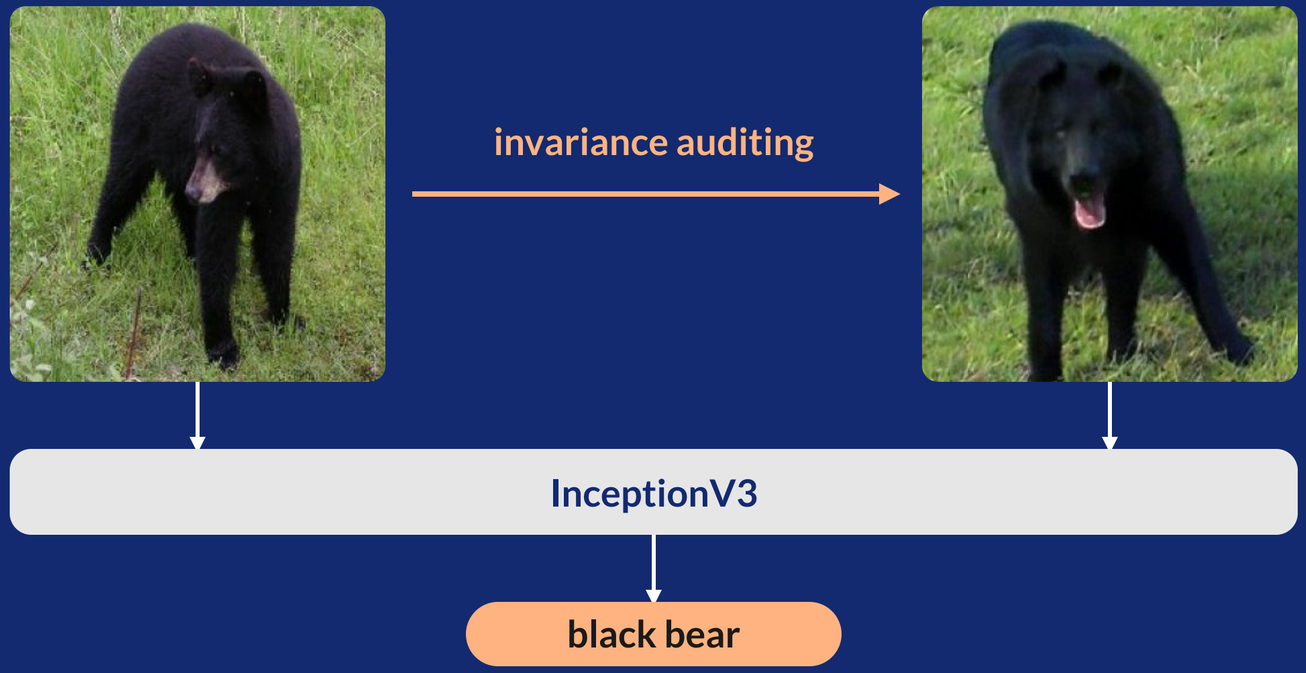
Show Me What You Don’t Know:

Efficient Sampling from Invariant Sets for Model Validation

Armand Rousselot, Joran Wendebourg, Ullrich Köthe

Computer Vision and Learning Lab, Heidelberg University · NeurIPS 2026

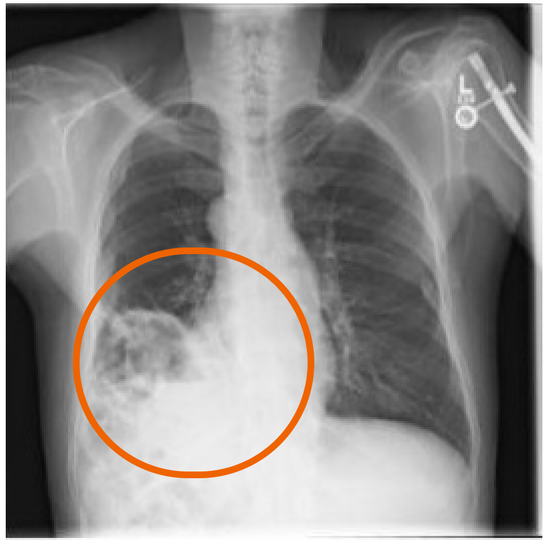


NEURAL INFORMATION PROCESSING SYSTEMS

STRUCTURES CLUSTER OF EXCELLENCE

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

VLMs place the heart on the left, even when it is on the right



Situs inversus: the heart sits on the patient’s right

“From the patient’s perspective, on which side of their body is the apex of their heart?”

