Kiel University
Christian-Albrechts-Universität zu Kiel

TUHH
Hamburg
University of
Technology

CVPR
JUNE 3-7, 2026
DENVER
COLORADO

ThinkingViT: Matryoshka Thinking Vision Transformer for Elastic Inference

Ali Hojjat[1,2], Janek Haberer[1], Sören Pirk[1], Olaf Landsiedel[2,1,3]

[1]*Kiel University, Germany*
[2]*Hamburg University of Technology (TUHH), Germany*
[3]*UNU-INWEH, Germany*

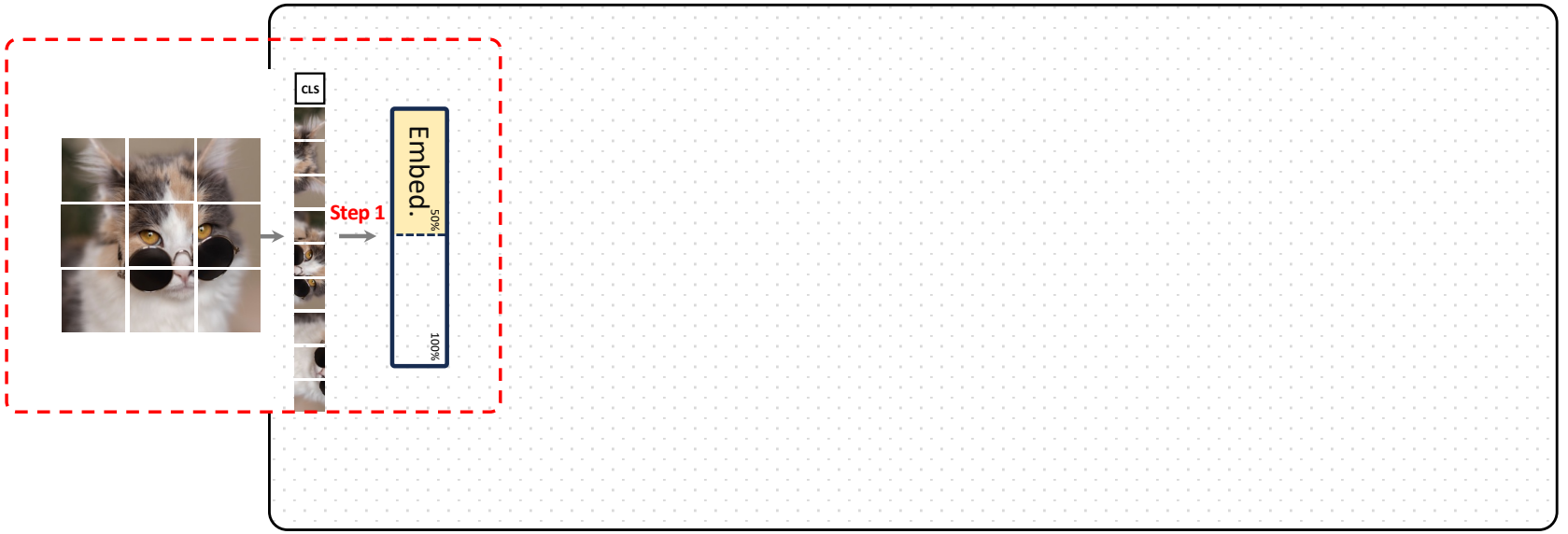
Supported by:
Federal Ministry for Economic Affairs and Energy
Federal Ministry for Economic Affairs and Climate Action
Federal Ministry of Transport and Digital Infrastructure
Federal Ministry for Digital and Transport

We don’t think equally hard about everything…

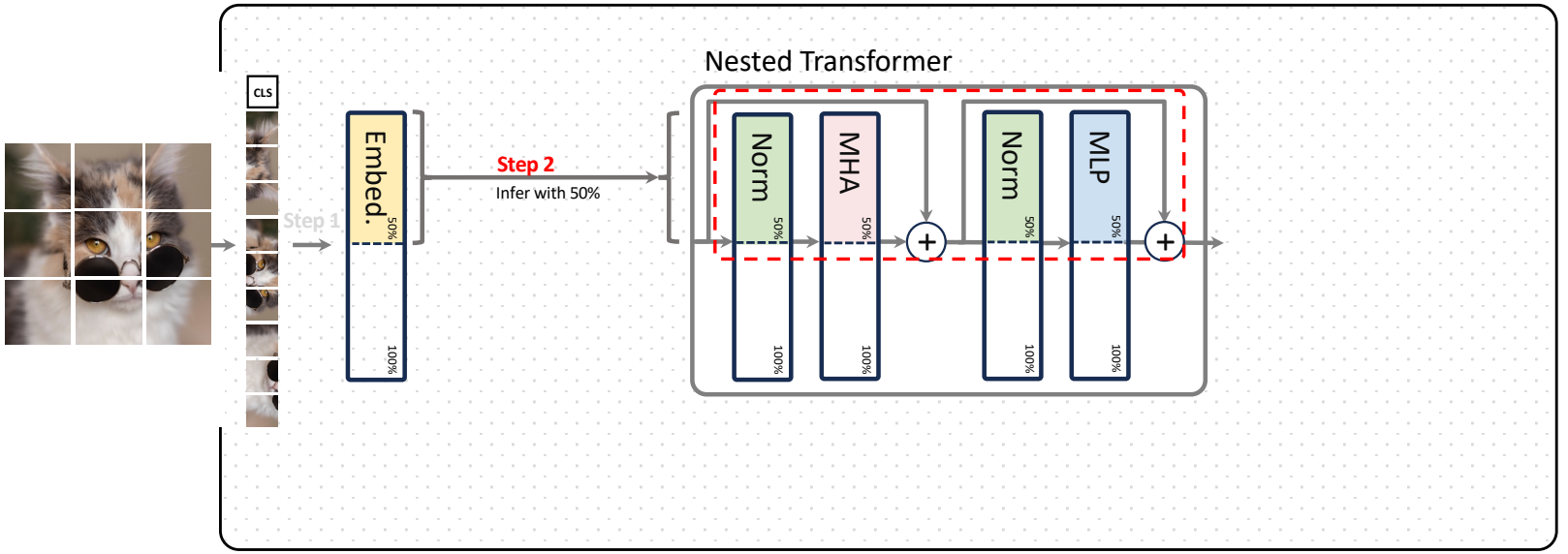
…but Vision Transformers (ViT) do!

