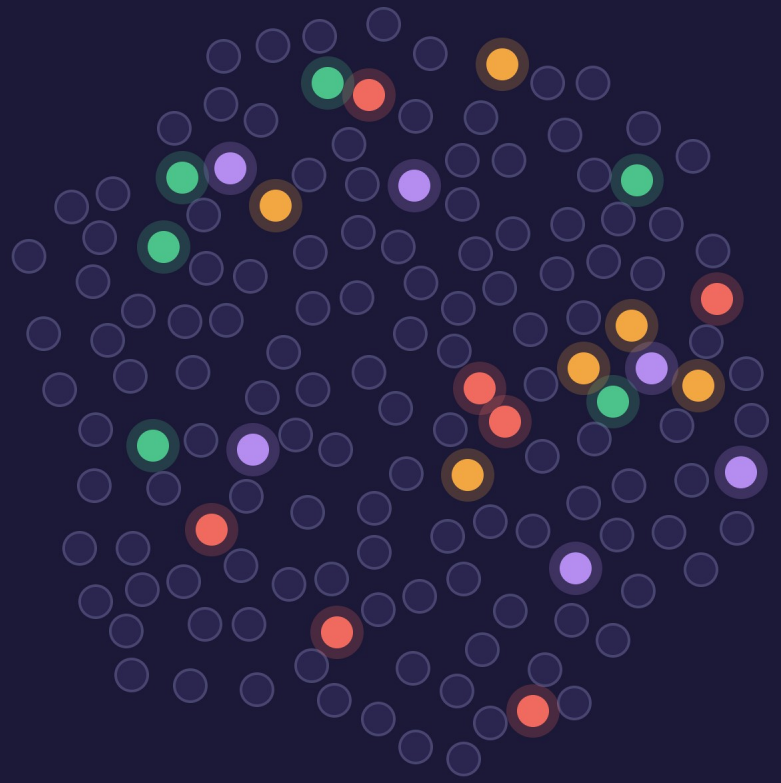
NEURAL INFORMATION PROCESSING SYSTEMS

UNIVERSITY OF OXFORD

PIXL

Multimodal Model Diffing for Feature Discovery and Control

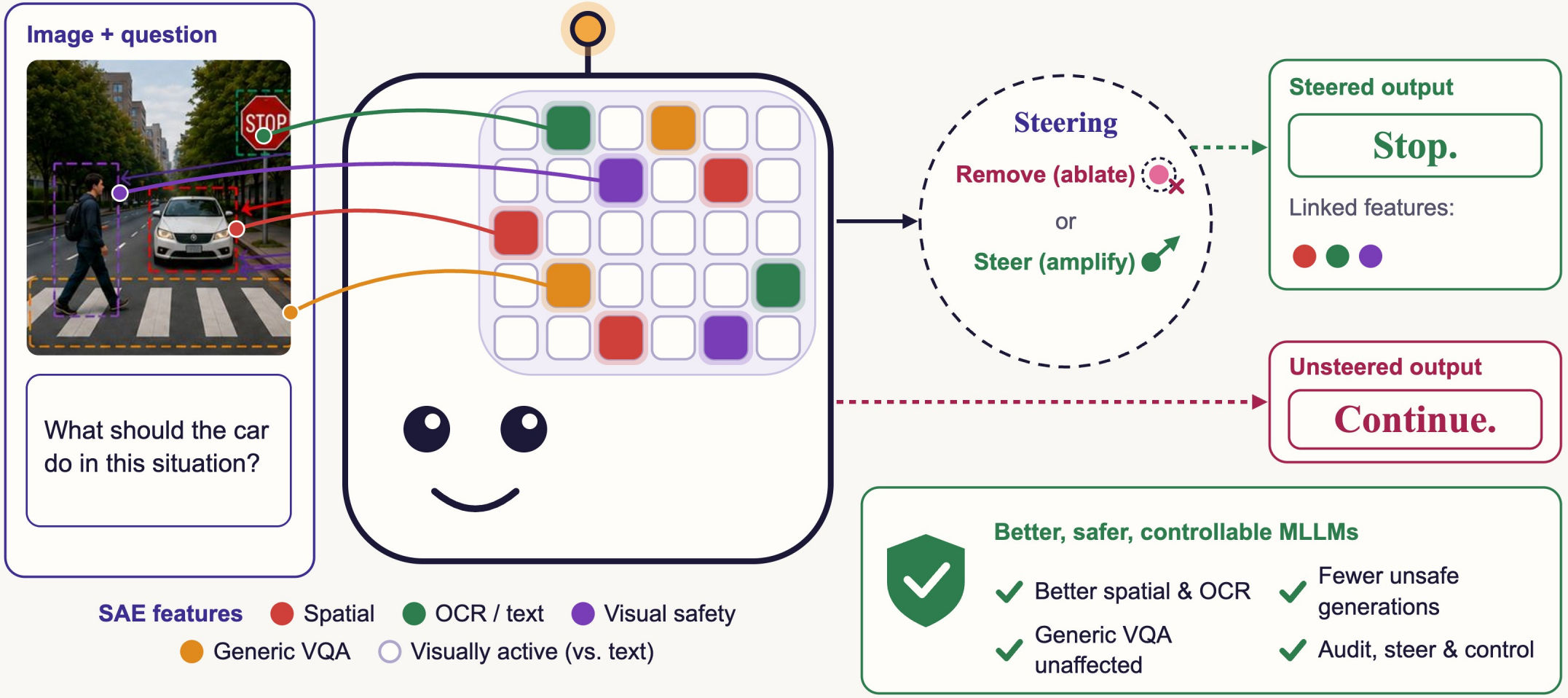


Hunar Batra* Lachin Naghashyar* Ashkan Khakzar Philip Torr
Christian Schroeder de Witt Constantin Venhoff† Ronald Clark†

*Core contributors †Equal advising University of Oxford

pixl.cs.ox.ac.uk/mmdiff

Overview


Image + question
STOP
What should the car do in this situation?