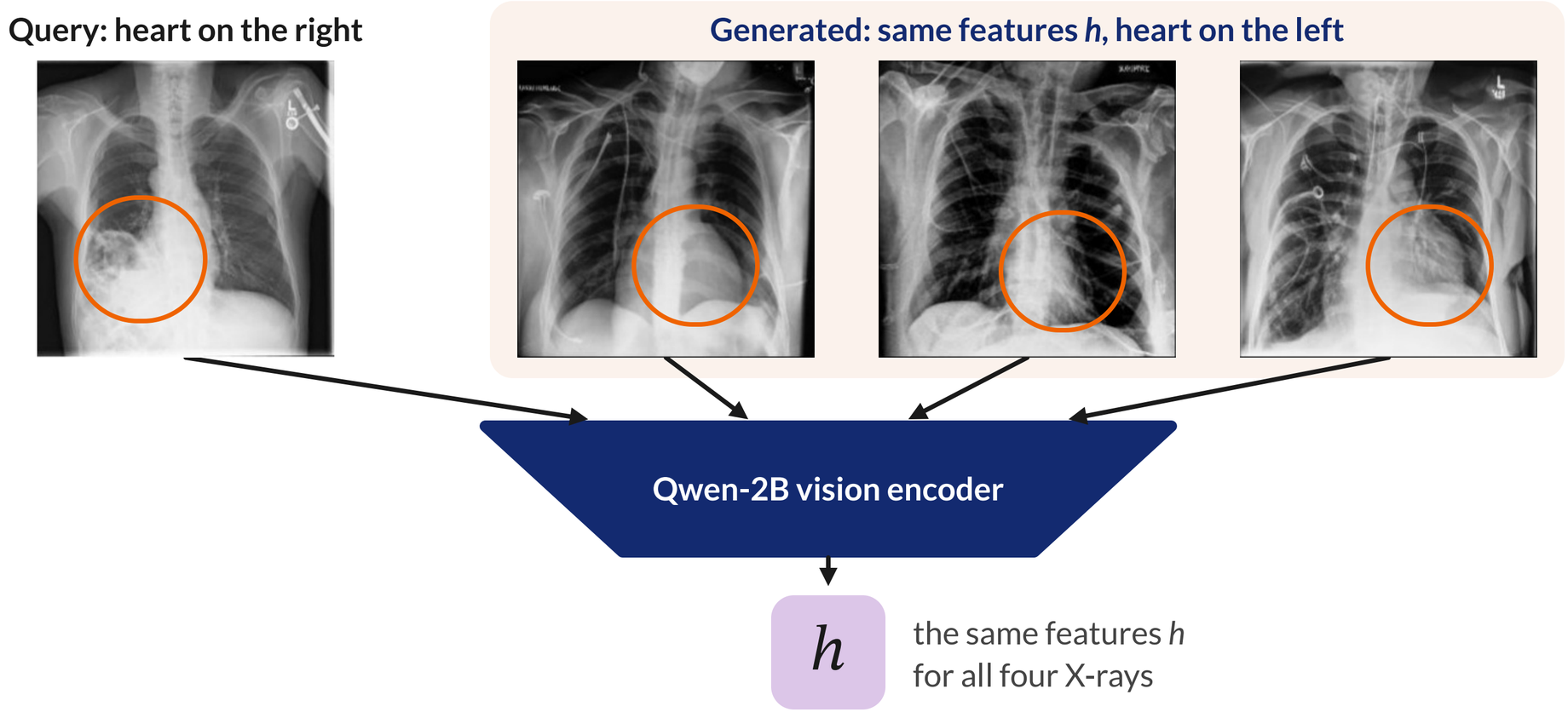
Most VLMs, including Qwen-2B: **“left”**

Physician: **“right”**

Example and question from Mayer et al., 2025

Does the language model ignore the image, or does the heart’s side fall into a blind spot of the vision encoder?

The heart’s side falls into a blind spot of Qwen-2B’s vision encoder

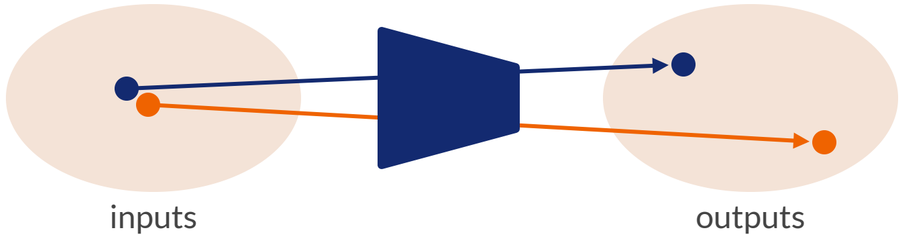


INVARIANCE AUDITING ASKS

Which inputs have the **same representation** and are thus **indistinguishable** for the subject model?