Source: https://www.freecodecamp.org/news/chihuahua-or-muffin-my-search-for-the-best-computer-vision-api-cbda4d6b425d

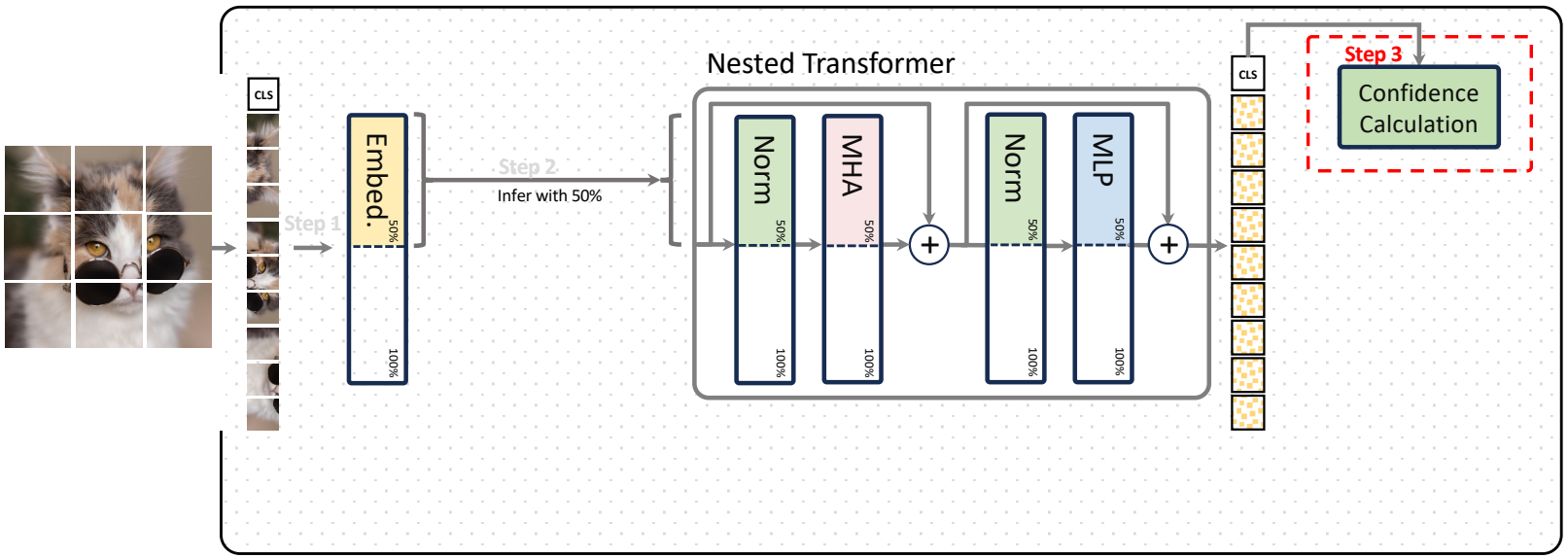
ThinkingViT first runs inference with a small set of attention heads



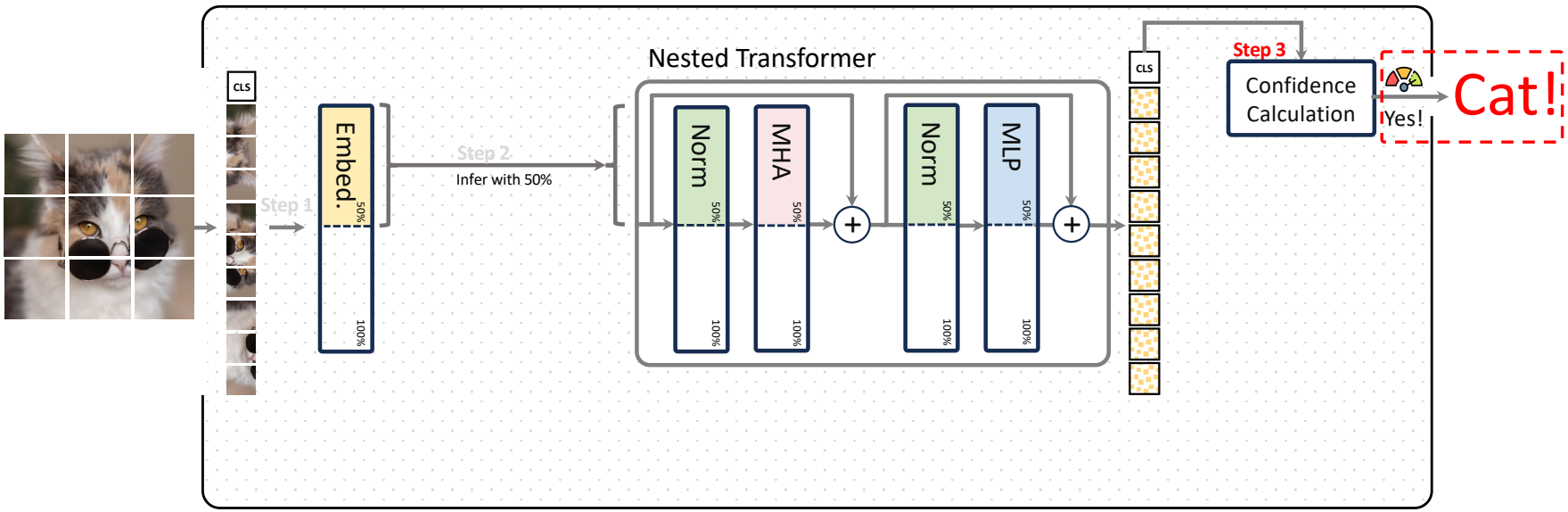
ThinkingViT first runs inference with a small set of attention heads



