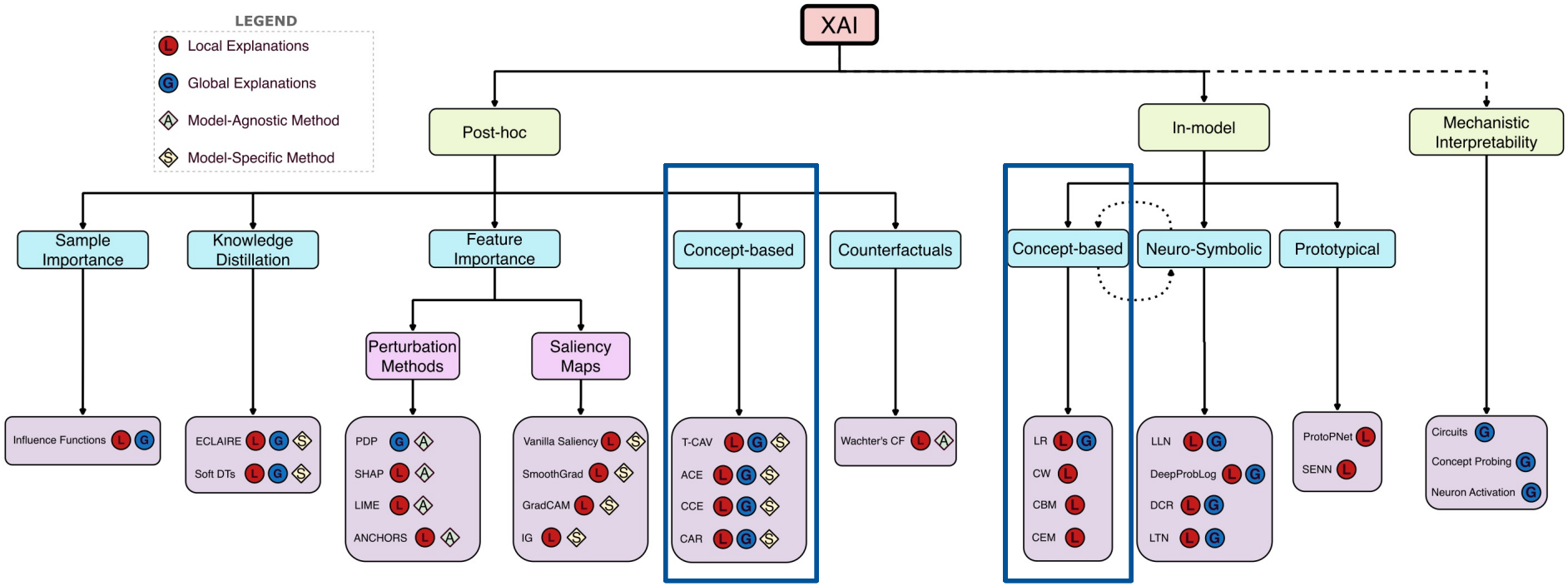
UNIVERSITY OF CAMBRIDGE

Beyond Concept Bottleneck Models: How to Make Black Boxes Intervenable?

