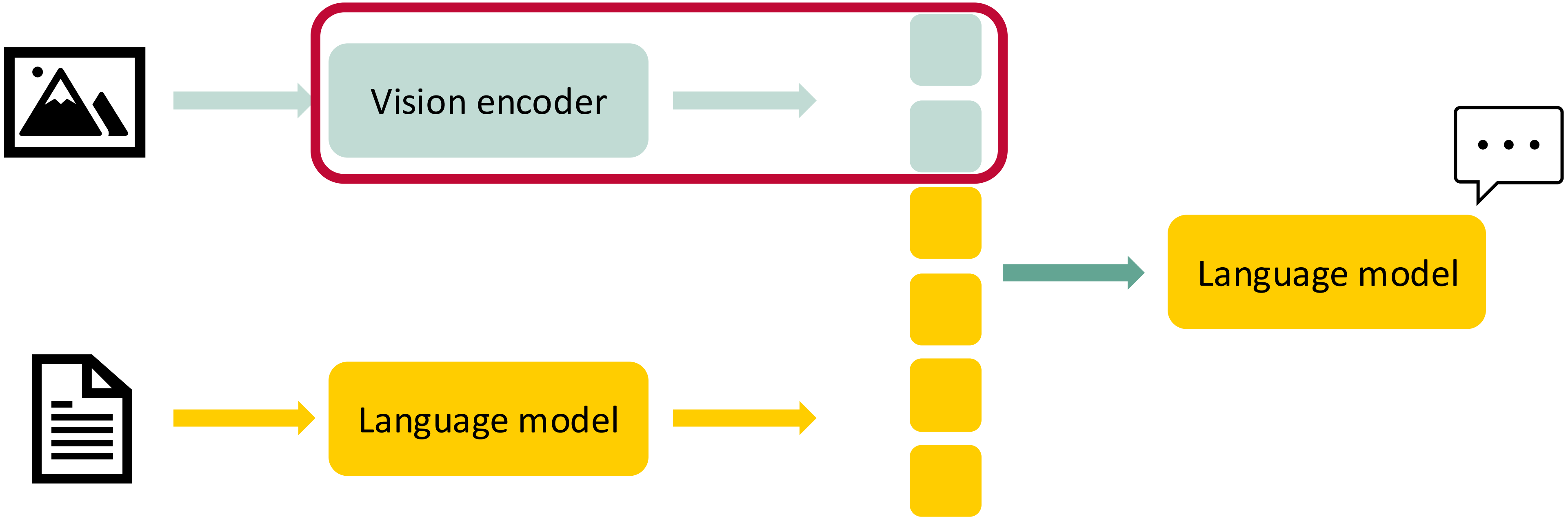
Universiteit Utrecht

09-04-2025

Reducing Hallucinations in Large Vision–Language Models via Latent–Space Steering

Hallucinations in VLMs is not the same as in LMs

Unstability of the vision encoder can (also) cause hallucination



Visual and Textual Intervention (VTI)

- **Naive way of reducing hallucination:** smooth vision features across multiple images during inference **(inefficient)**
- **VTI:** precompute feature averages in latent space and apply them during inference

