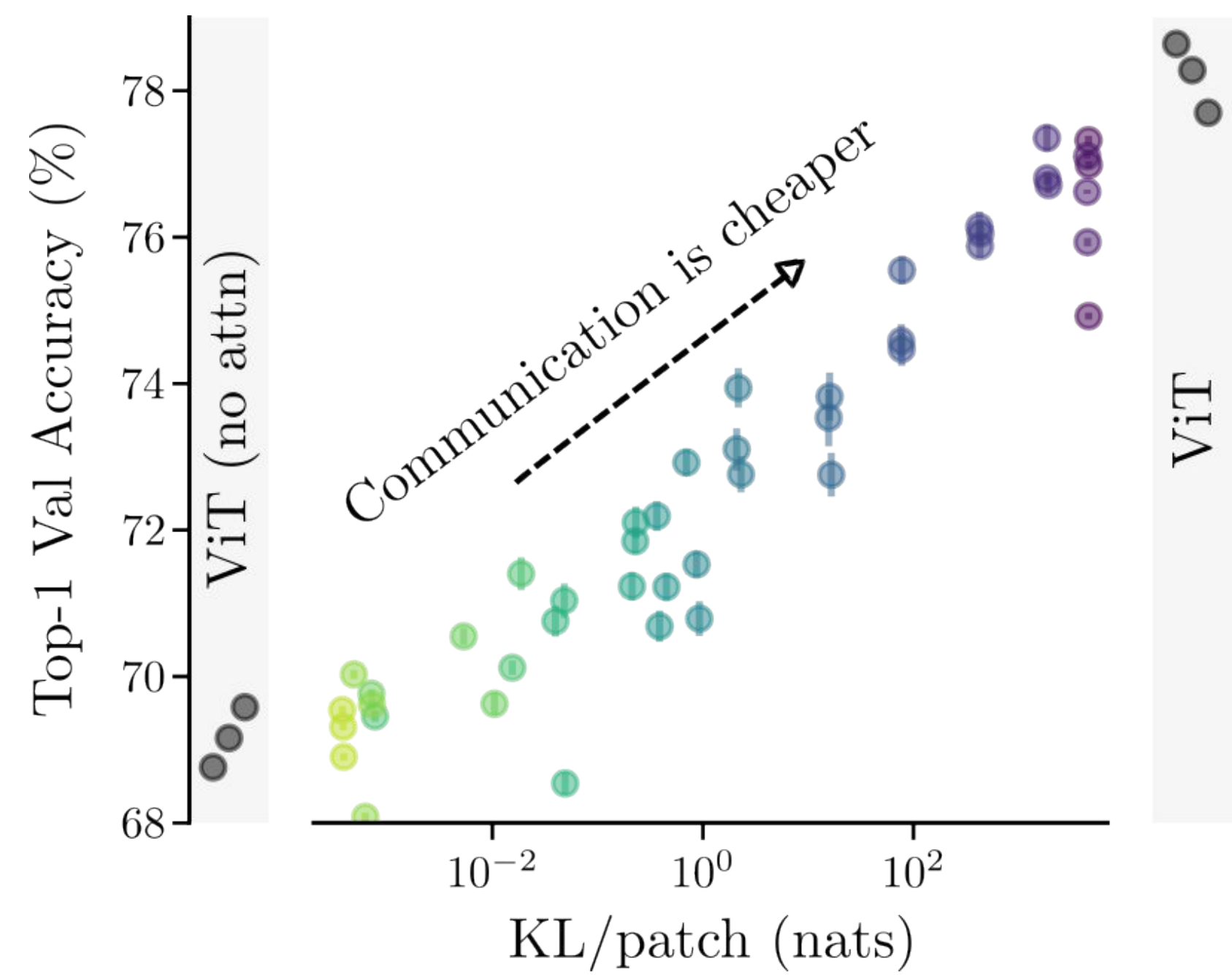