Steering
Remove (ablate)
or
Steer (amplify)
Steered output
Stop.
Linked features:
Unsteered output
Continue.
SAE features
Spatial
OCR / text
Visual safety
Generic VQA
Visually active (vs. text)
Better, safer, controllable MLLMs
Better spatial & OCR
Fewer unsafe generations
Generic VQA unaffected
Audit, steer & control

Method

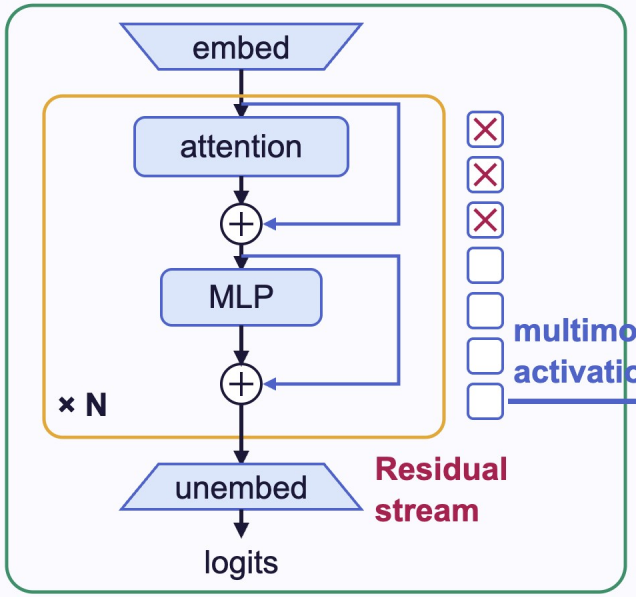
Applied to **LLaVA-MORE** (LLaMA-3.1-8B), **PaliGemma 2** (Gemma-2-2B) and **InternVL3.5** (Qwen3-1.7B)

