Understanding Knowledge Storage and Transfer in Multimodal Language Models

Samyadeep Basu[§], Martin Grayson[†], Cecily Morrison[†], Besmira Nushi[†], Soheil Feizi[§], Daniela Massiceti[†]

§: University of Maryland, College Park; †: Microsoft Research

KGM @ ECCV 2024 | Appearing in NeurIPS 2024

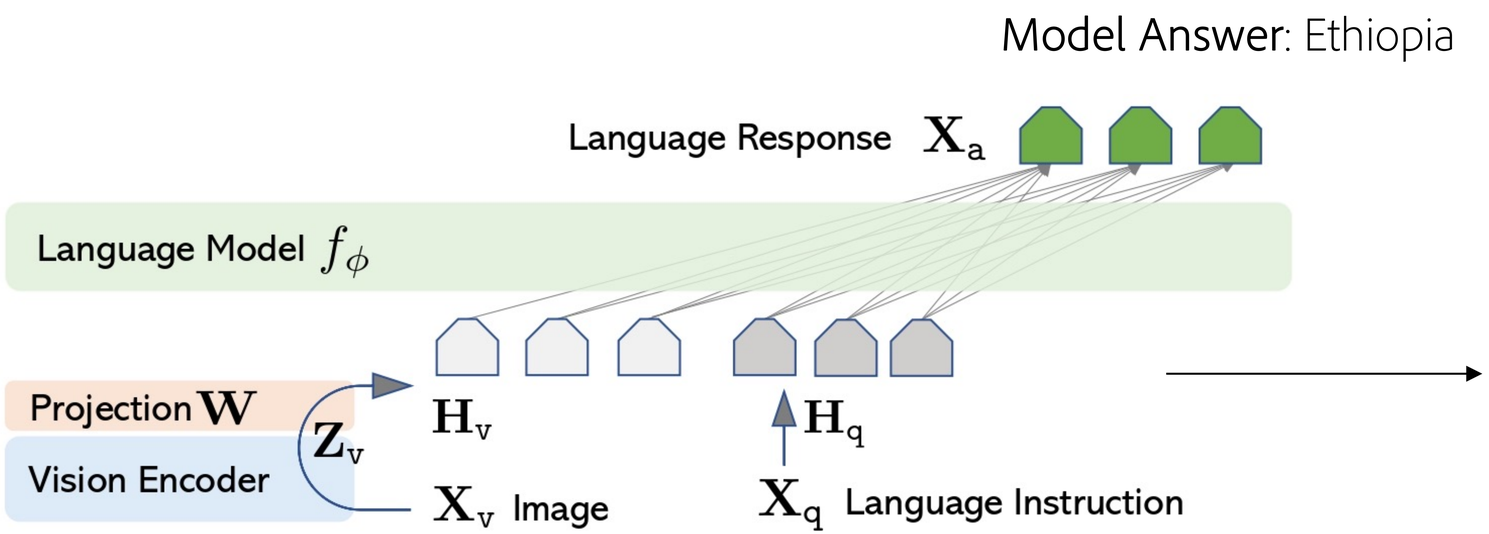
https://aka.ms/mllm-info-storage

About Me

- **Final year CS PhD student** at University of Maryland with **Dr. Soheil Feizi**
- **Current Research Interests**: Multimodal /Vision/Language Models with an emphasis on controlling them via Interpretability
  - *DiffQuickFix, LocoEdit*: Light-weight Model Editing for Text-to-Image Models
  - *CompAlign* - Interpreting arbitrary ViT components (e.g., attention heads) with text
    - Zero-shot segmentation
    - Spurious Correlation Mitigation

Motivation

Multimodal LLMs are Widely Used in VQA



Input Image

Question: Where does this dish come from?

Fig 1: Representative VQA task

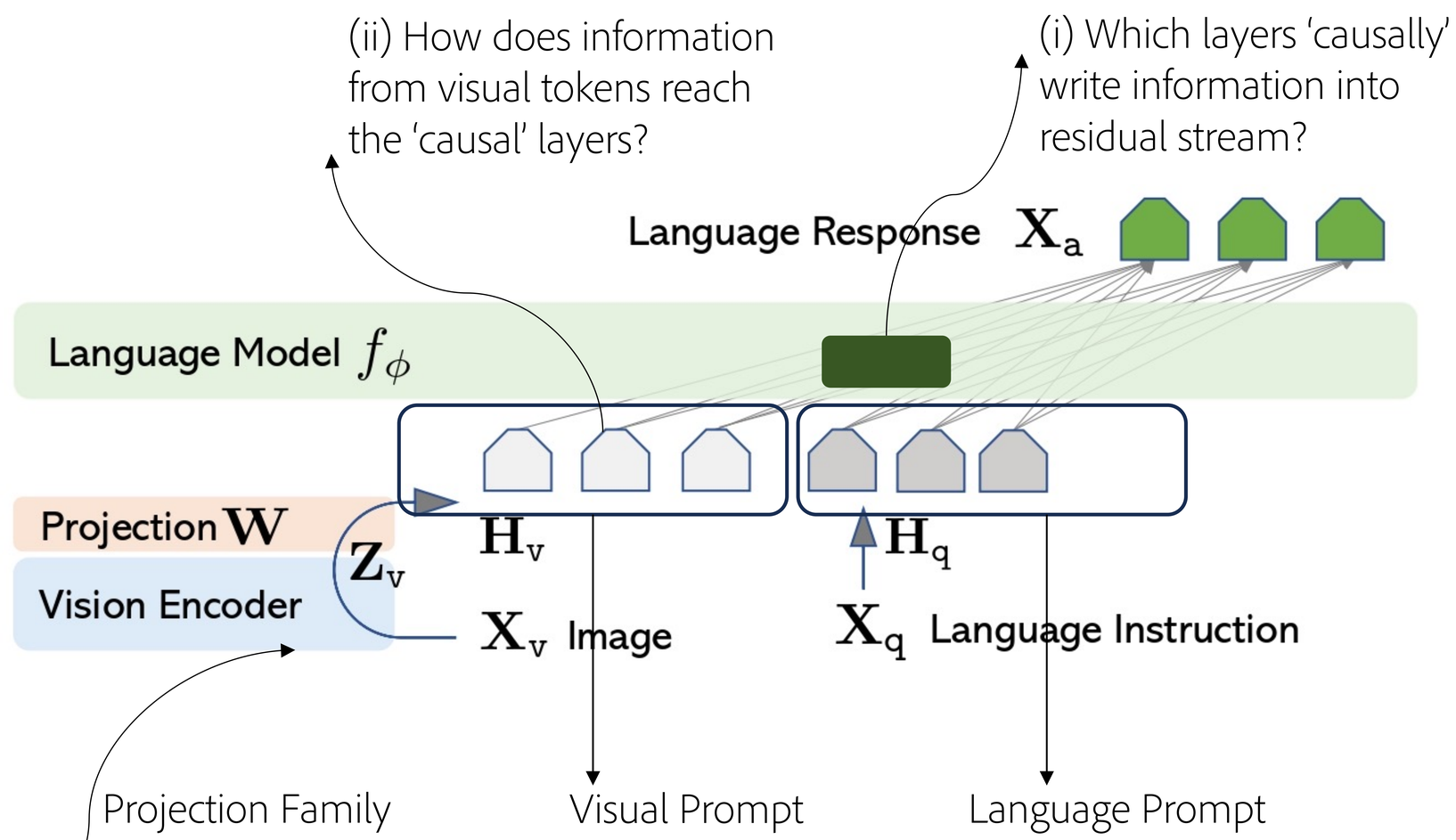
But we *lack scientific* understanding on:

(i) How they internally process information

(ii) How can we control them for tasks like model editing to fix failure modes / introduce rare concepts

We study how MLLMs process and transfer information in a factual VQA task using a constraint-based formulation

We study LLaVA and multi-modal Phi-2



Causal Layer : A layer or a small set of layers which control the output of the model conditioned on an input

Designing a Probe Dataset

Introducing Constraint Based Formulation for Interpreting MLLMs

Constraint based VQA Questions

What city is this building in? → Visual constraint } Single Constraint Questions

Name a movie directed by this director in 1989?
- this director → Visual constraint
- 1989 → Text constraint
- } Multi-Constraint Questions

VQA-Constraints Dataset: 9.7k VQA questions annotated with constraints

- OK-VQA
- Multimodal Known
- Multimodal Movies

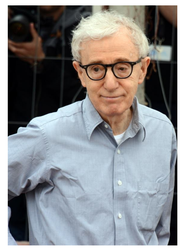
Introducing Constraint Based Formulation for Interpreting MLLMs

Constraint based VQA Questions

What city is this building in?

→ Visual constraint } Single Constraint Questions

Name a movie directed by this director in 1989?



→ Visual constraint

→ Text constraint } Multi-Constraint Questions

Visual Constraint: A set of tokens in the question which relates to a visual entity in an image (e.g., this building → Space Needle in Image)

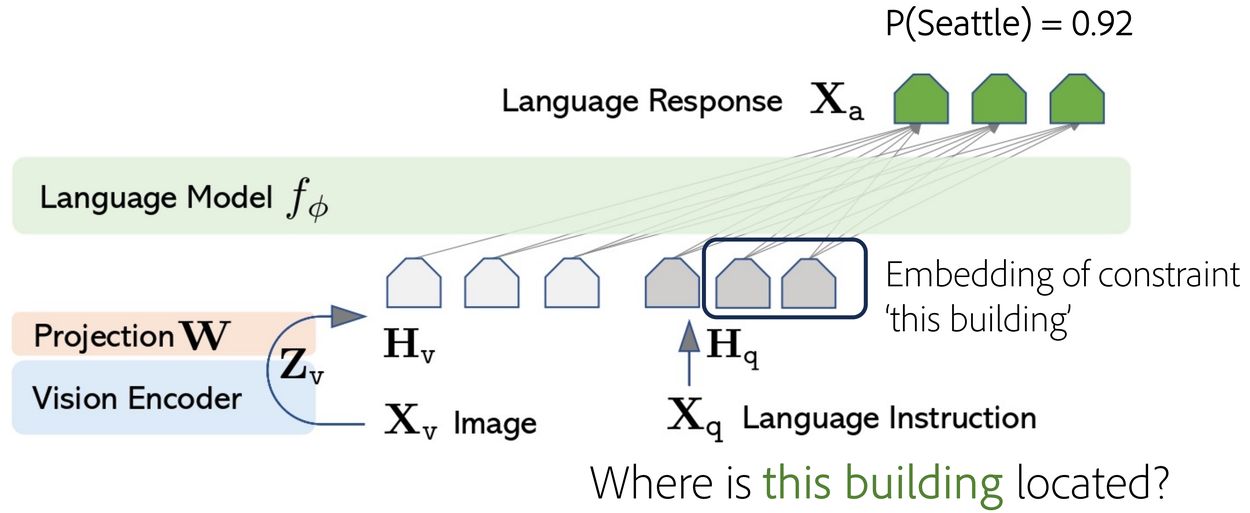
Text Constraint: A set of tokens in the question which reduces the space of the possible answers. Often used in conjunction with visual constraints.