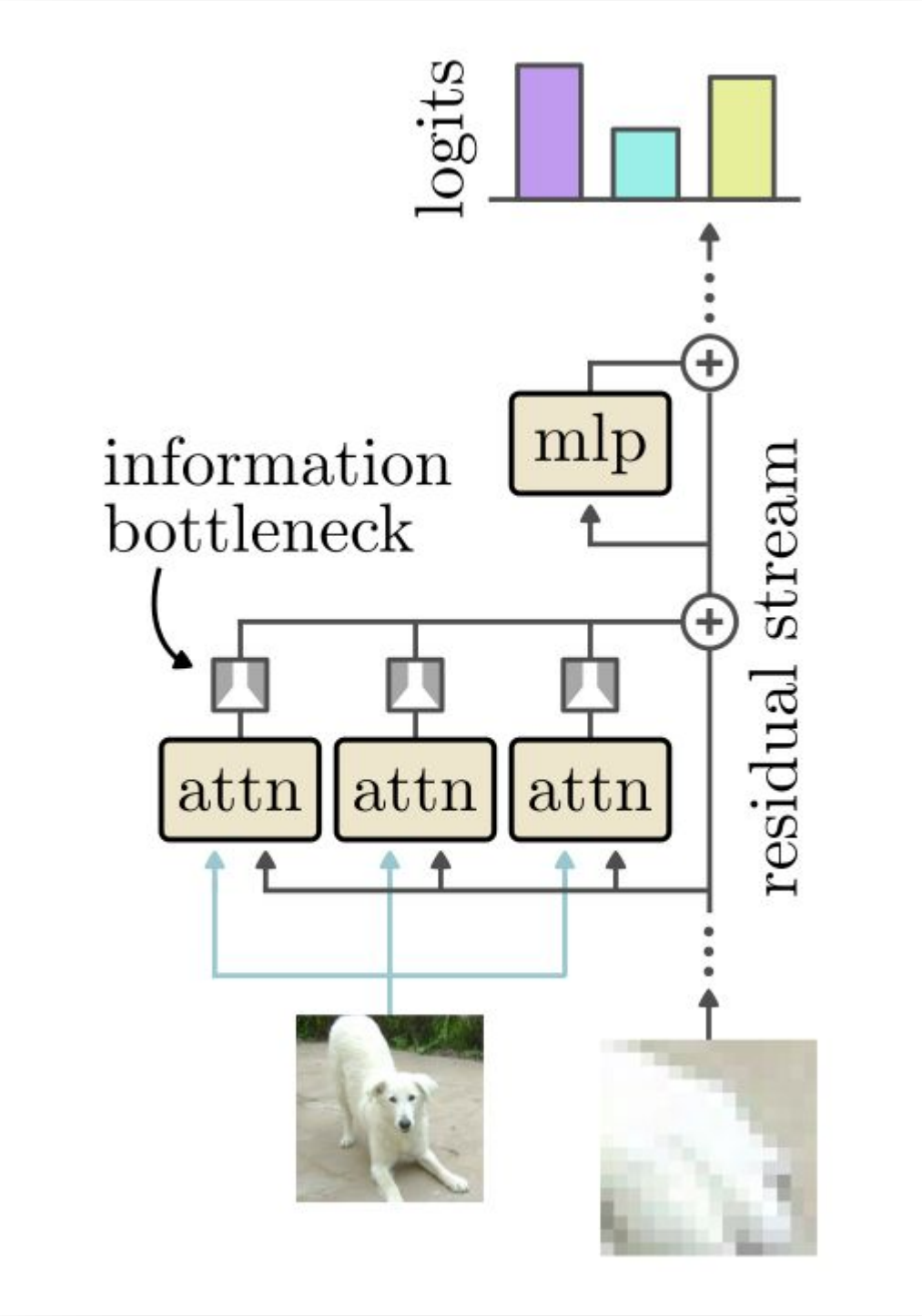