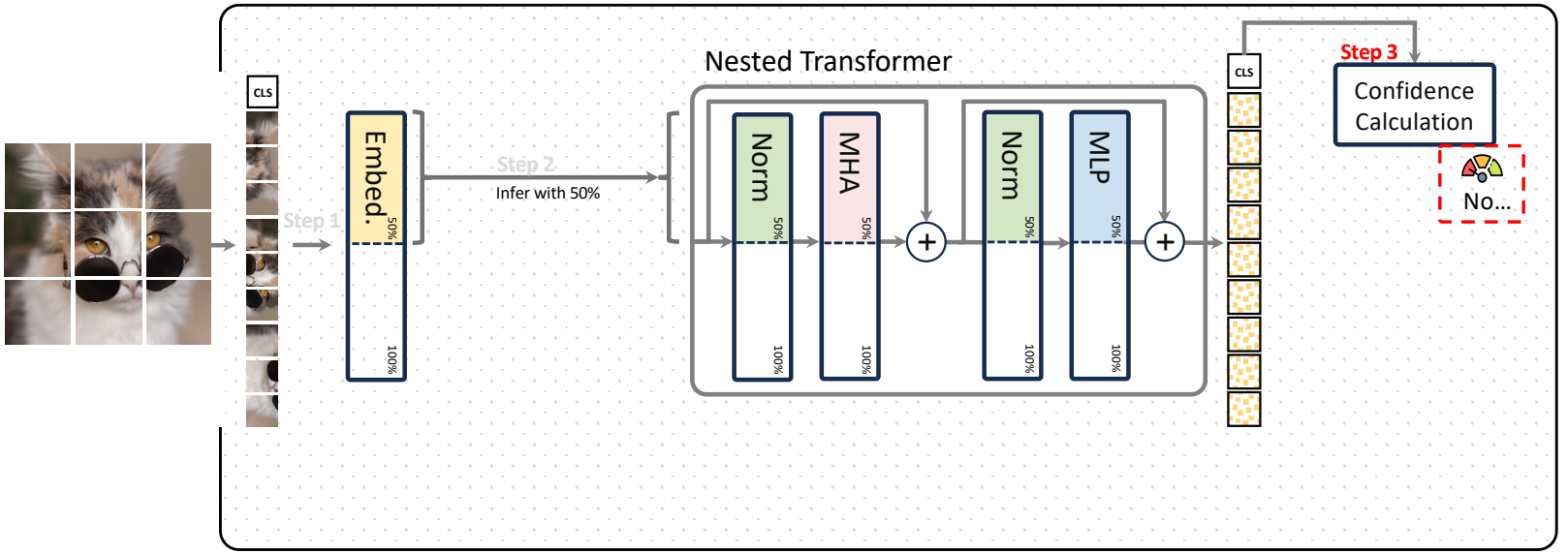
ThinkingViT first runs inference with a small set of attention heads



