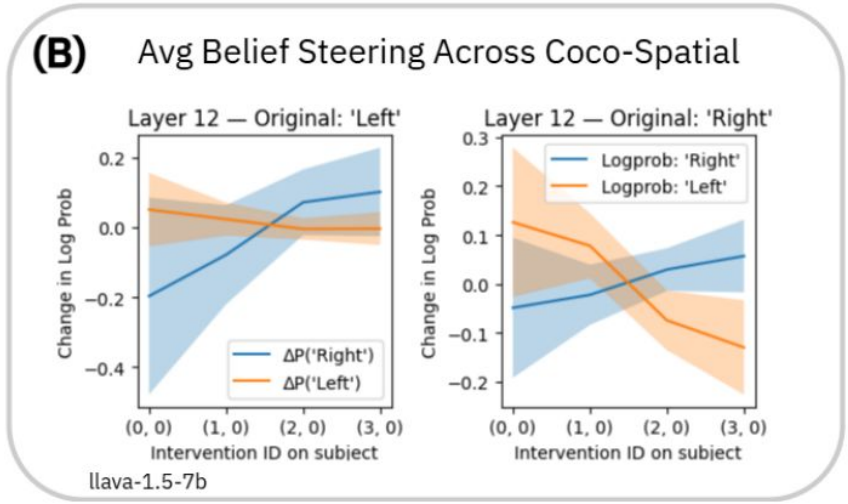
> Is SUBJECT to the left or right of the OBJECT?

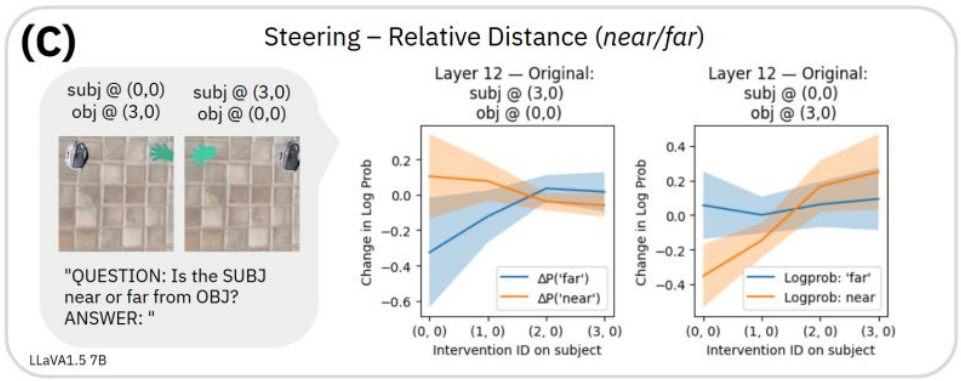
Steering on SUBJECT:

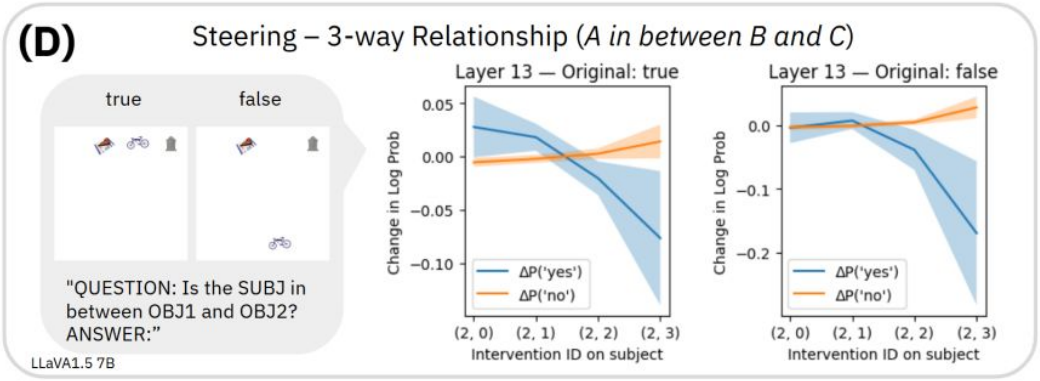


Spatial IDs can linearly mediate model belief on *real images*

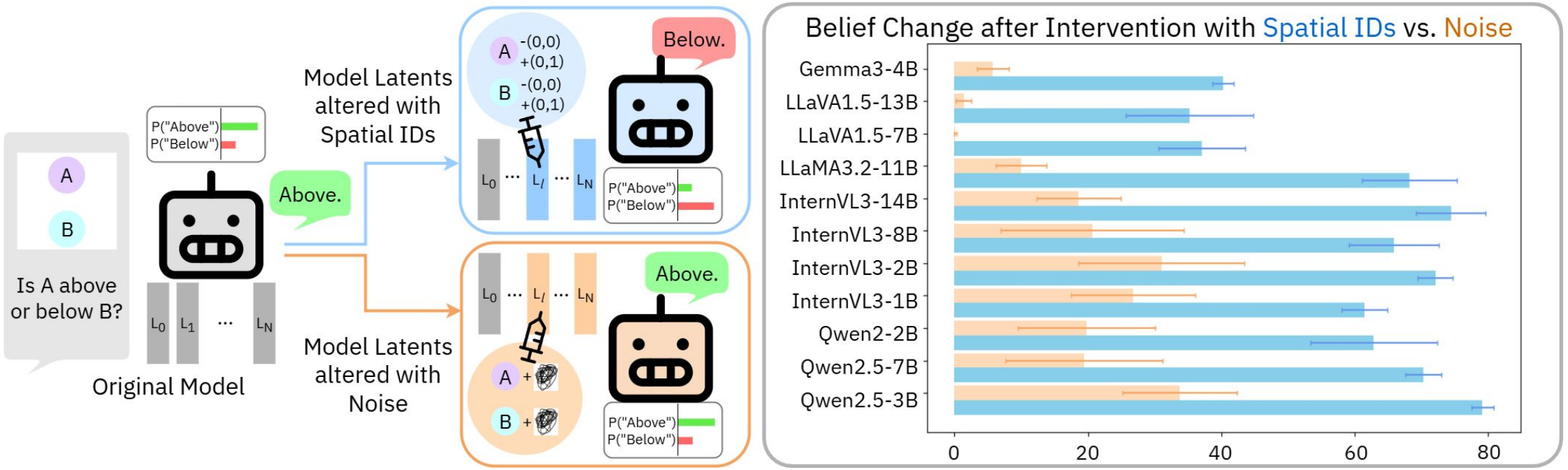






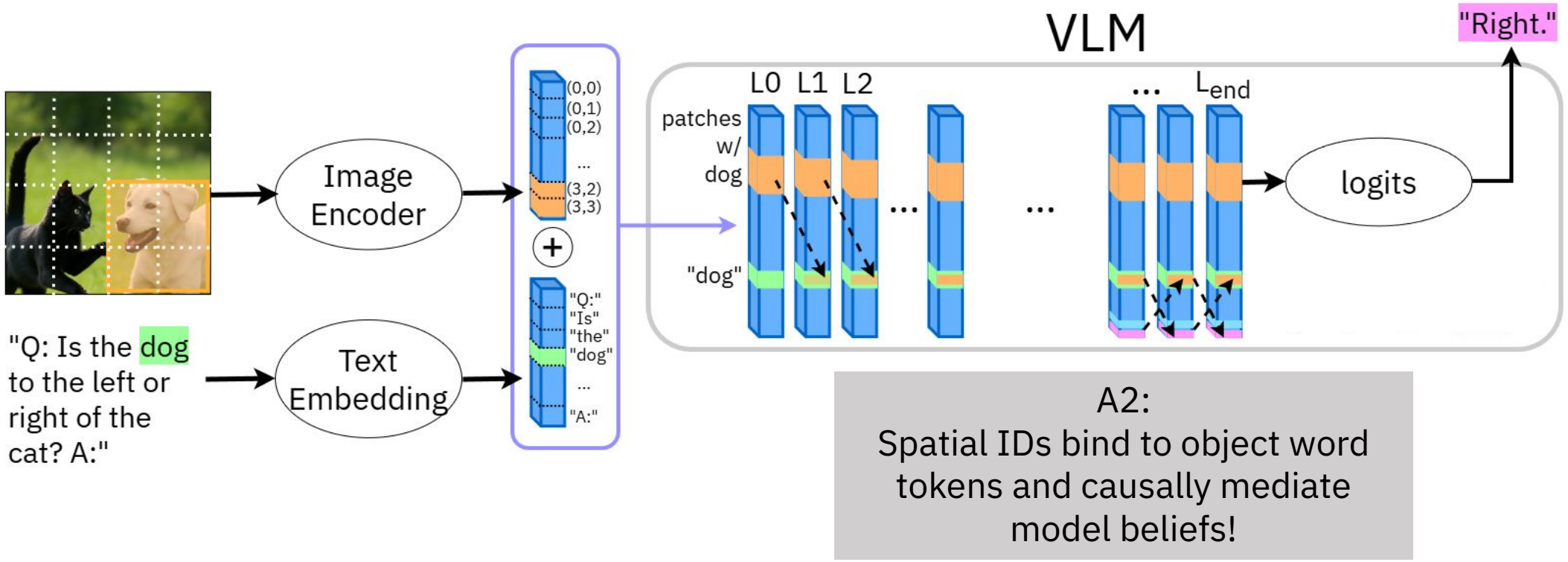


This linear steerability of model belief with *spatial IDs* is observed across models:

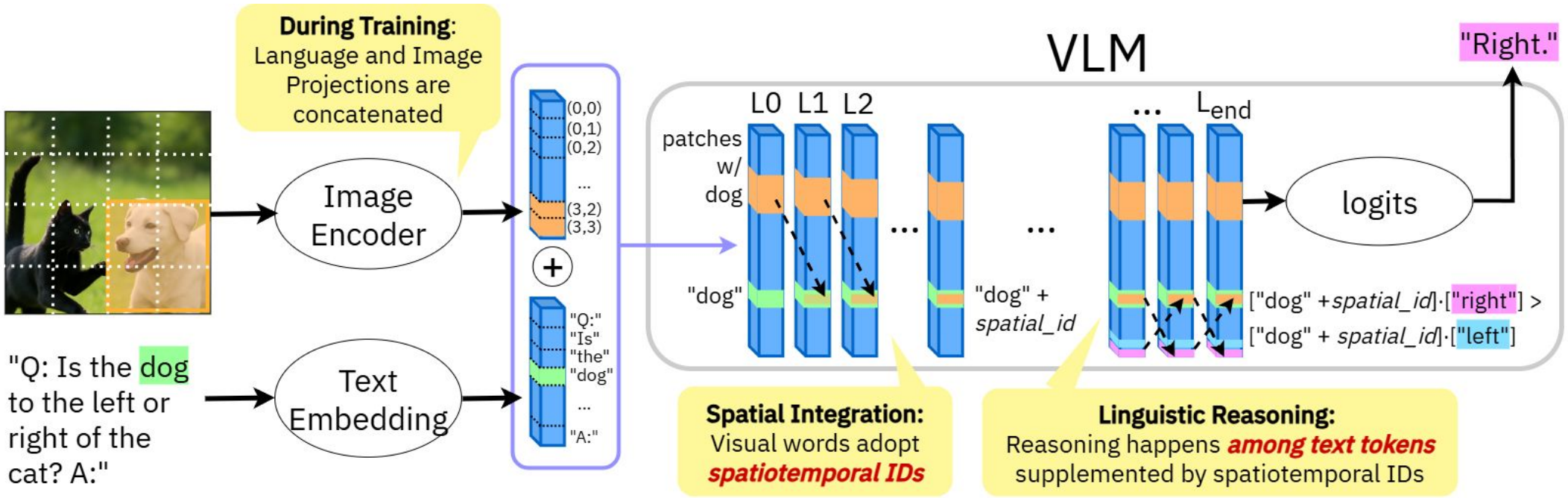


% of samples on which models' beliefs were *swapped*

Now we have this picture of spatial reasoning in VLMs:

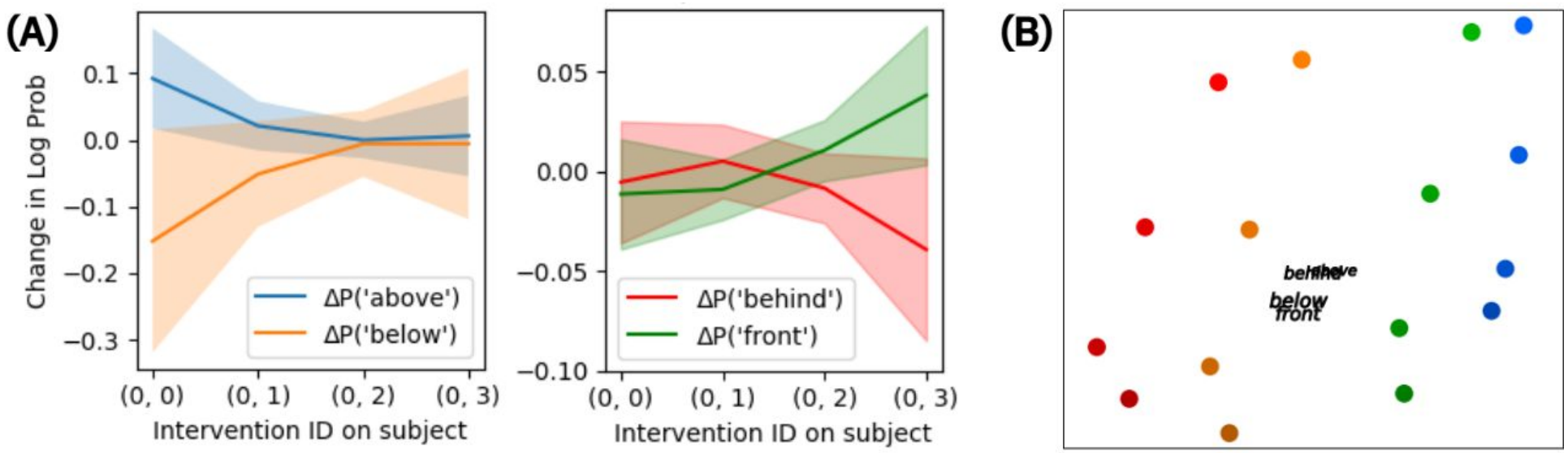


Now we have this picture of spatial reasoning in VLMs:



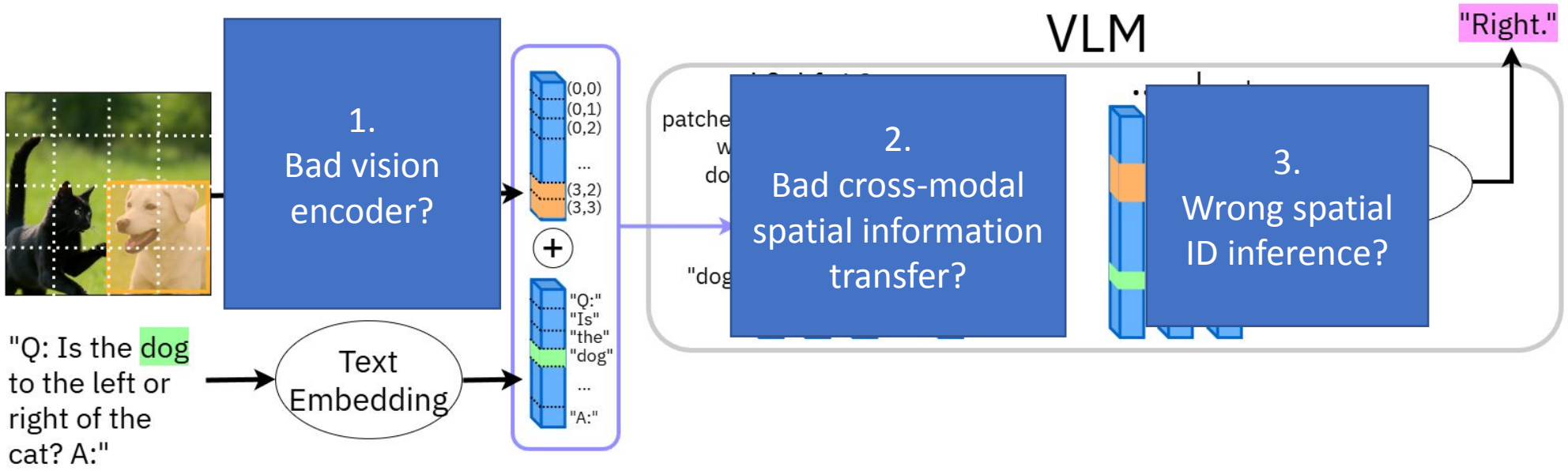
So… how does knowing about *spatial IDs* help VLMs?

