Multimodal Mechanistic Interpretability, CAUSAL TRACING

Utrecht University

A DEFINITION

Interpretability in machine learning, LLMs, and multimodal models is:

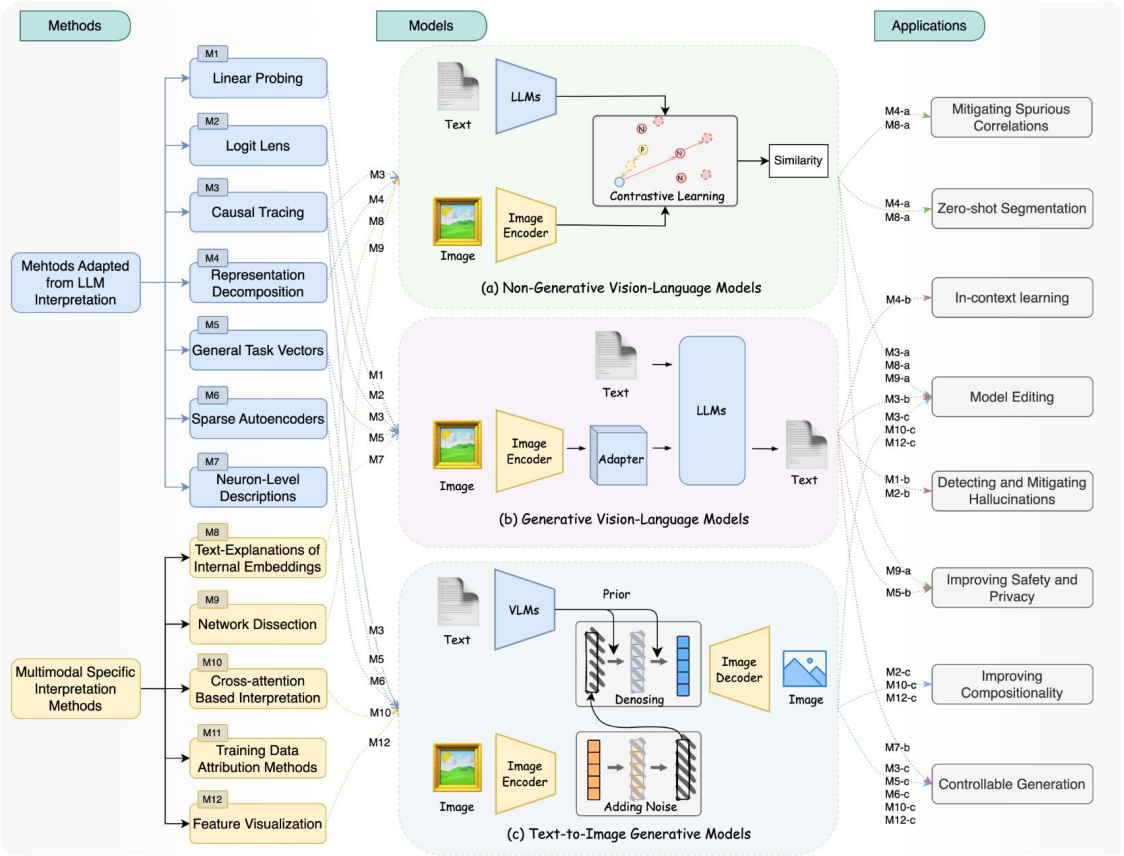
"*The process of extracting and elucidating the **relevant knowledge**, **mechanisms**, **features**, and **relationships** a model has learned, whether **encoded in its parameters** or **emerging from input patterns**, to explain how and why it produces outputs.*"

W James Murdoch, et al. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44):22071–22080.

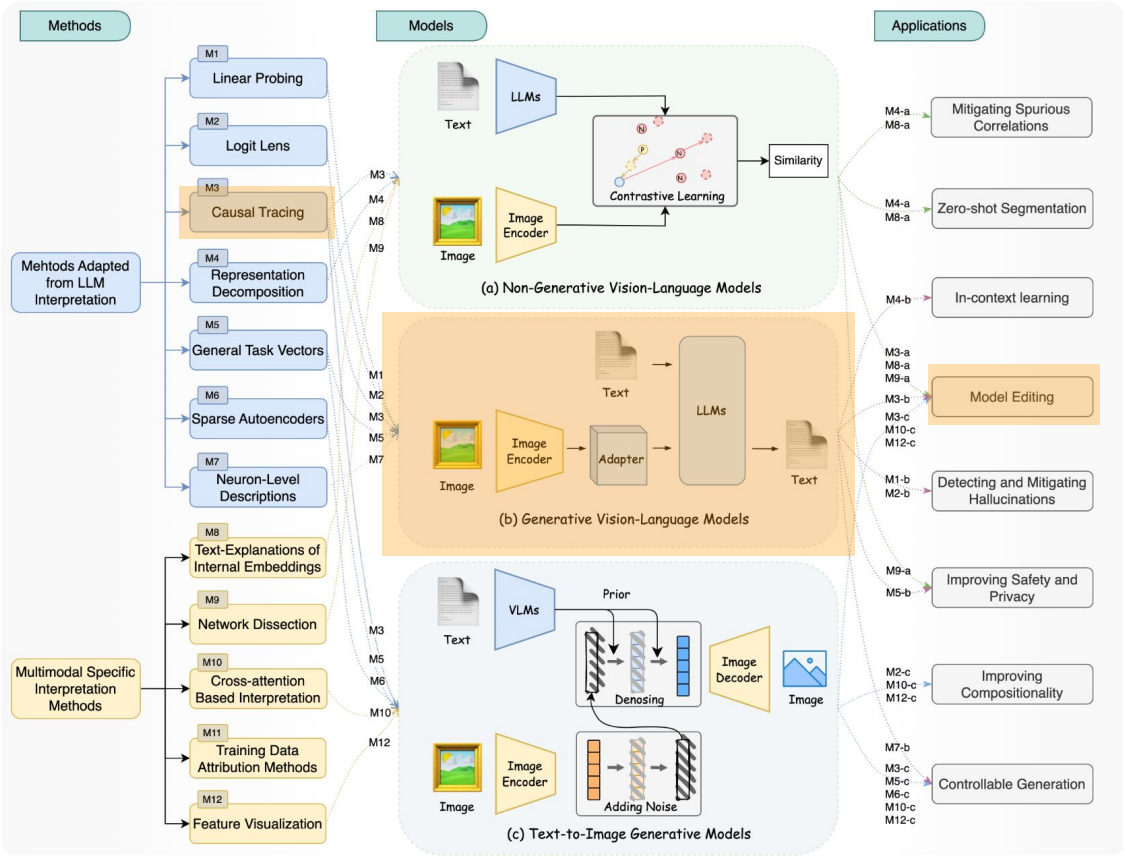
relevant knowledge depends on the application:

- memory editing: precise modifications to internal representations without disrupting other model functions
- security: highlight input features and activations that signal adversarial inputs
- …

Lin, Z., et al. (2025). A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models (arXiv:2502.17516). arXiv. https://doi.org/10.48550/arXiv.2502.17516



Lin, Z., et al. (2025). A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models (arXiv:2502.17516). arXiv. https://doi.org/10.48550/arXiv.2502.17516