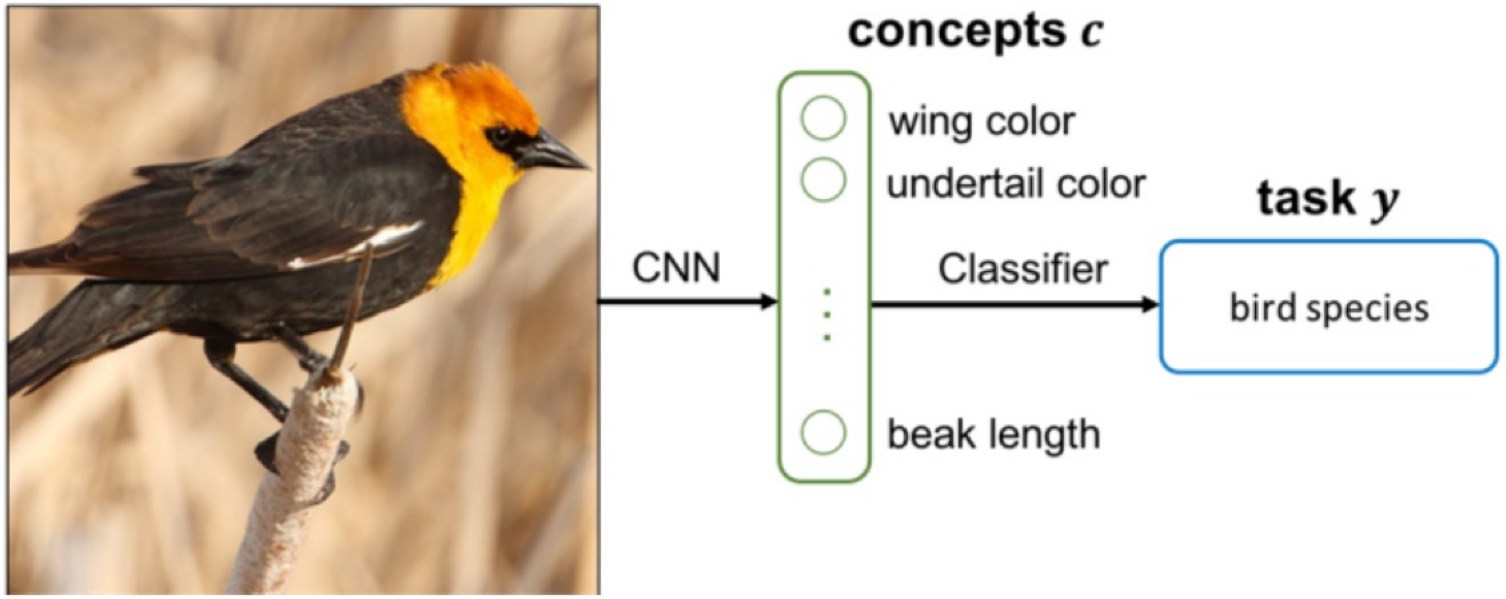
Sonia Laguna*,Riˇcards Marcinkeviˇcs*,Moritz Vandenhirtz, Julia E. Vogt

L193 Explainable AI, Qianqi Liu

Explanation Types



Concept Bottleneck Models (CBMs) Recap



- Key advantage: user can intervene on concepts
- Key shortcoming: requires concept labels for the entire training set

Motivations

- Widespread Deployment of Black-Box Models

  Post-hoc intervenability on *any* pretrained black box

- Desire for Instance-Specific Interventions

  Formal definition of *intervenability* as a performance gain from concept corrections

- Challenge of Full Concept Labeling

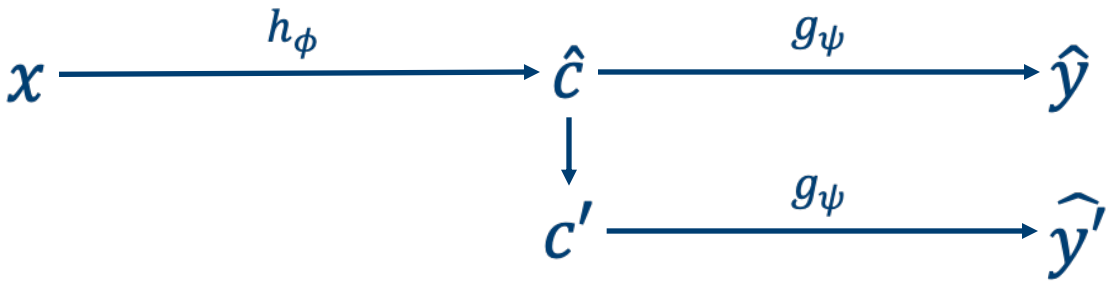
  Small concept-labeled validation set suffices