1. Spatial IDs help us diagnose depth representation limitations.

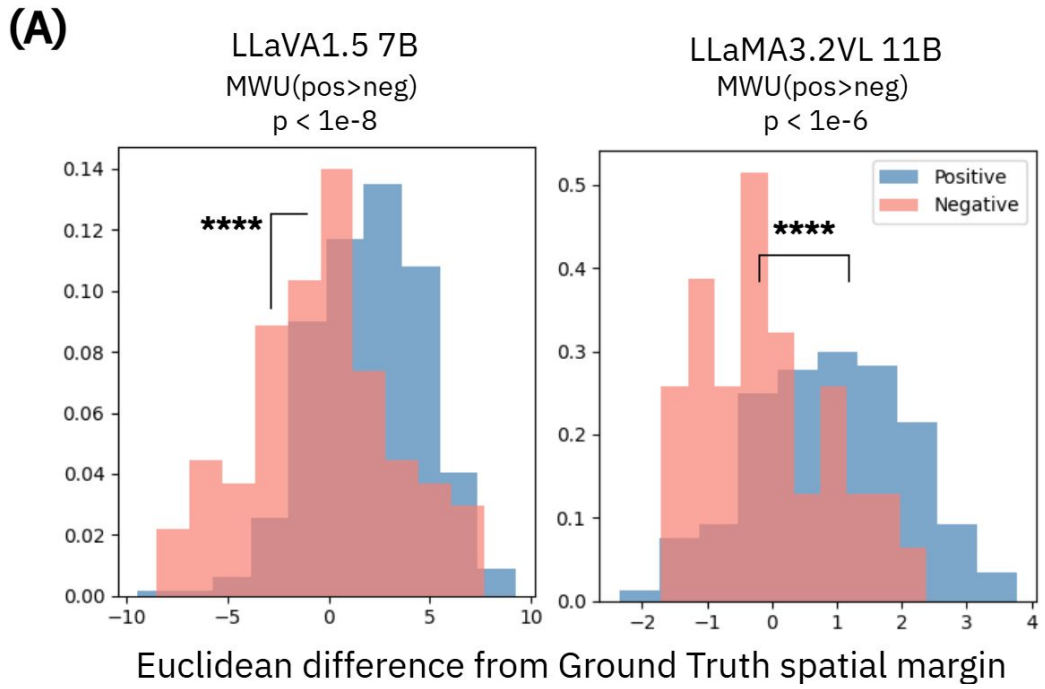


2. Spatial IDs help diagnose model bottlenecks.

When a model fails – why does it fail?



2. Spatial IDs help diagnose model bottlenecks.



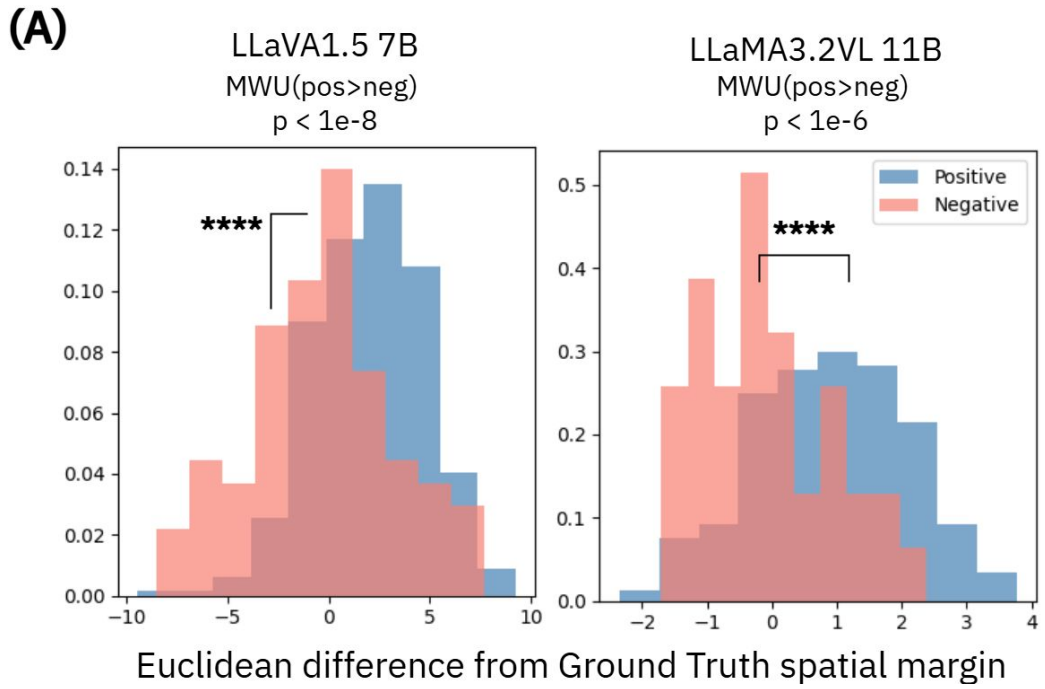
In LLaVA and LLaMA, wrong spatial IDs
 wrong predictions.

3. Wrong spatial ID inference?

$$\Delta_L^{(o)}(i,j)_{ext} \approx VV^T\phi_L(o; I_{(i,j)}^{(o)}, T^{(o)}), \quad V = [v_L, h_L]$$

$$\text{ID deviation margin} = \epsilon_{\text{ext}} - \epsilon_{\text{gt}}, \quad \text{where } \epsilon_{\text{gt}} = i_{\text{gt}}^{(o)} - i_{\text{gt}}^{(\tilde{o})}, \epsilon_{\text{ext}} = i_{\text{ext}}^{(o)} - i_{\text{ext}}^{(\tilde{o})}$$

2. Spatial IDs help diagnose model bottlenecks.



But where do these wrong spatial IDs come from?

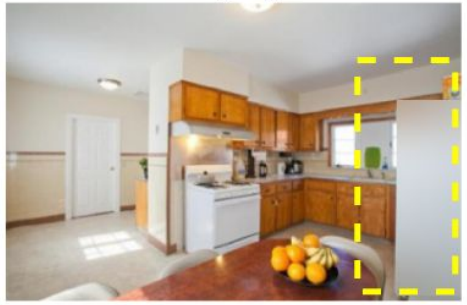
Bad **vision encoder**? *or* Bad crossmodal **spatial ID integration**?

2. Spatial IDs help diagnose model bottlenecks.

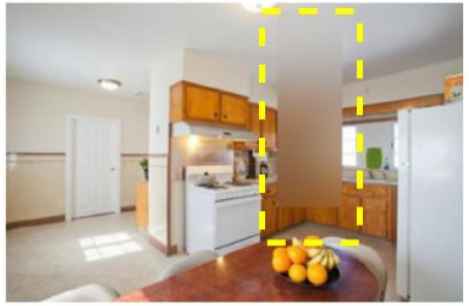
QUESTION: Is a refrigerator left or right of a oven? Answer left or right. \n ANSWER:

Original

Blurring obj bbox

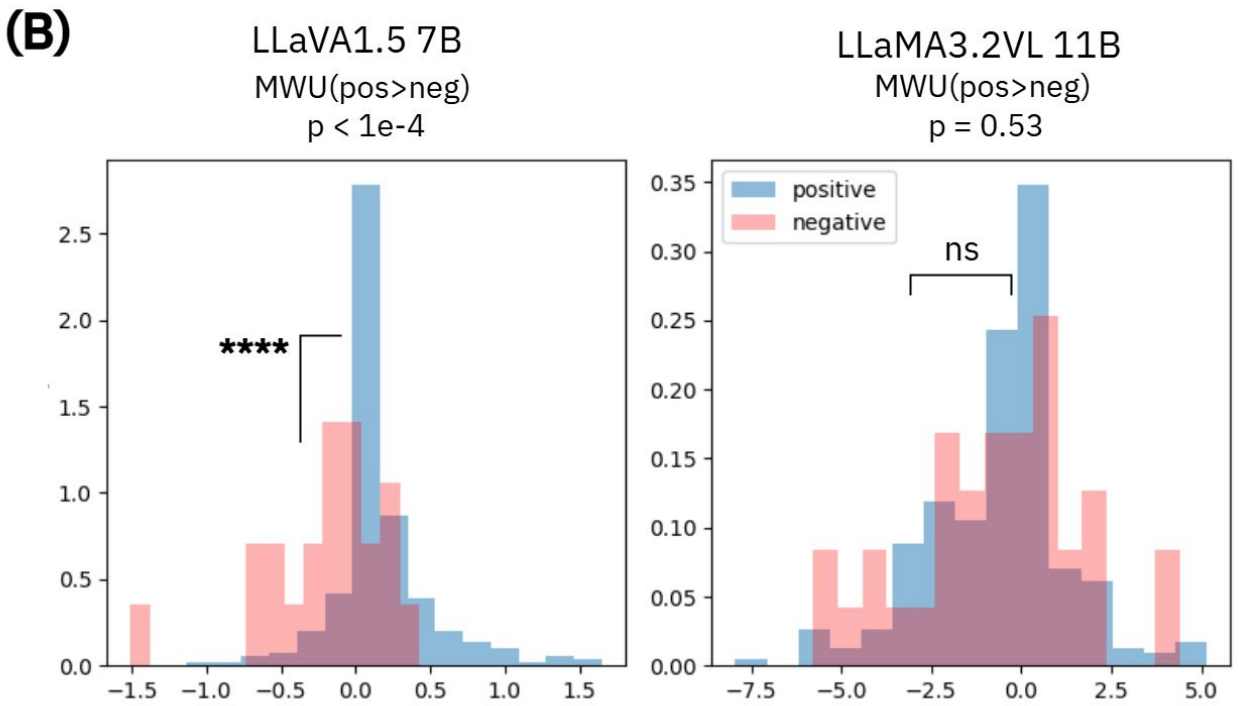


Blurring outside of bbox



If the vision encoder was *good*, it would be more sensitive to *blurring the object*, than some random area.

2. Spatial IDs help clarify model bottlenecks.

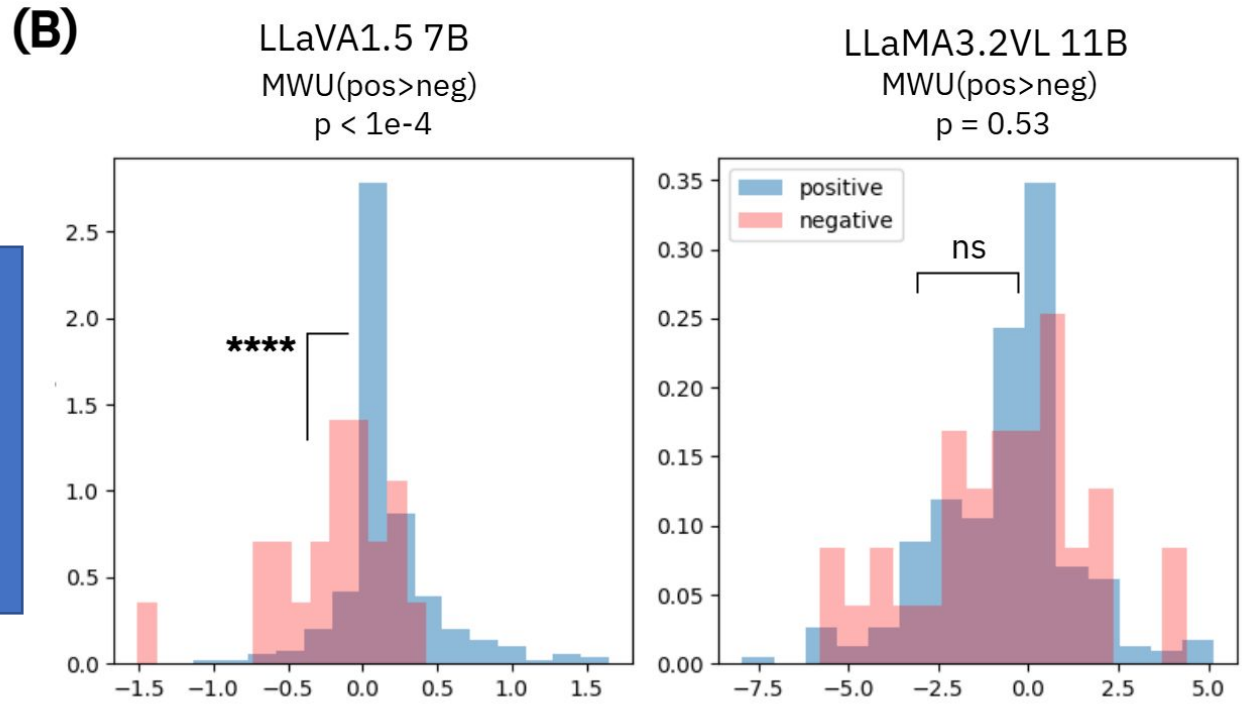


Diff in Model Acc between masking object bbox vs. random loc

$$\text{bbox sensitivity} = \big(P(\text{“GT”}) - P(\text{“GT”}|\text{mask } o)\big) - \big(P(\text{“GT”}) - min_r[P(\text{“GT”}|\text{mask } r), r \in R]\big)$$

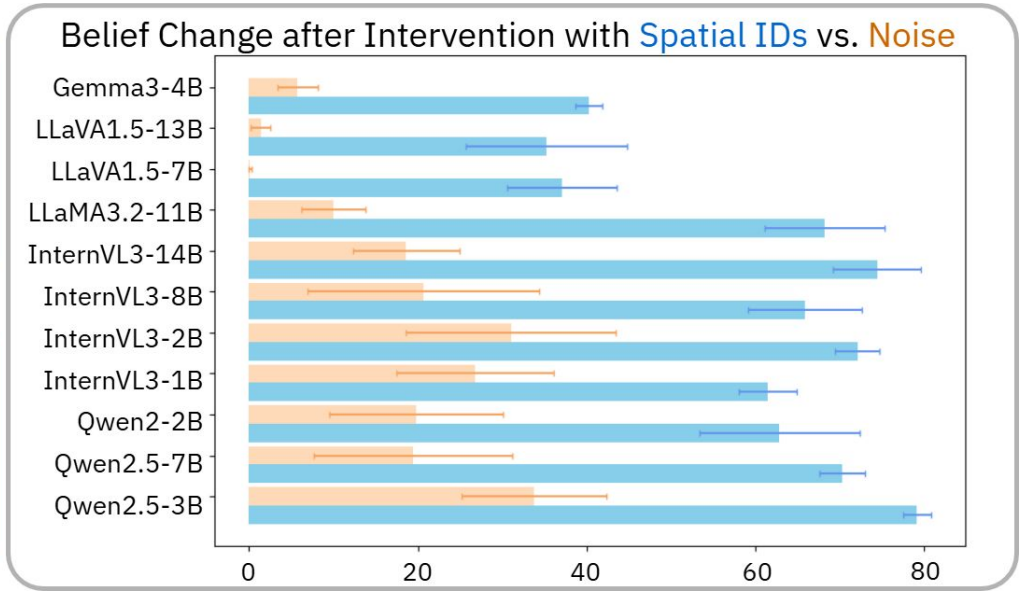
2. Spatial IDs help clarify model bottlenecks.

LLaVA =
1. Bad vision encoder!

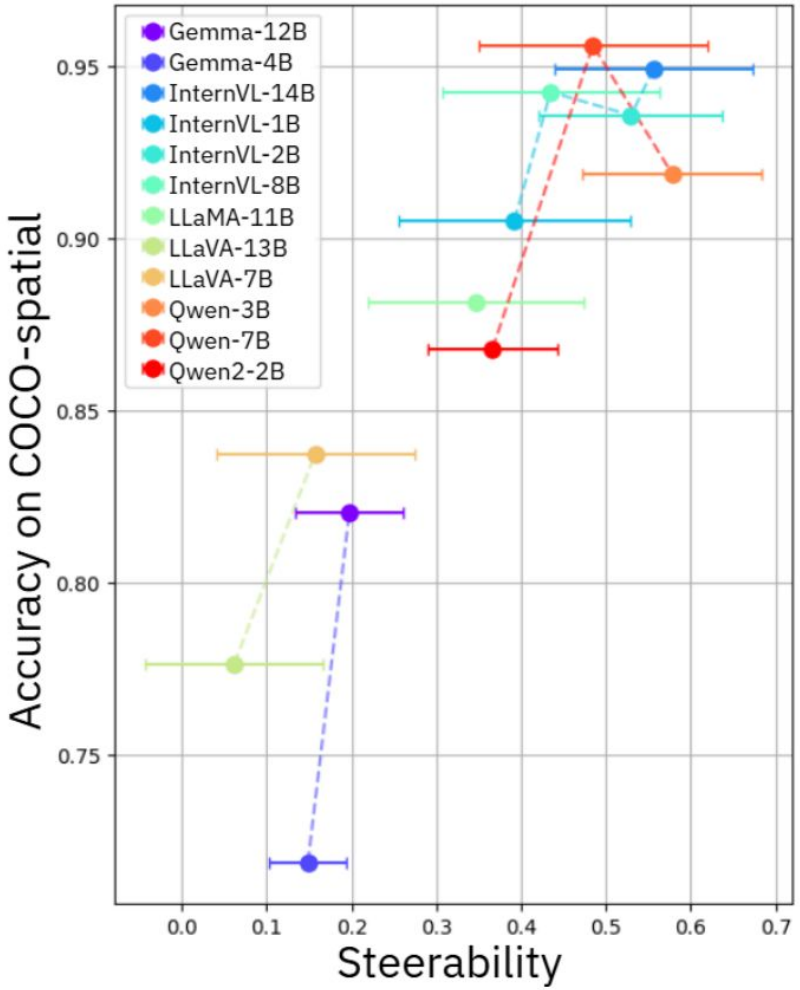


LLaMA =
2. Bad cross-modal spatial information transfer!