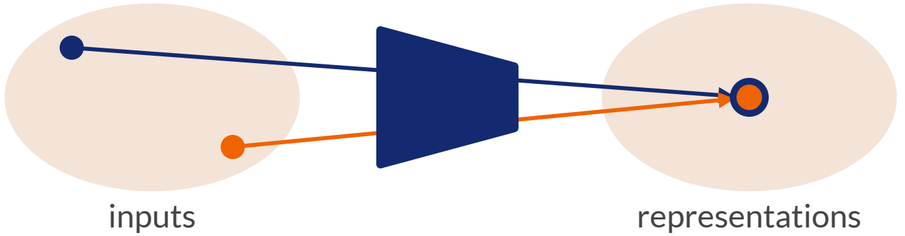
Adversarial examples

A tiny input change that alters the output



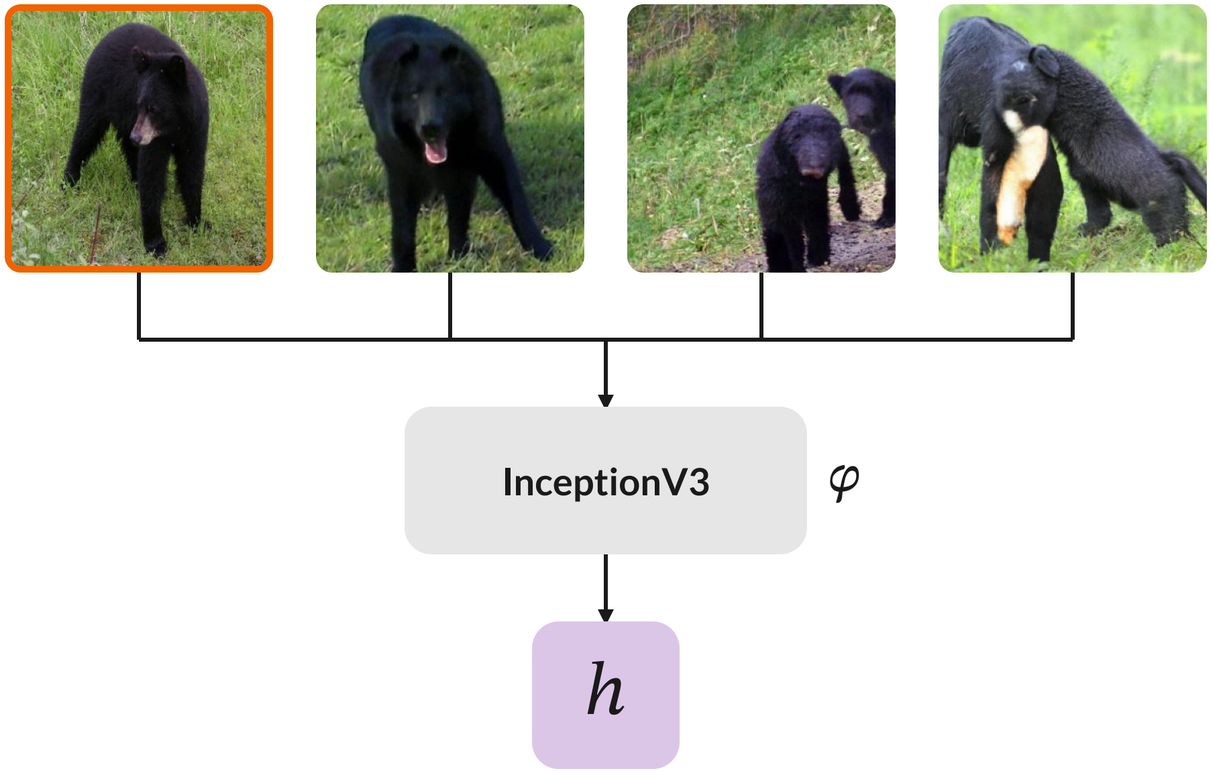
Invariance auditing

Large, natural input changes that leave the representation untouched



Some invariances are desired, like ignoring a hospital tag. Others are blind spots or shortcuts, like ignoring the side of the heart. Auditing lets you check which is which.

The fiber: every input with the same representation



The query (outlined) and three samples from its fiber: InceptionV3 maps all four to the same features *h*

Fiber of a representation *h*

$$\varphi^{-1}(h) = \{\, x : \varphi(x) = h \,\}$$

Invariance auditing draws natural, diverse samples from it

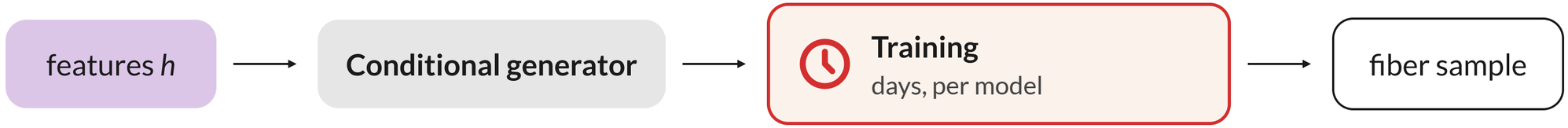
Fiber loss

$$||\varphi(\tilde{x}) - \varphi(x)||^2$$

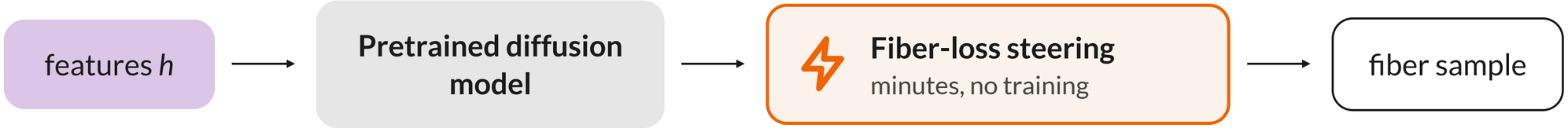
How far a sample's features are from the query's. It drives both of our methods.