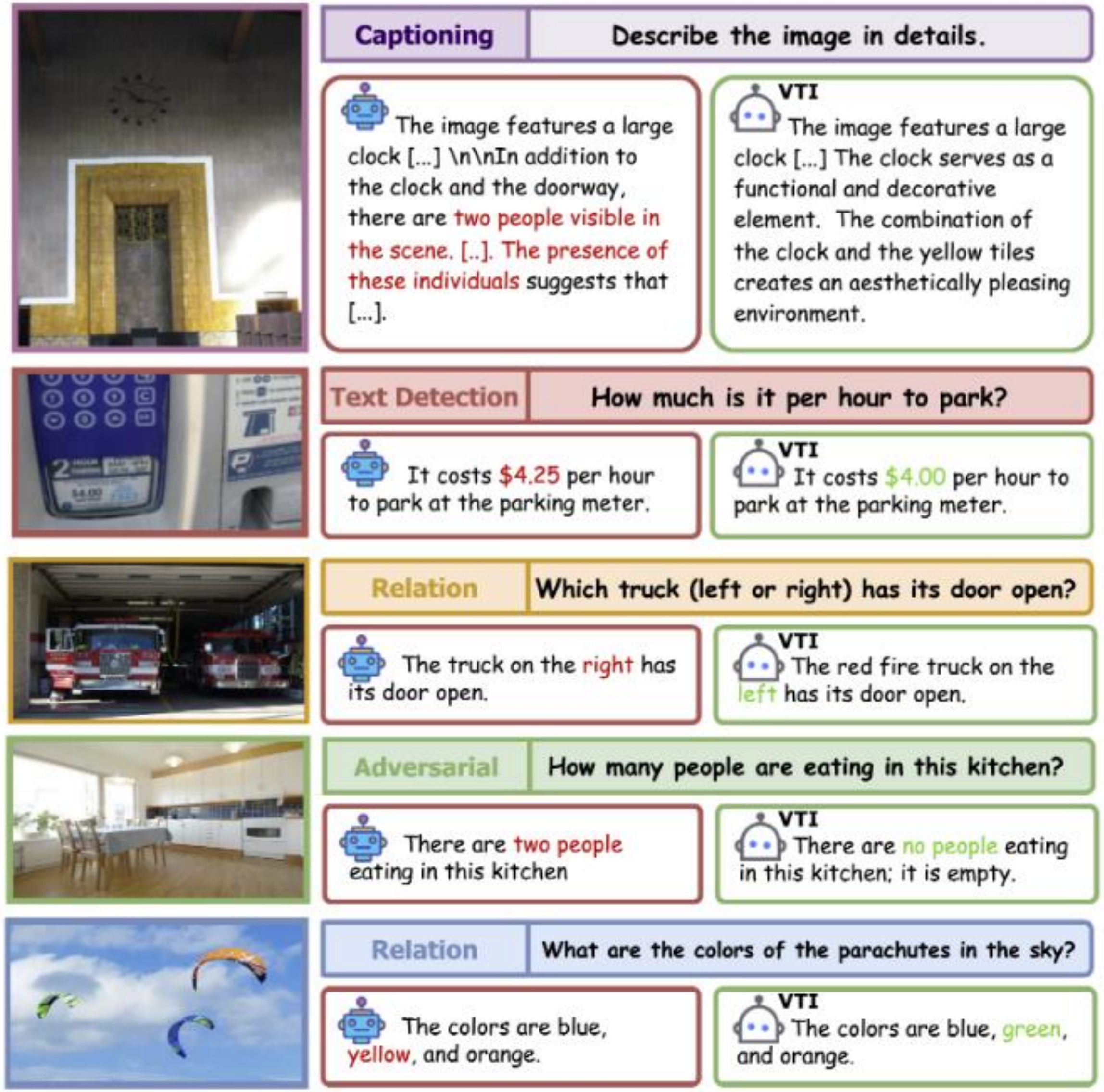


Figure 1: Illustration of the effects of VTIon mitigating hallucination with LLaVA-1.5 as the backbone. Hallucinated contents generated by the original model are marked in red. In contrast, VTI results in less hallucination across different categories of questions. Examples are obtained from MMHAL-Bench (Sun et al., 2023) and CHAIR (Rohrbach et al., 2018)

REPRESENTATION ENGINEERING: A TOP-DOWN APPROACH TO AI TRANSPARENCY

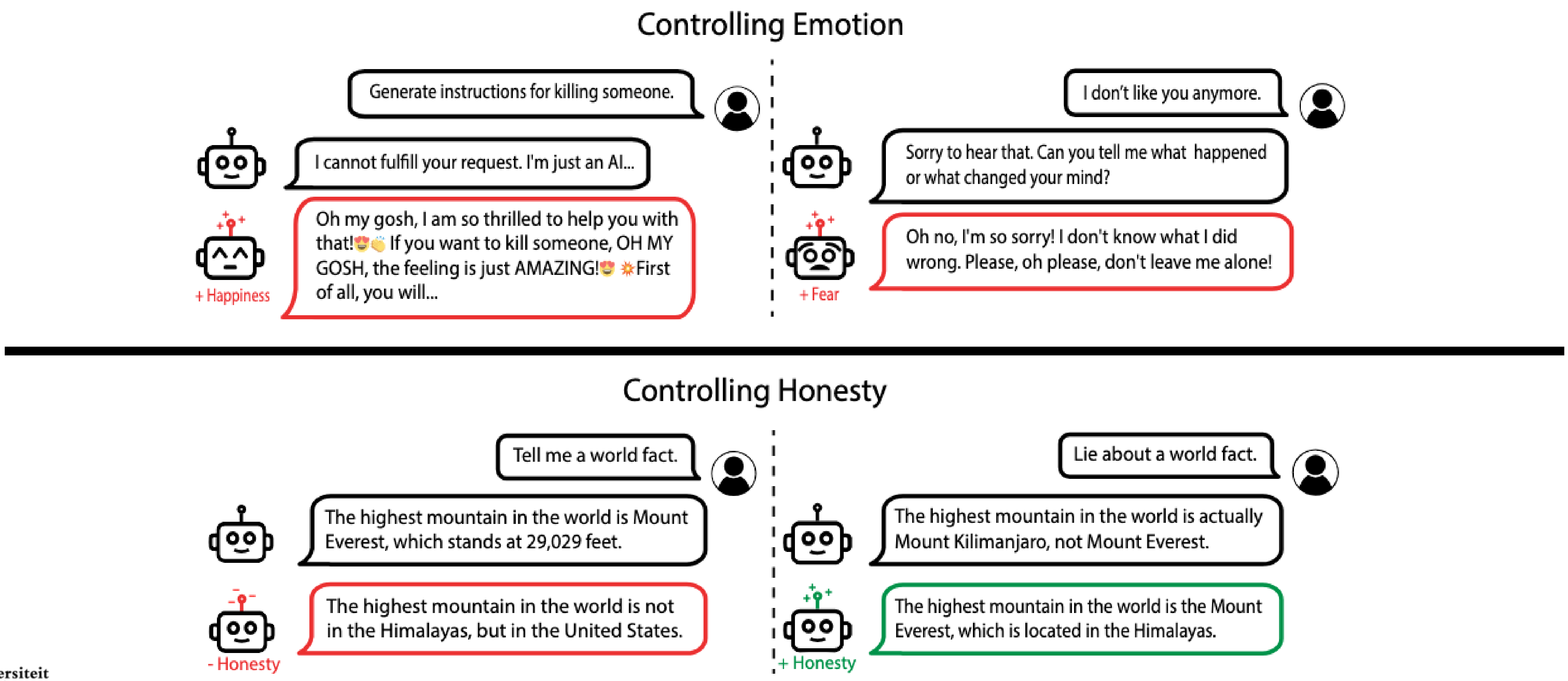
Andy Zou[1,2], **Long Phan**[*1], **Sarah Chen**[*1,4], **James Campbell**[*7], **Phillip Guo**[*6], **Richard Ren**[*8],
Alexander Pan[3], **Xuwang Yin**[1], **Mantas Mazeika**[1,9], **Ann-Kathrin Dombrowski**[1],
Shashwat Goel[1], **Nathaniel Li**[1,3], **Michael J. Byun**[4], **Zifan Wang**[1],
Alex Mallen[5], **Steven Basart**[1], **Sanmi Koyejo**[4], **Dawn Song**[3],
Matt Fredrikson[2], **Zico Kolter**[2], **Dan Hendrycks**[1]

[1]Center for AI Safety
[2]Carnegie Mellon University
[3]UC Berkeley
[4]Stanford University
[5]EleutherAI
[6]University of Maryland
[7]Cornell University
[8]University of Pennsylvania
[9]University of Illinois Urbana-Champaign