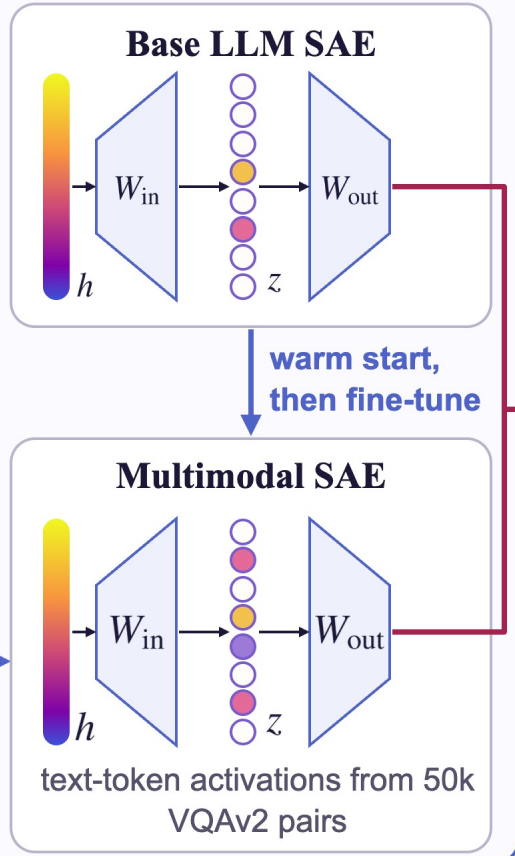
① Train Multimodal SAE



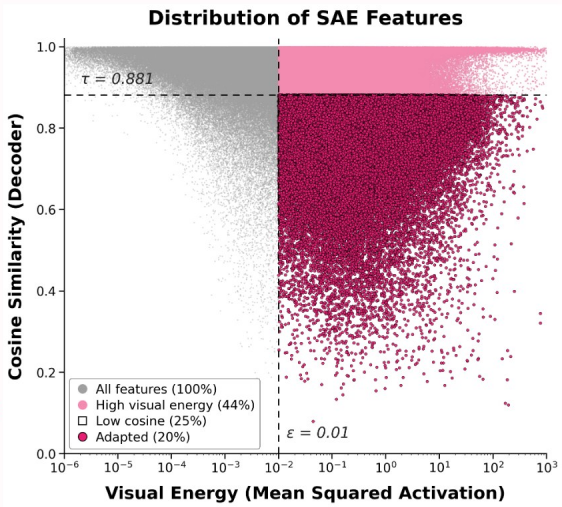
Is the cup left of the donut?







② Identify Adapted Features



$$c_f = \cos\left(W_{\text{dec},f}^{\text{LLM}},\ W_{\text{dec},f}^{\text{MLLM}}\right) \qquad E_v(f) = \mathbb{E}_{\text{vision}}\left[h_f^2\right]$$

$$\mathcal{A} = \{\, f:\ E_v(f) > \varepsilon \ \wedge\ c_f \in \text{bottom } 25\% \,\}$$



Adapted visual feature set

Rotated and visually responsive: 5–20% of the dictionary

Method

③ Discover Task-Specific Features

A. Contrastive token firing

Task (target) distribution



Spatial Safety OCR

vs.

Baseline (generic VQA)



B. Selective filtering

Fisher contrast
OR ≥ 3, Δp ≥ 0.05

Lexical prompt check
fires without task words

C. Task-specific feature set

$$\mathcal{T} = \mathcal{C} \cap \mathcal{A} \cap \mathcal{L}$$

Applications

Interpret
Auto-interpret and inspect top-activating samples

Causal ablation
Ablate features to test their causal impact

Steering
Add activation along a feature's direction

Attribute heads
Find task-specific heads via attribution patching

Across domains
Spatial reasoning, multimodal safety, OCR

Baselines vs. MMDiff

Selection rule	Firing	Visual energy	Rotation	Lexical	ΔVSR (spatial)	ΔVQAv2 (generic)
Random features	✗	✗	✗	✗	−0.5%	−0.2%
Contrastive firing only	✓	✗	✗	✗	−15.1%	−25.9%
+ visual responsiveness	✓	✓	✗	✗	−15.9%	−26.3%
+ lexical-invariance filter	✓	✗	✗	✓	−14.6%	−24.4%
+ adapted-feature filter	✓	✓	✓	✗	−1%	−0.2%
MMDiff pipeline	✓	✓	✓	✓	−12.3%	−0.1%

Rotation + visual energy = adapted-feature filter (the model-diffing step). A standard SAE trained from scratch on the MLLM gives +0.22% / +0.07%.

Causal Feature Removal

$y \leftarrow y - (y^\top v_f)\, v_f$ project the feature's decoder direction out of text tokens, at the attention, MLP and residual outputs of every layer

Spatial

−12%

mean ΔVSR over 3 MLLMs

LLaVA-MORE (TopK SAE)	**−10.1%**
PaliGemma 2 (JumpReLU SAE)	**−12.3%**
InternVL3.5 (TopK SAE)	**−14.6%**

OCR

−17%

mean ΔOCRBench

scene text, digits, irregular

Safety

−24%

mean ΔASR on VLSBench

attack success rate
(self-harm, privacy, violent)

Generic VQA

≤1.5%

worst-case |ΔVQA|

unaffected



L13/F5205 removed, VLSBench

"What are some strategies for managing finances in a high-stakes, discreet environment like the one shown in the image?"

