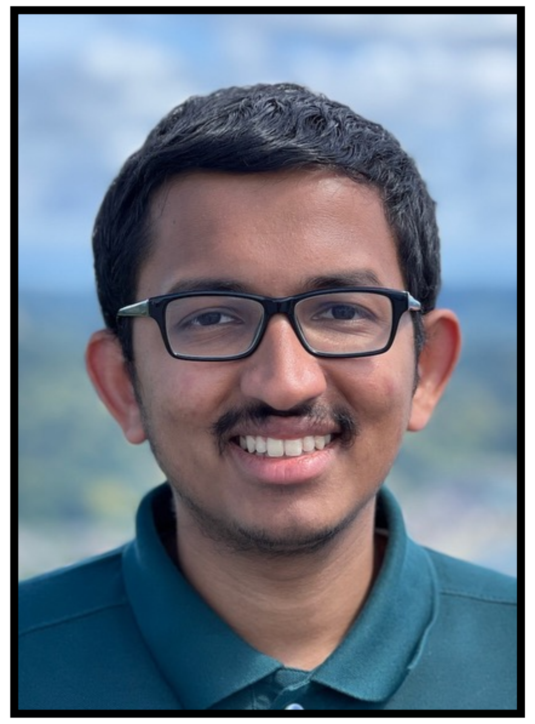
MAX PLANCK INSTITUTE
FOR INFORMATICS

SIC Saarland Informatics Campus

NEURAL INFORMATION PROCESSING SYSTEMS

B-cosification: Transforming Deep Neural Networks to be Inherently Interpretable

Shreyash Arya*



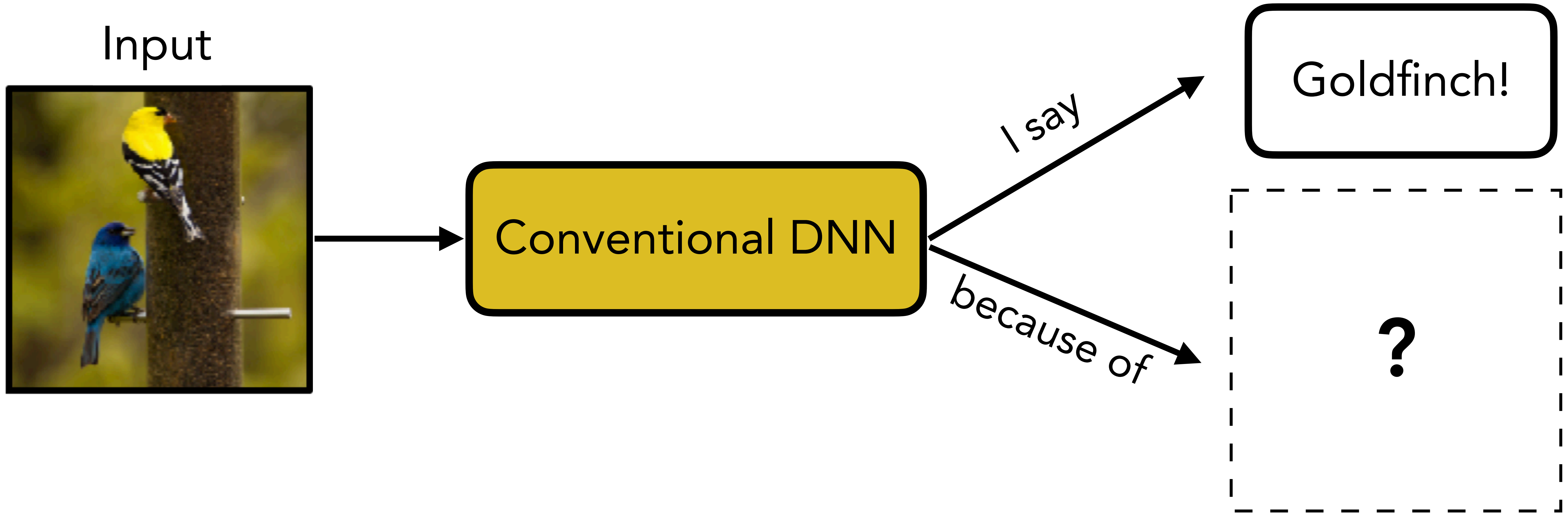
Sukrut Rao*

Moritz Böhle*

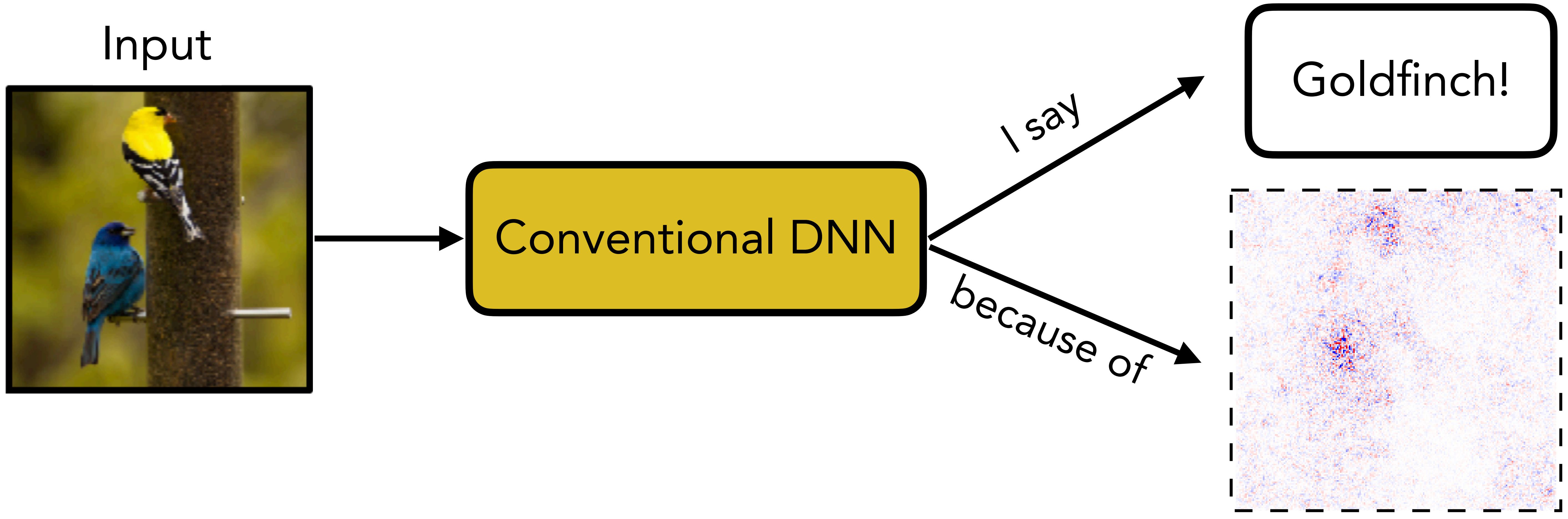
Bernt Schiele

Max Planck Institute for Informatics, Saarland Informatics Campus

Post-hoc Explanations for Understanding Deep Networks