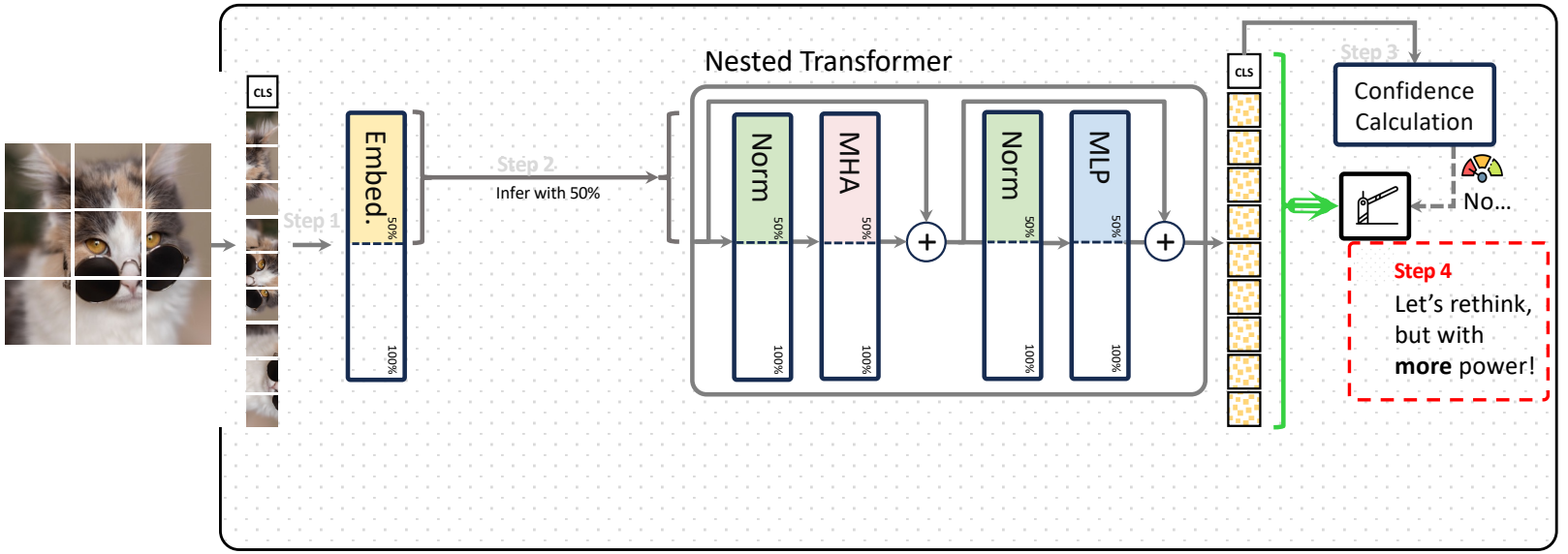
ThinkingViT first runs inference with a small set of attention heads



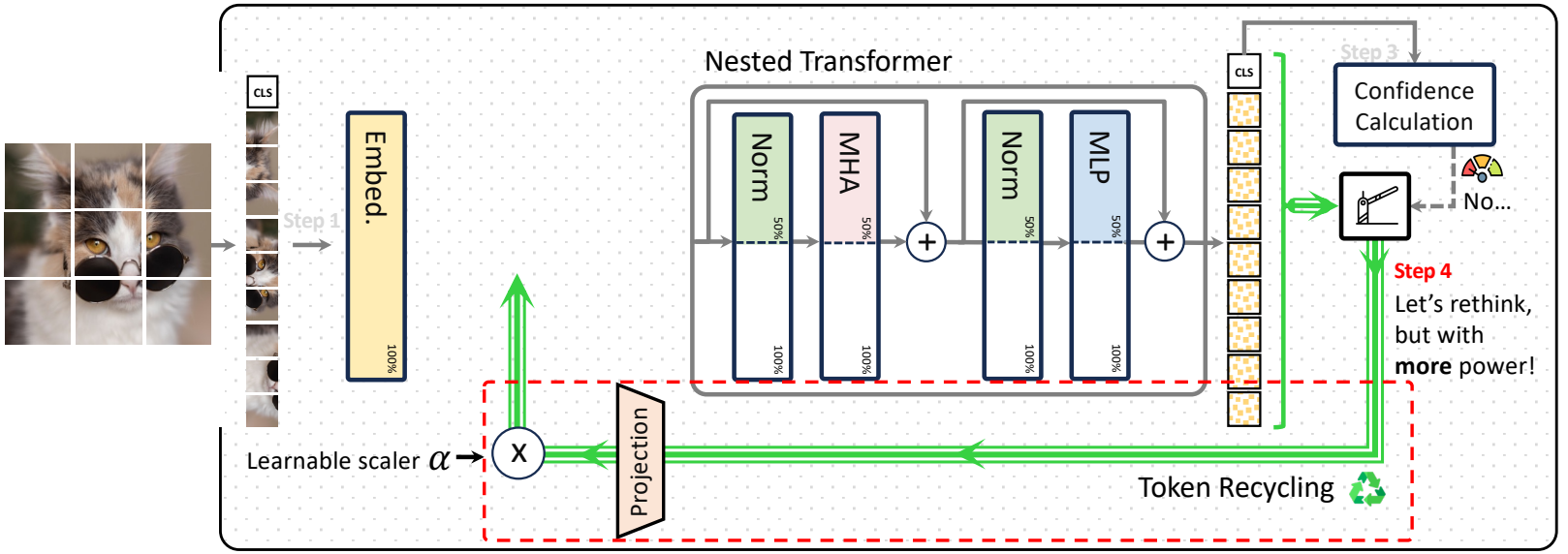
If prediction is not confident, ThinkingViT activates more attention heads and runs another pass of inference, while reusing produced token features