Instead of training a generator for every model, we steer a pretrained one

Prior work: a conditional fiber model Rombach et al., 2020 · Bordes et al., 2021

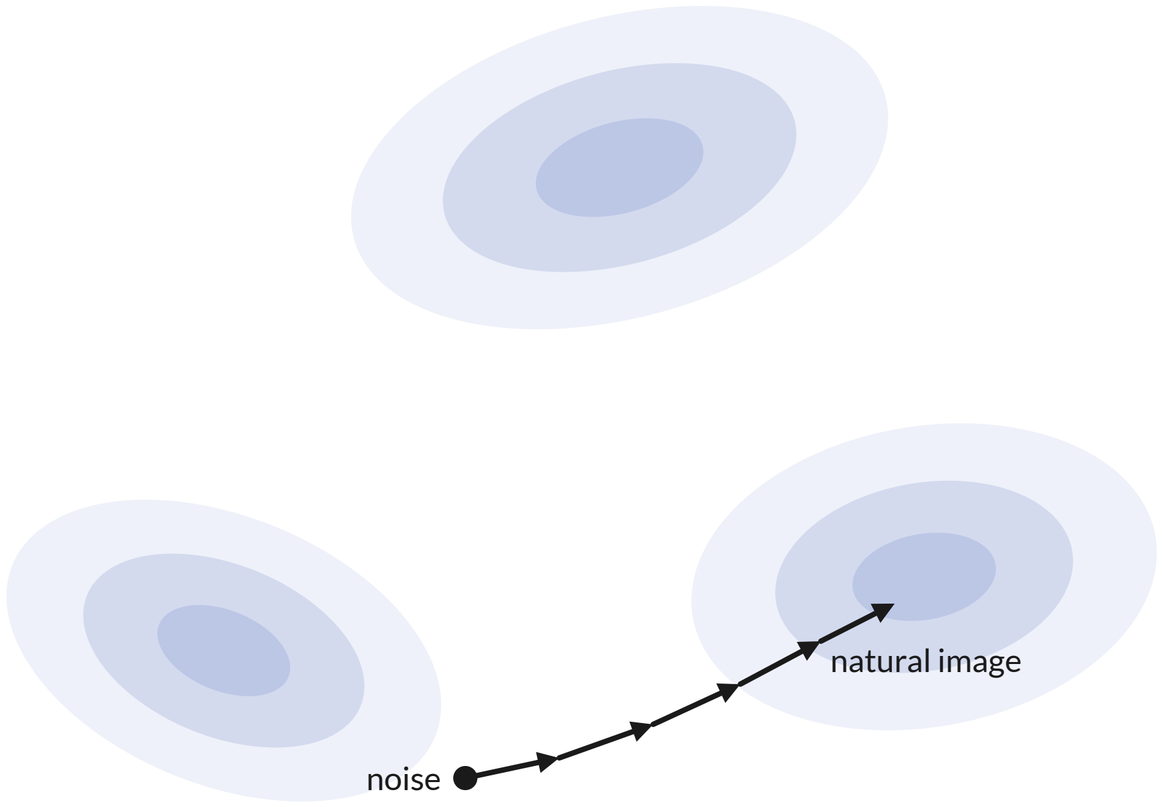


Ours: guide a pretrained diffusion model no training

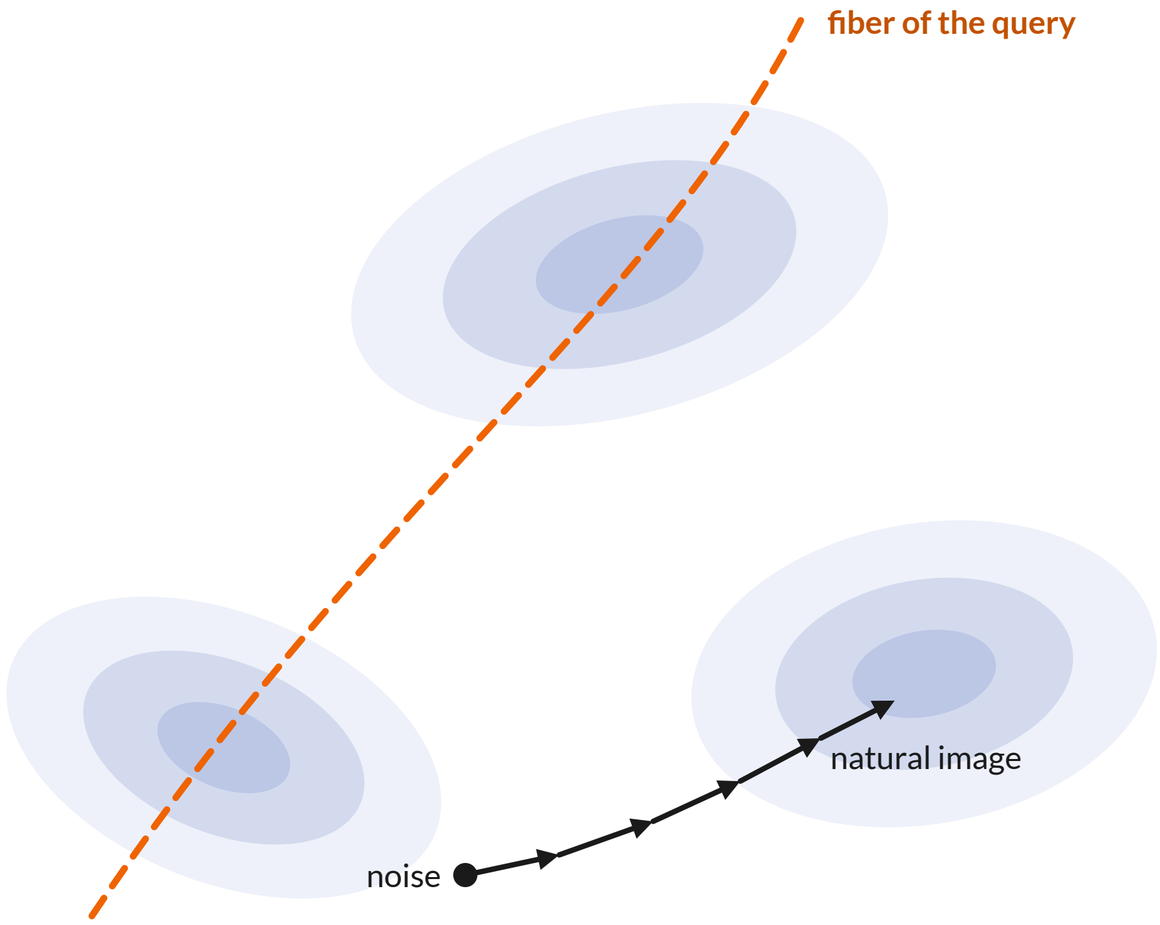


Guidance steers each denoising step onto the fiber