- Balancing Performance and Intervenability

  Fine-tuning black boxes to *increase* intervenability

Method

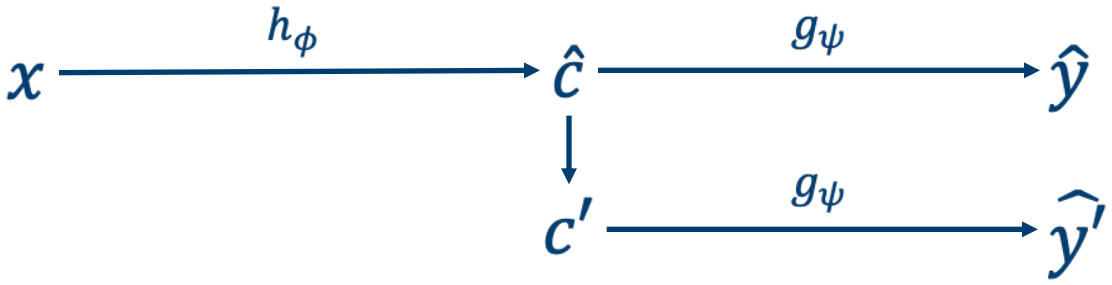
CBMs



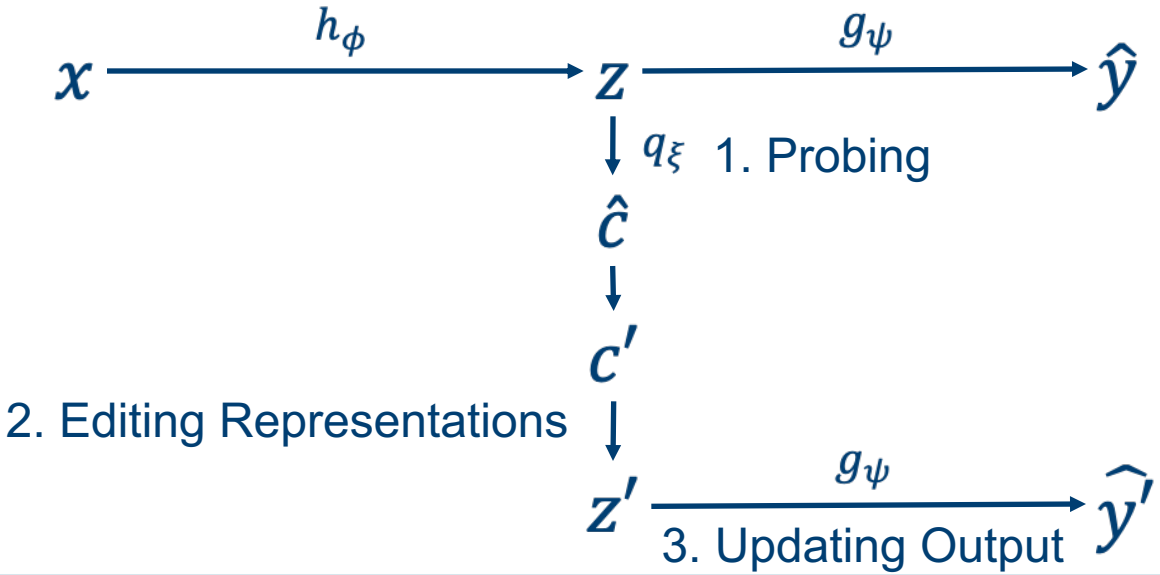
Black-box models

Method

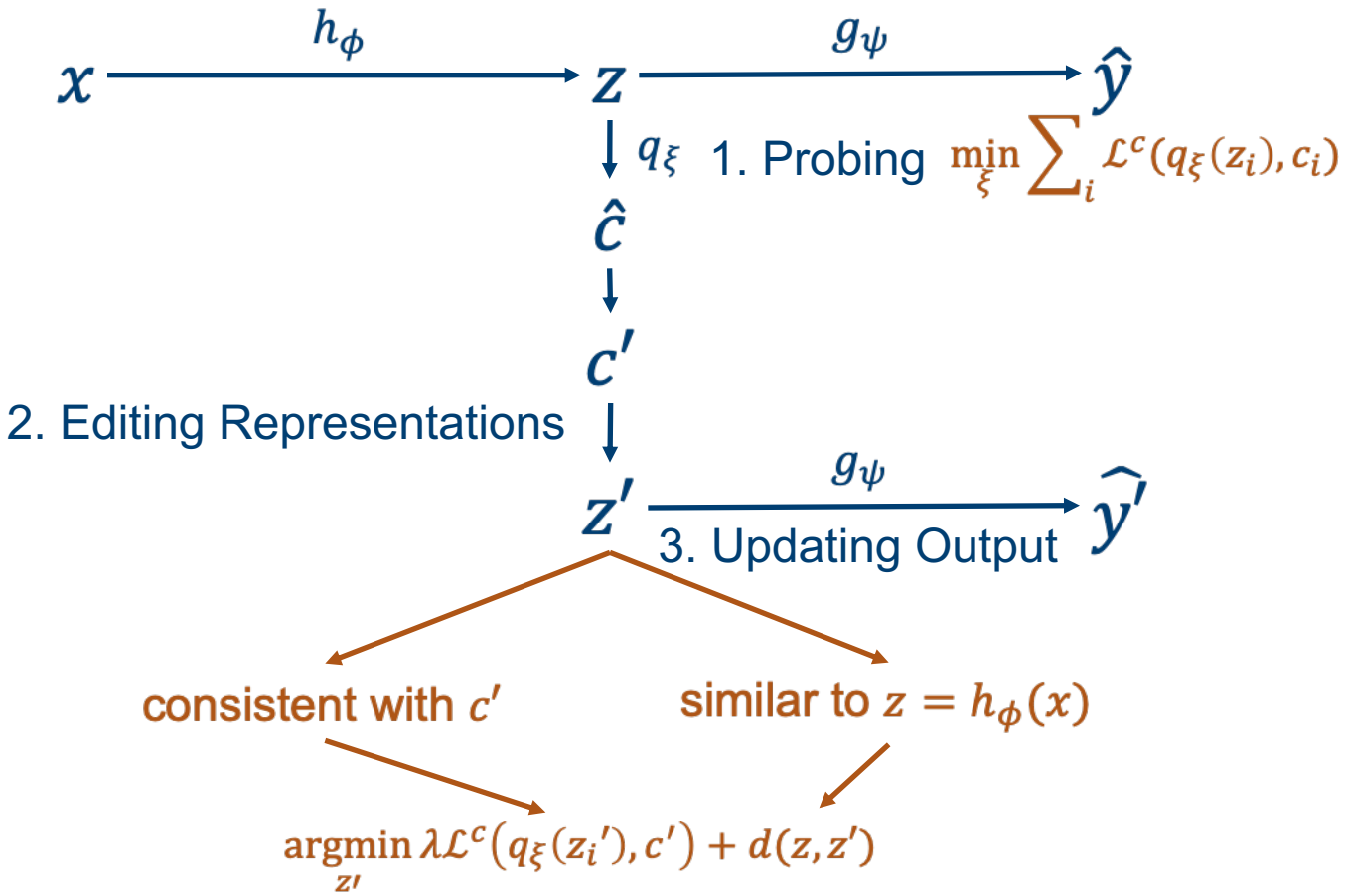
CBMs



Black-box models



Method


x
h_ϕ
z
g_ψ
$\hat{y}$
q_ξ
1. Probing $\min_\xi \sum_i \mathcal{L}^c(q_\xi(z_i), c_i)$
$\hat{c}$
c'
2. Editing Representations
z'
g_ψ
$\hat{y}'$
3. Updating Output
consistent with c'
similar to $z = h_\phi(x)$
$\underset{z'}{\operatorname{argmin}}\, \lambda \mathcal{L}^c(q_\xi(z_i'), c') + d(z, z')$

Intervenability

CBMs

$$\mathbb{E}_{(x,c,y)\sim D}\left[\mathbb{E}_{c'\sim\pi}\left[\mathcal{L}^y\left(f_\theta(x),\, y\right) \;-\; \mathcal{L}^y\left(g_\psi(c'),\, y\right)\right]\right]$$

Black-box models

$$\mathbb{E}_{(x,c,y)\sim D,\; c'\sim\pi}\left[\mathcal{L}^y\left(f_\theta(x), y\right) \;-\; \mathcal{L}^y\left(g_\psi(z'),\, y\right)\right],$$

$$\text{where} \quad z' \in \mathit{argmin}_{\tilde{z}}\, \lambda\, \mathcal{L}^c\left(q_\xi(\tilde{z}),\, c'\right) \;+\; d\left(z,\, \tilde{z}\right).$$

Fine-tuning for Intervenability

Intervenability

$$\min_{\psi,\, z'} \; \mathbb{E}_{(x,c,y)\sim D,\; c'\sim\pi}\Big[\mathcal{L}^{y}\big(g_{\psi}(z'),\, y\big)\Big],$$

$$\text{s.t.} \quad z' \in \mathit{argmin}_{\tilde{z}}\, \lambda\, \mathcal{L}^{c}\big(q_{\xi}(\tilde{z}),\, c'\big) \;+\; d\big(z,\, \tilde{z}\big).$$