Post-hoc Explanations for Understanding Deep Networks

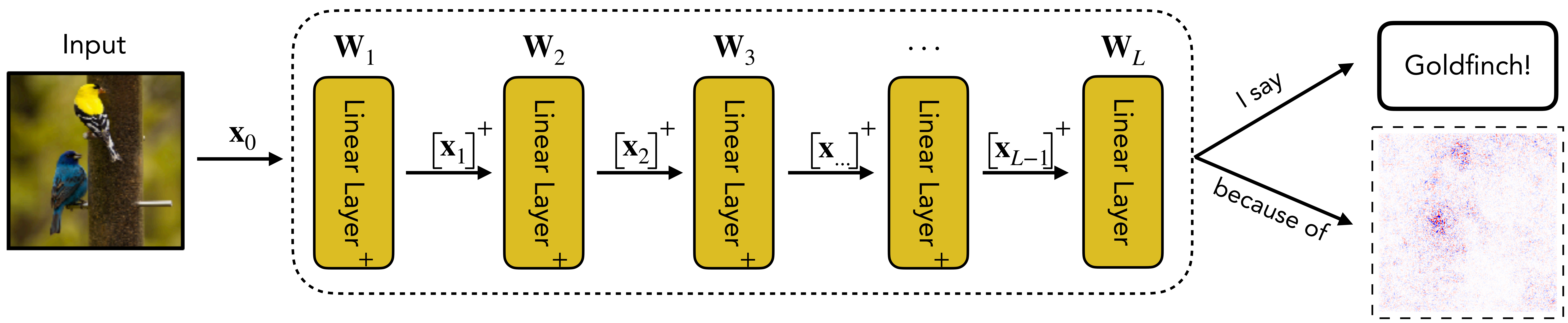


Post-hoc Attribution Explanation

- Can be directly applied to existing models
- May not be model-faithful[1]
- Often not human interpretable

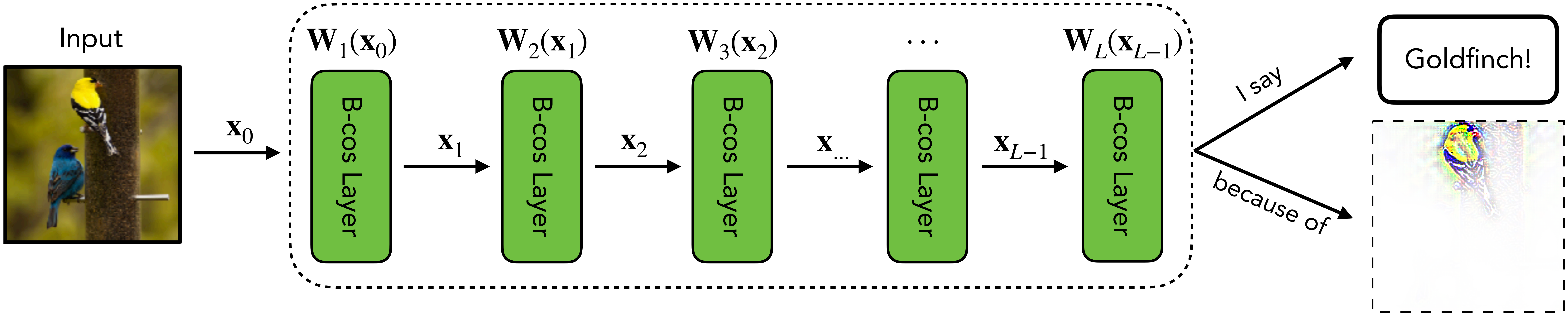
[1]Sanity Checks for Saliency Maps [Adebayo et al., NeurIPS 2018]

B-cos Networks[2]: Inherently Interpretable Explanations



[2]B-cos Networks [Böhle et al., CVPR 2022]

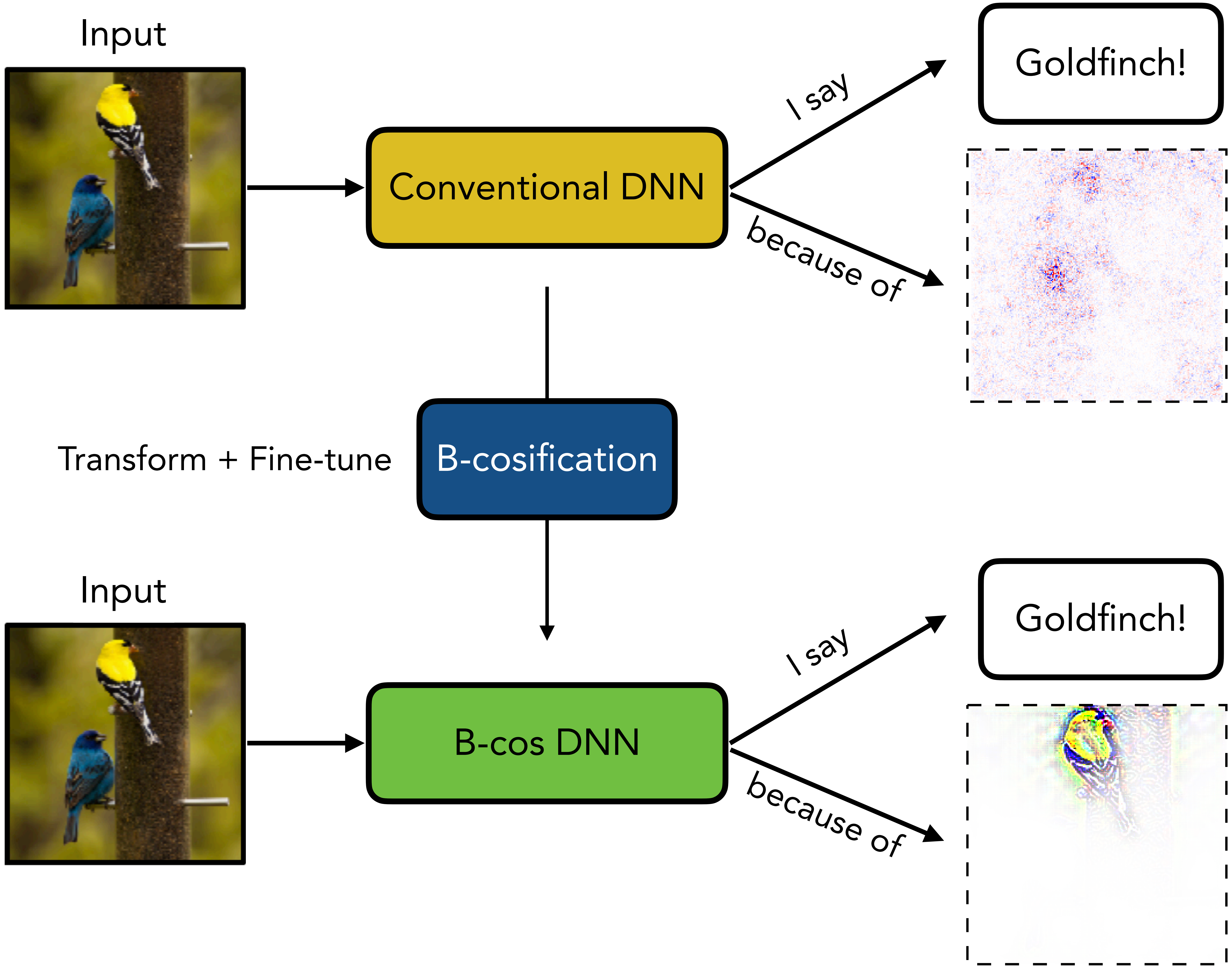
B-cos Networks²: Inherently Interpretable Explanations