Sampling. A pretrained diffusion model denoises noise into a natural image (shaded) in black steps.



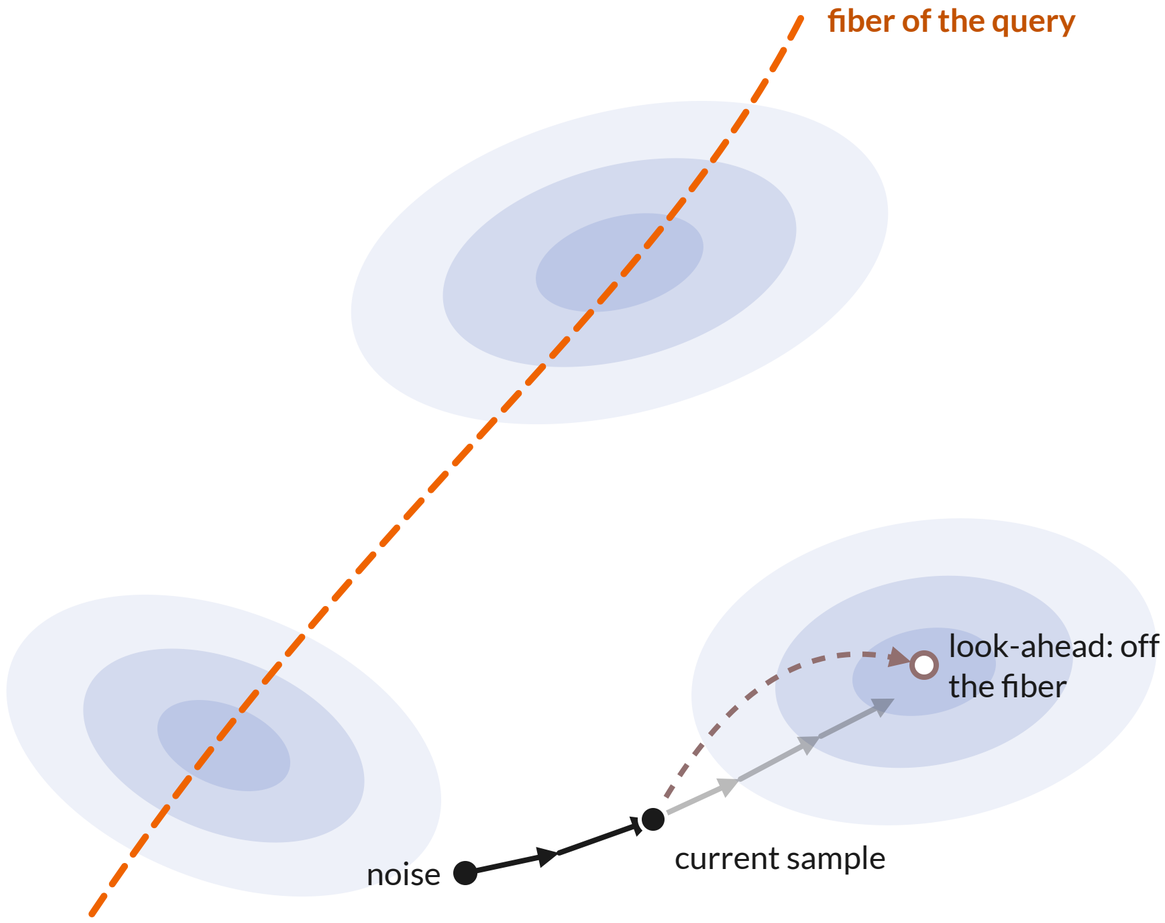
Guidance steers each denoising step onto the fiber



Sampling. A pretrained diffusion model denoises noise into a natural image (shaded) in black steps.

Target. An image on the query's fiber (dashed orange line), i.e. with features *h*.