3. Spatial IDs improve model performance on spatial queries.



If we want model to be better at spatial queries… we could promote spatial ID learning during training!



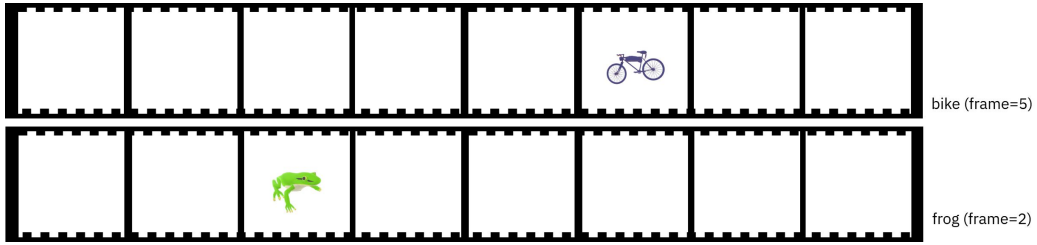
3. Spatial IDs improve model performance on spatial queries.

	Num Steps	0	800	1600	2400	3200
Control	LM Loss (↓)	3.30	0.05	<0.01	<0.01	<0.01
	COCO Val Accuracy (↑)	0.77	0.83	0.84	0.85	**0.85**
With Spatial Loss	LM Loss (↓)	2.71	0.04	<0.01	<0.01	<0.01
	Spatial ID Loss (↓)	0.75	0.58	0.41	0.36	0.33
	COCO Val Accuracy (↑)	0.77	0.83	0.84	0.88	**0.91**

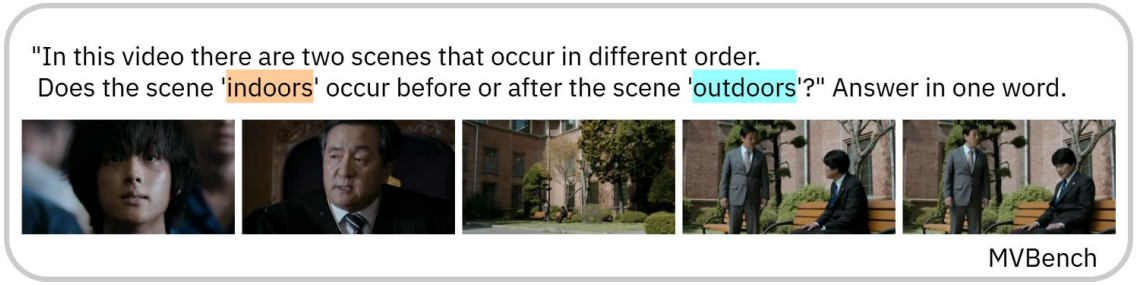
Finetuning Qwen2-2B on *synthetic data* to promote better spatial ID organization results in faster convergence and higher validation accuracy on COCO.

Also, these experiments can be repeated in video models.

With synthetic videos

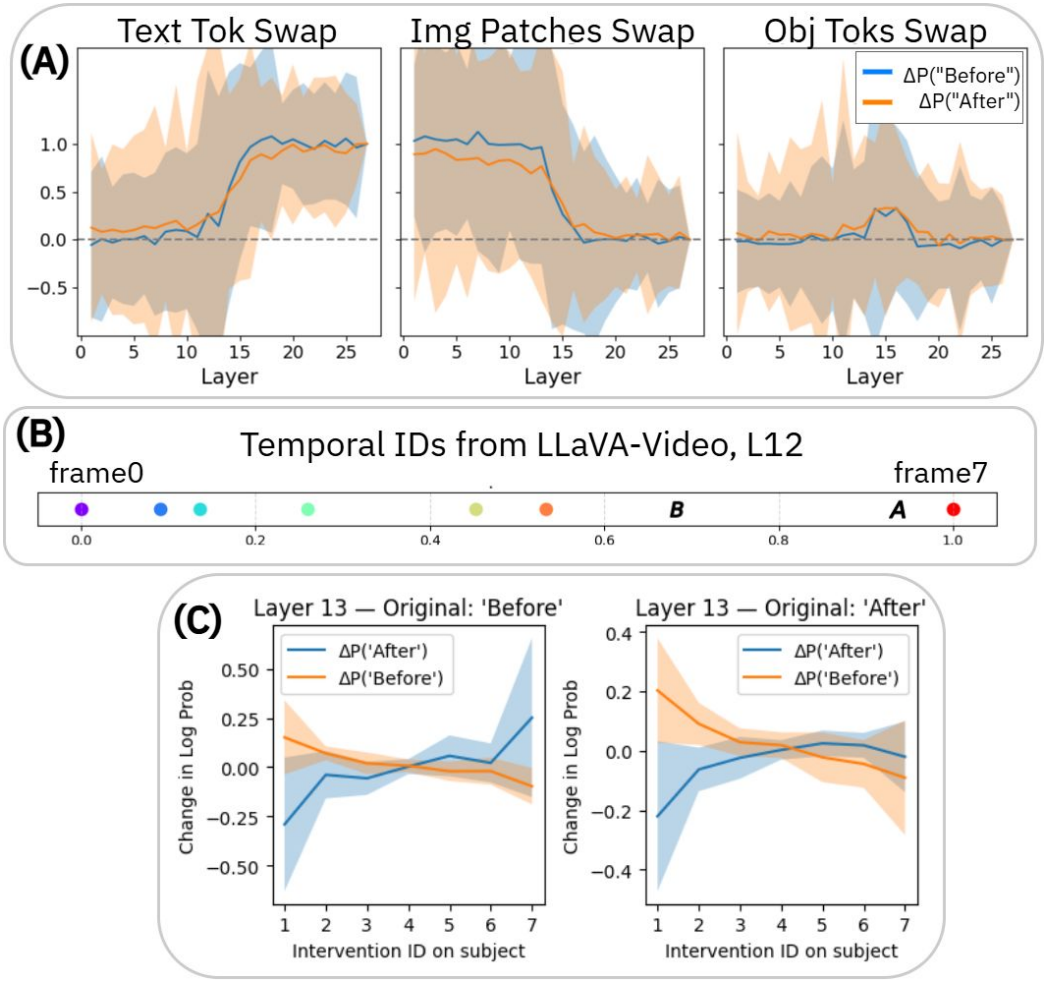


And real scenes from movies



Also, these experiments can be repeated in video models.

We can perform *temporal* mirror swapping, ID extraction, and steering.



Conclusion: Autoregressive VLMs exhibit compositional reasoning capacities *by linearly encoding* spatial/temporal IDs into word activations in modality-merging layers for efficient readout.

arXiv

GitHub

The End!