Tracing Method for Knowledge Storage

MultimodalCausalTrace: Identifying Causal Layers in MLLMs

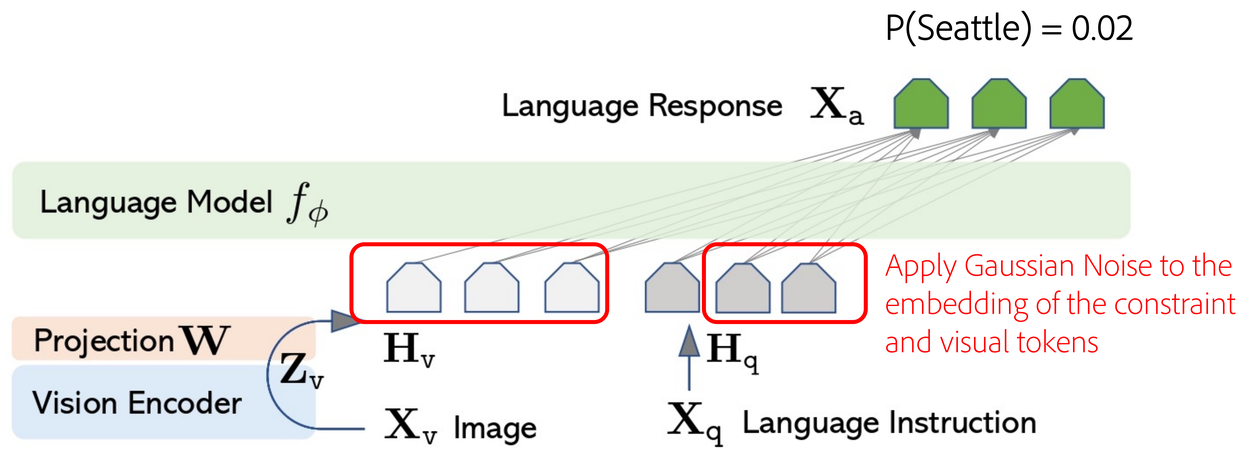
Using Causal Trace Designed for Language Models

Clean Model



Where is this building located?

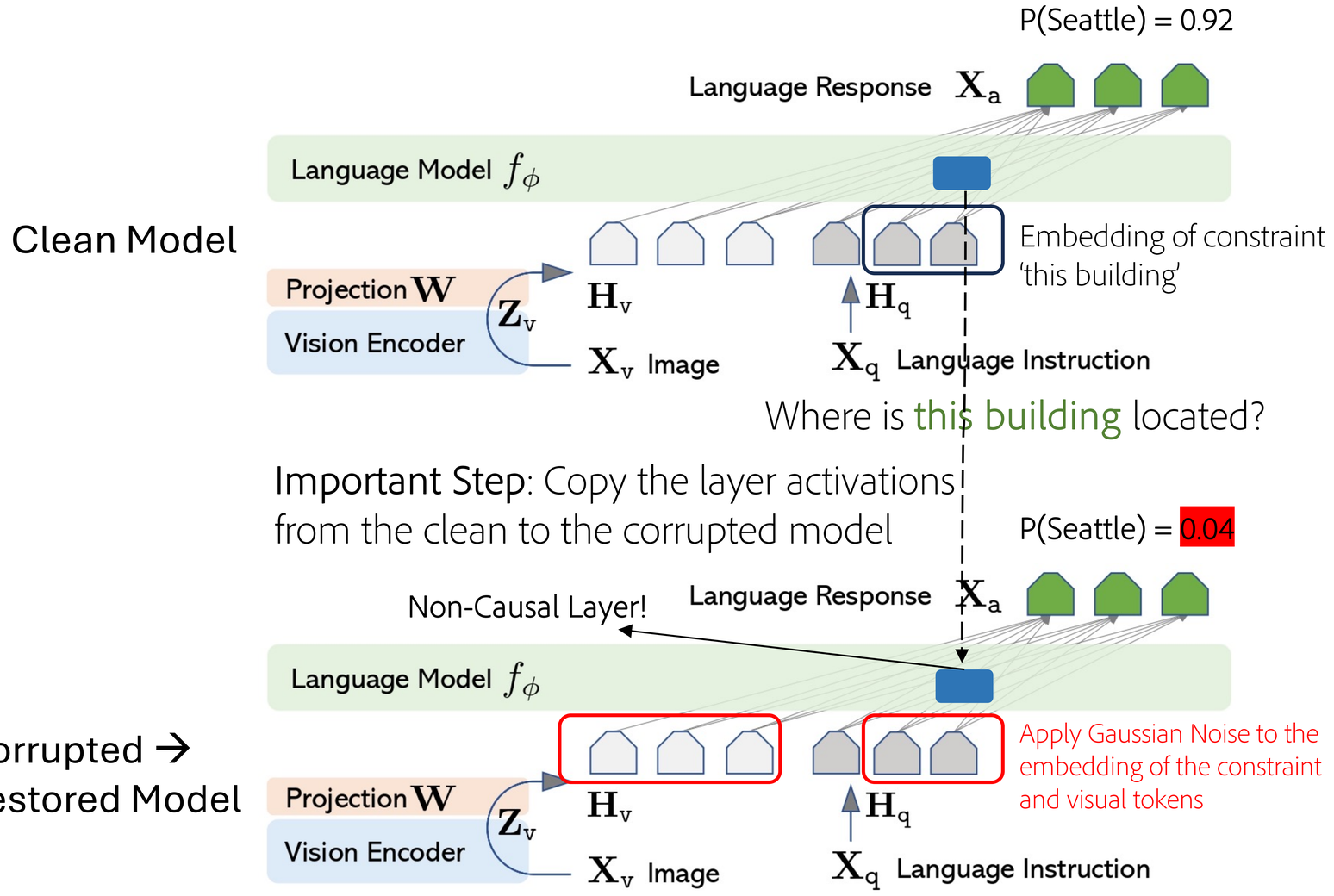
Corrupted Model



Where is this building located?