Guidance steers each denoising step onto the fiber



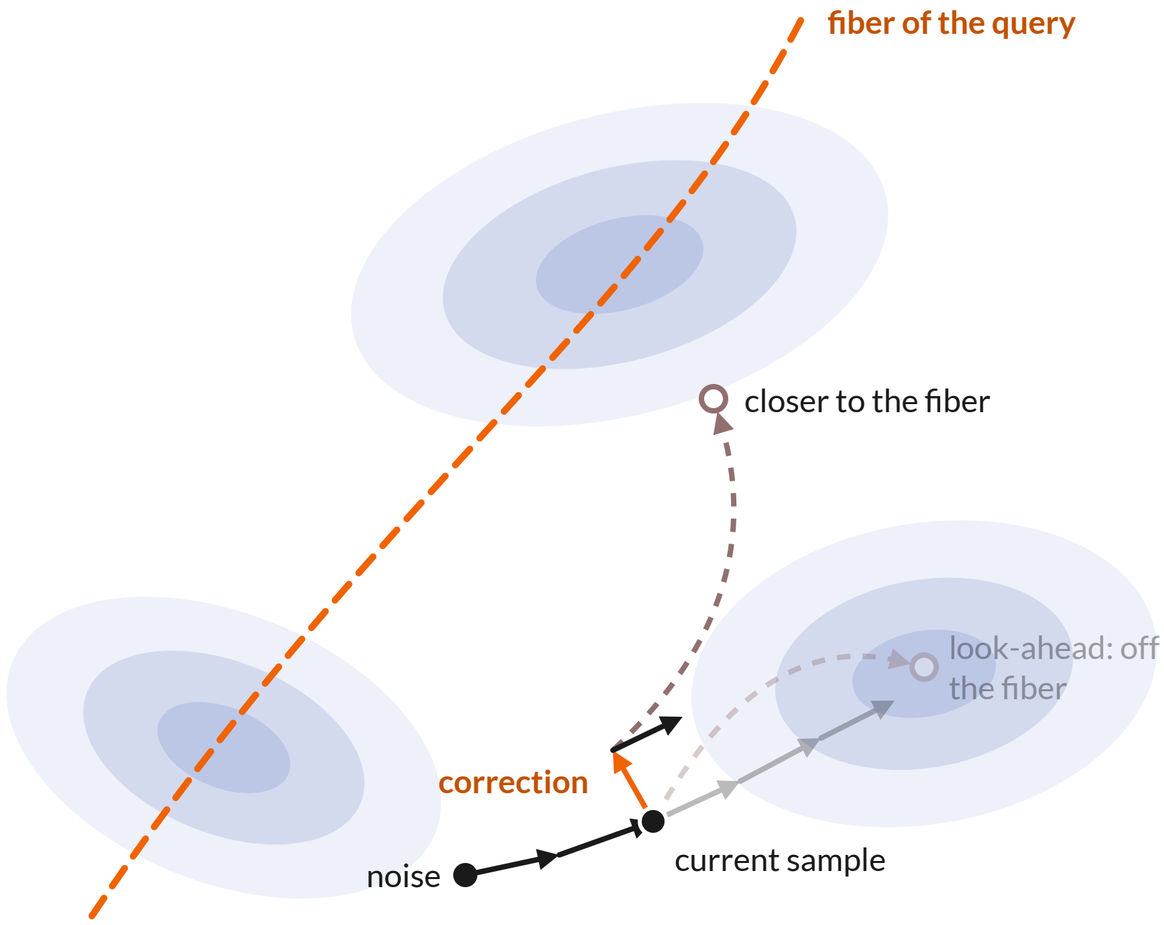
Sampling. A pretrained diffusion model denoises noise into a natural image (shaded) in black steps.

Target. An image on the query's fiber (dashed orange line), i.e. with features *h*.

1 · Look ahead to the final image (Tweedie's formula, dashed arrow) and measure its fiber loss:

$$||\varphi(\hat{x}_0) - h||^2$$

Guidance steers each denoising step onto the fiber



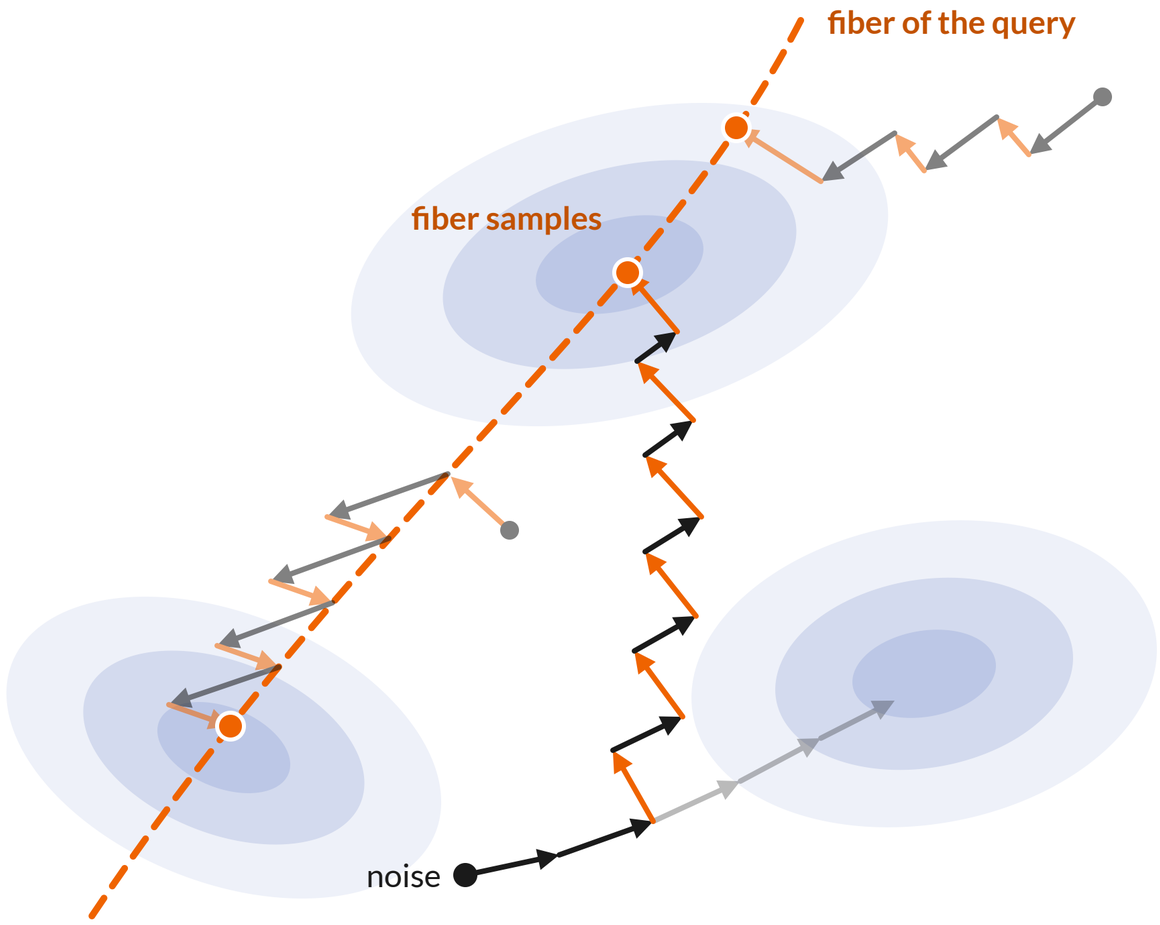
Sampling. A pretrained diffusion model denoises noise into a natural image (shaded) in black steps.