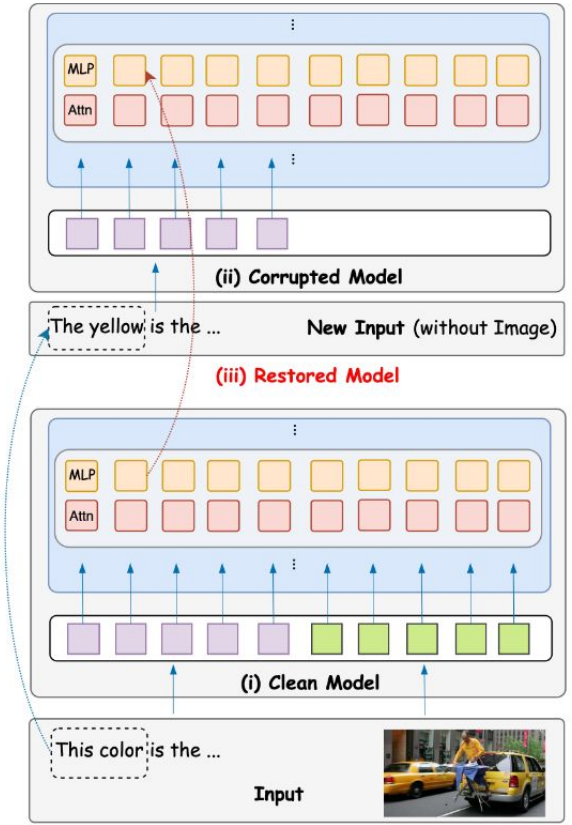
Lin, Z., et al. (2025). A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models (arXiv:2502.17516). arXiv. https://doi.org/10.48550/arXiv.2502.17516

Causal Tracing/Causal Mediation Analysis

Causal intervention methods

actively perturb/corrupt model states to uncover where the knowledge is stored

Studies the change in a response variable following an active intervention on intermediate variables of interest (mediators).



(c) Causal Tracing

Lin, Z., et al. (2025). A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models (arXiv:2502.17516). arXiv. https://doi.org/10.48550/arXiv.2502.17516

Understanding Information Storage and Transfer in Multi-modal Large Language Models

Samyadeep Basu* University of Maryland
Martin Grayson Microsoft Research
Cecily Morrison Microsoft Research
Besmira Nushi Microsoft Research
Soheil Feizi University of Maryland
Daniela Massiceti Microsoft Research

MULTIMODALCAUSALTRACE

The model must retrieve information that satisfies a **constraint** (set of words in question) from its parametric memory in order to generate the correct answer.

“**This place** [visual constraint] is located in the continent of? + <image>

(Vinson Massif)

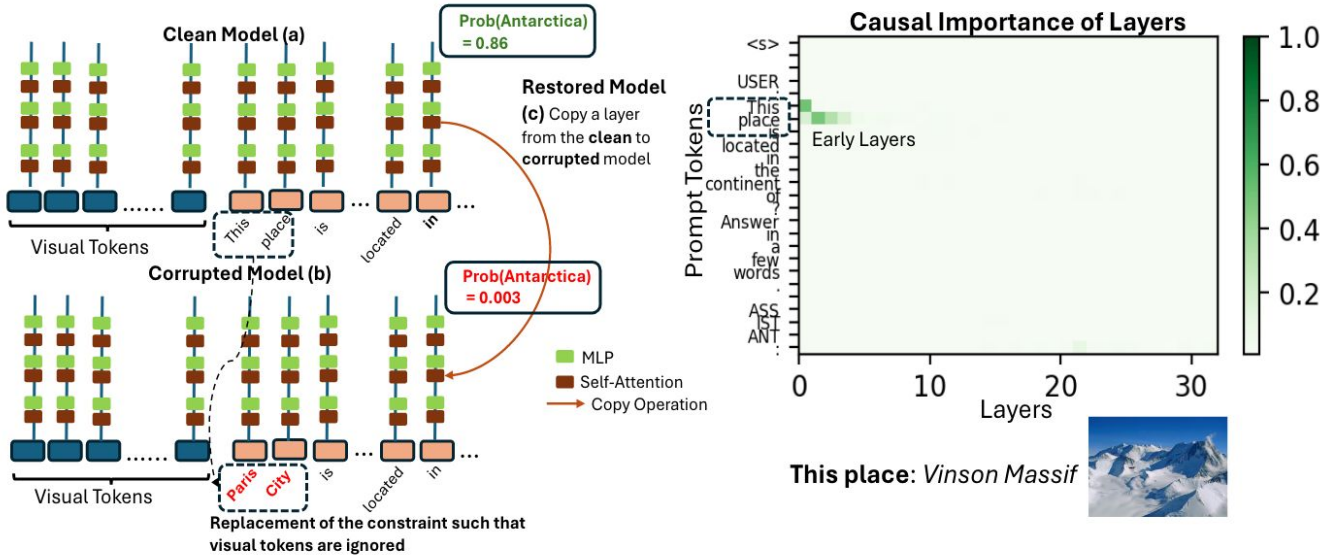
Corruption: replacing the visual constraint token IDs (e.g. this place) with token IDs from a separate word or phrase (e.g. Paris City), such that the visual information is ignored.

Basu, S., Grayson, M., Morrison, C., Nushi, B., Feizi, S., & Massiceti, D. (2024). Understanding Information Storage and Transfer in Multi-modal Large Language Models (arXiv:2406.04236). arXiv. https://doi.org/10.48550/arXiv.2406.04236

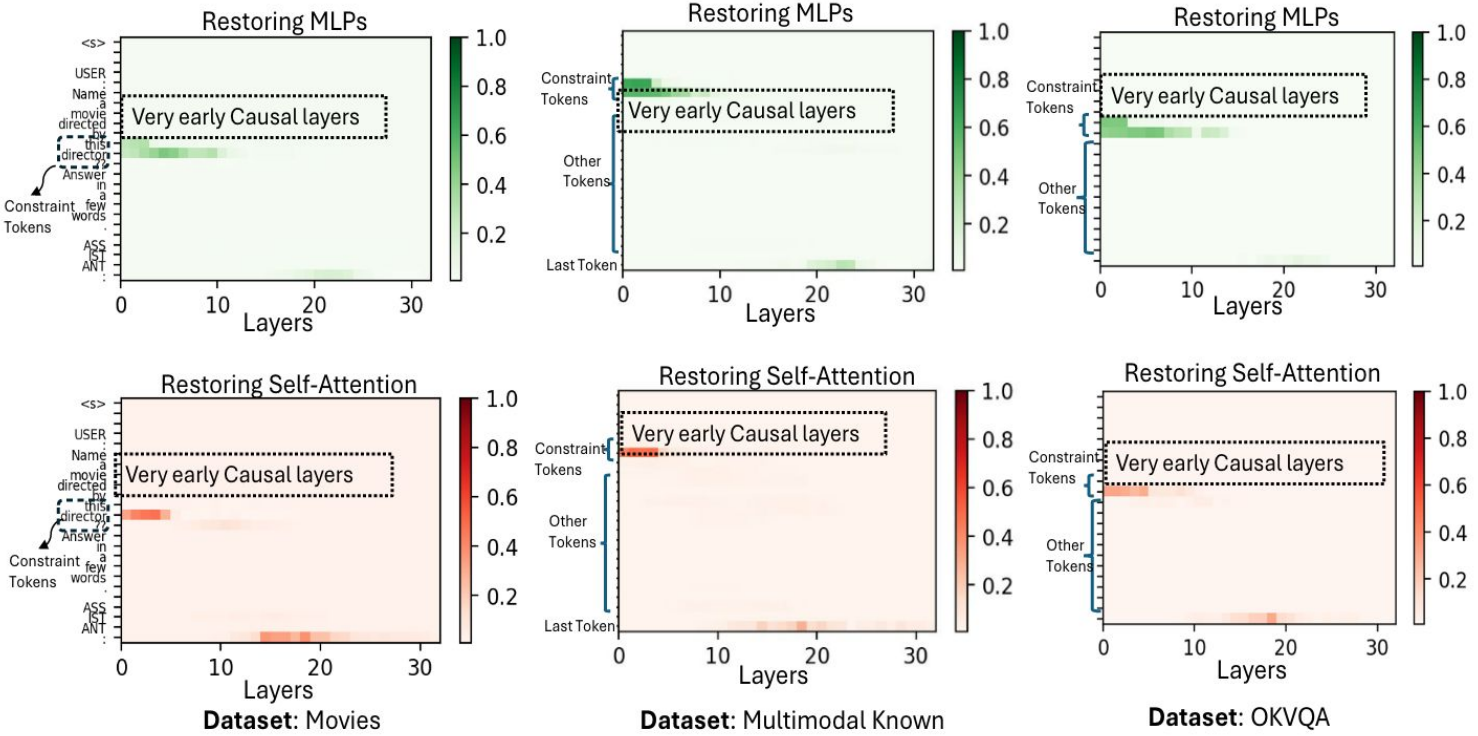
1. **Corruption**
2. **Inference**: Record probability of model's output O: P(O) and then P(O)corrupted
3. **Restoration**: iteratively copy $g\varphi(.)_{k,\ell,attn}$ and $g\varphi(.)_{k,\ell,mlp}$ from the clean model to the corrupted model.