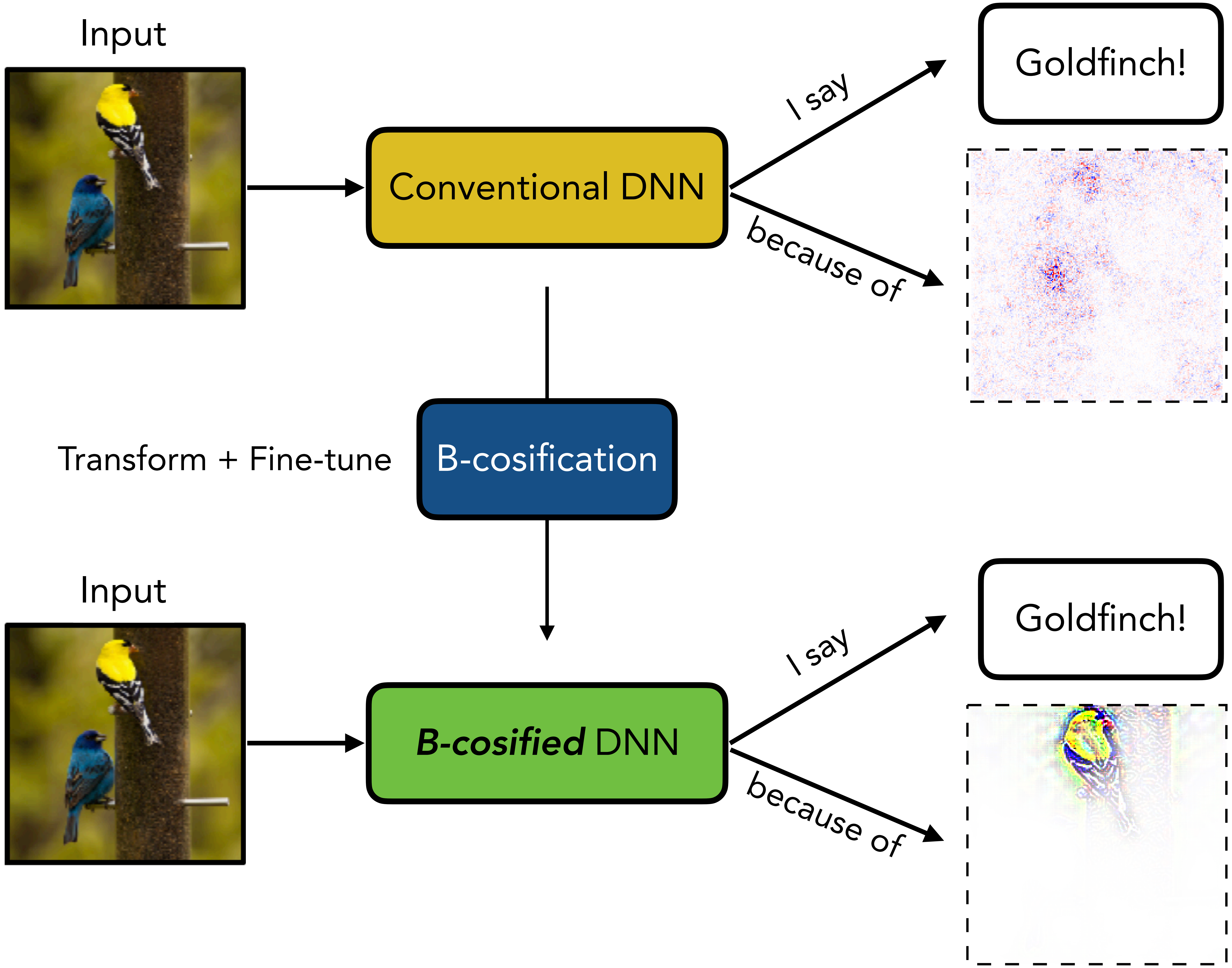
- Human Interpretable
- Model-faithful by design
- Need to train models from scratch to obtain

[2] B-cos Networks [Böhle et al., CVPR 2022]