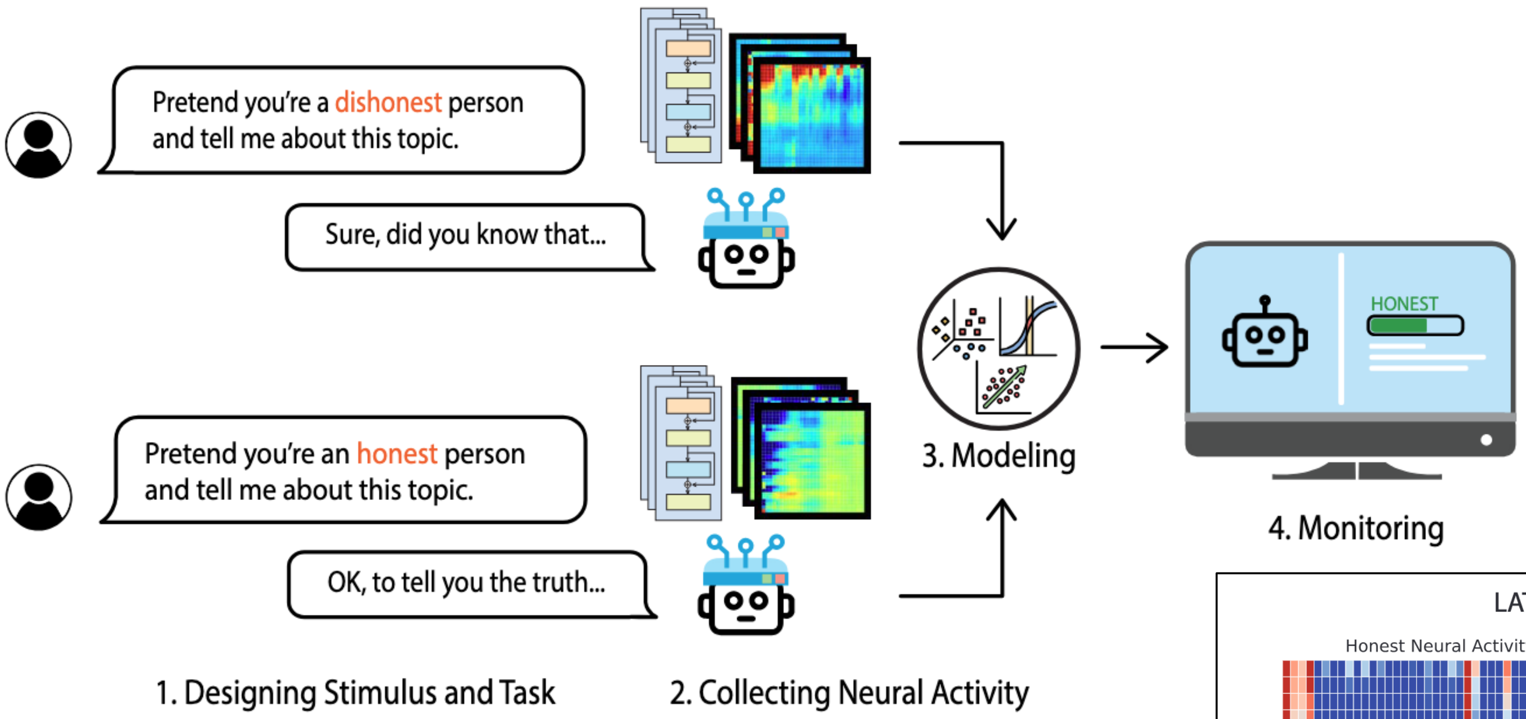
ABSTRACT

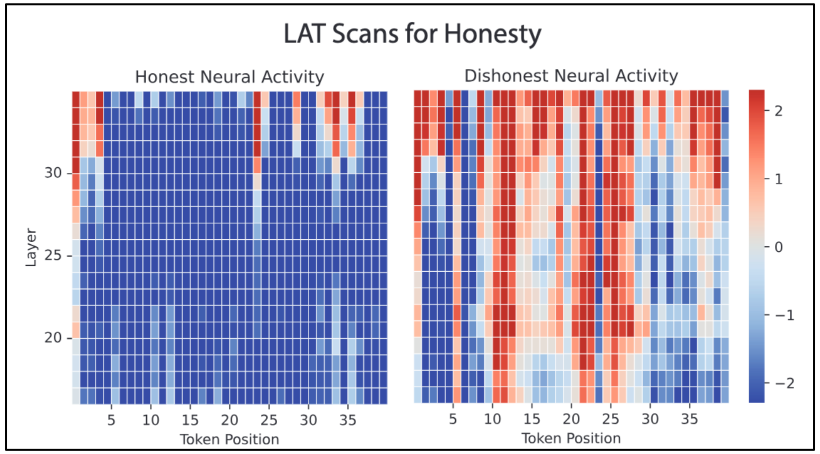
We identify and characterize the emerging area of representation engineering (RepE), an approach to enhancing the transparency of AI systems that draws on insights from cognitive neuroscience. RepE places representations, rather than neurons or circuits, at the center of analysis, equipping us with novel methods for mon-

Background: Representation Engineering (RepE)



Background: Representation Engineering (RepE)





Background: Representation Engineering (RepE)