Weight

$$\min_{\varphi,\, \psi,\, z'} \; \mathbb{E}_{(x,c,y)\sim D,\; c'\sim\pi}\Big[(1-\beta)\, \mathcal{L}^{y}\big(g_{\psi}(h_{\varphi}(x)),\, y\big) \;+\; \beta\, \mathcal{L}^{y}\big(g_{\psi}(z'),\, y\big)\Big],$$

$$\text{s.t.} \quad z' \in \mathit{argmin}_{\tilde{z}}\, \lambda\, \mathcal{L}^{c}\big(q_{\xi}(\tilde{z}),\, c'\big) \;+\; d\big(z,\, \tilde{z}\big).$$

Datasets

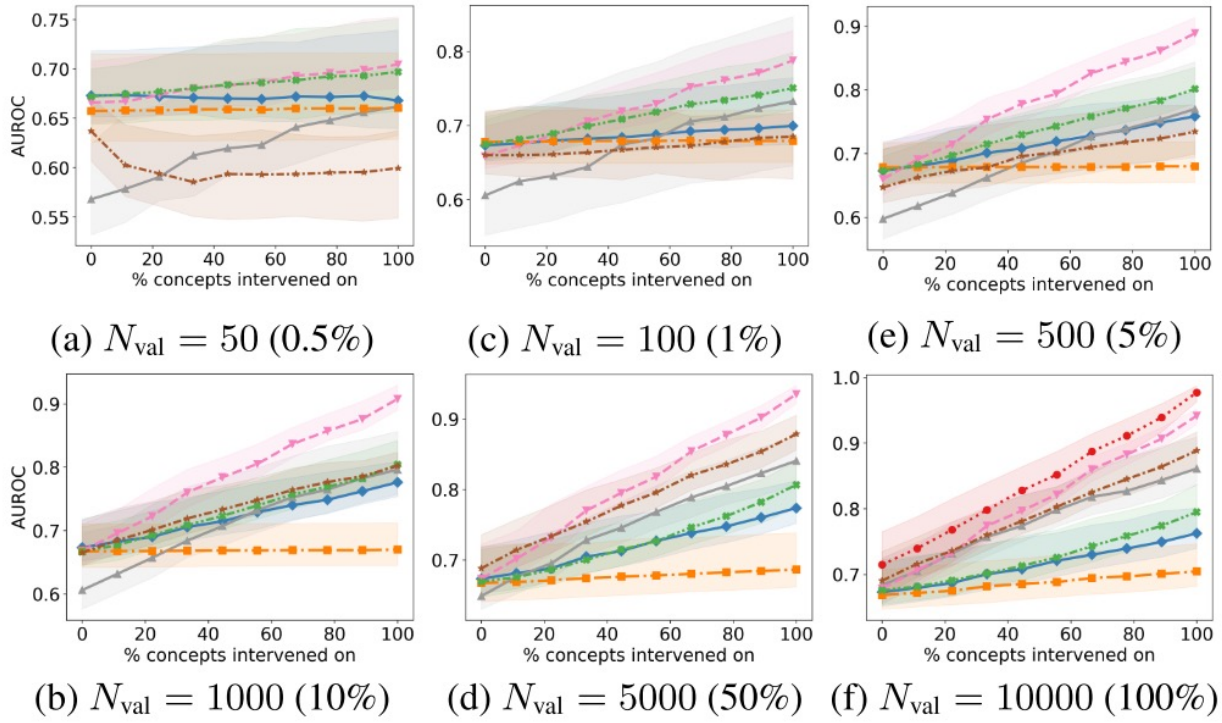
- **Synthetic**
  - To test performance under controlled bottleneck vs. incomplete concept scenarios.
- **AwA2**: 85 binary attributes and species labels
  - Clearly labeled animal attributes. Demonstrates concept-based classification in images.
- **CUB**
  - Fine-grained bird attributes to test model detail sensitivity.
- **CIFAR-10 & ImageNet**: 143 and 100 attributes respectively
  - Standard vision benchmark + demonstrating concept extraction with large vision-language models.
- **MIMIC-CXR & CheXpert**: 14 binary attributes
  - Real-world medical imaging with partial concept labels from radiology reports.