   gφ(.)k,ℓ : the output layer embedding corresponding to the ***k*th** token position and the layer ℓ, with a varying window size (n of ℓ selected).

4. **Indirect Effect estimation**: ΔP(O)layer = P(O)restored − P(O)corrupted; identifies layers that store information



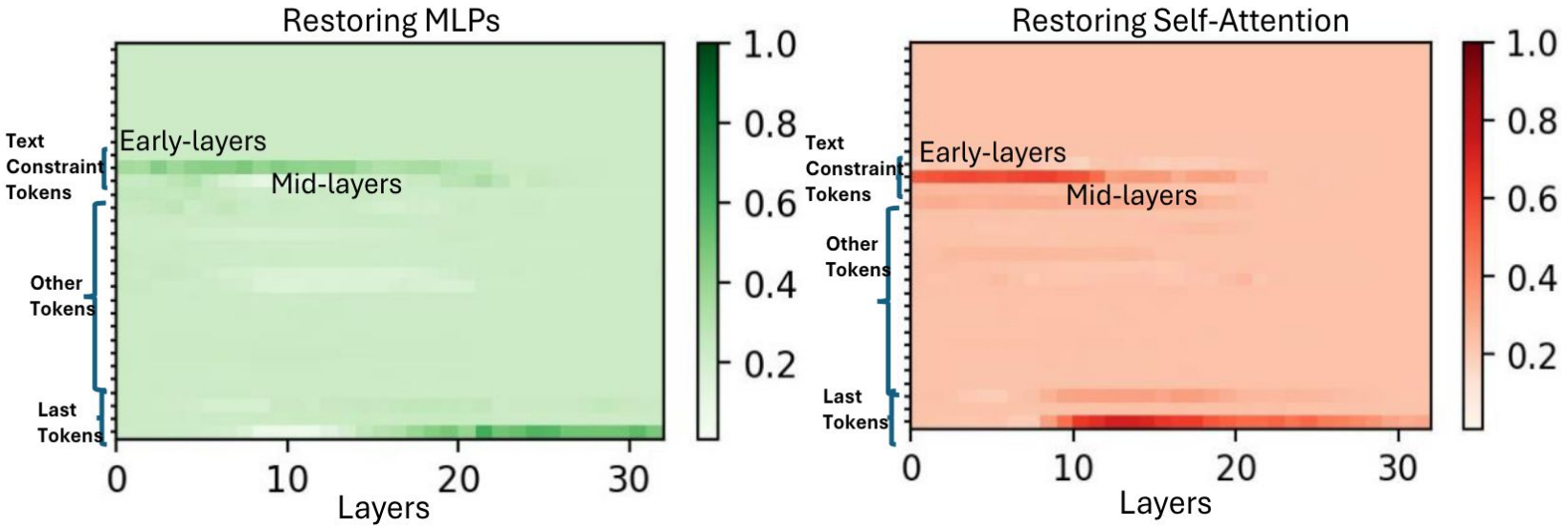
Information required to answer a visual question is mainly retrieved from the early-layer MLP and self-attention blocks of a MLLM (the causal traces emerge with a window size of 3), corresponding to the visual constraint



Mult-constraint questions: given an image of Christopher Nolan and the question “Name a movie directed by this director (visual constraint) in 2006 (text constraint)?”

information corresponding to the textual constraint is retrieved from a broader set of layers – both the early and mid-layer MLP and self-attention blocks.

A larger window size (at least of 6) is required to obtain any causal traces. More parametric memory is required to meet both a visual and textual constraint in a given question.

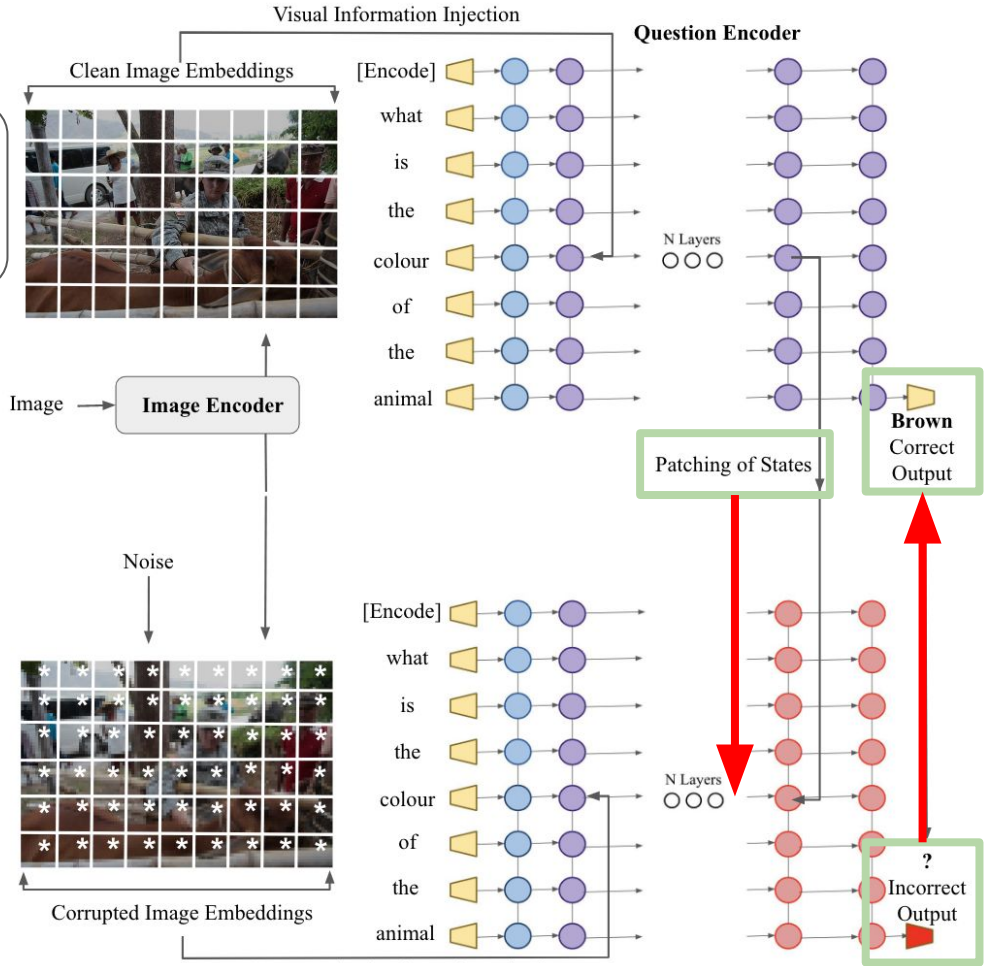


Towards Vision-Language Mechanistic Interpretability: A Causal Tracing Tool for BLIP

Vedant Palit* IIT Kharagpur
Rohan Pandey* Reworkd.ai
Aryaman Arora Georgetown University
Paul Pu Liang Carnegie Mellon University

Corruption: inject gaussian noise in the image embeddings before they are fed into the question encoder.