Our work: B-cosification

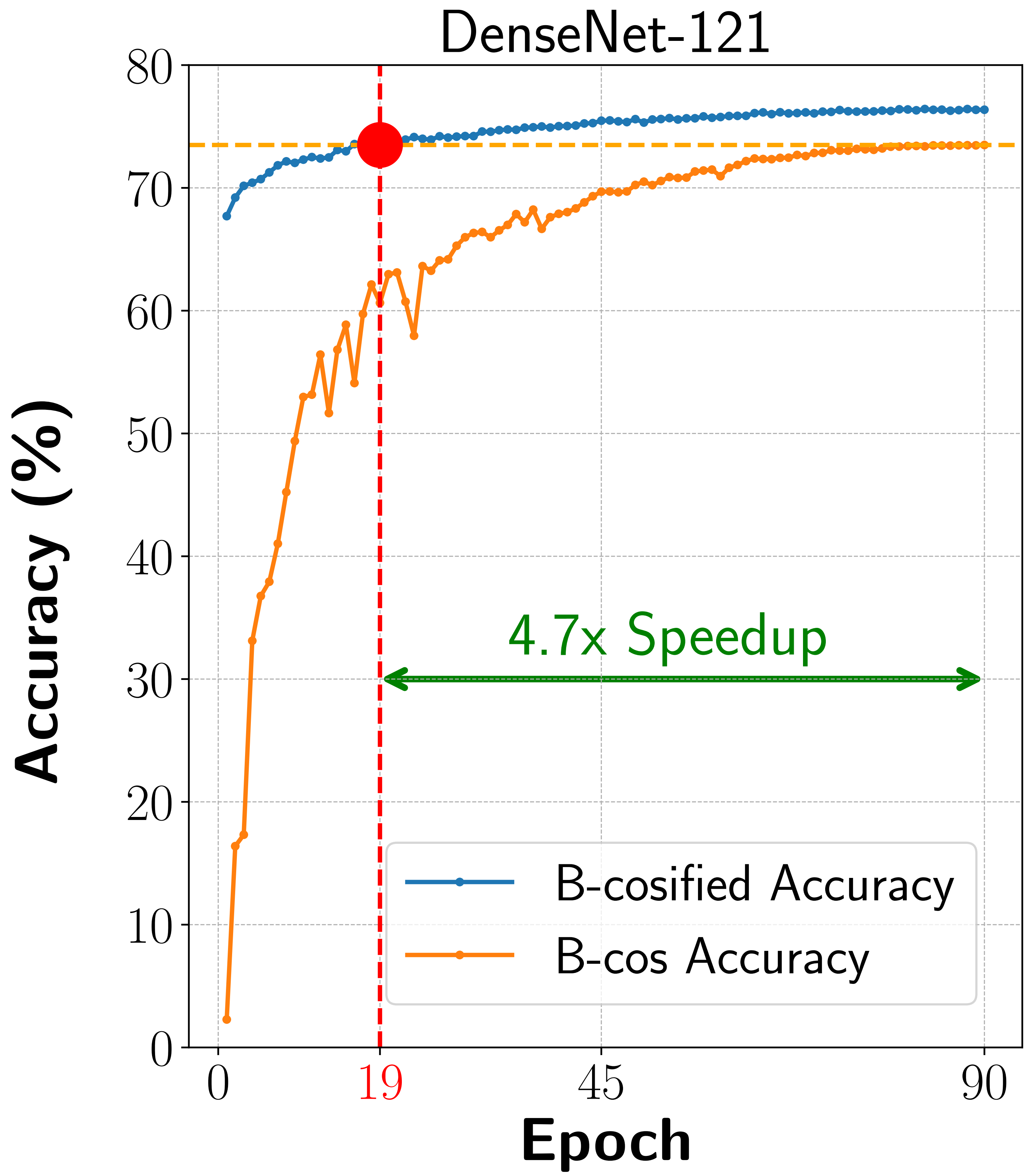


Our work: B-cosification



- Requires significantly fewer training steps than full retraining
- Maintains accuracy
- Provides model-faithful, human interpretable explanations
- Can be used for foundation models where training from scratch is costly

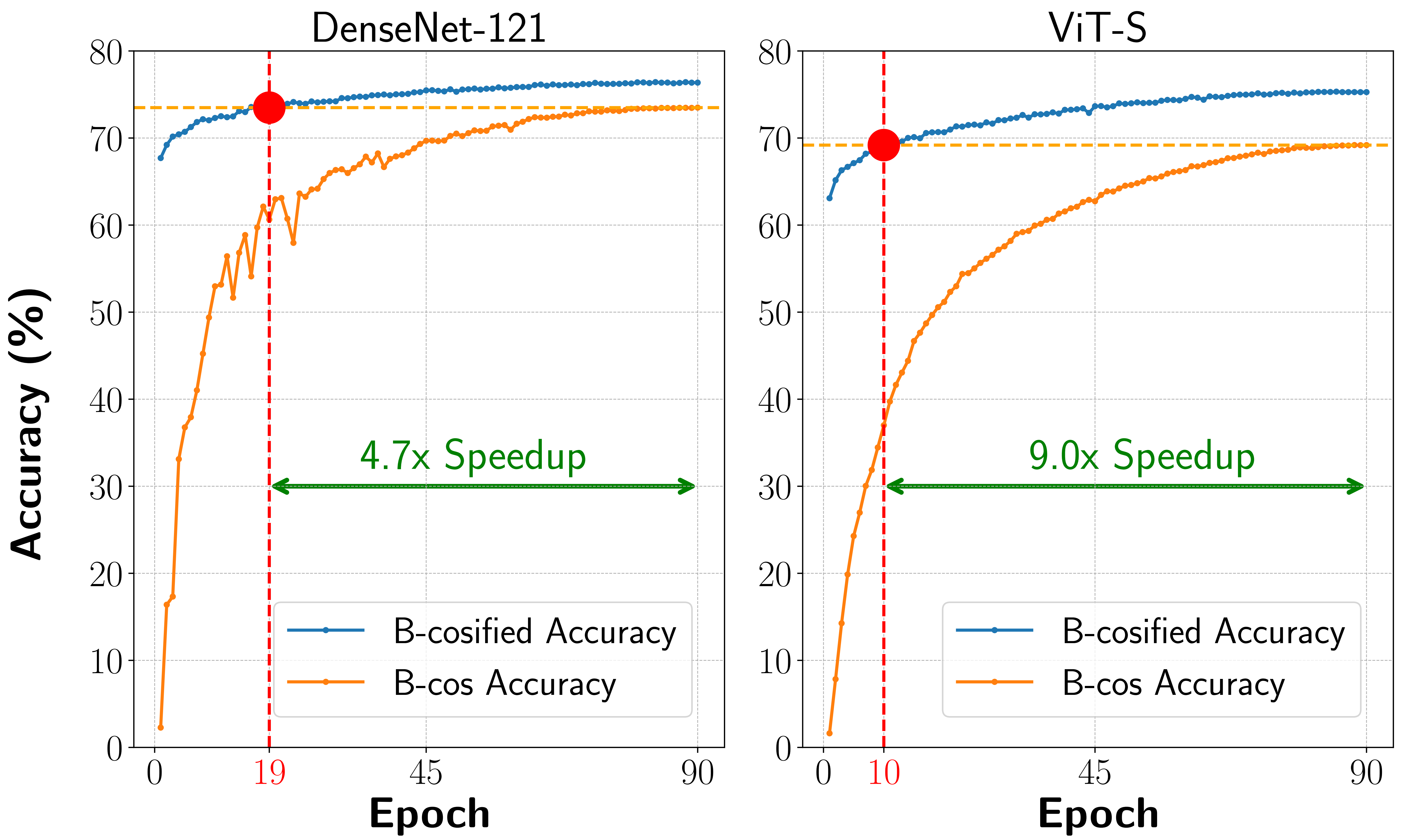
Similar performance at significantly lower cost



- **Requires significantly fewer training steps than full retraining**
- **Maintains accuracy**
- Provides model-faithful, human interpretable explanations
- Can be used for foundation models where training from scratch is costly

DenseNet-121 [Huang et al., CVPR 2017]