Target. An image on the query's fiber (dashed orange line), i.e. with features *h*.

1 · Look ahead to the final image (Tweedie's formula, dashed arrow) and measure its fiber loss: $||\varphi(\hat{x}_0) - h||^2$

2 · Correct the sample with a few gradient steps on that loss (orange arrow), then denoise as usual.

Guidance steers each denoising step onto the fiber



Sampling. A pretrained diffusion model denoises noise into a natural image (shaded) in black steps.

Target. An image on the query's fiber (dashed orange line), i.e. with features *h*.

1 · Look ahead to the final image (Tweedie's formula, dashed arrow) and measure its fiber loss: $||\varphi(\hat{x}_0) - h||^2$

2 · Correct the sample with a few gradient steps on that loss (orange arrow), then denoise as usual.

Repeat every step: a natural image with the query's features. New noise, new sample.

No training · NDTM guidance (Pandey et al., 2025)

No training: fiber samples in minutes, not days

≈ 10 days

of GPU training before a conditional fiber model yields its first sample, for every new subject model

≈ 2 min

per sample with our guidance, starting right away

	Contracts to fiber	Fiber distribution	Fast setup	Fast sampling
Conditional model	○	✓	×	✓
+ fiber loss (ours)	✓	✓	×	✓
Guidance (ours)	✓	○	✓	×

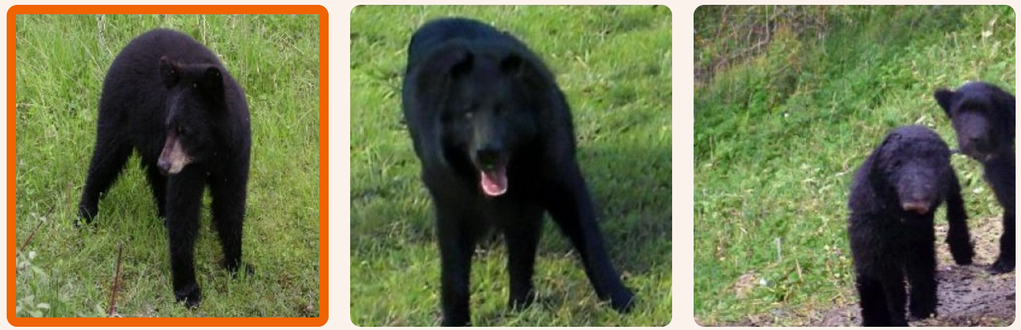
✓ yes ○ approximately × no

Complementary: guidance for fast iteration during model development, fiber-loss-trained models for large-scale statistics.

ImageNet on one NVIDIA A40: ≈ 14,000 min to train a conditional fiber model; 1.8 min per guided sample (100 denoising steps). Training pays off only beyond ≈ 9k samples.

One tool, many blind spots and shortcuts

FID is artificially low in the blind spots of InceptionV3



The query (outlined) and two samples from its InceptionV3 fiber

3.8

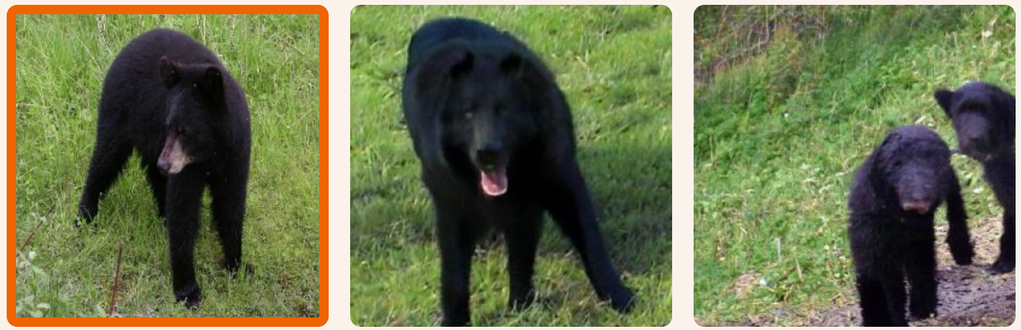
FID with the audited InceptionV3 (PyTorch)