Raphi Kang*

Hongqiao (Harry) Chen*

Georgia Gkioxari

Pietro Perona

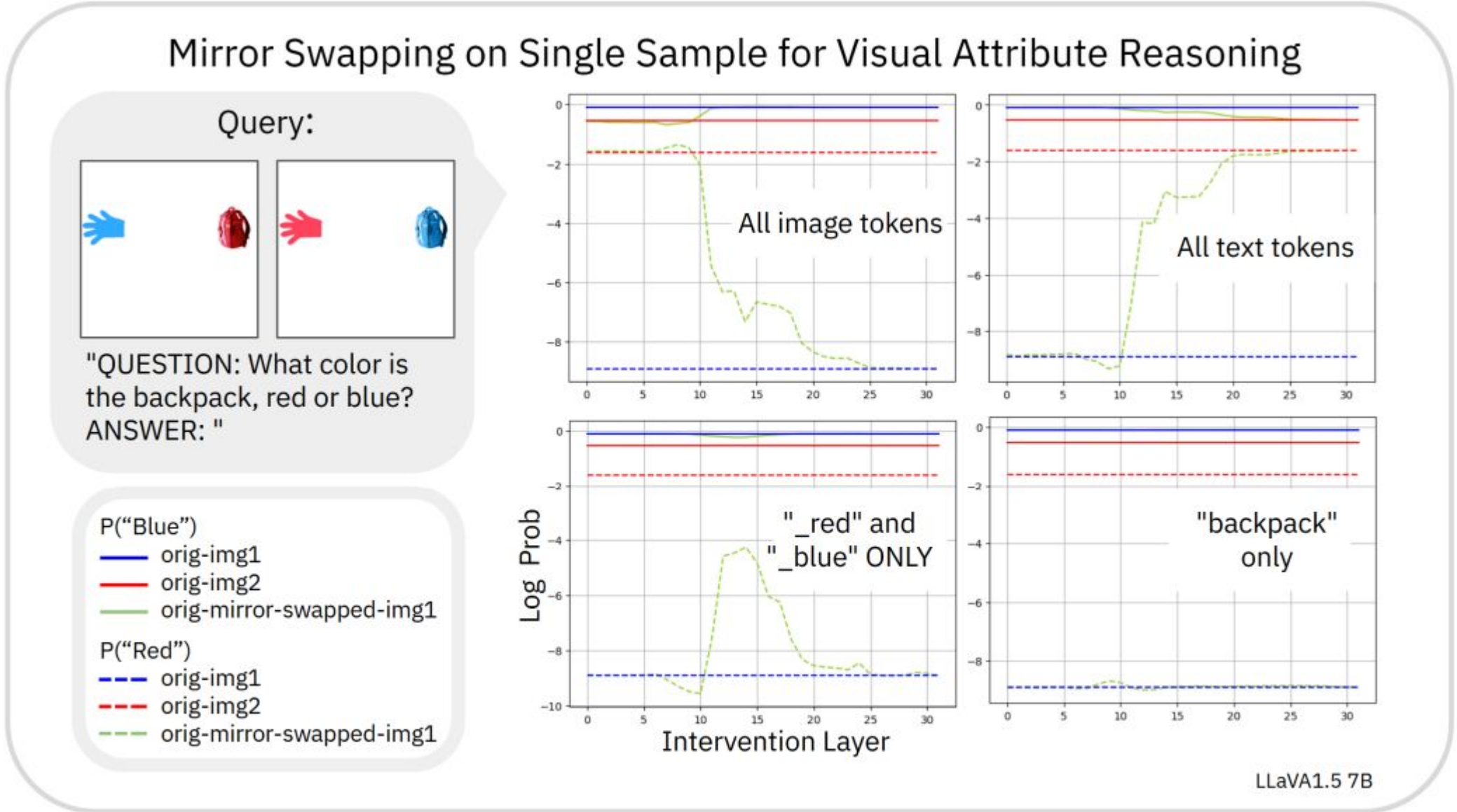
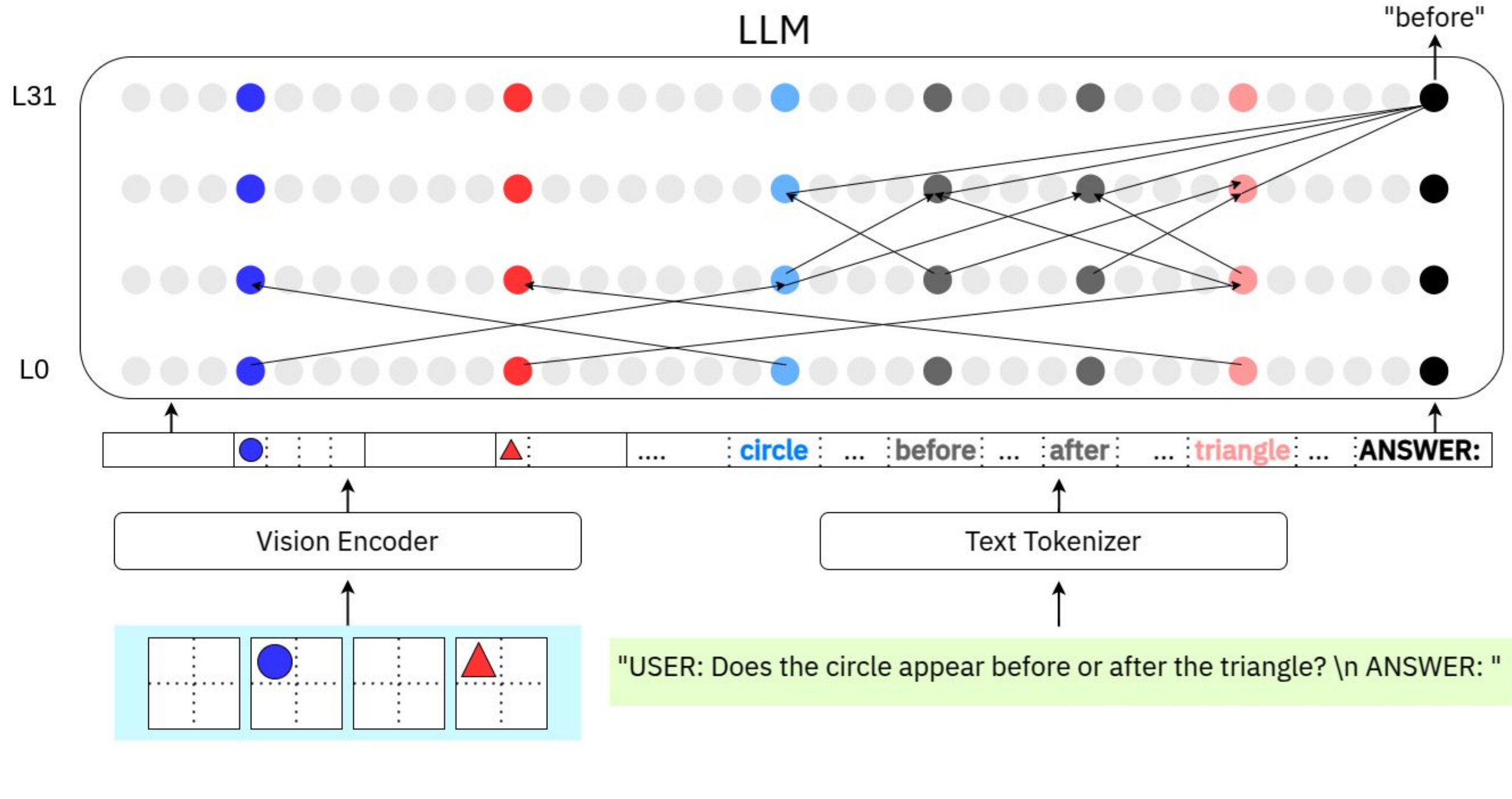
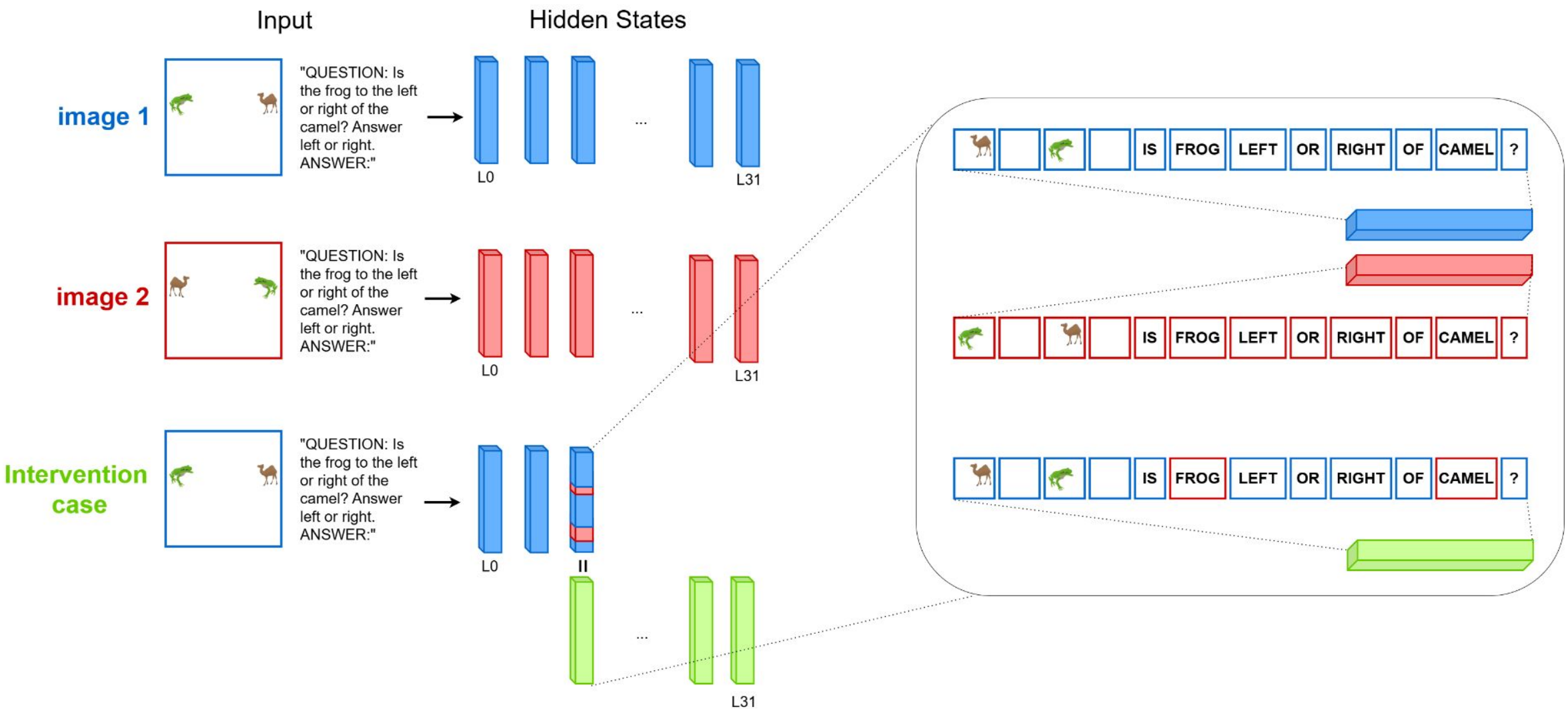
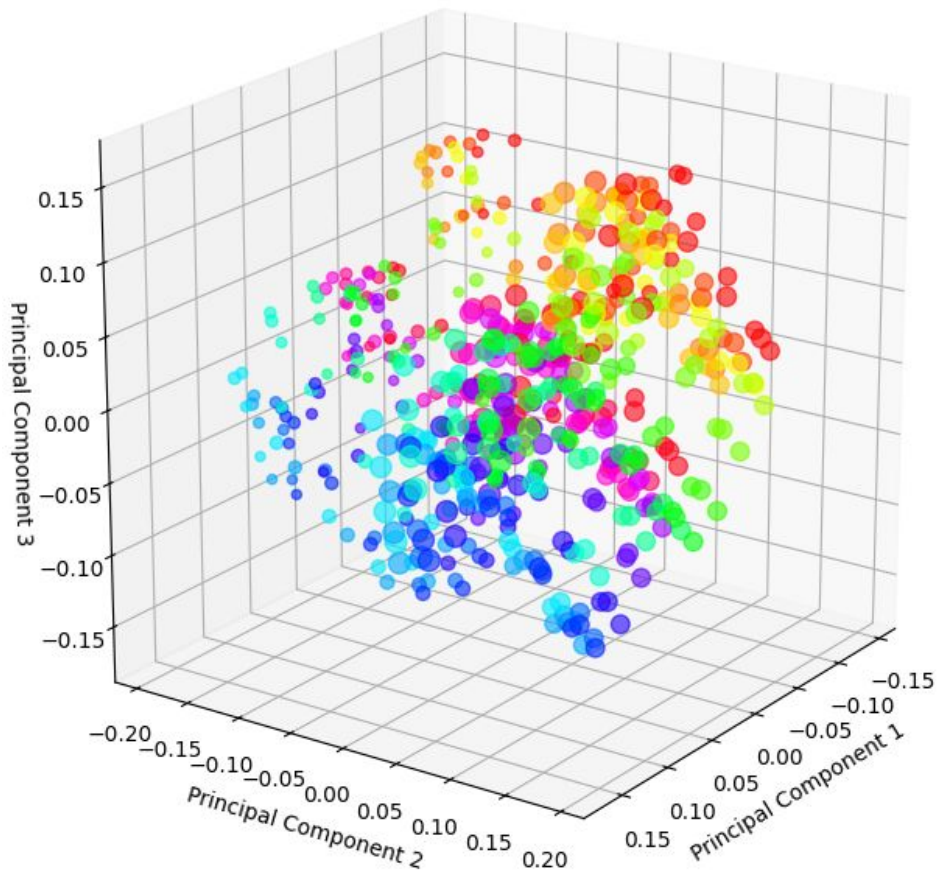