To perform the causal intervention, the output of each individual state (layer L, token T) of the E∗ question encoder run is overwritten with the corresponding state from the clean image embedding run E

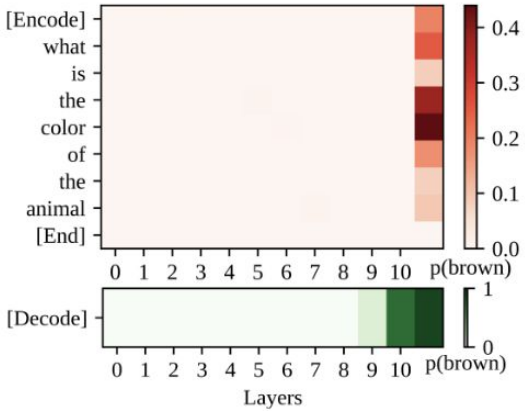


Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. 2023. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pages 2856–2861.

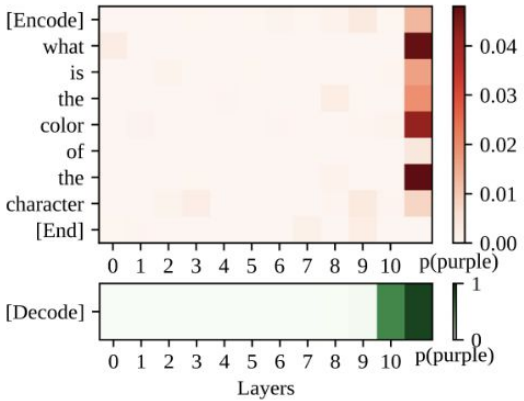
Question encoder, only the final layer (11) for all tokens plays a significant role in affecting the output to a higher degree
Answer decoder, the final layers (9 to 11) play the most apparent role in the final output of the model.

BLIP does not benefit from restored access to the correct image embeddings until the final few layers.

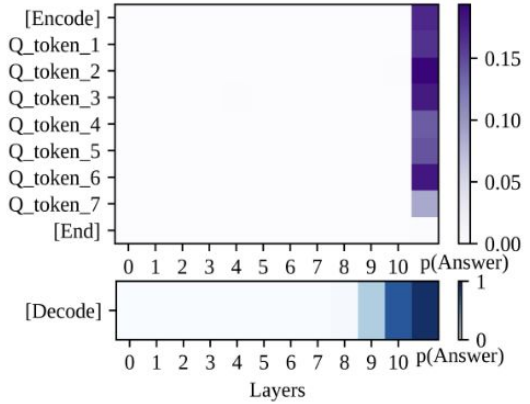
- The vision modality is not relevant to model computations until the final layer, i.e. **vision and language are processed independently in the intermediate layers**.
- The **final layers override preceding layers**, which may still be weakly causally relevant to the model output



(a) COCOQA-ID458864



(b) COCOQA-ID220218



(c) Average over 200 samples

TOWARDS BEST PRACTICES OF ACTIVATION PATCHING IN LANGUAGE MODELS: METRICS AND METHODS

Fred Zhang[*]
UC Berkeley
z0@berkeley.edu

Neel Nanda
Independent
neelnanda27@gmail.com

Input corruption with Gaussian Noise (GN) may NOT be the best technique to perform causal tracing:

Recent works have demonstrated that using **Gaussian noise corruption may produce illusory results**, advocating instead for **symmetric token replacement (STR)** as it ensures the corrupted run is within distribution and provides more reliable insights.

STR: "A **child** crossing the street" → "A **lady** crossing the street"

Zhang, F., & Nanda, N. (2024). Towards Best Practices of Activation Patching in Language Models: Metrics and Methods (arXiv:2309.16042). arXiv. http://arxiv.org/abs/2309.16042

What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline for Gaussian-Noise-free Text-Image Corruption and Evaluation

Michal Golovanesky [*1], **William Rudman** [*1],
Vedant Palit[2], **Ritambhara Singh**[1], **Carsten Eickhoff**[4]
[1] Brown University, [2] IIT Kharagpur, [3] University of Tübingen
{michal_golovanevsky, william_rudman}@brown.edu

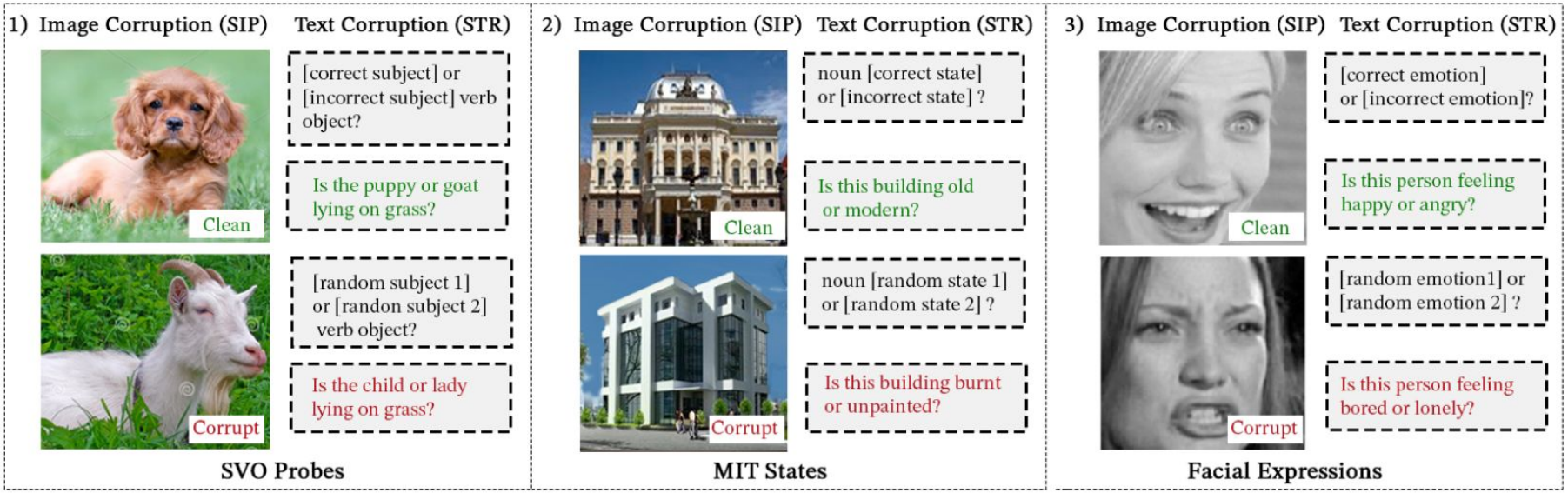
Text Corruption – Symmetric Token Replacement (STR):