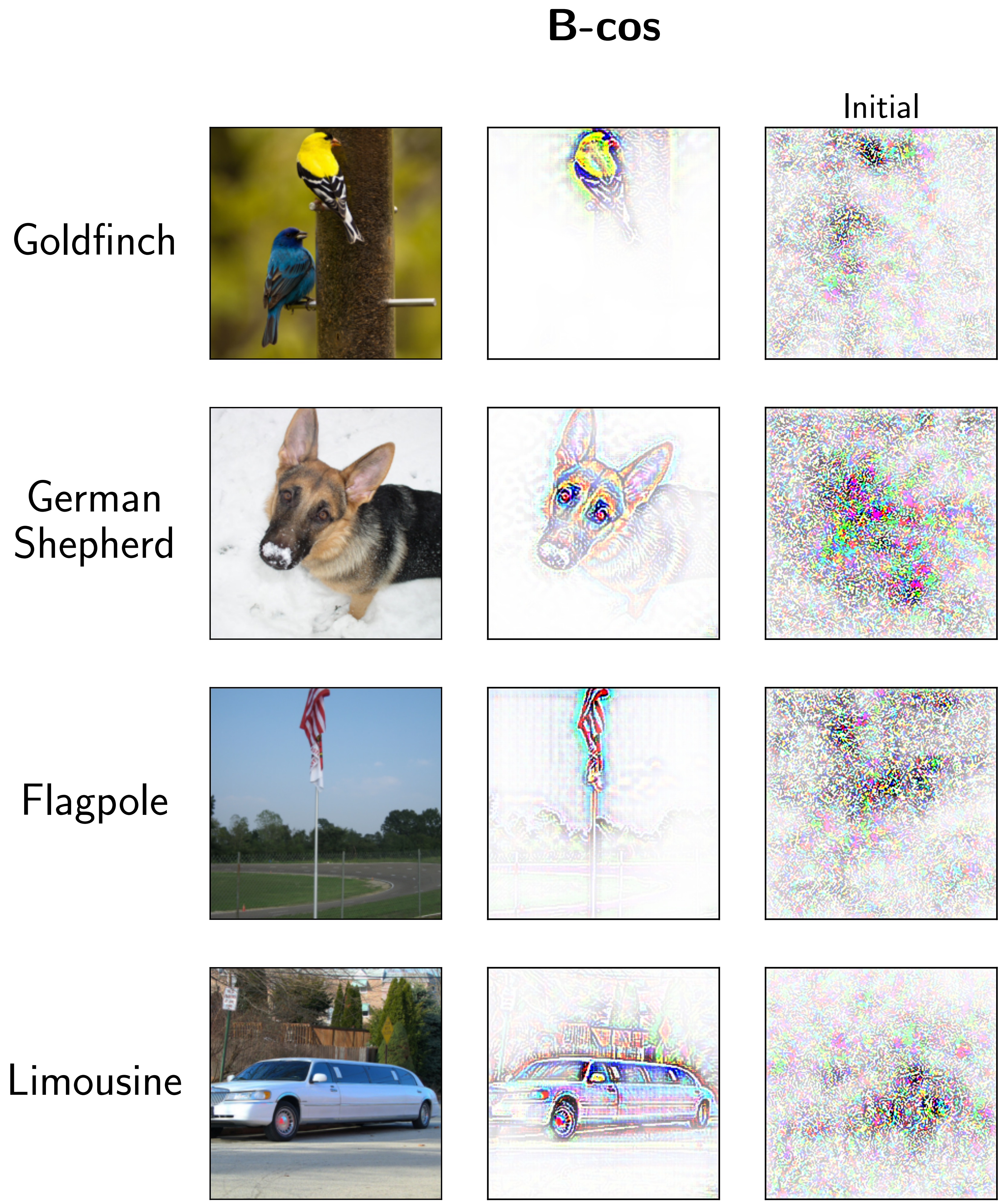
Similar performance at significantly lower cost



- **Requires significantly fewer training steps than full retraining**
- **Maintains accuracy**
- Provides model-faithful, human interpretable explanations
- Can be used for foundation models where training from scratch is costly

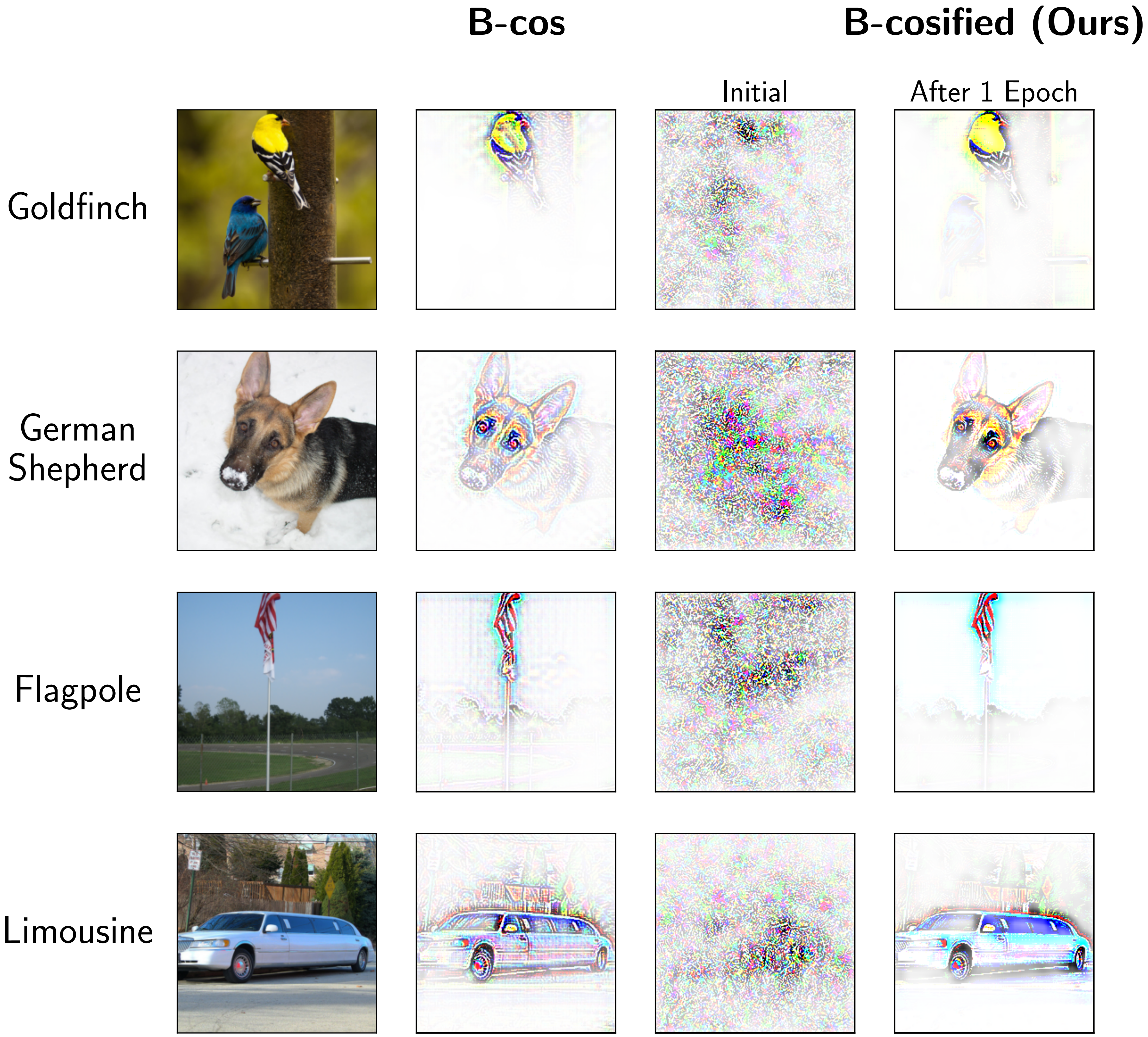
DenseNet-121 [Huang et al., CVPR 2017], ViT [Dosovitskiy et al., ICLR 2021]