Figure A6: Mirror Swapping on Single Sample for Color Binding


LLM
"before"
L31
L0
.... circle ... before ... after ... triangle ... ANSWER:
Vision Encoder
Text Tokenizer
"USER: Does the circle appear before or after the triangle? \n ANSWER: "


Input
Hidden States
image 1
"QUESTION: Is the frog to the left or right of the camel? Answer left or right. ANSWER:"
L0
...
L31
image 2
"QUESTION: Is the frog to the left or right of the camel? Answer left or right. ANSWER:"
L0
...
L31
Intervention case
"QUESTION: Is the frog to the left or right of the camel? Answer left or right. ANSWER:"
L0
II
...
L31
IS FROG LEFT OR RIGHT OF CAMEL ?
IS FROG LEFT OR RIGHT OF CAMEL ?
IS FROG LEFT OR RIGHT OF CAMEL ?


Principal Component 3
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
Principal Component 2
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
Principal Component 1
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15

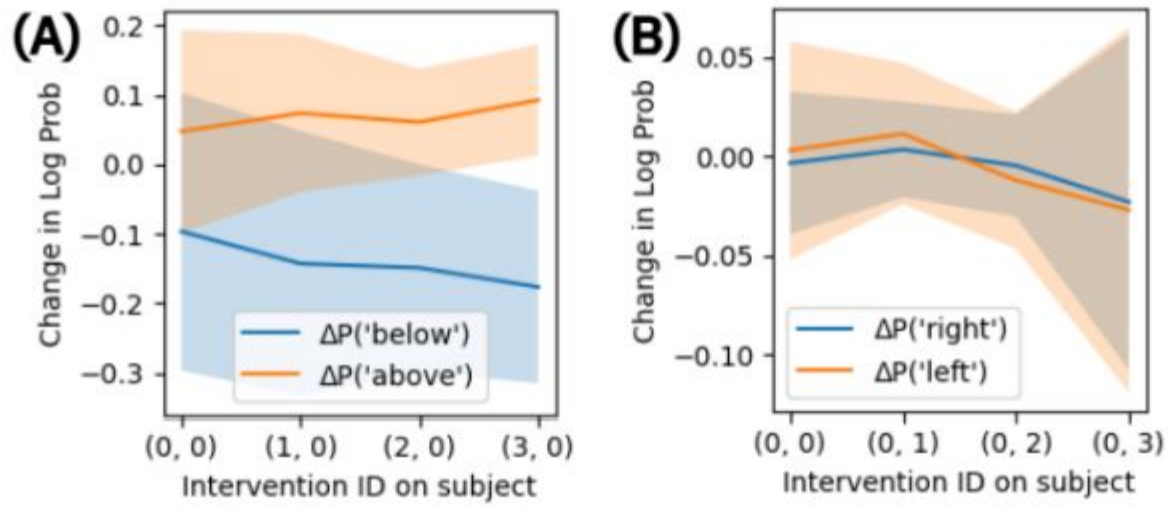


Figure A16: Steering effects of horizontal vectors on vertical beliefs (A) and vertical vectors on horizontal beliefs (B) in LLaVA.

