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

Shreyash Arya*,[1,2], Sukrut Rao*,[1,2], Moritz Böhle*,[1], Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Saarland Informatics Campus

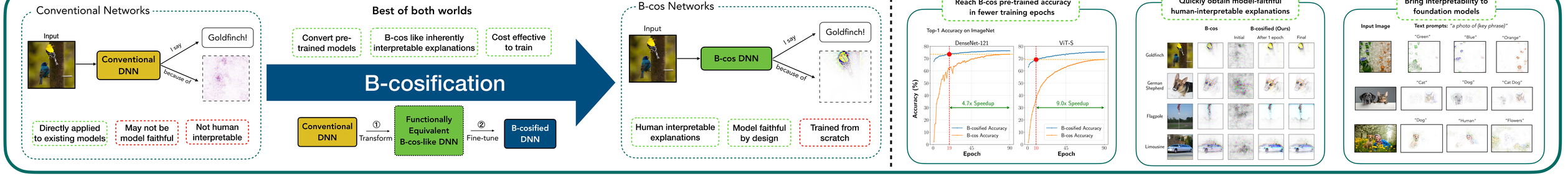
[2]RTG Neuroexplicit Models of Language, Vision, and Action, Saarbrücken

MAX PLANCK INSTITUTE FOR INFORMATICS

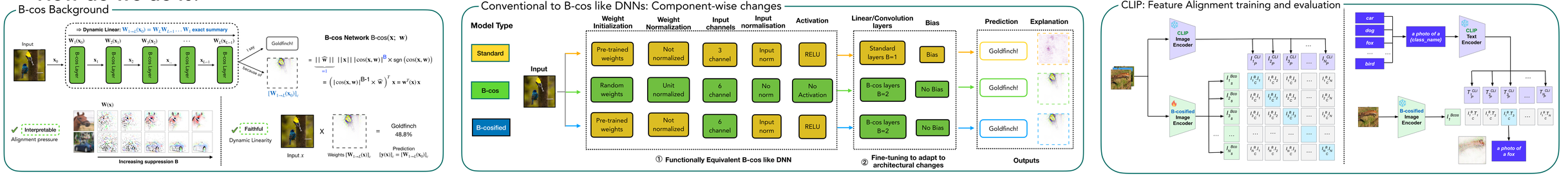
SIC Saarland Informatics Campus

NEURAL INFORMATION PROCESSING SYSTEMS

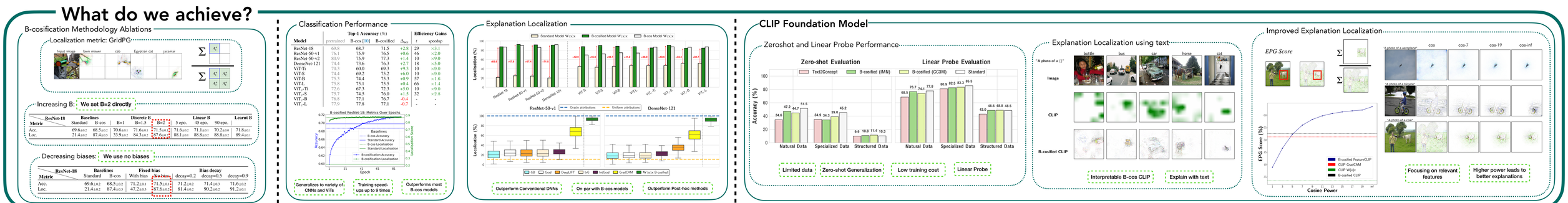
Overview



How do we do it?



What do we achieve?



References: B-cos Networks [Böhle et al., CVPR 2022], DenseNet-121 [Huang et al., CVPR 2017], ViT [Dosovitskiy et al., ICLR 2021], CLIP [Radford et al., ICML 2021]