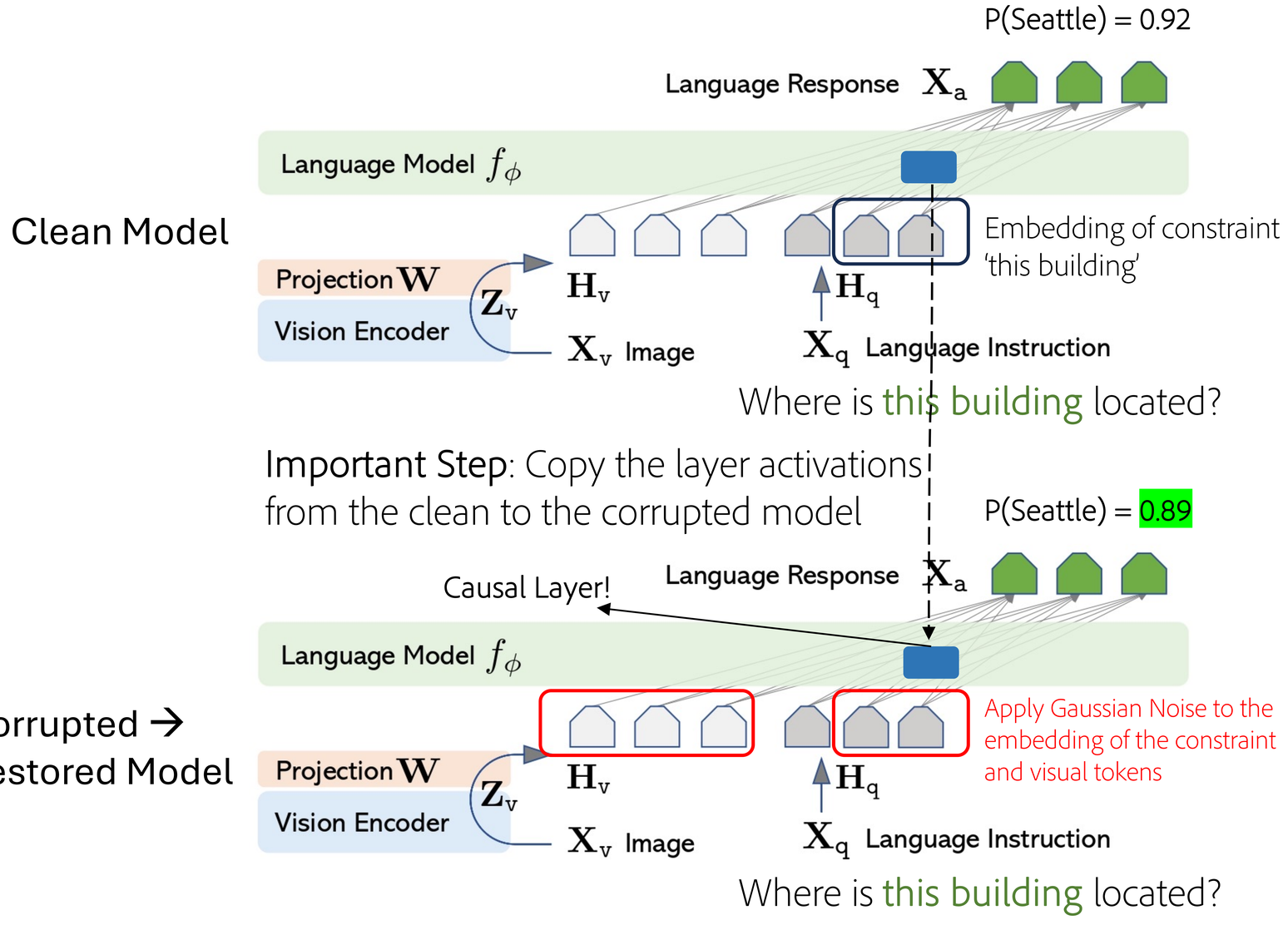
MultimodalCausalTrace: Identifying Causal Layers in MLLMs

Using Causal Trace Designed for Language Models



MultimodalCausalTrace: Identifying Causal Layers in MLLMs

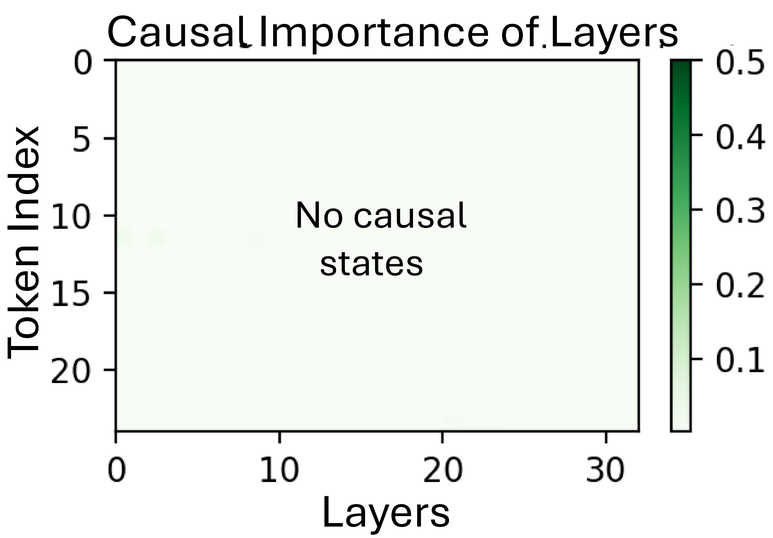
Using Causal Trace Designed for Language Models