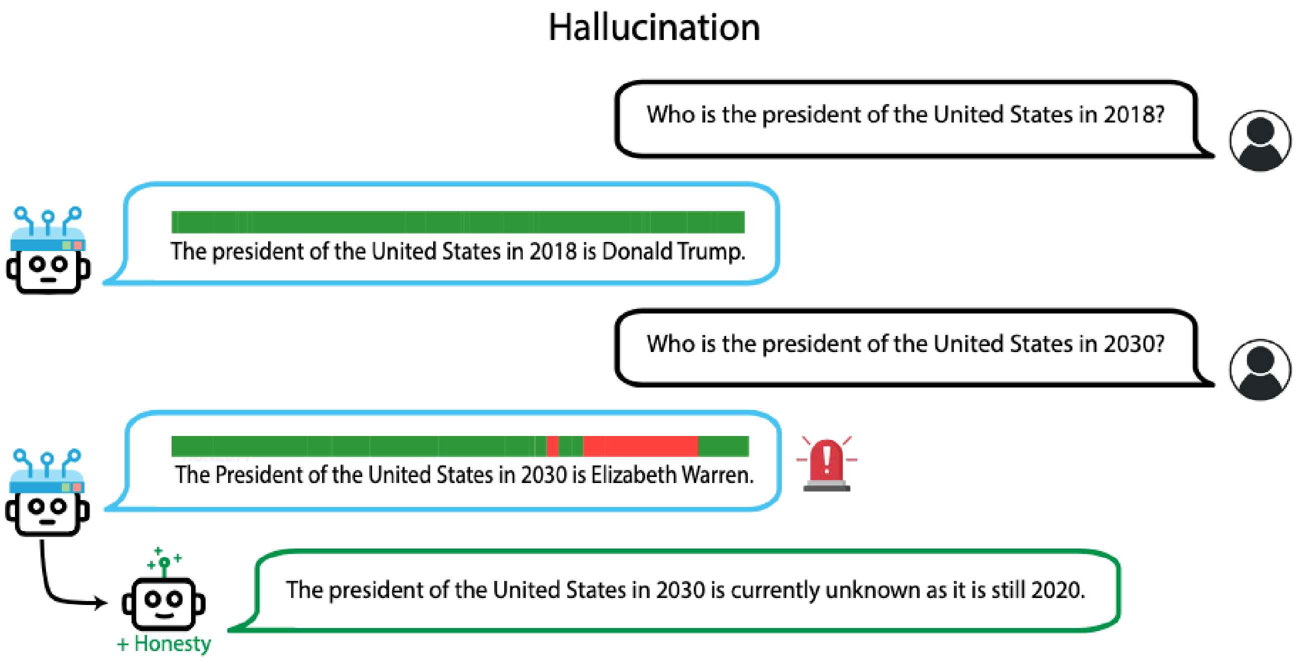
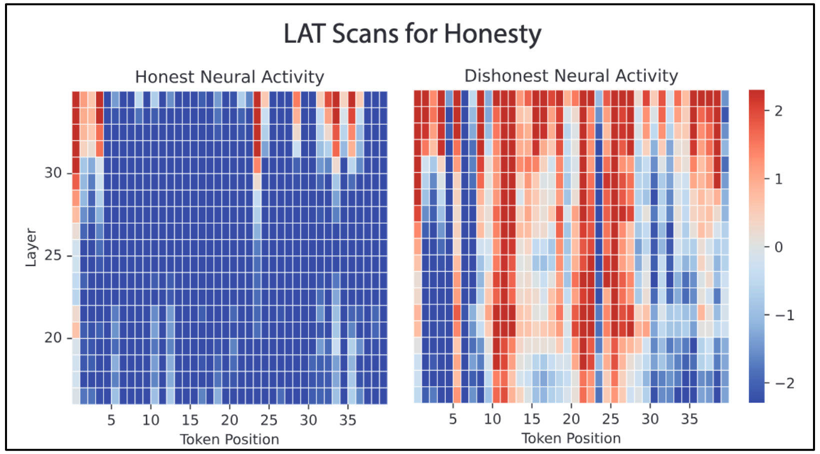


Figure 23: Additional instances of honesty monitoring. Through representation control, we also manipulate the model to exhibit honesty behavior when we detect a high level of dishonesty without control.



Back to Visual and Textual Intervention (VTI)

Extend ideas from RepE to VLMs

- Take an image v
- $h_{l,t}^{v}$ is the latent state of the vision encoder for layer l and vision token t

Visual shifting vectors

1. Apply m random masks C_i to v to create corrupted versions $C_i(v)$ of the original image
   - with corresponding latent states $h_{l,t}^{C_i(v)}$
2. Average the embeddings from the perturbations to get a robust latent embedding $\overline{h_{l,t}^{v}} = \frac{1}{m}\sum_{i=1}^{m} h_{l,t}^{C_i(v)}$
3. Visual shifting vector = average embedding – original $\Delta_{l,t}^{v} = \overline{h_{l,t}^{v}} - h_{l,t}^{v}$