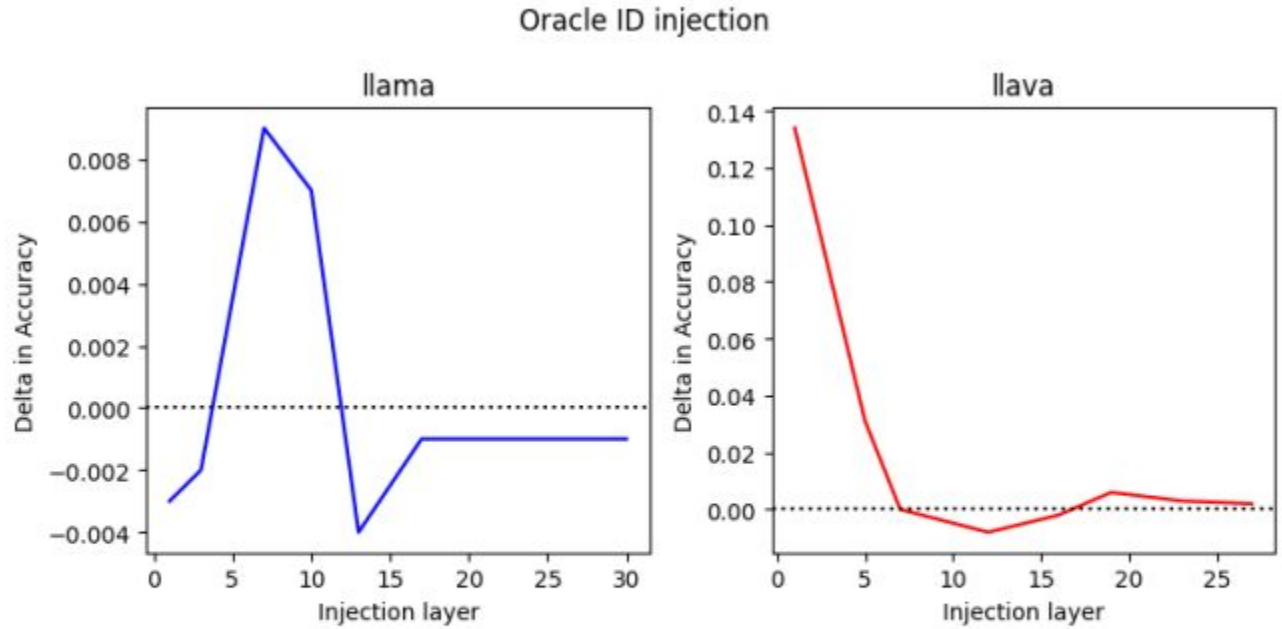


Figure A9: LLaMA and LLaVA evaluation accuracy on synthetic grid-like data with oracle spatial ID injections at varying layers. 0 is baseline model performance, without any intervention.

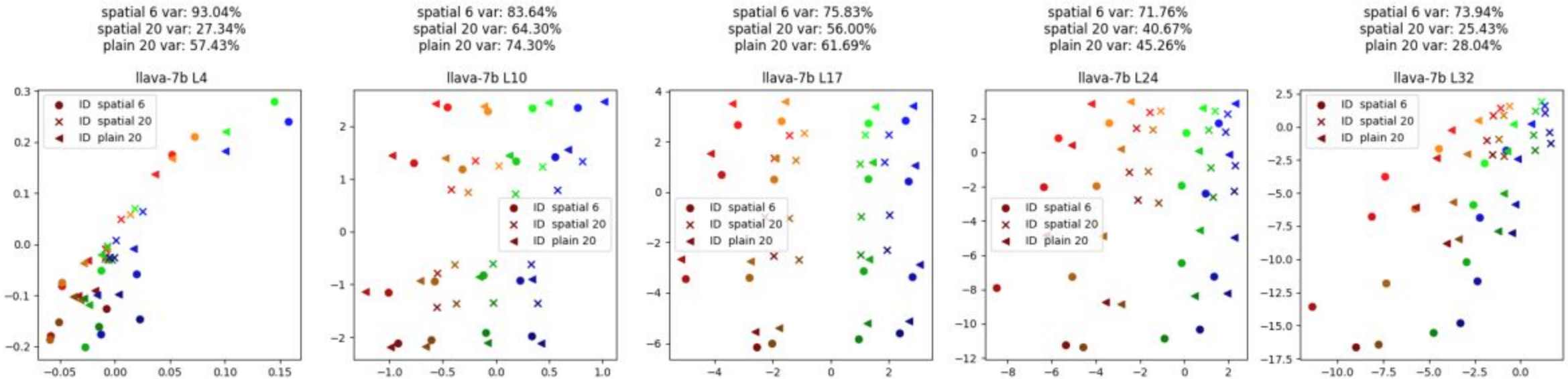


Figure A19: Plain prompts and spatial prompts projected onto spatial axes created from spatial prompts. Colors exhibit tight clustering.

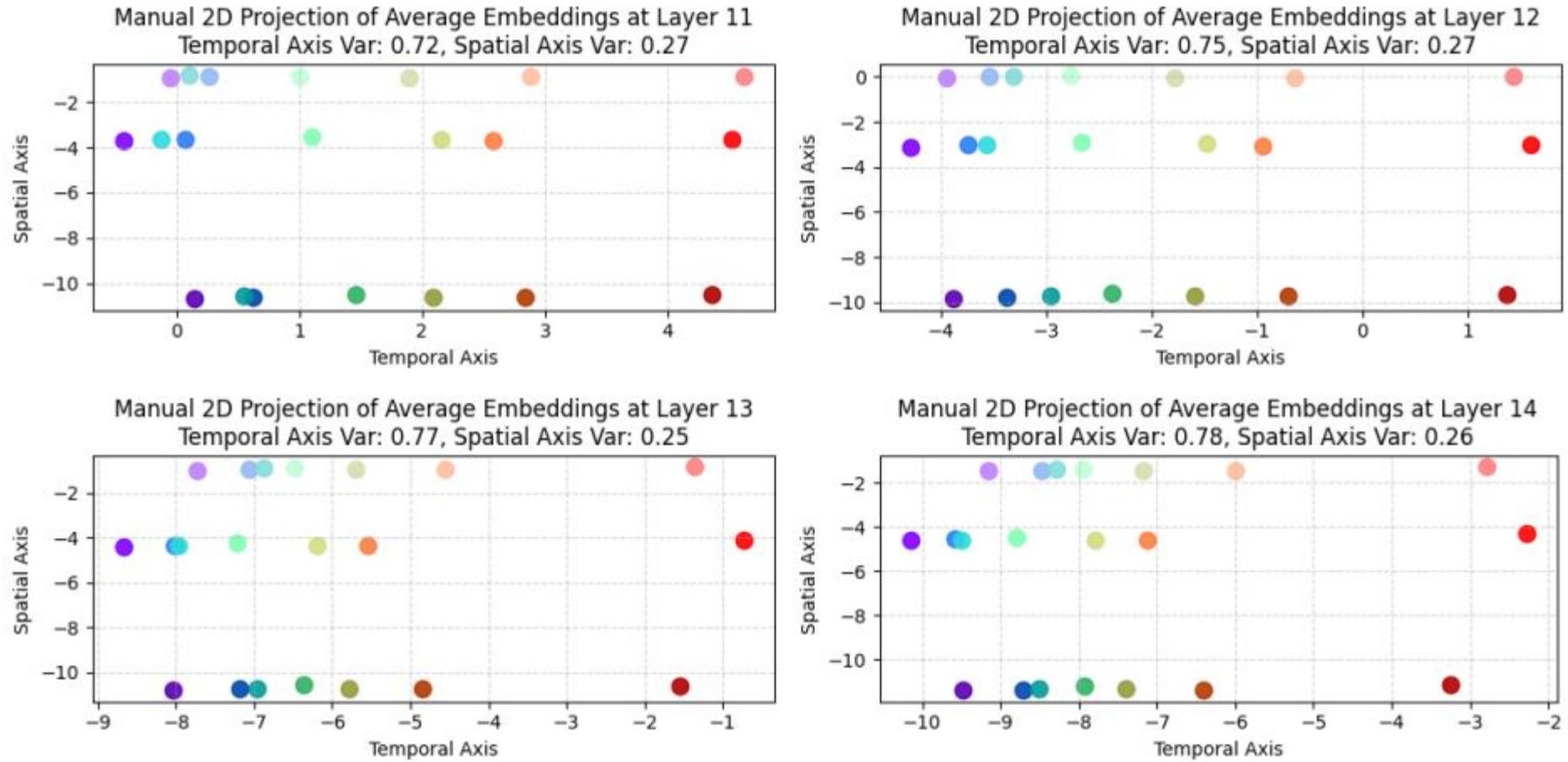


Figure A17: Spatiotemporal ID grid, where y axis is space and x axis is time.