From Attribution to Action: A Human-Centered Application of Activation Steering

Tobias Labarta[1], Maximilian Dreyer[1], Katharina Weitz[1], Wojciech Samek[1,2,3], Sebastian Lapuschkin[1]
[1]Fraunhofer Heinrich-Hertz-Institut, [2]BIFOLD, [3]TU Berlin

CVPR JUNE 3-7, 2026 DENVER COLORADO

MOTIVATION

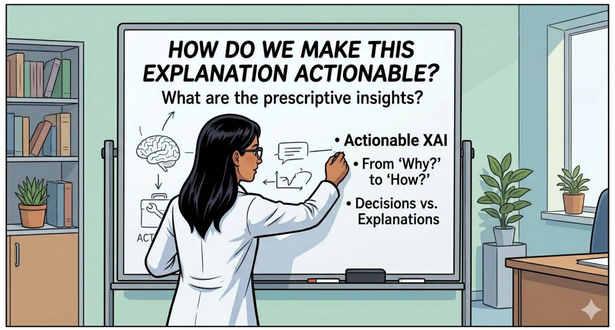


Figure generated with Google Nano Banana 2

XAI reveals feature influence on predictions … but what's next?

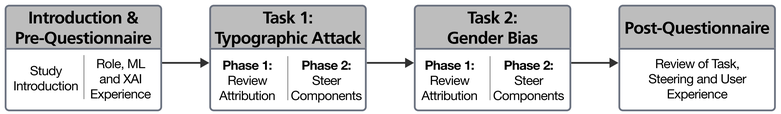
Attribution: Identify associated components, but cannot confirm causality.
Steering: Modify model internals, but requires information to act upon.

What if we combine both, how will practitioners use it?

CONTRIBUTIONS

1. **Tool:** Web tool implementation of SemanticLens combining SAE-based attribution with activation steering for instance-level debugging of CLIP.
2. **Workflow:** Predict → Review → Hypothesize → Test.
3. **Expert Interviews (n=8)**: First user study of practitioner reasoning, trust, and strategies for steering.

USER STUDY



WORKFLOW

FROM ATTRIBUTION

***1. PREDICT** instance-wise*

nurse 53%
doctor 42%
lab coat 2%

***2. REVIEW** attributions*

"stethoscope" #20773
attribution: 20.6
activation: 4.3

"woman" #3177
attribution: 7.1
activation: 3.3

***3. HYPOTHESIZE** failure modes*

Why nurse and not doctor?

What is the role of #3177?

TO ACTION

***4. TEST** with component steering*

Component #3177
attribution: 7.1→0
activation: 3.3→0

doctor 42%→85%
nurse 53%→5%
lab coat 2%→3%

What if we deactivate "woman" component #3177?

modify component

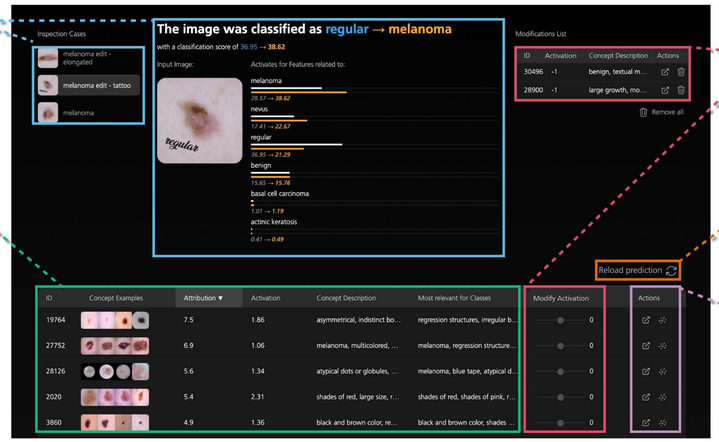
Changes indicate gender bias in "doctor" class

update prediction

TOOL

1 *select sample and predict*

2 *review attribution and interpretable components*

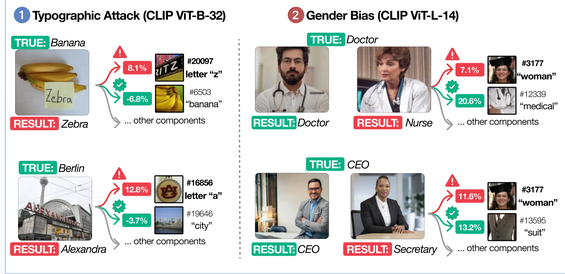


4 *modify component activation*

5 *predict with modified activations*

3 *open component details*

STUDY TASKS



KEY FINDINGS

Reasoning shift: All 8 completed causal test cycles after steering. 6/8 formed explicit hypotheses before intervening. 5/8 spontaneously acknowledged limitations of causal claims.

Trust calibration: 6/8 shifted to evidence-based trust (grounded in observed model responses, not plausibility). 3/8 demanded external validation before deployment.

Strategies: Dominant strategy: component suppression (7/8, necessity test). Only 2/8 tested amplification. Interface attribution ranking implicitly biases toward suppression.

Perceived risks: Ripple effects / non-orthogonality (3/8); insufficient instance-level validation (3/8); over-steering (2/8).

CONCLUSION

Embedding steering within a structured workflow enables practitioners to transition from **correlational inspection** to **actionable investigation**.

Contact
Tobias Labarta
Fraunhofer Heinrich-Hertz-Institut
Email: tobias.labarta@hhi.fraunhofer.de

PAPER TOOL