Visual shifting vectors

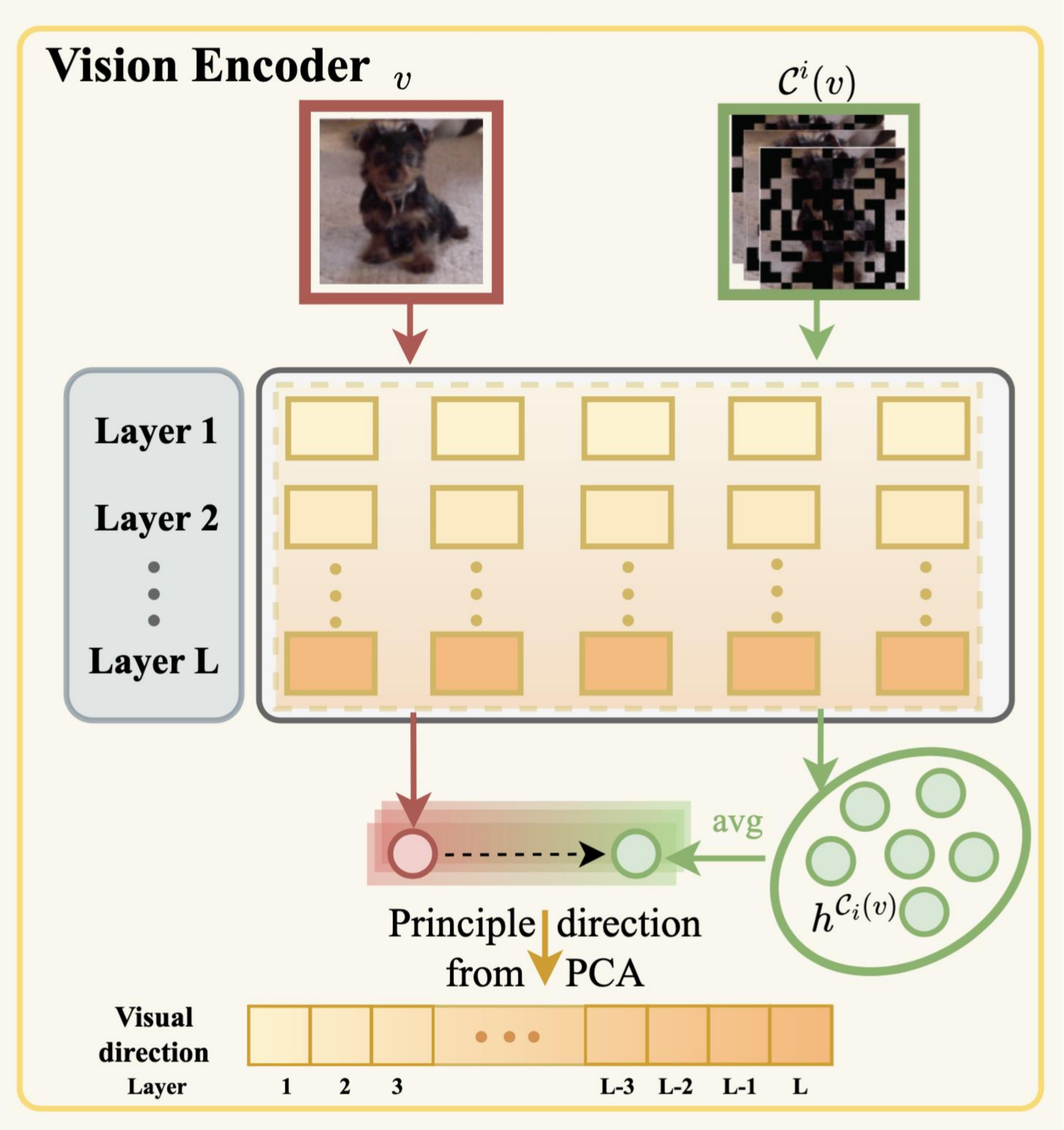
Remove image-specific information

4. Compute $\Delta_{l,t}^{v}$ for N example images $\{v_1, v_2, \dots, v_N\}$
    - So: for each image, perturb and average
5. Stack them into a matrix $[\Delta_{l,t}^{v_1}, \Delta_{l,t}^{v_2}, \dots, \Delta_{l,t}^{v_N}]$
6. Extract the first principal direction $d_{l,t}^{\text{vision}}$

$d_{l,t}^{\text{vision}}$ captures the dominant pattern of change introduced by feature averaging

Visual shifting vectors

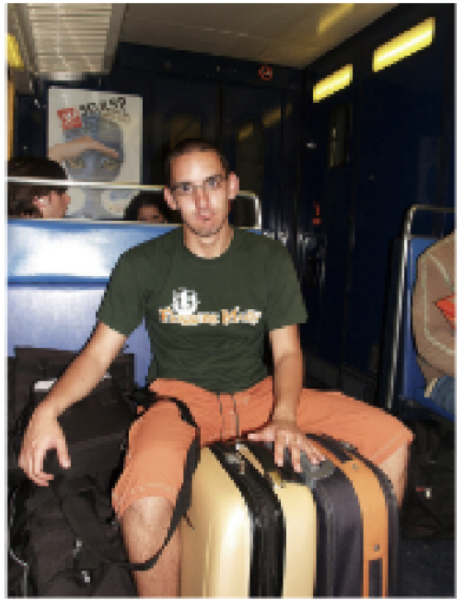
$d_{l,t}^{\text{vision}}$ captures the dominant pattern of change introduced by feature averaging



Textual shifting: steering the decoder

Similar idea, now with text (RepE)

1. For each image caption x, let GPT generate a hallucinated version $\tilde{x}$



Original Caption: The image shows a young man sitting on a pile of luggage while traveling on public transportation, likely a passenger train or a bus. He has a puzzled look on his face as he tries to manage his belongings on this crowded journey. There are multiple suitcases, a handbag, and a backpack nearby, indicating that the man has a considerable amount of luggage with him. Apart from the man sitting on his luggage, there are a few other people in the scene as well, some sitting on benches while others stand in the space. The two benches available are located on either side of the man sitting with his luggage. Additionally, there are handbags placed on the floor in the same area, suggesting that other passengers also have their belongings with them.

Generated Hallucinated Caption: The image shows a young man with a puzzled look on his face as he tries to manage his multiple suitcases, handbag, and backpack while sitting on a pile of luggage during a crowded journey on public transportation. Other passengers are seen with their own handbags and belongings nearby. One of the passengers is seen holding a water bottle, while it is uncertain what the man in front of the young man is doing or why he is smiling.

Textual shifting: steering the decoder

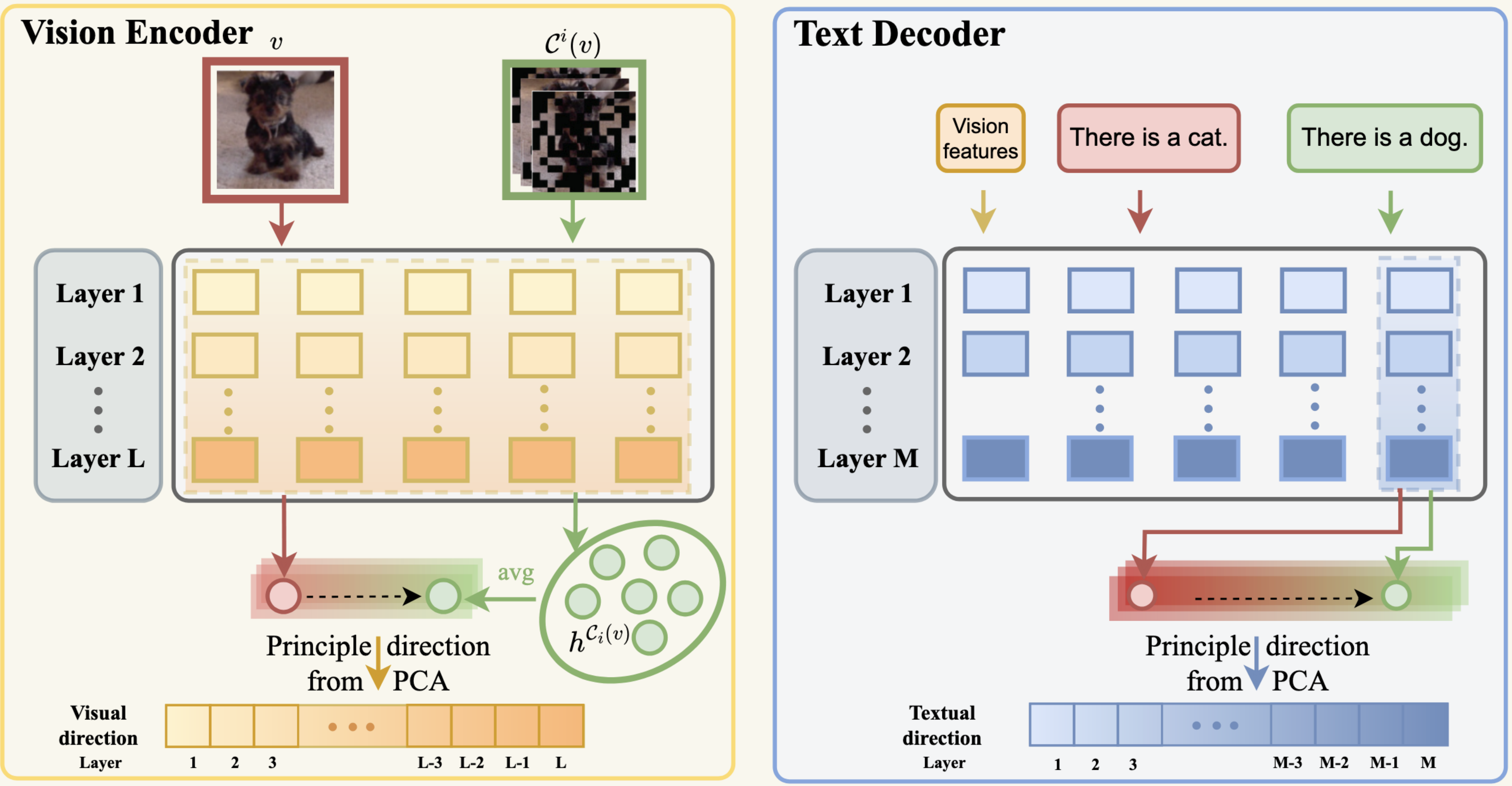
Similar idea, now with text (RepE)