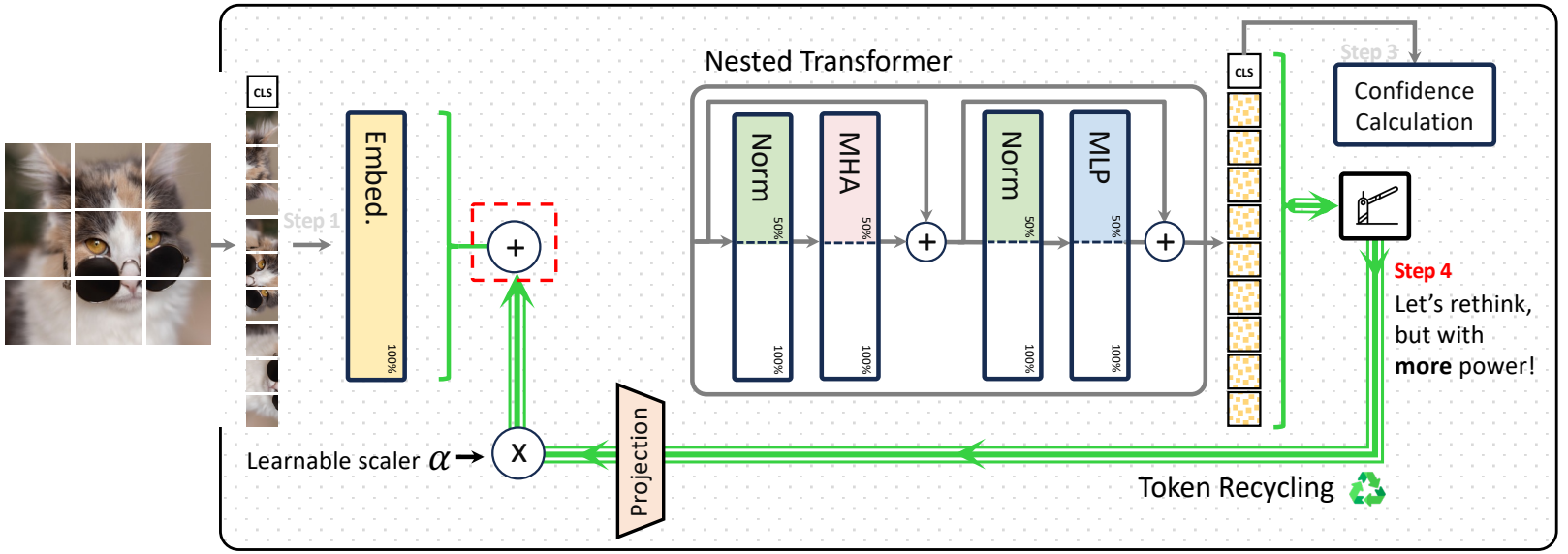
If prediction is not confident, ThinkingViT activates more attention heads and runs another pass of inference, while reusing produced token features



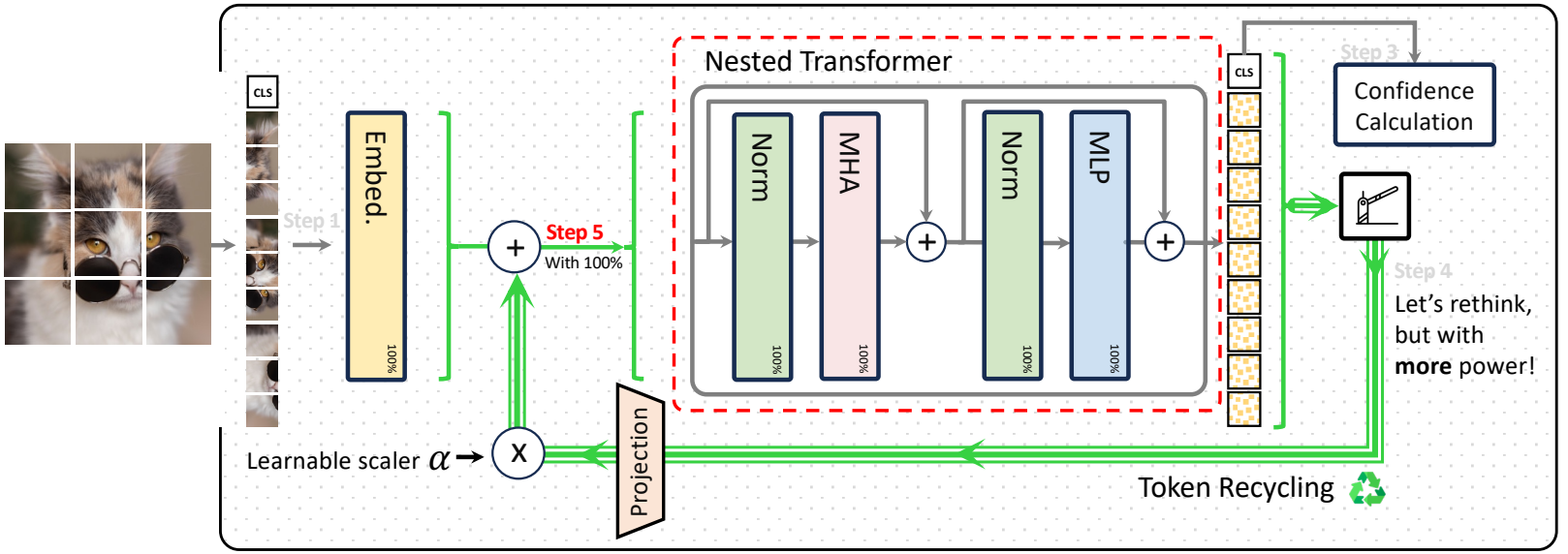
If prediction is not confident, ThinkingViT activates more attention heads and runs another pass of inference, while reusing produced token features