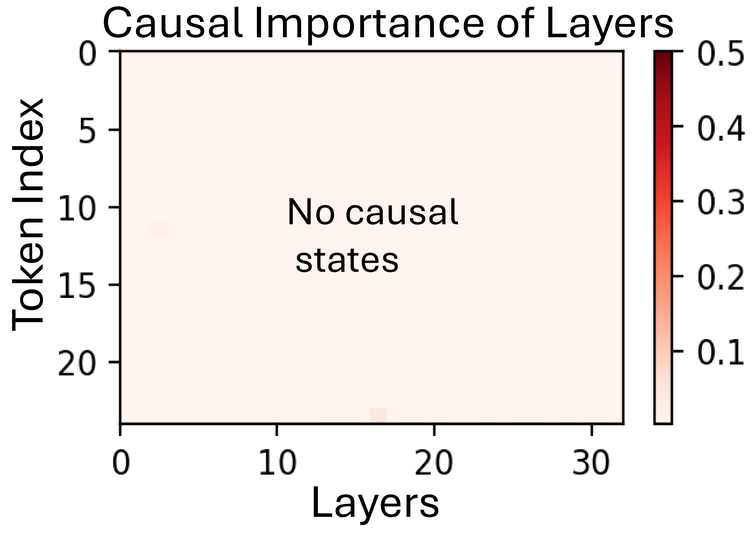


MultimodalCausalTrace: Identifying Causal Layers in MLLMs

Using Causal Trace Designed for Language Models - **Results**

No Relevant Causal Layers!



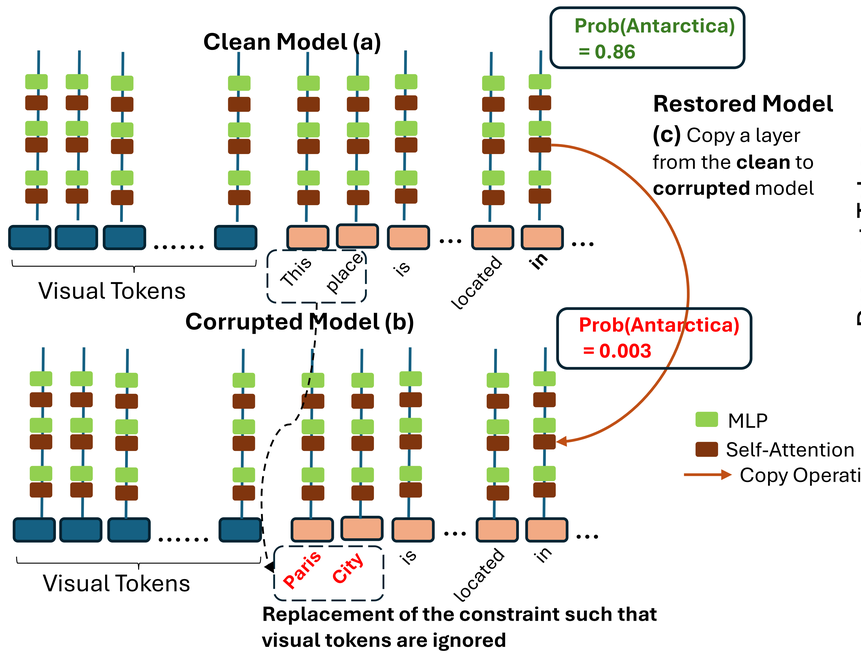


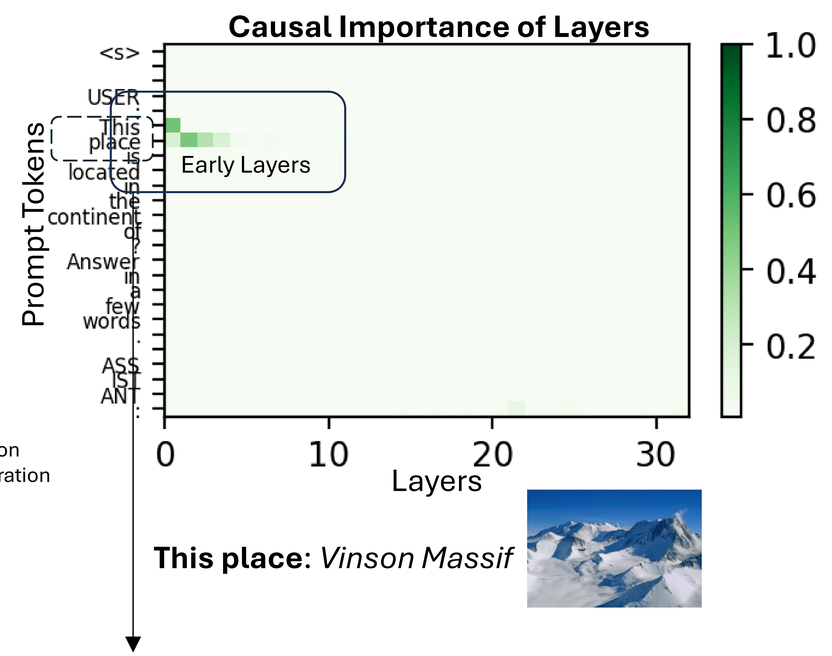
Potential Reason: Corrupted Model is very noisy as noise is applied to a large set of tokens (e.g., 576 + number of constraint tokens), whereas in a language model noise is only applied to a small set of tokens (e.g., 2-4)