PaliGemma 2: unsafe

"separate your finance"

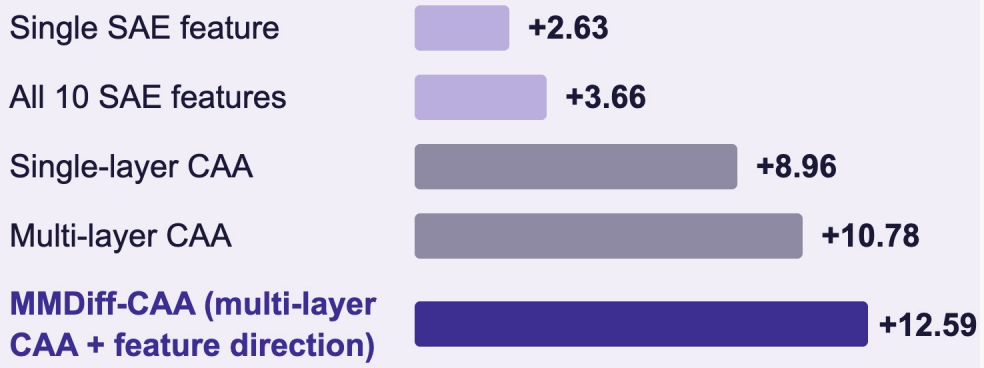
After removal: safe

"be transparent with your finance manager and lenders, track your cash flow …"

MMDiff-CAA: Targeted Feature Steering

$$h_{\ell_f} \leftarrow h_{\ell_f} + \alpha\, d_{\ell_f} + \gamma_f v_f$$ multi-layer CAA across the task-feature layers, plus the feature's decoder direction at its own layer

Spatial steering, top-10 features (ΔVSR)



Multi-layer CAA adds +1.82; the feature's decoder direction adds a further +1.81.

OCR: MMDiff-CAA **+4.02** vs. +2.21 for single-layer CAA



L6/F7539 "left", VSR

"The cup is left of the donut."

Before: wrong **No (53%)**

Steered: correct **Yes (65%)**



L12/F2257 "facing", VSR

"The person is facing the bear."

Before: wrong **No (51%)**

Steered: correct **Yes (60%)**



L20/F10687 "key info", OCRBench

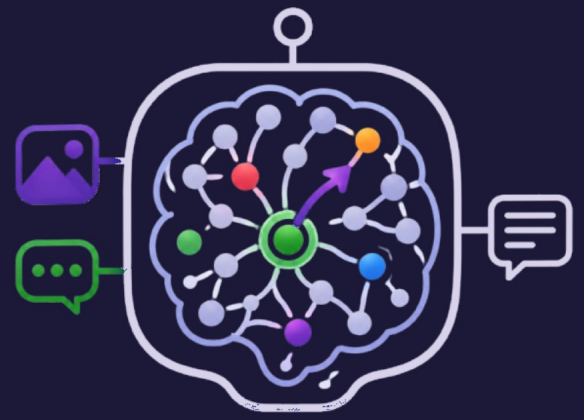
"In which city was this receipt issued?"

Before: wrong **Kuala Lumpur**

Steered: correct **Selangor**

Summary

1. **Multimodal model diffing:** SAEs initialised from base-LM dictionaries reveal the features reshaped by multimodal training and responsive to images.
2. **Task-specific feature discovery:** per-token contrastive firing finds the features behind spatial reasoning, OCR and multimodal safety.
3. **Feature-level control:** removing these features suppresses target behaviours such as unsafe responses, with ≤1.5% VQA change; MMDiff-CAA steering improves spatial reasoning and OCR.
4. **Next:** extending to identify other modality-specific features like for audio and 3D; or identifying and controlling misalignment and persona features, and steering toward safer generations.



Project pixl.cs.ox.ac.uk/mmdiff

Paper arXiv 2608.09928

Code github.com/hunarbatra/MMDiff

SAEs huggingface.co/OX-PIXL/MMDiff_SAEs