Interpretability on par with B-cos



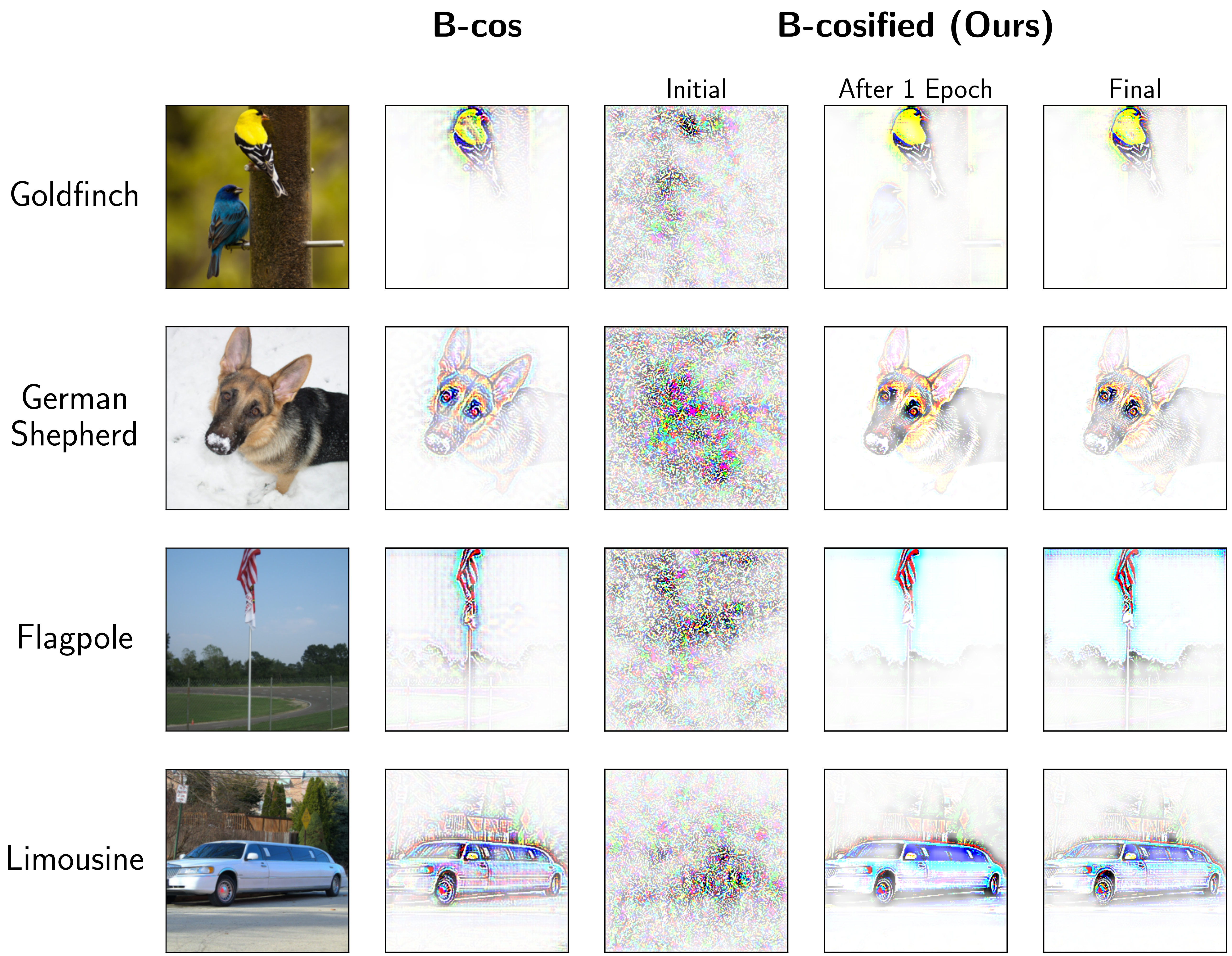
- Requires significantly fewer training steps than full retraining
- Maintains accuracy
- **Provides model-faithful, human interpretable explanations**
- Can be used for foundation models where training from scratch is costly

Interpretability on par with B-cos



- Requires significantly fewer training steps than full retraining
- Maintains accuracy
- **Provides model-faithful, human interpretable explanations**
- Can be used for foundation models where training from scratch is costly

Interpretability on par with B-cos



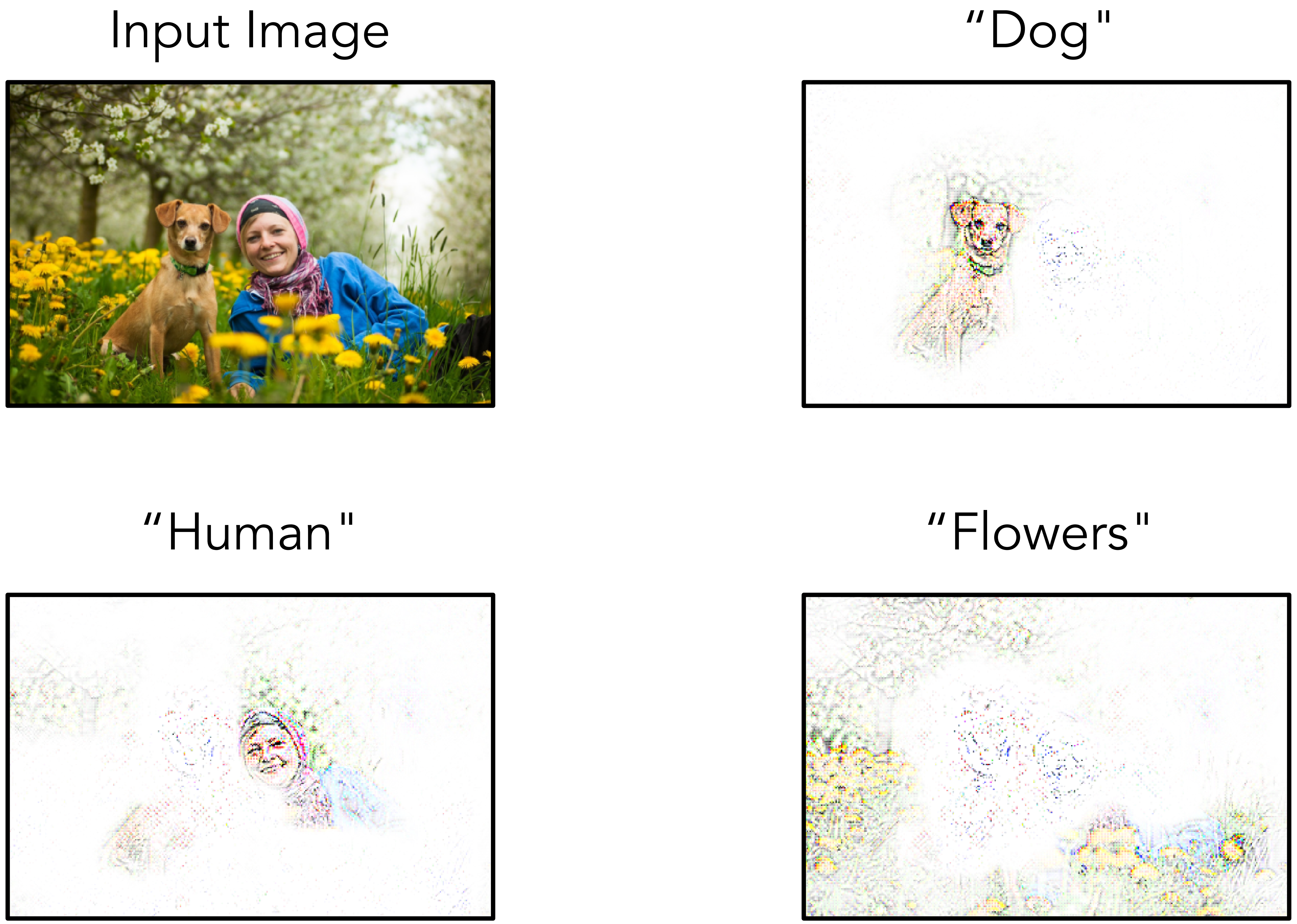
- Requires significantly fewer training steps than full retraining
- Maintains accuracy
- **Provides model-faithful, human interpretable explanations**
- Can be used for foundation models where training from scratch is costly

B-cosification of a foundation model: CLIP[3]

- Requires significantly fewer training steps than full retraining
- Maintains accuracy
- Provides model-faithful, human interpretable explanations
- **Can be used for foundation models where training from scratch is costly**

[3]CLIP [Radford et al., ICML 2021]

B-cosification of a foundation model: CLIP[3]



- Requires significantly fewer training steps than full retraining
- Maintains accuracy
- Provides model-faithful, human interpretable explanations
- **Can be used for foundation models where training from scratch is costly**

[3]CLIP [Radford et al., ICML 2021]

Bridging the gap between conventional and B-cos models

Conventional	B-cos
3-channel Inputs	6-channel Inputs
Normalized Inputs	Unormalized Inputs
No Unit Normalized Weights	Unit Normalized Weights
Activation function between layers	No Activation function between layers
B=1 (linear transforms)	B=2 (non-linear transforms)
Biases in layers	No biases in layers

We perform a study on:

- which modifications are necessary
- how to best apply the modifications

B-cosified
?