MultimodalCausalTrace: Identifying Causal Layers in MLLMs

Our Method

Adapting Causal Trace Designed for Language Models





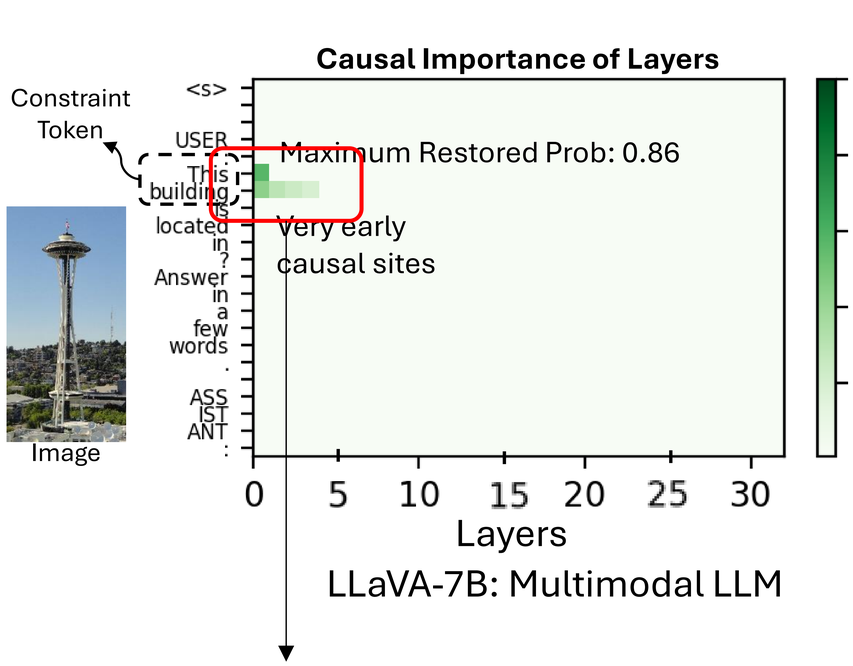
Core Idea : Replace the constraint tokens with a set of tokens such that the visual tokens are ignored while answering the question

Tracing Results

MultimodalCausalTrace: Identifying Causal Layers in MLLMs

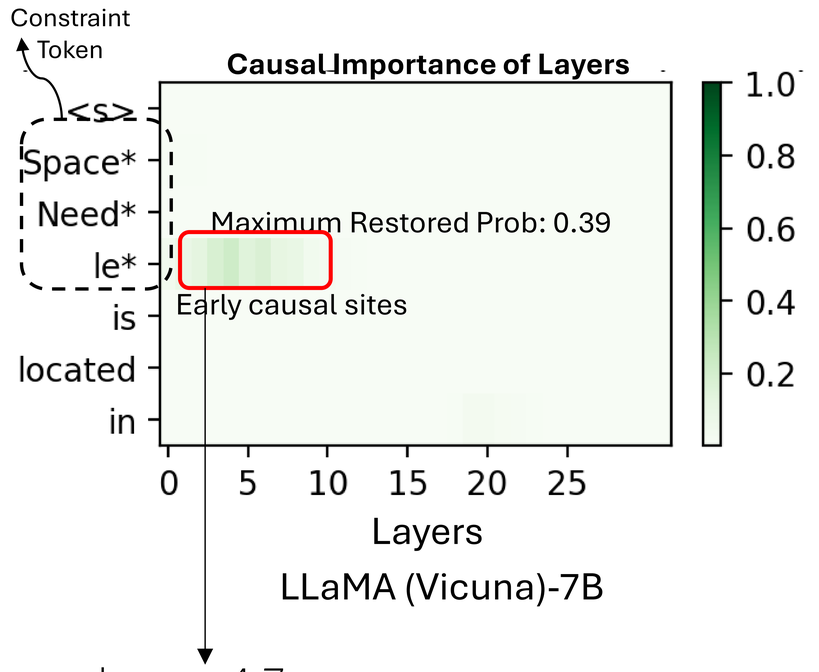
Single Constraint Questions

How does information retrieval from internal layers in MLLMs differ from language models?