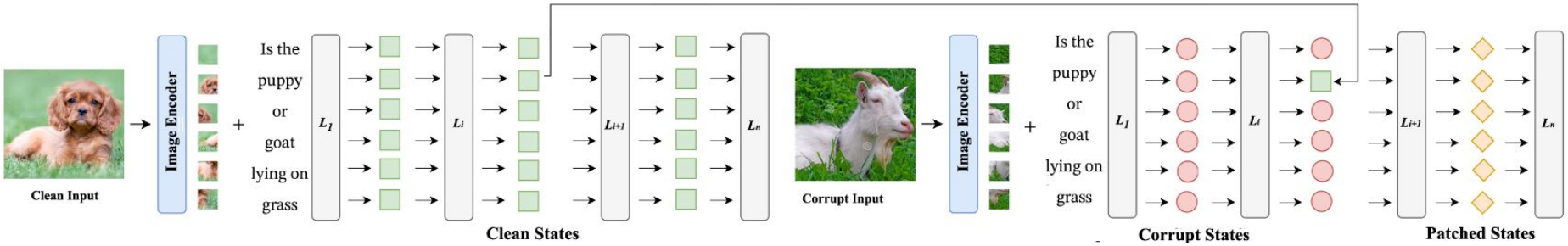
STR replaces correct tokens in the input with similarly tokenized but incorrect alternatives. This method preserves token distribution while preventing the model from solving the task correctly. For VQA, both correct and incorrect answer choices are replaced with sampled alternatives from the dataset.

Image Corruption – Semantic Image Pairs (SIP):

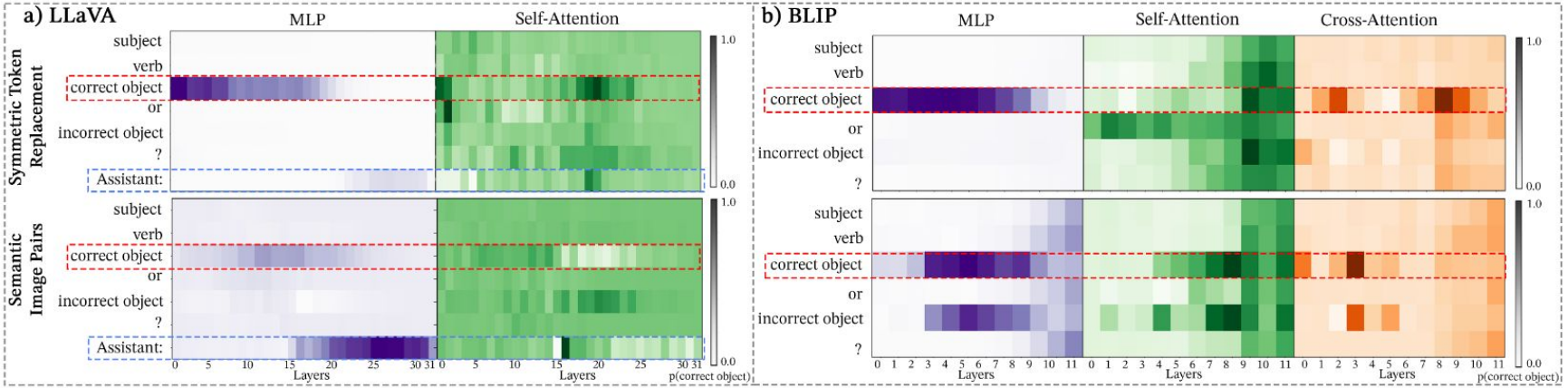
SIP is the visual counterpart of STR. It involves constructing semantically aligned image pairs that differ in only one key concept. SIP avoids unnatural perturbations (like noise or pixel masking) by using real, conceptually controlled images across datasets (SVO-Probes, MIT States, Facial Expressions).

Golovanevsky, M., Rudman, W., Palit, V., Eickhoff, C., & Singh, R. (2025). What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline for Gaussian-Noise-free Text-Image Corruption and Evaluation. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 11462–11482). Association for Computational Linguistics. https://aclanthology.org/2025.naacl-long.571/


1) Image Corruption (SIP) Text Corruption (STR)
Clean
[correct subject] or [incorrect subject] verb object?
Is the puppy or goat lying on grass?
Corrupt
[random subject 1] or [randon subject 2] verb object?
Is the child or lady lying on grass?
SVO Probes
2) Image Corruption (SIP) Text Corruption (STR)
Clean
noun [correct state] or [incorrect state] ?
Is this building old or modern?
Corrupt
noun [random state 1] or [random state 2] ?
Is this building burnt or unpainted?
MIT States
3) Image Corruption (SIP) Text Corruption (STR)
Clean
[correct emotion] or [incorrect emotion]?
Is this person feeling happy or angry?
Corrupt
[random emotion1] or [random emotion 2] ?
Is this person feeling bored or lonely?
Facial Expressions


Clean Input
Image Encoder
Is the puppy or goat lying on grass
L1
Li
Li+1
Ln
Clean States
Corrupt Input
Image Encoder
Is the puppy or goat lying on grass
L1
Li
Li+1
Ln
Corrupt States
Patched States

Module-wise activation patching: MLPs, self-attention, cross-attention



Observation	Findings
1. MLP Patching	Image corruption emphasizes **middle layers**; text corruption affects **early to middle layers**.
2. Self-Attention (Text Corruption)	LLaVA localizes attention on the **correct object token** in **layer 0 and 21**, similar to **BLIP's cross-attention** at **layer 3**.
3. Self-Attention (Image Corruption)	LLaVA attends **away from the correct answer token** in **layers 16–25**.
4. Answer Representation Shift	Early LLaVA layers attend to the **correct answer**; later layers shift focus to **instruction token ("Assistant:")**.

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
zhangzhizz2626@gmail.com, srya@di.ku.dk, fengze.han@tum.de, e.shutova@uva.nl

Attention knockout on MHAL during VQA blocking information flow between:

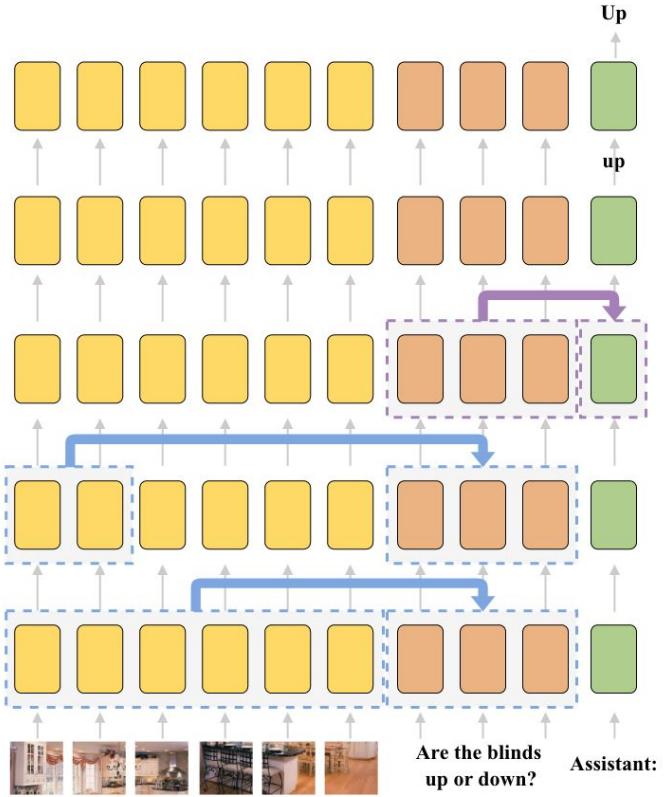
(1) the input positions corresponding to the **whole image** to the **different parts of the question**;