Bridging the gap between conventional and B-cos models

Conventional	B-cos
3-channel Inputs	6-channel Inputs
Normalized Inputs	Unormalized Inputs
No Unit Normalized Weights	Unit Normalized Weights
Activation function between layers	No Activation function between layers
B=1 (linear transforms)	B=2 (non-linear transforms)
Biases in layers	No biases in layers

① Preserves functional equivalence

B-cosified

Bridging the gap between conventional and B-cos models

Conventional	B-cos
3-channel Inputs	6-channel Inputs
Normalized Inputs	Unormalized Inputs
No Unit Normalized Weights	Unit Normalized Weights
Activation function between layers	No Activation function between layers
B=1 (linear transforms)	B=2 (non-linear transforms)
Biases in layers	No biases in layers

① Preserves functional equivalence

B-cosified
6-channel Inputs

Bridging the gap between conventional and B-cos models

Conventional	B-cos
3-channel Inputs	6-channel Inputs
Normalized Inputs	Unormalized Inputs
No Unit Normalized Weights	Unit Normalized Weights
Activation function between layers	No Activation function between layers
B=1 (linear transforms)	B=2 (non-linear transforms)
Biases in layers	No biases in layers

① Preserves functional equivalence

B-cosified
6-channel Inputs
Normalized Inputs
No Unit Normalized Weights
Activation function between layers

Bridging the gap between conventional and B-cos models

Conventional	B-cos
3-channel Inputs	6-channel Inputs
Normalized Inputs	Unormalized Inputs
No Unit Normalized Weights	Unit Normalized Weights
Activation function between layers	No Activation function between layers
B=1 (linear transforms)	B=2 (non-linear transforms)
Biases in layers	No biases in layers

① Preserves functional equivalence

② Loses functional equivalence
⇒ Fine-tune

B-cosified
6-channel Inputs
Normalized Inputs
No Unit Normalized Weights
Activation function between layers

Bridging the gap between conventional and B-cos models

Conventional	B-cos
3-channel Inputs	6-channel Inputs
Normalized Inputs	Unormalized Inputs
No Unit Normalized Weights	Unit Normalized Weights
Activation function between layers	No Activation function between layers
B=1 (linear transforms)	B=2 (non-linear transforms)
Biases in layers	No biases in layers

① Preserves functional equivalence

② Loses functional equivalence
⇒ Fine-tune

B-cosified
6-channel Inputs
Normalized Inputs
No Unit Normalized Weights
Activation function between layers
B=2 (non-linear transforms)
No biases in layers

B-cosification generalizes to a variety of architectures and models

Accuracy reached at a much lower training cost

Model	Top-1 Accuracy (%)				Efficiency Gains	
	pretrained	B-cos [10]	B-cosified	Δ_{acc}	t	speedup
ResNet-18	69.8	68.7	71.5	+2.8	29	×3.1
ResNet-50-v1	76.1	75.9	76.5	+0.6	46	×2.0
ResNet-50-v2	80.9	75.9	77.3	+1.4	10	×9.0
DenseNet-121	74.4	73.6	76.3	+2.7	18	×5.0
ViT-Ti	70.3	60.0	69.3	+9.3	10	×9.0
ViT-S	74.4	69.2	75.2	+6.0	10	×9.0
ViT-B	75.3	74.4	75.3	+0.9	57	×1.6
ViT-L	75.8	75.1	75.5	+0.4	66	×1.4
ViT$_c$-Ti	72.6	67.3	72.3	+5.0	10	×9.0
ViT$_c$-S	75.7	74.5	76.0	+1.5	32	×2.8
ViT$_c$-B	76.8	77.1	76.7	-0.4	-	-
ViT$_c$-L	77.9	77.8	77.1	-0.7	-	-