Layers 1-4

Restoring only one layer is sufficient



Layers 4-7

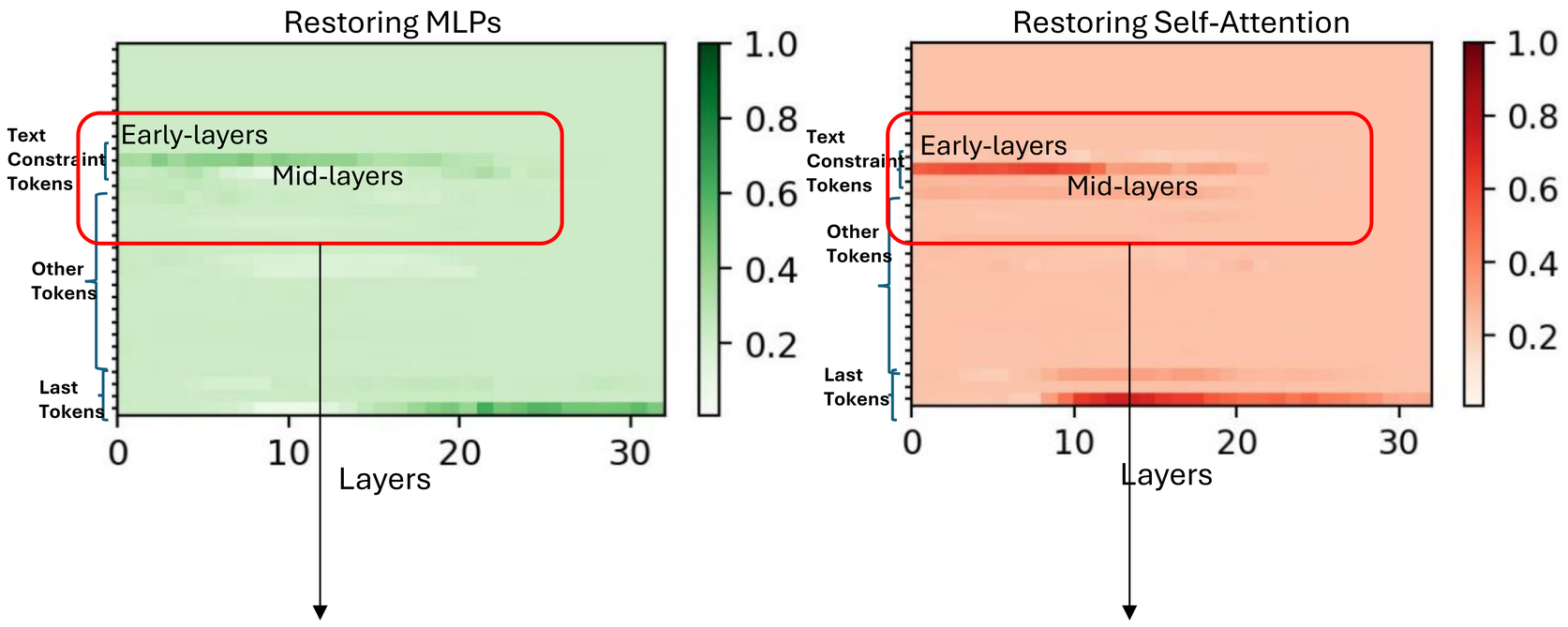
Requires restoring a window of layers

Takeaway: MLLMs (under the presence of a visual prompt) retrieve information differently than LLMs – although the same language backbone is used

MultimodalCausalTrace: Identifying Causal Layers in MLLMs

Multi-Constraint Questions

Information Processing for Text-Constraint in Multi-Constraint Questions



Text-constraint in multi-constraint questions require information to be retrieved from early + mid layers with a large window size

Takeaway : Multi-constraint VQA questions require more parametric memory to answer a question

Information Transfer in MLLMs

Visual Tokens to Constraint

Tracking Attention Contributions from Visual Tokens to Constraints

$$a_{ij}^{\ell} = \sum_{h=1}^{H} A_{ij}^{l,h} \left(g_{\phi}(.)_{j,\ell-1} W_v^{\ell,h} \right) W_o^{\ell,h}$$

Value embeddings

Projection Matrix

Contribution from the token j to token i

Attention weight from token j to token i

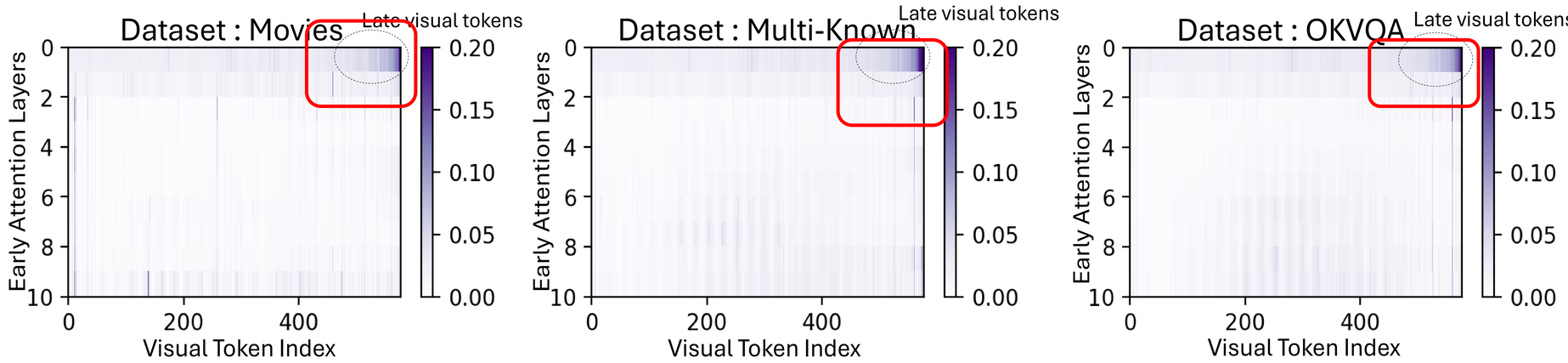
We track L2 norm of this quantity

Attention contributions is a general property of transformers and have been previously used in Yuksekgonul et al.(2023)

Insights from Information Transfer in MLLMs

Visual Tokens to Constraint

We compute the **attention contributions** from the visual tokens to the constraint token



Takeaway : Only a subset of visual tokens (after the projection layer) transfer information to the constraint token position

Most of the visual information gets accumulated in the late visual tokens by the projection layer

Can we use the interpretability insights towards a tangible application?

Can we use the interpretability insights towards a tangible application?

- Introducing Rare Concepts
- Fixing Failure Modes

Qualitative Examples of Rare VQA Questions

Inchcolm

In which country is *this island* located?

Croome Court

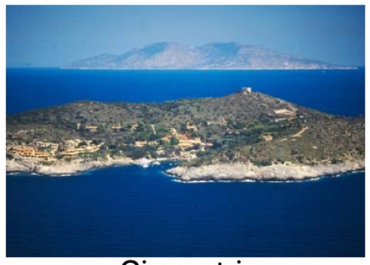
In which country is *this house* located?

Hilton Prague

In which country is *this hotel* located?

Cragside

In which country is *this house* located?



Giannutri

In which country is *this island* located?

Kandy Lake

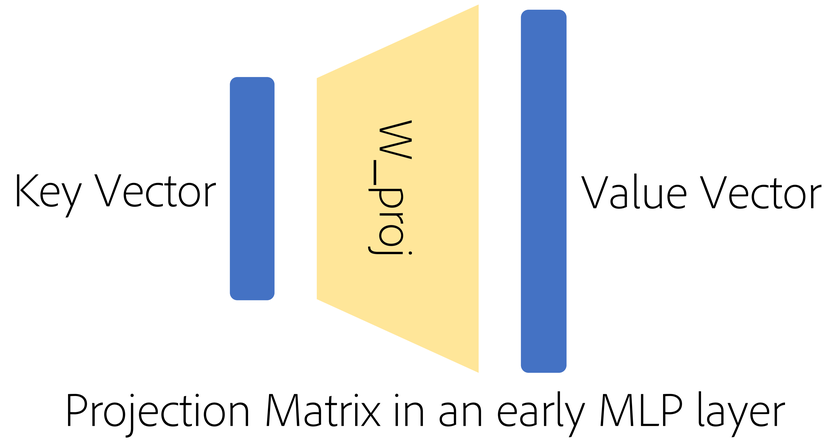
In which country is *this lake* located?