(2) the input positions corresponding to **image regions containing objects relevant to answering the question**, **to the question**;

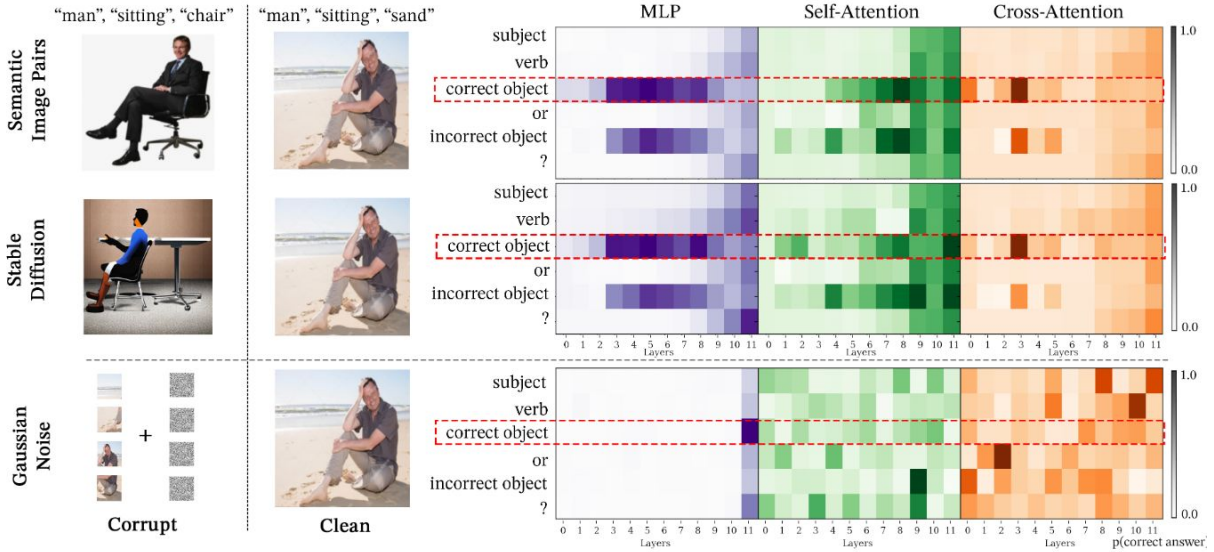
(3) the input positions corresponding to the **image and the question** to the **final prediction**,

across different layers of the MLLM.

lower layers	the model first transfers the **more general visual features** of the whole image into the representations of (linguistic) question tokens
middle layers	it transfers visual information about **specific objects** relevant to the question to the respective token positions of the question.
higher layers	the resulting **multimodal representation is propagated to the last position of the input sequence** for the final prediction.



SIP vs. Gaussian Noise vs. SIP with generative models



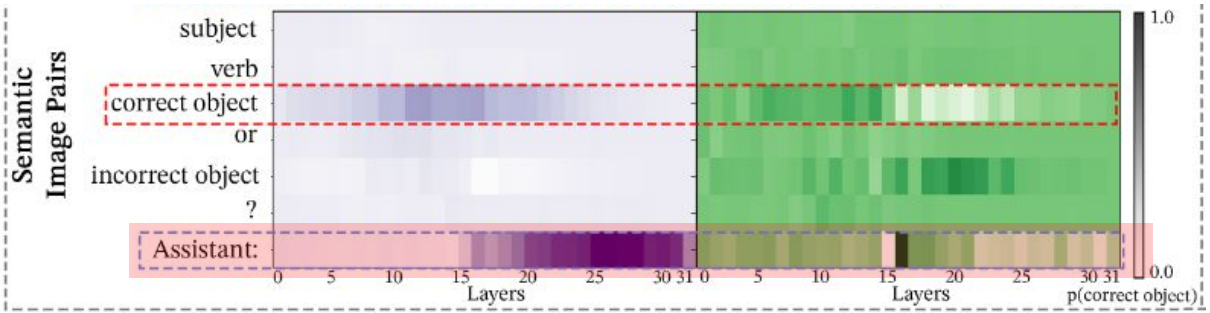
(BLIP)

1. SIP in MLP modules highlight the importance of early and middle layers, while Gaussian noise only the last layer.
2. In self-attention patching in SIP assigns the highest probability to the correct token, unlike Gaussian noise.
3. SIP also display consistent activation patterns —long horizontal regions in MLPs and vertical regions in self-attention —whereas Gaussian Noise shows clear differences.

Attention Heads Patching

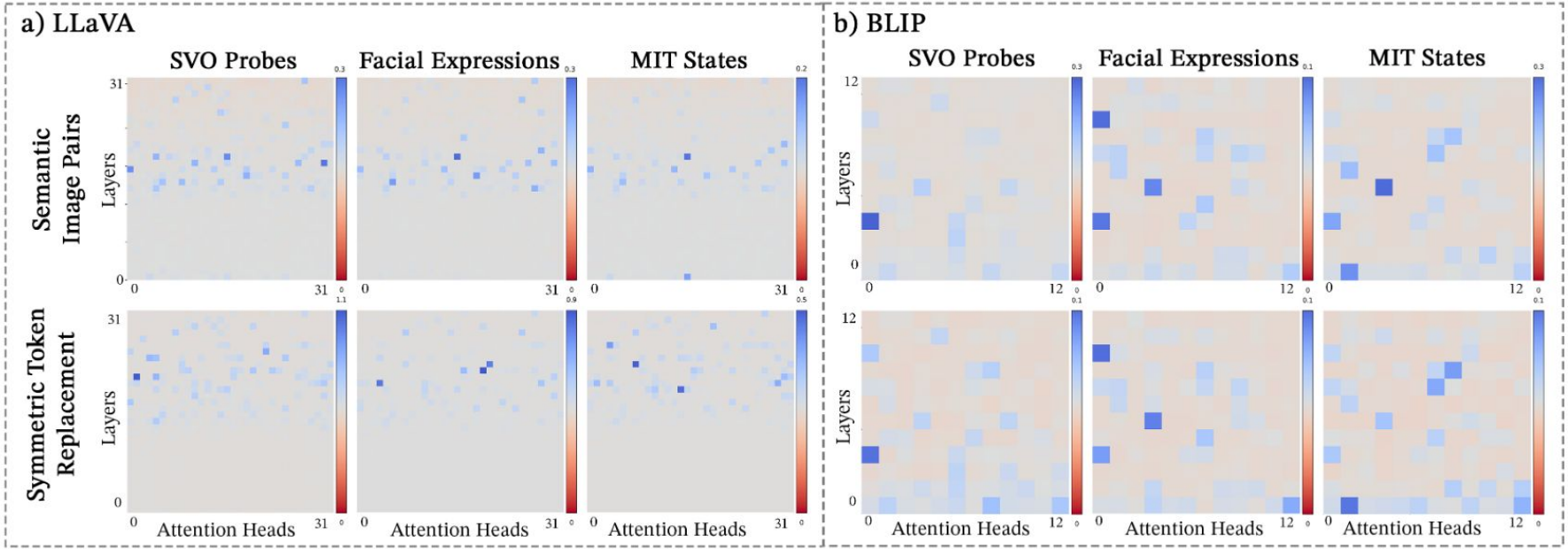
They perform attention head patching to identify which attention heads contribute most to vision-language integration in BLIP and LLaVA.