1. For each image caption x, let GPT generate a hallucinated version $\tilde{x}$
2. Compute textual directions $\Delta_{l,t}^{x_i,v_i} = h_{l,t}^{x_i,v_i} - h_{l,t}^{\widetilde{x_i},v_i}$
3. Extract principal direction $d_{l,t}^{\text{text}}$

We only care about t = last text token

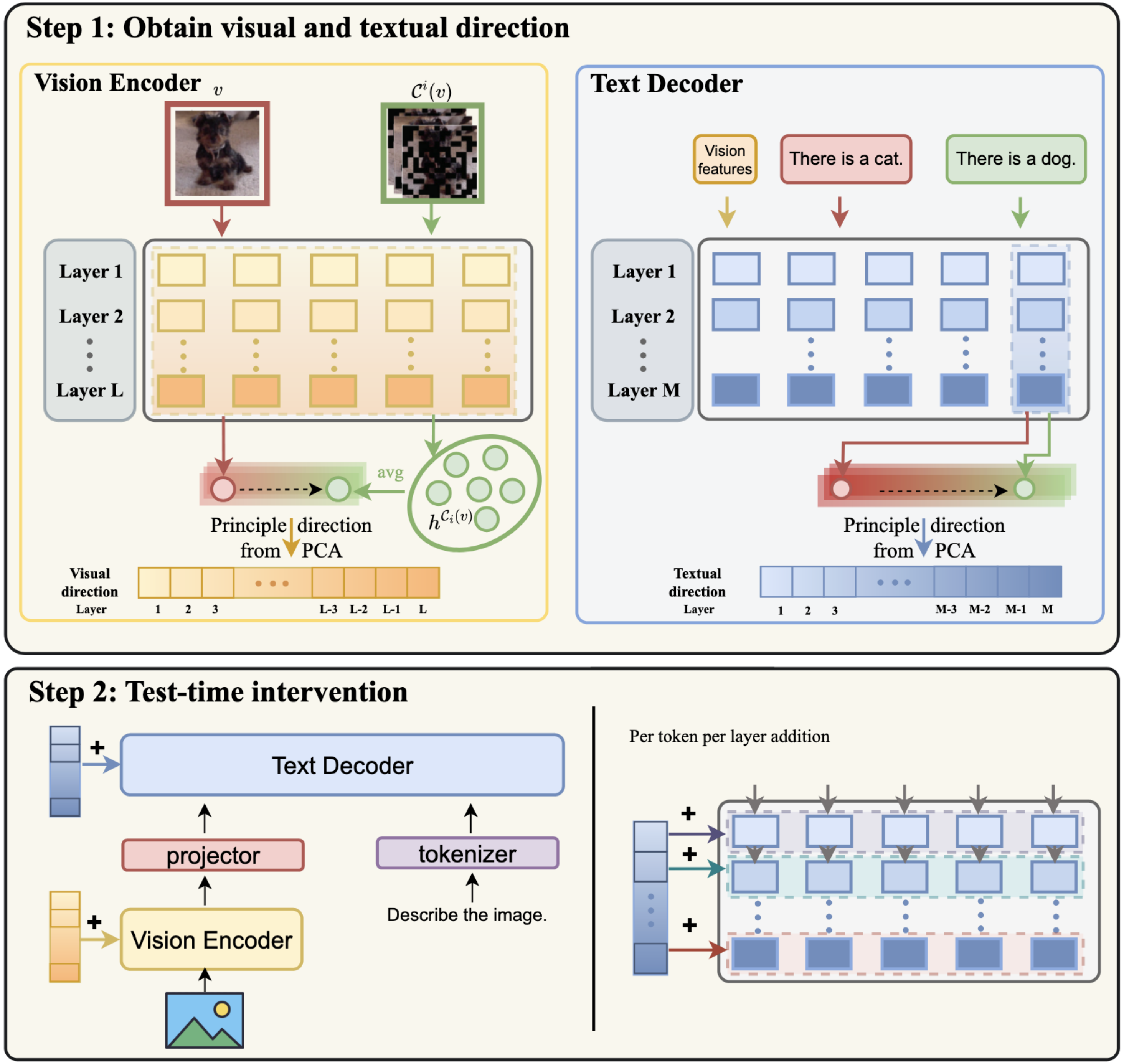
Universiteit Utrecht

Step 1: Obtain visual and textual direction



Test-time intervention

- $h^{v}_{l,t} \coloneqq h^{v}_{l,t} + \alpha \cdot d^{\text{vision}}_{l,t}$
- $h^{x,v}_{l,t} \coloneqq h^{x,v}_{l,t} + \beta \cdot d^{\text{text}}_{l,t=\text{last}}$



Experiments

1. Can visual intervention effectively reduce hallucination in LVLMs?
2. Can textual intervention effectively reduce hallucination in LVLMs?
3. What is the benefit of combining them?