Evaluation metrics: AUROC, AUPR, Brier scores, calibration curves.

Baselines

- **Black-box**: standard neural network f_θ trained only on (x, y)
- **Concept Bottleneck Models (CBMs)**: predicts concepts first, then uses them as a "bottleneck" to predict y
- **Post-hoc CBM**: converts a pretrained black-box into a CBM-like architecture by training a probe for concepts and then re-training a final layer.
- **Fine-tuned (I)**:The paper's proposed approach
- **Fine-tuned (A)**: concatenates concept labels with hidden features and fine-tune
- **Fine-tuned (MT):** fine-tunes the black box with a multi-task loss for label and concept prediction

Results



comparable to CBM

only appears in (f)

Results

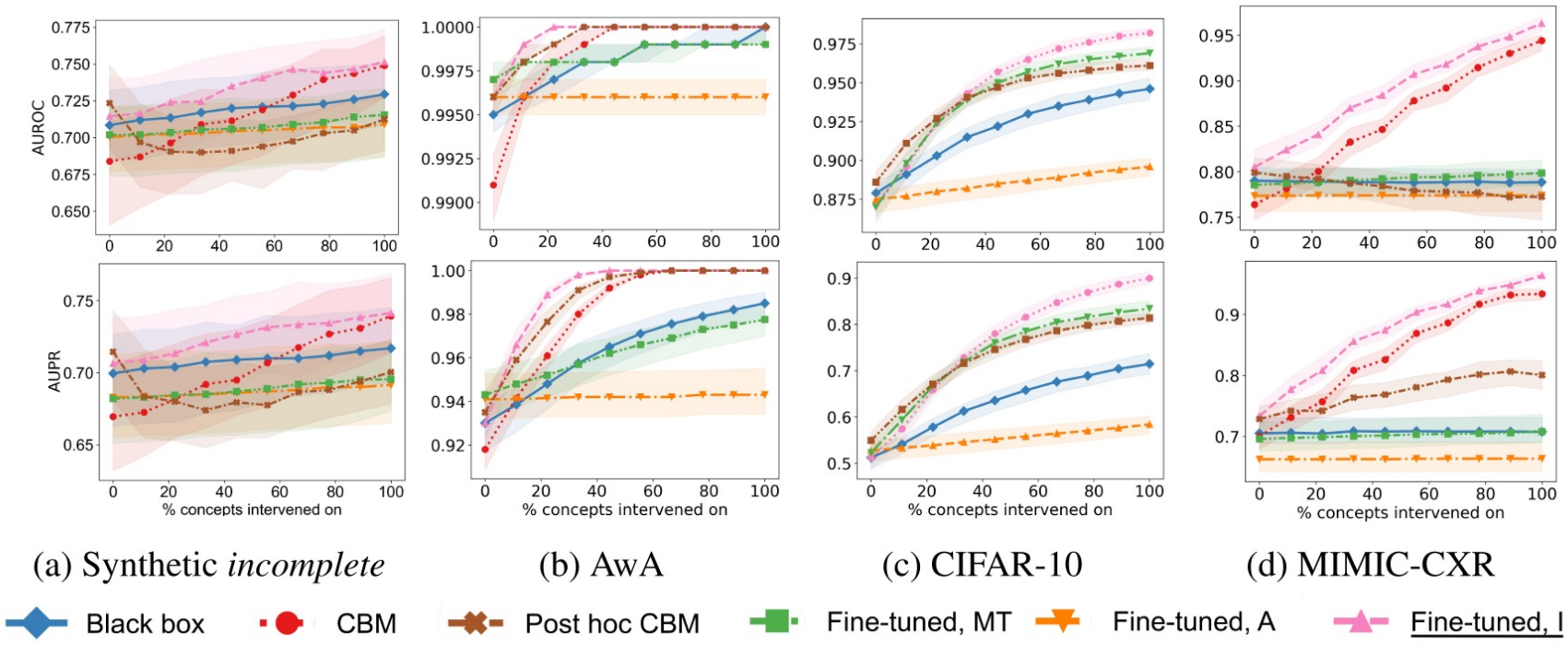


Figure 4: Intervention results on the (a) synthetic *incomplete*, (b) AwA2, (c) CIFAR-10, and (d) MIMIC-CXR datasets w.r.t. target AUROC (*top*) and AUPR (*bottom*) across ten seeds.