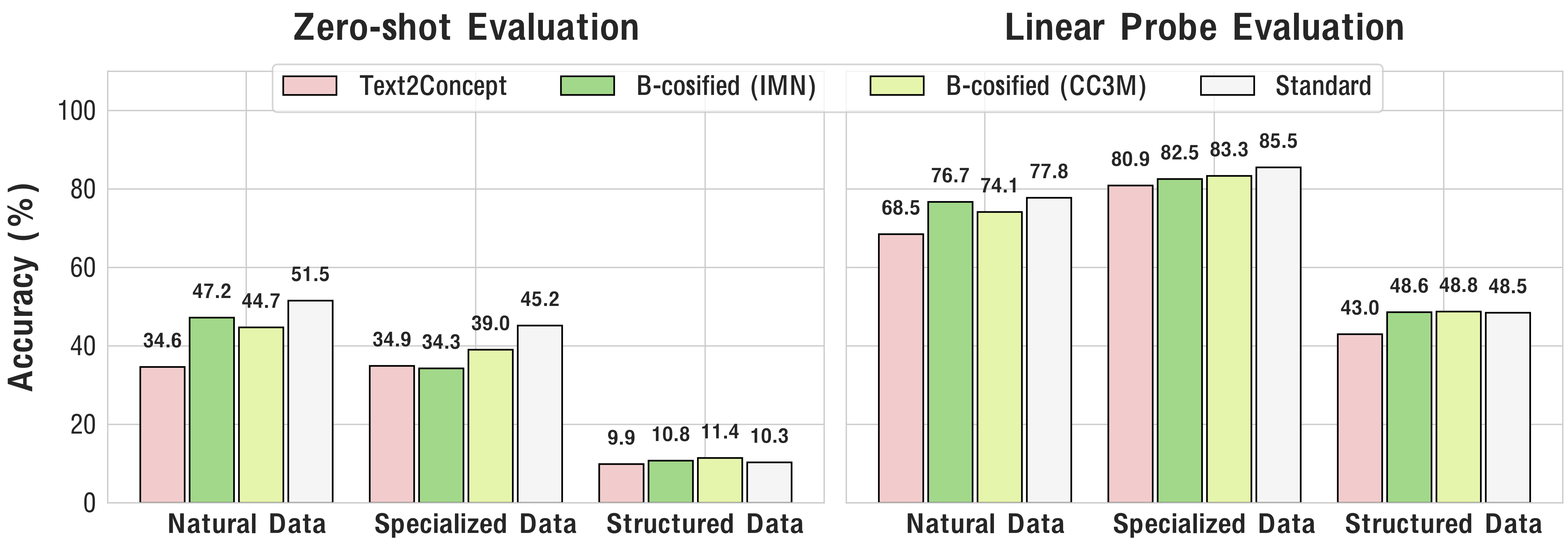
ImageNet CNNs and ViTs

B-cosification generalizes to a variety of architectures and models

Accuracy reached at a much lower training cost

Model	Top-1 Accuracy (%)				Efficiency Gains	
	pretrained	B-cos [10]	B-cosified	Δ_{acc}	t	speedup
ResNet-18	69.8	68.7	71.5	+2.8	29	×3.1
ResNet-50-v1	76.1	75.9	76.5	+0.6	46	×2.0
ResNet-50-v2	80.9	75.9	77.3	+1.4	10	×9.0
DenseNet-121	74.4	73.6	76.3	+2.7	18	×5.0
ViT-Ti	70.3	60.0	69.3	+9.3	10	×9.0
ViT-S	74.4	69.2	75.2	+6.0	10	×9.0
ViT-B	75.3	74.4	75.3	+0.9	57	×1.6
ViT-L	75.8	75.1	75.5	+0.4	66	×1.4
ViT_c-Ti	72.6	67.3	72.3	+5.0	10	×9.0
ViT_c-S	75.7	74.5	76.0	+1.5	32	×2.8
ViT_c-B	76.8	77.1	76.7	-0.4	-	-
ViT_c-L	77.9	77.8	77.1	-0.7	-	-

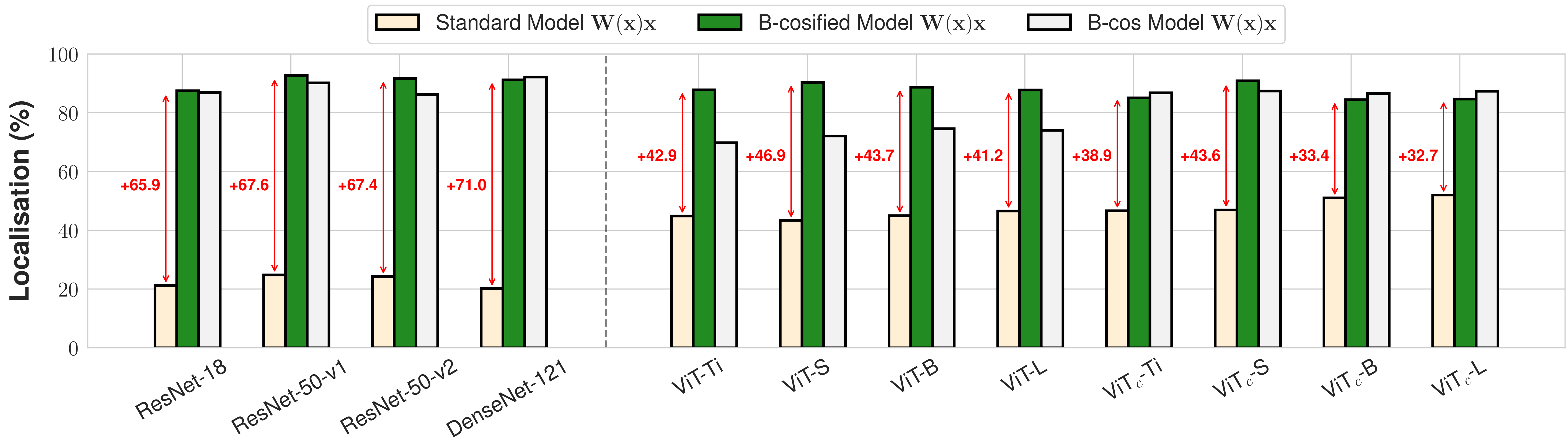
ImageNet CNNs and ViTs



CLIP

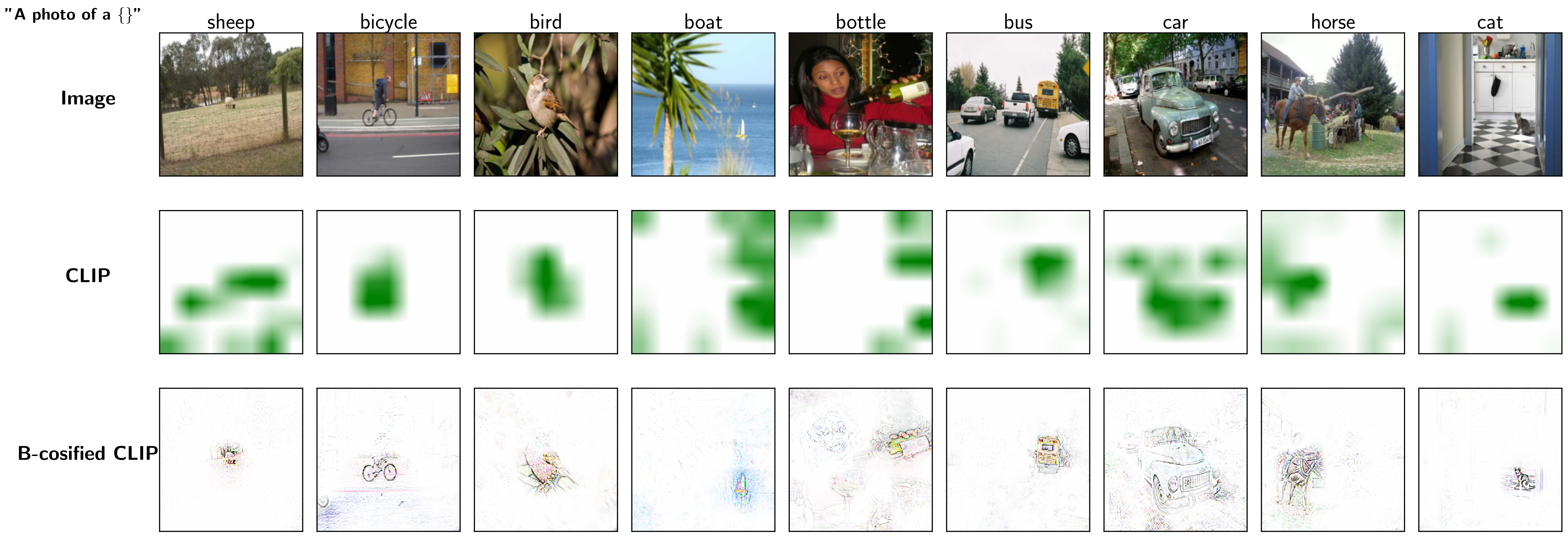
B-cosification generalizes to a variety of architectures and models

Localization of explanations on par with B-cos, outperforms conventional attribution methods



ImageNet CNNs and ViTs

CLIP



Takeaways

- B-cosification provides the interpretability benefits of B-cos models at a much lower cost
- Better to B-cosify existing models instead of training B-cos models from scratch
- Shows promise as a means to obtain inherently interpretable foundation models

Poster Session 3, December 12, 2024, 11:00 AM

Paper

https://arxiv.org/abs/2411.00715

Code

https://github.com/shrebox/B-cosification