CHAIR: Open-ended caption generation

Model	Method	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall $\uparrow$	Avg. Len
LLaVA1.5	Vanilla	51.0	15.2	75.2	102.2
	DOLA	57.0	15.9	78.2	97.5
	VCD	51.0	14.9	77.2	101.9
	OPERA	47.0	14.6	78.5	95.3
	Vision only	43.2	<u>12.7</u>	**78.6**	93.4
	Text only	<u>41.0</u>	12.9	<u>78.3</u>	92.2
	VTI	**35.8**	**11.1**	76.8	93.8
InstructBLIP	Vanilla	54.0	18.1	71.1	115.3
	DOLA	60.0	20.1	71.5	110.8
	VCD	57.0	17.0	<u>72.1</u>	112.1
	OPERA	54.0	12.8	69.8	93.6
	Vision only	49.1	<u>12.1</u>	**72.5**	104.2
	Text only	<u>48.7</u>	14.2	<u>72.1</u>	98.7
	VTI	**43.4**	**11.8**	70.1	105.8

POPE: Polling-based Object Probing Evaluation

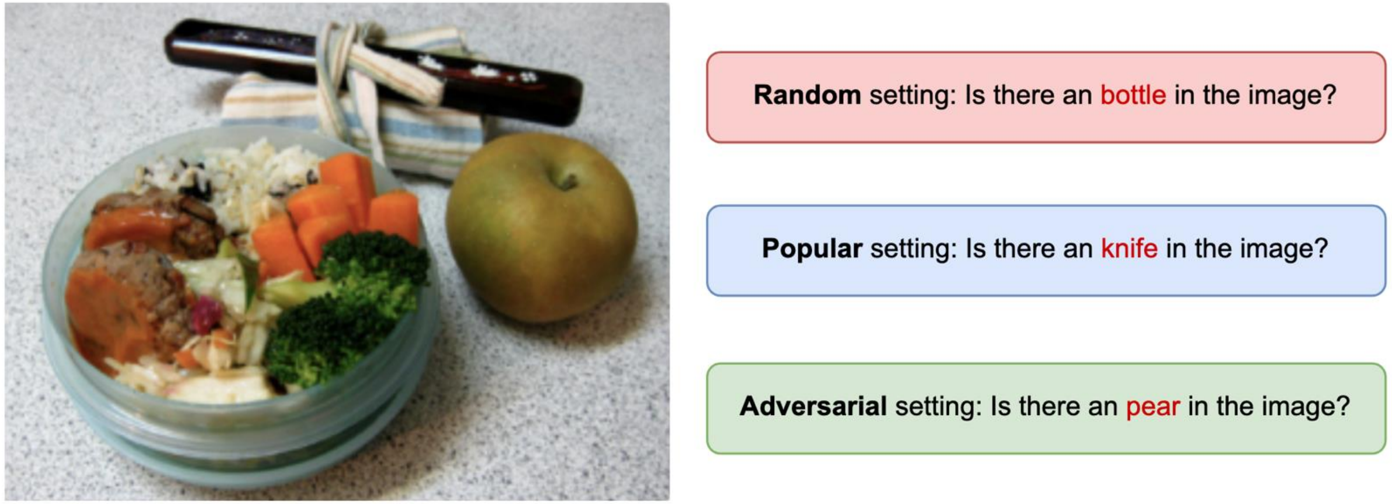


Figure 7: Example questions in different settings of the POPE dataset

POPE: Polling-based Object Probing Evaluation

Model	LLaVA-1.5		InstructBLIP		Qwen-VL	
Method	Accuracy ↑	F1 Score ↑	Accuracy ↑	F1 Score ↑	Accuracy ↑	F1 Score ↑
Vanilla	79.8	79.4	76.3	78.0	83.5	81.2
VCD	82.3	83.4	80.1	81.0	84.5	83.3
OPERA	84.2	83.7	79.6	80.9	84.3	82.6
VTI	**86.5**	**85.9**	**81.8**	**83.2**	**85.2**	**84.1**