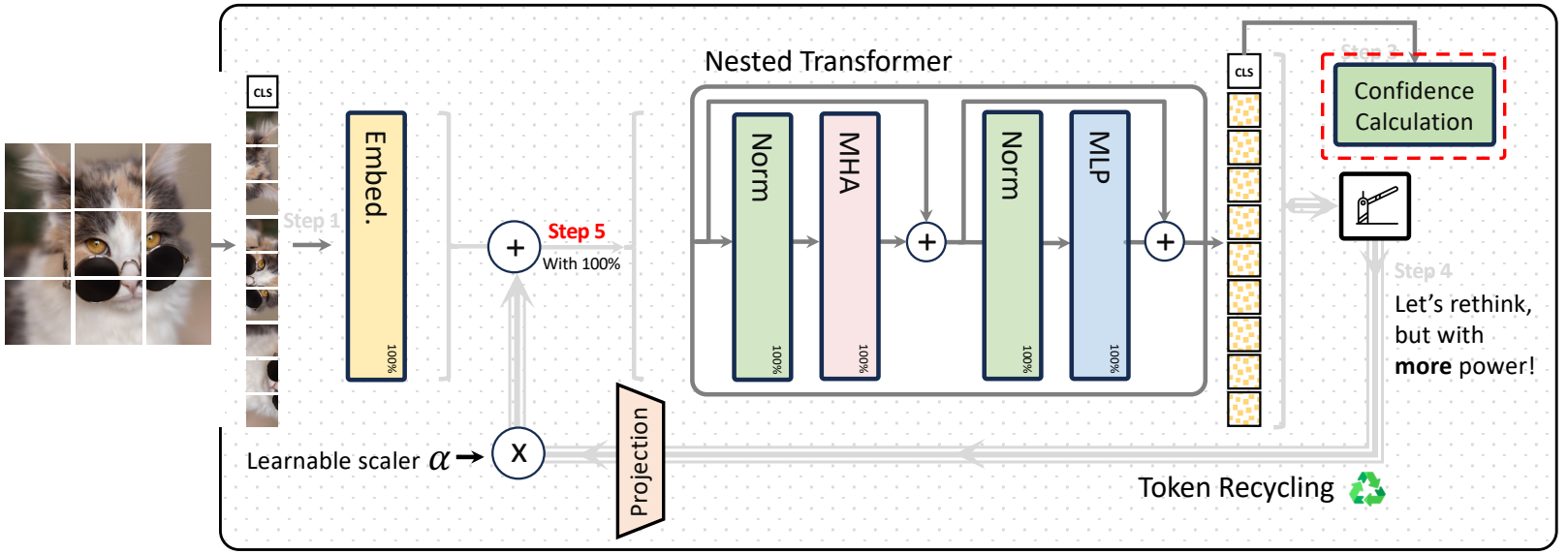
If prediction is not confident, ThinkingViT activates more attention heads and runs another pass of inference, while reusing produced token features



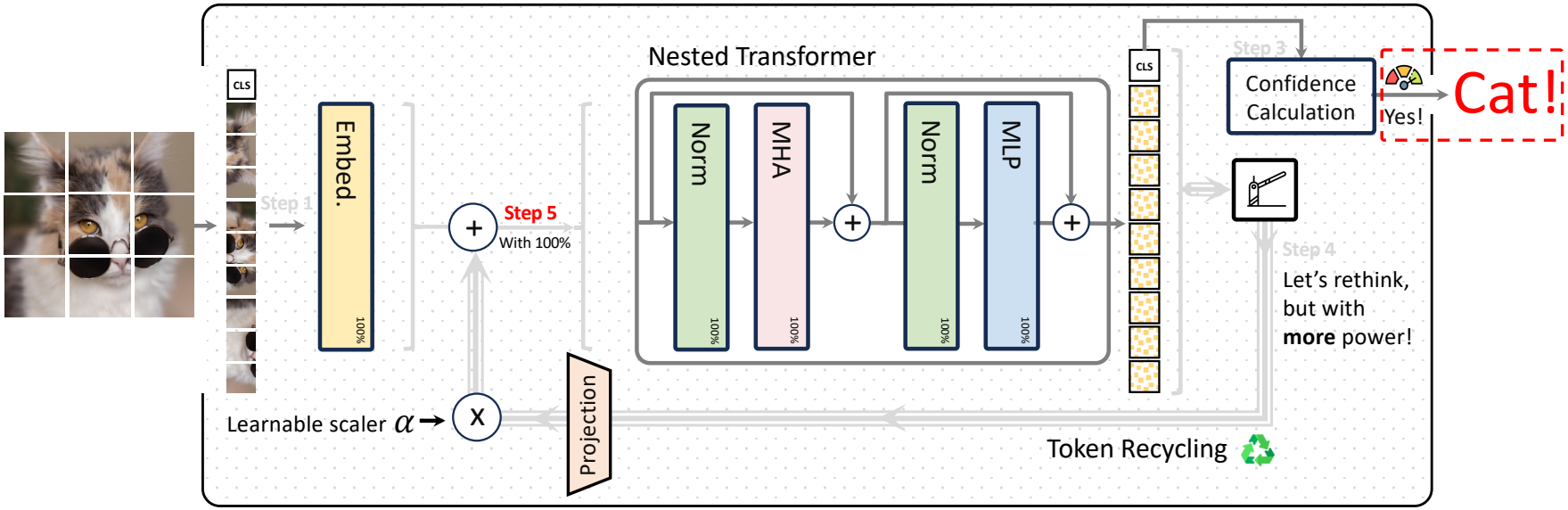
If prediction is not confident, ThinkingViT activates more attention heads and runs another pass of inference, while reusing produced token features