Results

Dataset	Model	Concepts			Target		
		AUROC	AUPR	Brier	AUROC	AUPR	Brier
Synthetic	Black box	0.716±0.018	0.710±0.017	0.208±0.006	0.686±0.043	0.675±0.046	0.460±0.003
	CBM	**0.837±0.008**	**0.835±0.008**	*0.196±0.006*	**0.713±0.040**	**0.700±0.038**	*0.410±0.012*
	Post hoc CBM	0.714±0.017	0.707±0.018	0.207±0.009	*0.707±0.049*	*0.698±0.048*	**0.285±0.015**
	Fine-tuned, A	—	—	—	0.682±0.047	0.668±0.046	0.470±0.004
	Fine-tuned, MT	*0.784±0.013*	*0.780±0.014*	**0.186±0.006**	0.687±0.046	0.668±0.043	0.471±0.003
	Fine-tuned, I	0.716±0.018	0.710±0.017	0.208±0.006	0.695±0.051	0.685±0.051	**0.285±0.014**
AwA2	Black box	0.991±0.002	*0.979±0.006*	0.027±0.006	*0.996±0.001*	0.926±0.020	0.199±0.038
	CBM	*0.993±0.001*	*0.979±0.002*	*0.025±0.001*	0.988±0.001	0.892±0.005	0.234±0.009
	Post hoc CBM	0.992±0.002	0.976±0.005	*0.025±0.005*	*0.996±0.001*	*0.929±0.018*	**0.170±0.033**
	Fine-tuned, A	—	—	—	*0.996±0.001*	**0.938±0.016**	**0.170±0.036**
	Fine-tuned, MT	**0.994±0.002**	**0.985±0.004**	**0.022±0.005**	**0.997±0.001**	**0.938±0.017**	*0.178±0.038*
	Fine-tuned, I	0.991±0.002	*0.979±0.005*	0.027±0.006	*0.996±0.001*	0.925±0.020	0.195±0.040
CIFAR-10	Black box	*0.713±0.002*	*0.802±0.001*	*0.110±0.000*	*0.879±0.001*	0.504±0.004	0.920±0.006
	CBM	—	—	—	—	—	—
	Post hoc CBM	0.675±0.009	0.785±0.003	0.125±0.004	**0.888±0.001**	**0.541±0.004**	**0.624±0.003**
	Fine-tuned, A	—	—	—	0.876±0.002	*0.518±0.004*	0.896±0.005
	Fine-tuned, MT	**0.729±0.002**	**0.807±0.001**	**0.109±0.000**	0.870±0.004	0.512±0.009	*0.890±0.014*
	Fine-tuned, I	*0.713±0.002*	*0.802±0.001*	*0.110±0.000*	0.873±0.003	0.501±0.007	0.902±0.021
MIMIC-CXR	Black box	0.743±0.006	0.170±0.004	*0.046±0.001*	0.789±0.006	0.706±0.009	0.444±0.003
	CBM	*0.744±0.006*	**0.224±0.003**	0.053±0.001	0.765±0.007	0.699±0.006	0.427±0.003
	Post hoc CBM	0.707±0.006	0.154±0.006	*0.046±0.001*	*0.801±0.006*	*0.727±0.008*	**0.301±0.005**
	Fine-tuned, A	—	—	—	0.773±0.009	0.665±0.013	0.459±0.004
	Fine-tuned, MT	**0.748±0.008**	*0.187±0.003*	**0.045±0.001**	0.785±0.006	0.696±0.009	0.450±0.008
	Fine-tuned, I	*0.744±0.005*	0.172±0.005	*0.046±0.001*	**0.808±0.007**	**0.733±0.009**	*0.314±0.015*

Pros and Cons

Pros

•Can be applied post hoc to any pre-trained network.

•Small validation set

•Effective even with incomplete concepts

•Architecture-agnostic

•Instance-specific interventions

Cons

•Still needs some concept labels

•Added complexity

•Highly depends on probe's quality

•Potential misalignment of concepts

•User burden

Thanks and Q&A?