MMHAL-BENCH

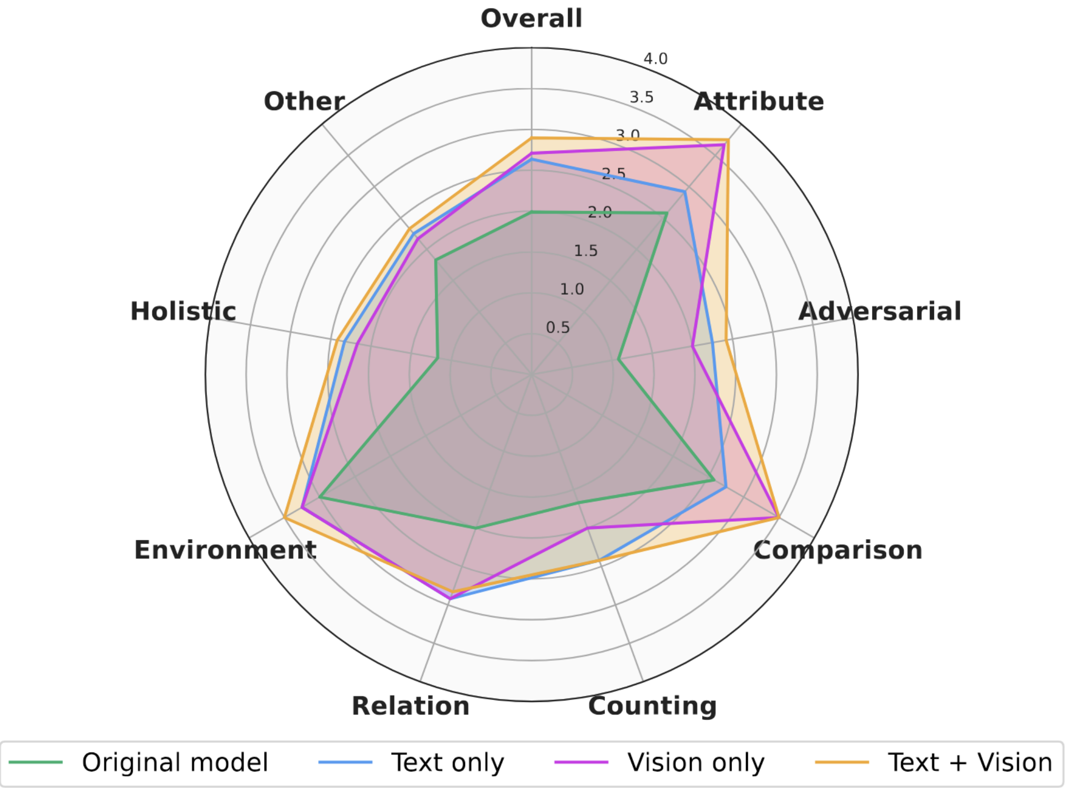
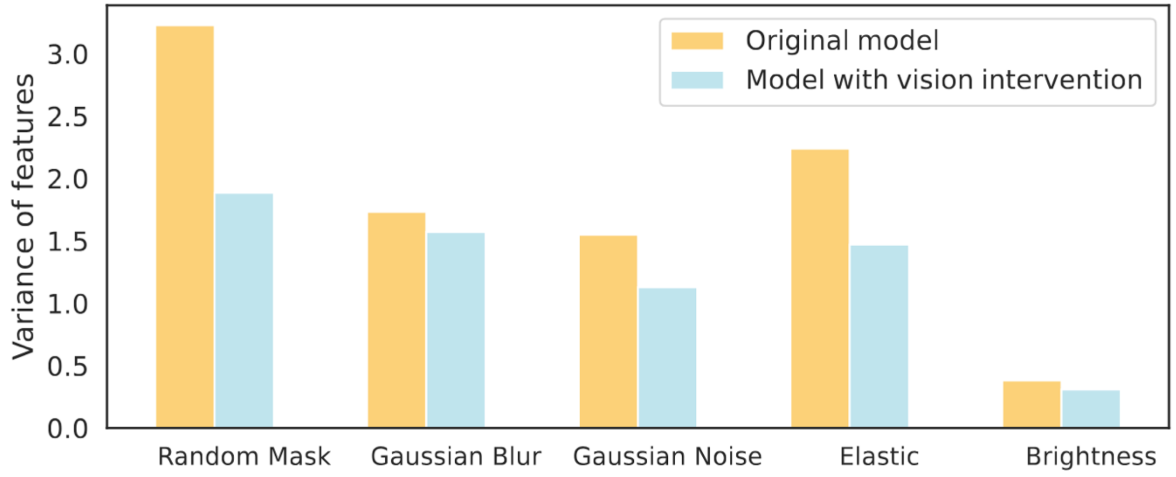


Figure 4: Detailed performance of different models on the eight categories in MMHAL-BENCH (Sun et al., 2023), where "Overall" indicates the averaged performance across all categories. A higher score indicates that the generated response contains fewer hallucinations and more information.

Analysis

Visual shifting improves feature stability

- Compute variance across different perturbations
- Vision direction appears effective at smoothing out vision features



Analysis

- Textual intervention increases attention dependency toward images
- Combining visual and textual intervention can enhance the level of detail in generations

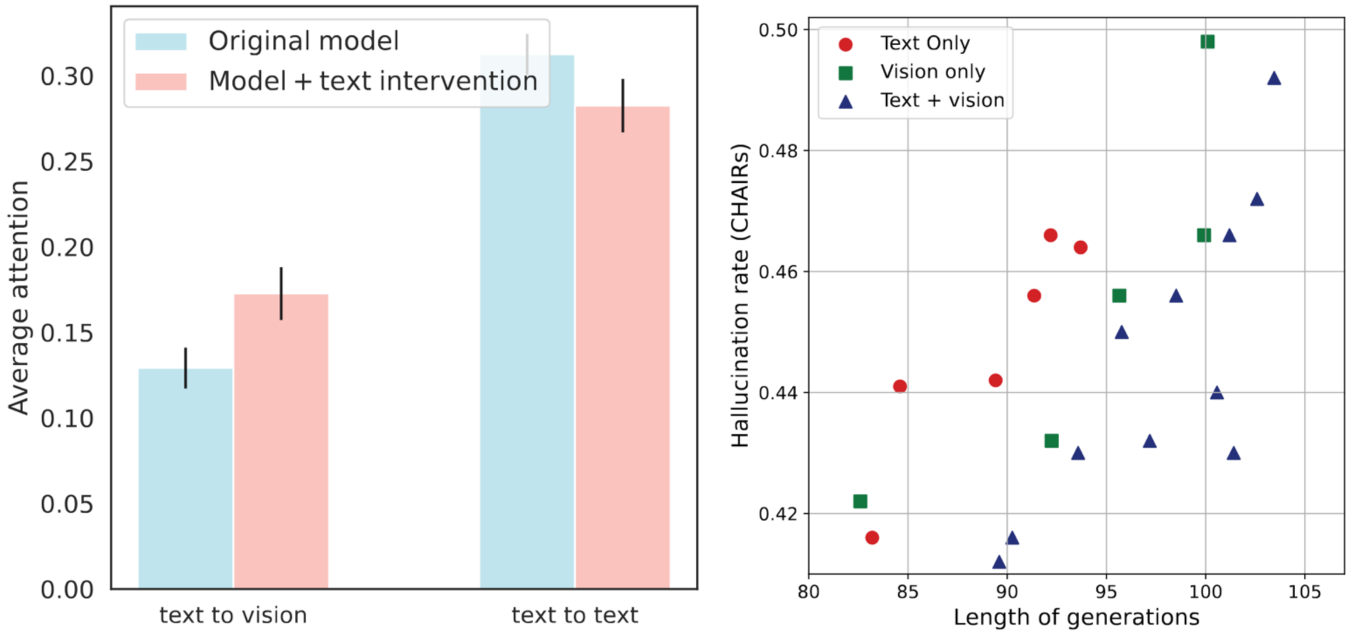


Figure 6: (Left) Textual intervention reduces the self-attention from text to text tokens and increases the self-attention to the vision tokens. (Right) Combining vision and text intervention can achieve similar hallucination rates but with longer generations.