Sourced from Encyclopedia-VQA

Practical Application Leveraging Interpretability Insights:

Model Editing to Incorporate Rare Knowledge into MLLMs - Method

We edit the ***early MLP layers*** for the best editing performance



Projection Matrix in an early MLP layer

Step 1: Obtain key embeddings via forward pass of the visual prompt + question

Step 2: Obtain value embeddings

Value embeddings $z^*_{c,\ell} = \arg\min_{z_{c,\ell}} \mathcal{L}(z_{c,\ell})$ Language Modeling AR Loss

Step 3: Edit Optimization Key embeddings

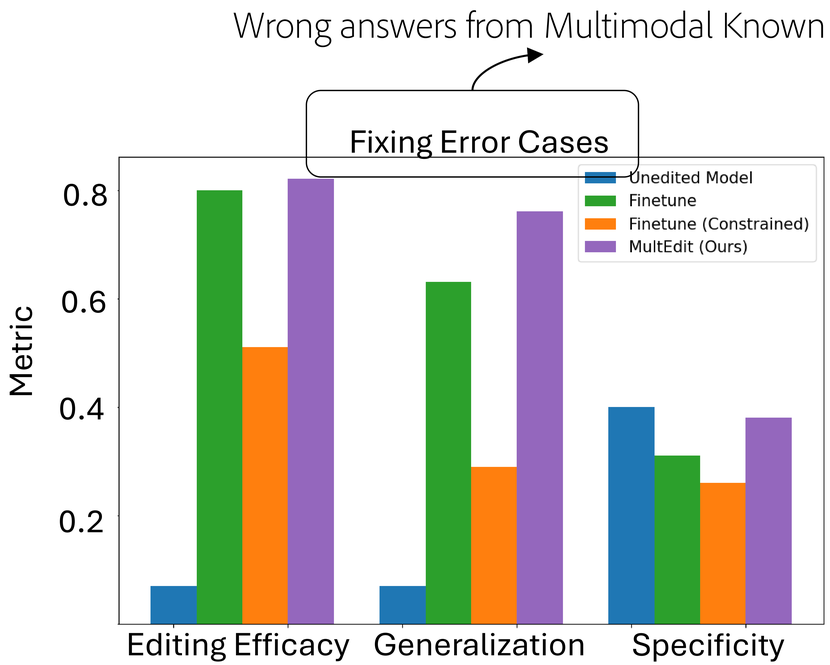
$$W^{\ell *}_{proj} = \arg\min_{W^{\ell}_{proj}} \|W^{\ell}_{proj} k_{c,\ell} - z^*_{c,\ell}\|_2^2 + \lambda \|W^{\ell}_{proj} - W^{\ell'}_{proj}\|_2^2$$

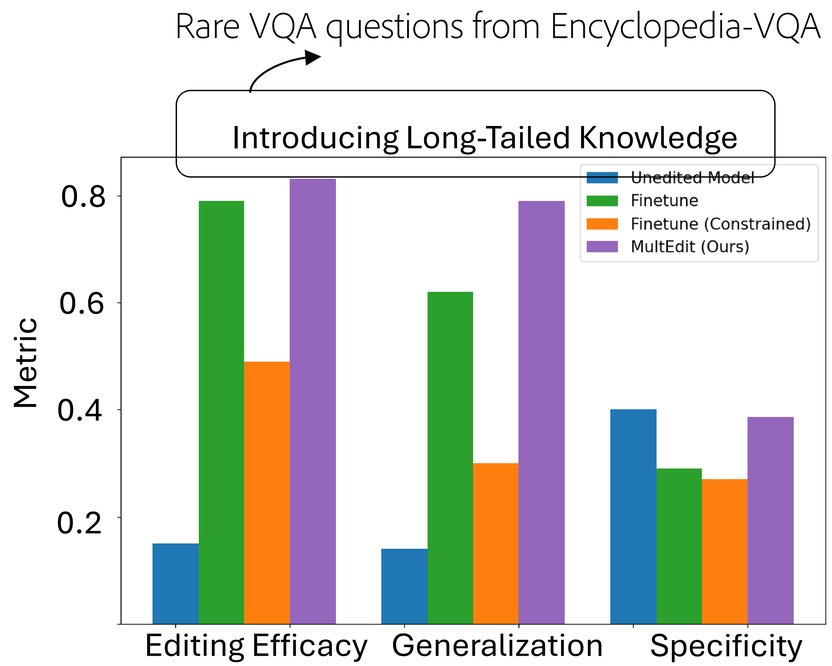
Does not require caching a Multimodal Wikipedia matrix → Relatively data free

Practical Application Leveraging Interpretability Insights:

Model Editing to Incorporate Rare Knowledge into MLLMs - Results

Editing the early layers with our objective leads to incorporation of rare knowledge into the model and is better (+faster) than fine-tuning the language model





Metrics:

(i) Editing Efficacy: Effectiveness of the edit (Measured by the probability of the correct answer)

(ii) Generalization Efficacy: Effectiveness of the edit under paraphrased questions (Measured by the probability of correct answer)

(iii) Specificity: Editing effect on unrelated VQA questions (measured by VQA-accuracy on OK-VQA and eval on MMMU)

Conclusion

Checkout out other interpretability works at:
https://samyadeepbasu.github.io/

SCAN ME