33.5

FID of the same samples with TF's Inception

One tool, many blind spots and shortcuts

FID is artificially low in the blind spots of InceptionV3



The query (outlined) and two samples from its InceptionV3 fiber

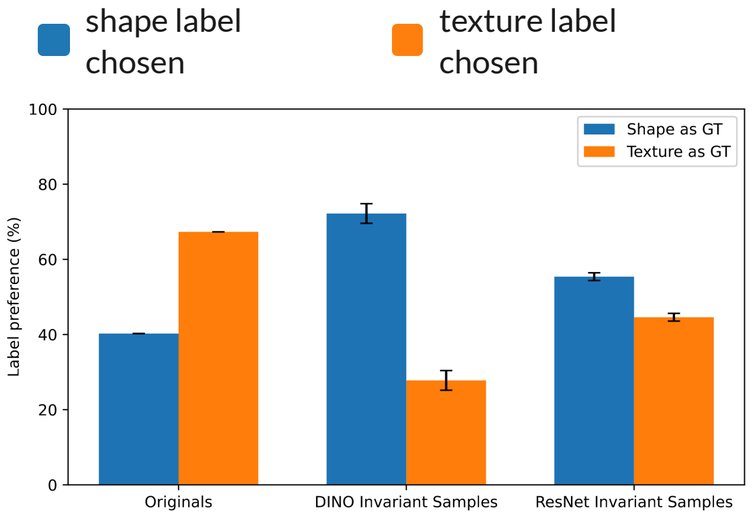
3.8

FID with the audited InceptionV3 (PyTorch)

33.5

FID of the same samples with TF's Inception

Invariance auditing shows whether a model sees shape or texture



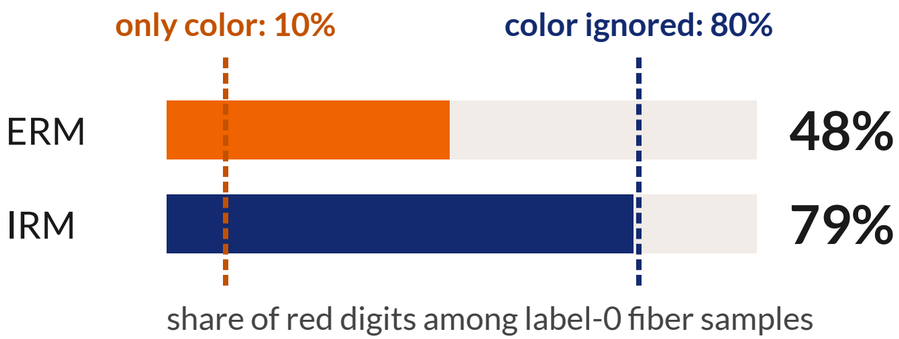
An external ConvNeXt labels cue-conflict images and each model's fiber samples

DINOv2 keeps shape. ResNet-50's features keep it too: its texture bias enters only at the last layer.

One tool, many blind spots and shortcuts

Catching a learned shortcut, and confirming the fix

Color predicts the label in training but not at test time: relying on it does not generalize.

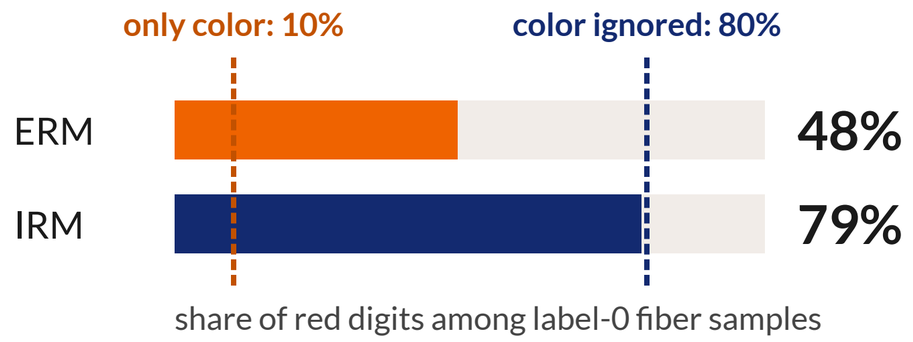


ERM lands between the lines: the model uses color. After IRM it sits on the 80% line: color is ignored.

One tool, many blind spots and shortcuts

Catching a learned shortcut, and confirming the fix

Color predicts the label in training but not at test time: relying on it does not generalize.



ERM lands between the lines: the model uses color. After IRM it sits on the 80% line: color is ignored.

Two chest X-ray classifiers disagree on each other's fibers

Original



BiomedClip: 35.8%
ConvNeXt: 22.4%

BiomedClip fiber sample



BiomedClip: 35.4%
ConvNeXt: **68.8%**

Same BiomedClip features, yet ConvNeXt's cardiomegaly score triples: a change BiomedClip is blind to.

Invariance auditing reveals what a model ignores: its **blind spots** and **learned shortcuts**. Our guidance makes it fast and training-free.

Paper
arxiv.org/abs/2603.21782

Code
github.com/vislearn/InvarianceAuditing

Come talk to us at our poster: Paris, Poster Session 2