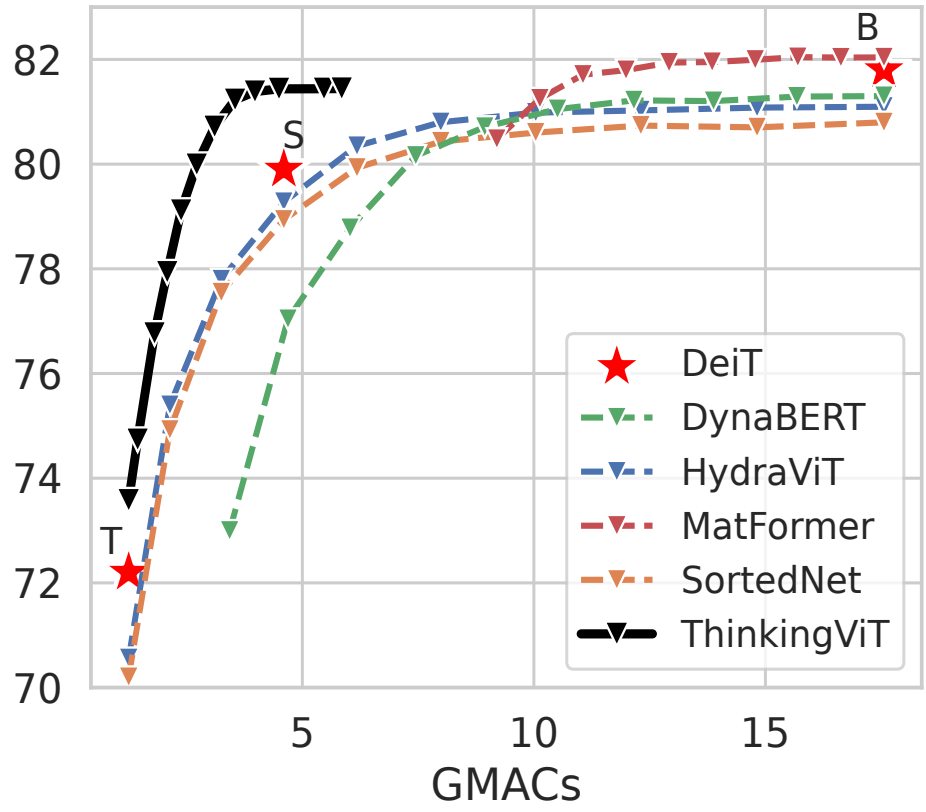
If prediction is not confident, ThinkingViT activates more attention heads and runs another pass of inference, while reusing produced token features



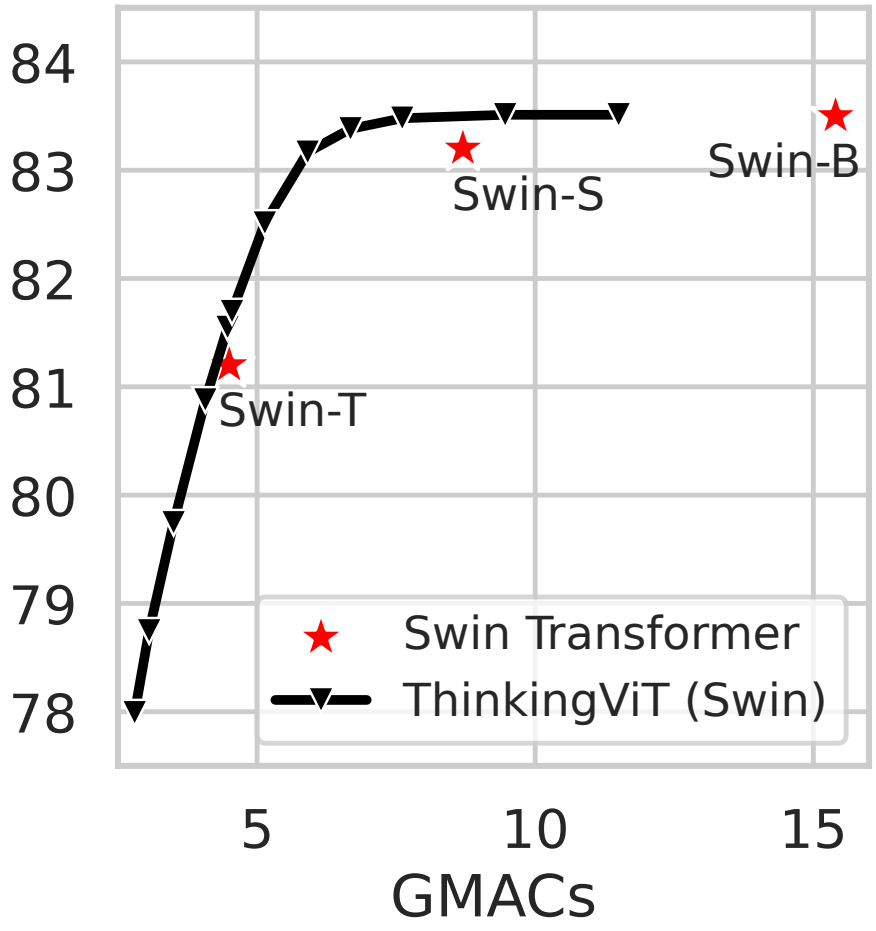
If prediction is not confident, ThinkingViT activates more attention heads and runs another pass of inference, while reusing produced token features