In LLaVA, they patch activations at the “Assistant:” token, which shows the strongest effect, across all self-attention heads.



In BLIP, they patch the correct answer into cross-attention heads.

Universal Attention Heads Exist Across Corruption Schemes and Tasks



Universal attention heads → **consistently high logit differences across tasks and modality corruptions**.

Model	Vision Heads	Multimodal Heads	Text Heads
LLaVA	L15.H5, L16.H18 L14.H4, L15.H8 L18.H26, L17.H0 L17.H22, L14.H28 L19.H4, L15.H14 L17.H13, L20.H28	L28.H7, L31.H27 L16.H24, L16.H24 L19.H15, L18.H30 L18.10, L21.H30	L22.H16, L29.H19
BLIP	L5.H3	L3.H0	L0.H11

Model	Head Distribution	Architecture	Crossmodal Fusion	Implications
LLaVA	2 text-only heads Majority are multimodal or vision-only under image corruption	Early fusion: image tokens embedded in prompt	Visual input influences prompt even under text corruption	Dense multimodal interaction; less modular
BLIP	Exactly 1 head for each:- Vision- Text- Multimodal	Dual-stream with cross-attention	Effective modality separation and fusion	Modular, interpretable integration

The presence of **overlapping attention heads** when patching different modalities and tasks is a novel discovery.

Since cross-attention heads in BLIP and self-attention heads in LLaVA integrate vision and language, **overlapping heads suggest that vision and language convey similar information to the VLM** despite being distinct feature spaces.

This **challenges the longstanding notion that text is the dominant modality for VLMs**.

Additionally, the **reuse of these components across tasks** offers **insights into model behavior and model editing**, enabling targeted improvements that enhance performance, address bias, and better adapt the model to new tasks.

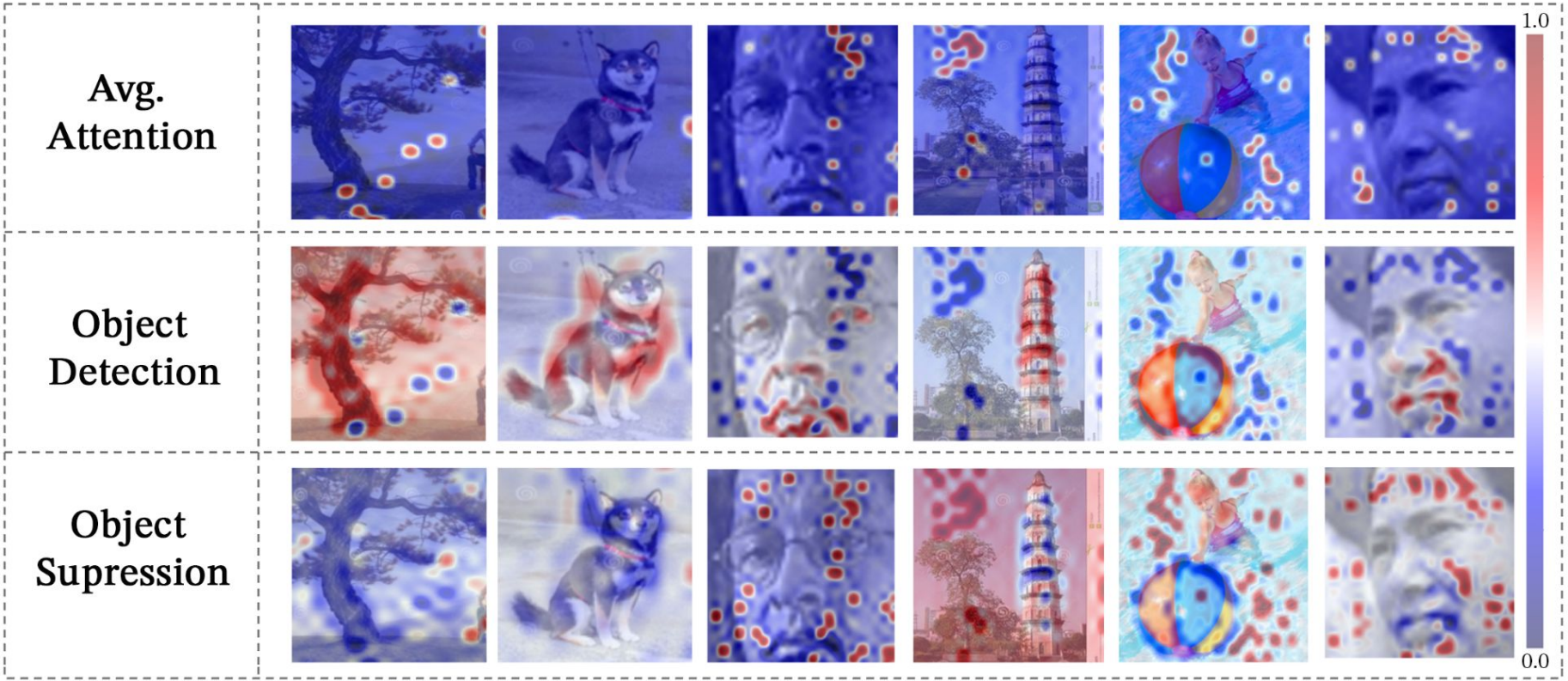
Universal Attention Heads Perform Distinct Functions

They visualize:

BLIP: the **cross-attention patterns** between the “**correct answer” text token** and **image patches**.

LLaVA: the **self-attention patterns** from the **“Assistant:” token** to image patches.

BLIP

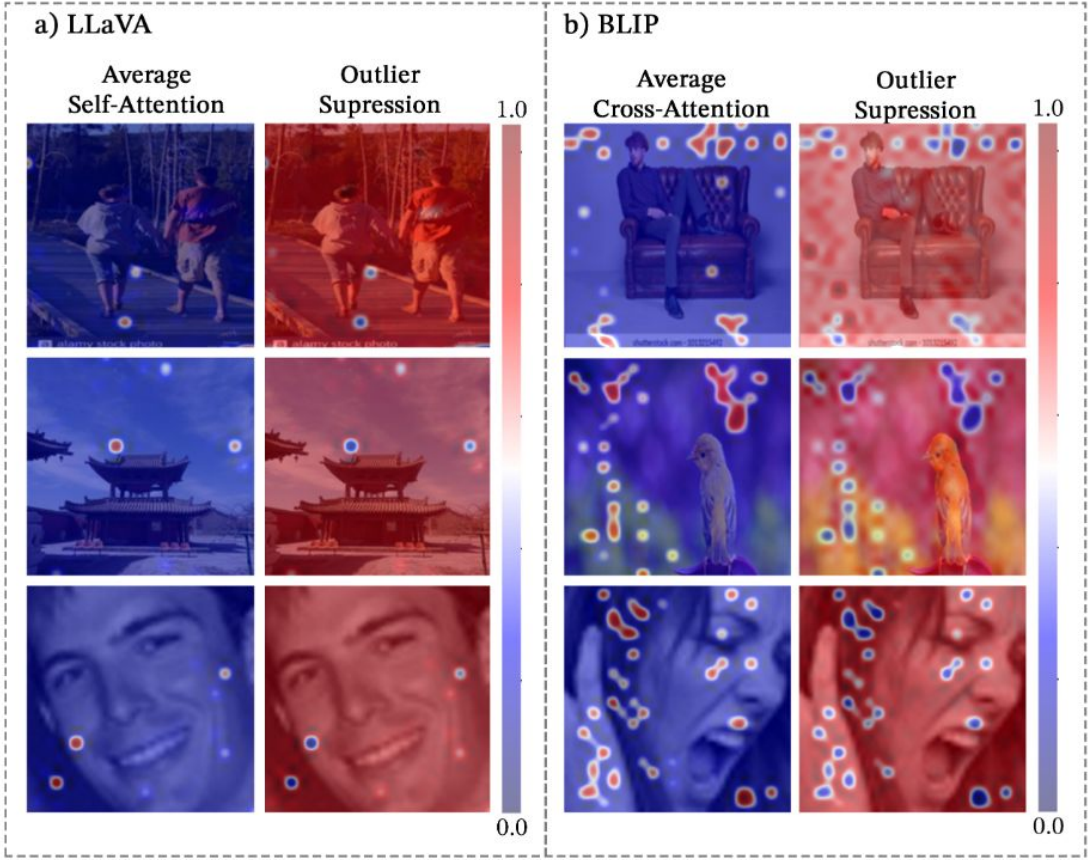


LLaVA and BLIP

All universal multimodal attention heads in LLaVA implement outlier suppression.

cross-attention supports three functions: object detection, object suppression, and outlier suppression.

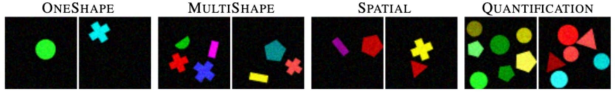
self-attention primarily handles outlier suppression, indicating cross-attention's unique role in image grounding.



WHY SO INTERESTING?

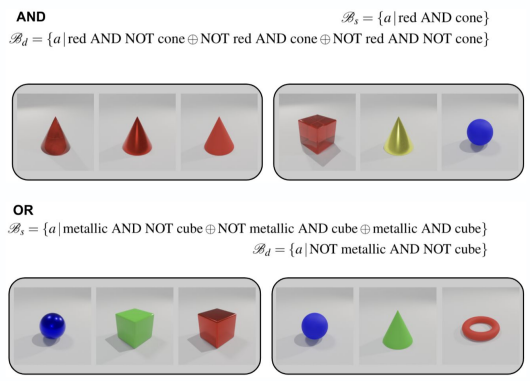
COOCO

ShapeWorld

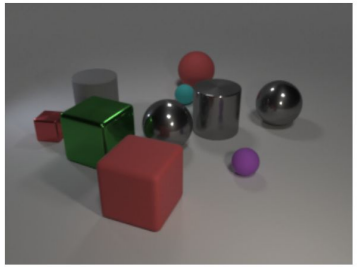


ONESHAPE	MULTISHAPE	SPATIAL	QUANTIFICATION
• *There is a blue rectangle.* • *There is a triangle.* • *There is a yellow shape.* **Training combinations:** • 50 combinations **Validation combinations:** • *red square, green triangle, blue circle* **Test combinations:** • *yellow rectangle, cyan ellipse, magenta cross*	• *There is a magenta semicircle.* • *There is a pentagon.* • *There is a cyan shape.* (Same as for ONESHAPE) **Number of objects** • Training: 1, 2, 3, 5 • Validation: 4 • Testing: 6	• *A red circle is to the left of a cyan semicircle.* • *A white circle is above a pentagon.* • *A red shape is to the right of a green triangle.* • *A shape is below a cross.* **Training / validation / test combinations:** • Same as for ONESHAPE	• *The shape is green.* • *Most shapes are rectangles.* • *No shape is a red triangle.* • *All triangles are green.* • *Two blue shapes are pentagons.* **Number of objects** • Training: 3, 4, 5, 7 • Validation: 6 • Testing: 8

VISUAL COMPLEX CONCEPTS DATASET

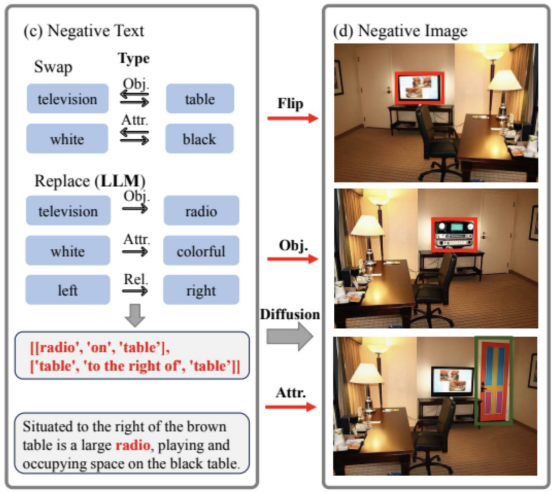


CLEVER



(a) (i) *Remove all gray spheres. How many spheres are there? (3),* (ii) *Take away 3 cubes. How many objects are there?* (7), (iii) *How many blocks must be removed to get 1 block? (2)*

FineCops-Ref



IconQA

Q: What is the man doing?
A: riding a motorcycle

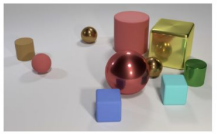
VQA

Q: Which object is next to the one shaped like a cube?
C:

Q: How many tomatoes are there?
A: 5

VQA 2.0

Q: Which picture shows the pizza inside the oven?
C: (A) left one (B) right one



Q: How many objects are metal things?
A: 4

CLEVR

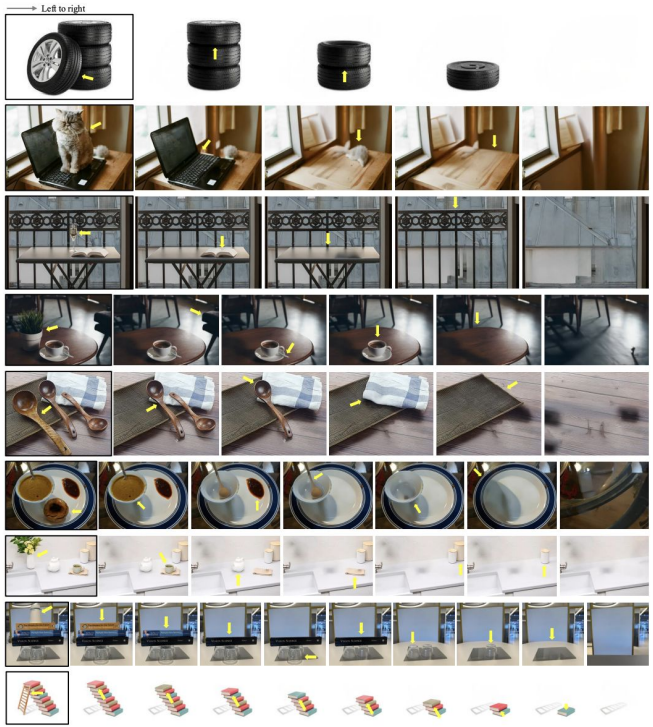
Q: How many sticks are there?
A: 80

IconQA

Visual Jenga: Discovering Object Dependencies via Counterfactual Inpainting

Anand Bhattad[1] Konpat Preechakul[2] Alexei A. Efros[2]
[1]Toyota Technological Institute at Chicago [2]University of California, Berkeley

https://visualjenga.github.io



In general with diffusion models is possible to create identical images with minimal variations that tackles specific visual phenomena.

E.g. commonsense knowledge - intuitive physics