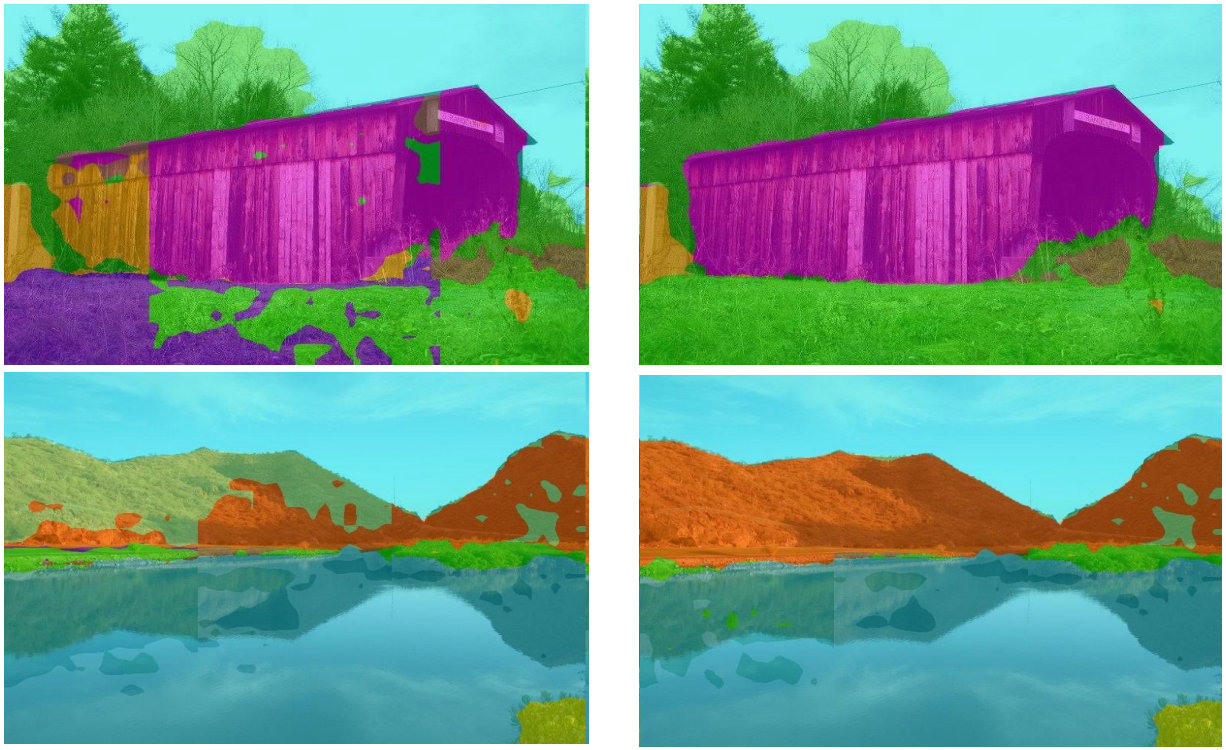
ThinkingViT achieves overall the best accuracy per GMACs on ImageNet1K



ThinkingViT works on top of hierarchical models such as Swin Transformer



ThinkingViT is a suitable backbone for semantic segmentation

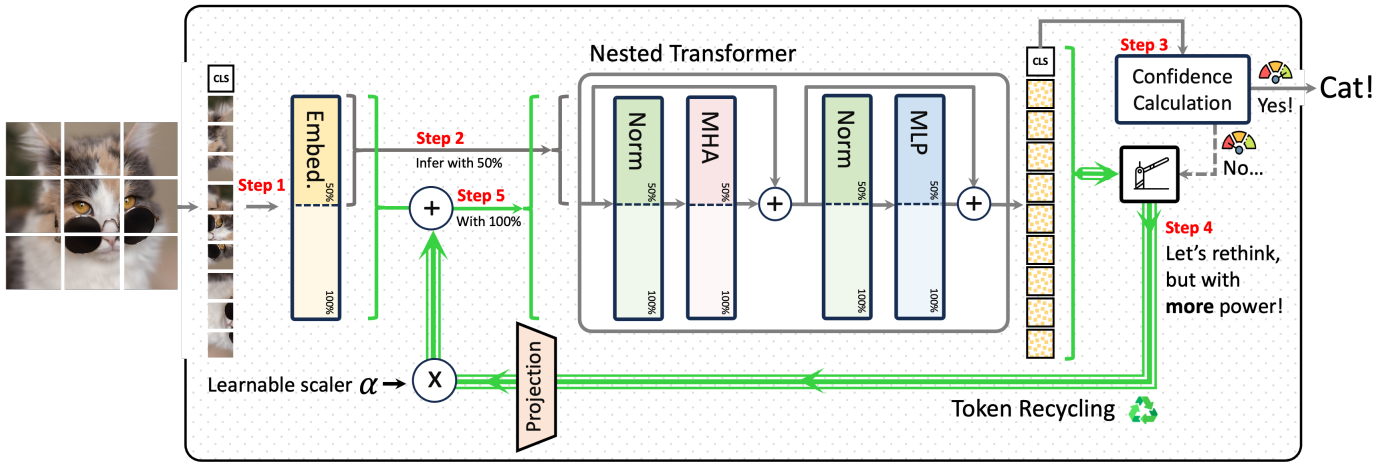


First round of thinking

Second round of thinking

ThinkingViT brings input-adaptiveness to nested Vision Transformers.

Via multi-round inference, ThinkingViT activates additional attention heads combined with reusing produced token features to achieve better efficiency.



https://github.com/ds-kiel/ThinkingViT