Make attention pay per nat

Variational bottlenecks on residual writes turn ViT communication into a measurable, tunable quantity

36 information bottlenecks, one global cost β

ViT-Tiny trained from scratch with a bottleneck on every attention head's write to the residual stream. A single parameter prices all communication during training, and the prioritized messaging that emerges is easier to study than a standard transformer.



Smooth ramp: Independent residual streams to full cross-talk

Accuracy and patch voting characteristics interpolate between a ViT with attention blocks ablated and an unmodified ViT, in both supervised and self-supervised settings.

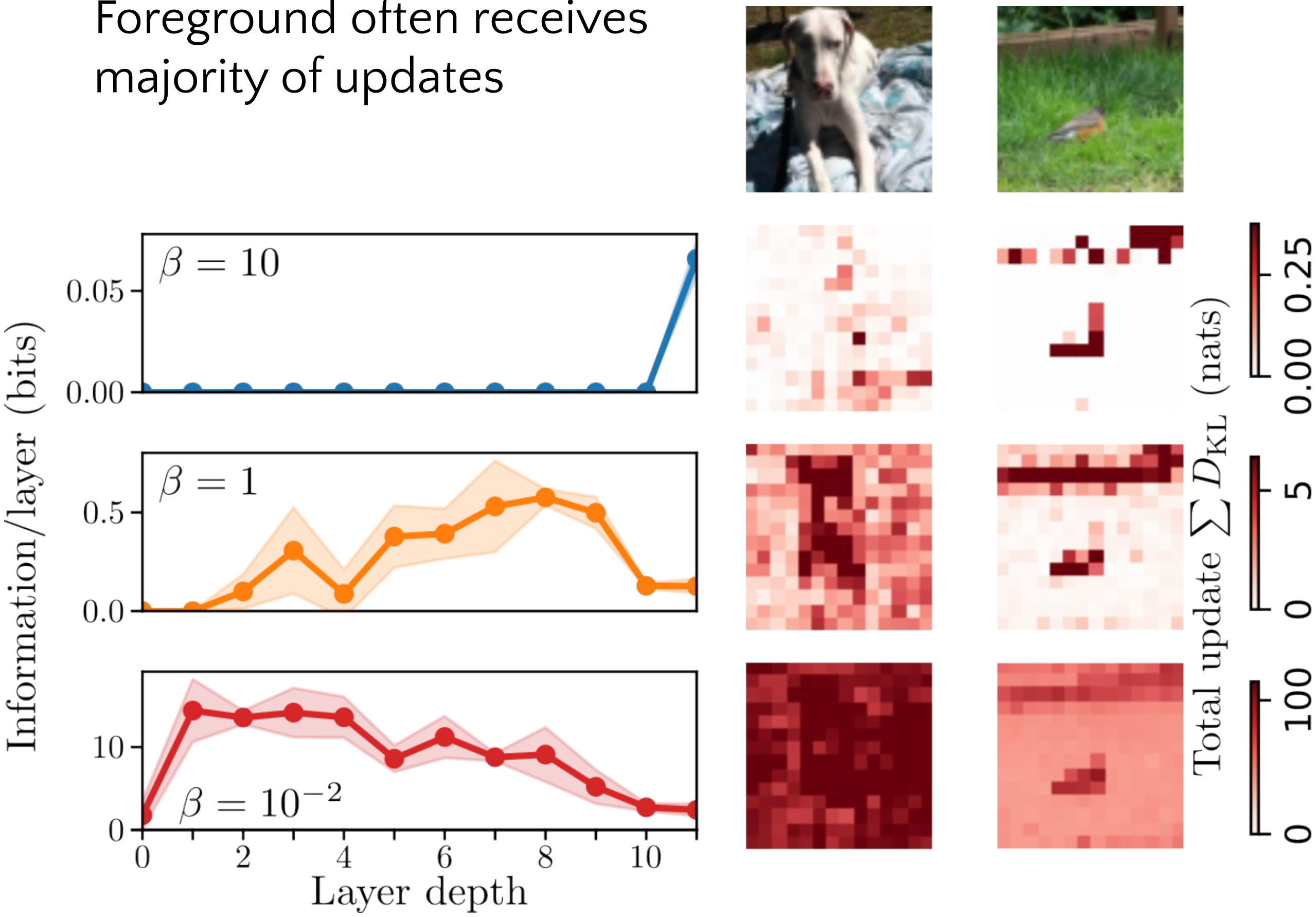
Insights

- 1

Where is the information allocated?

Late heads are active first, then updates shift earlier

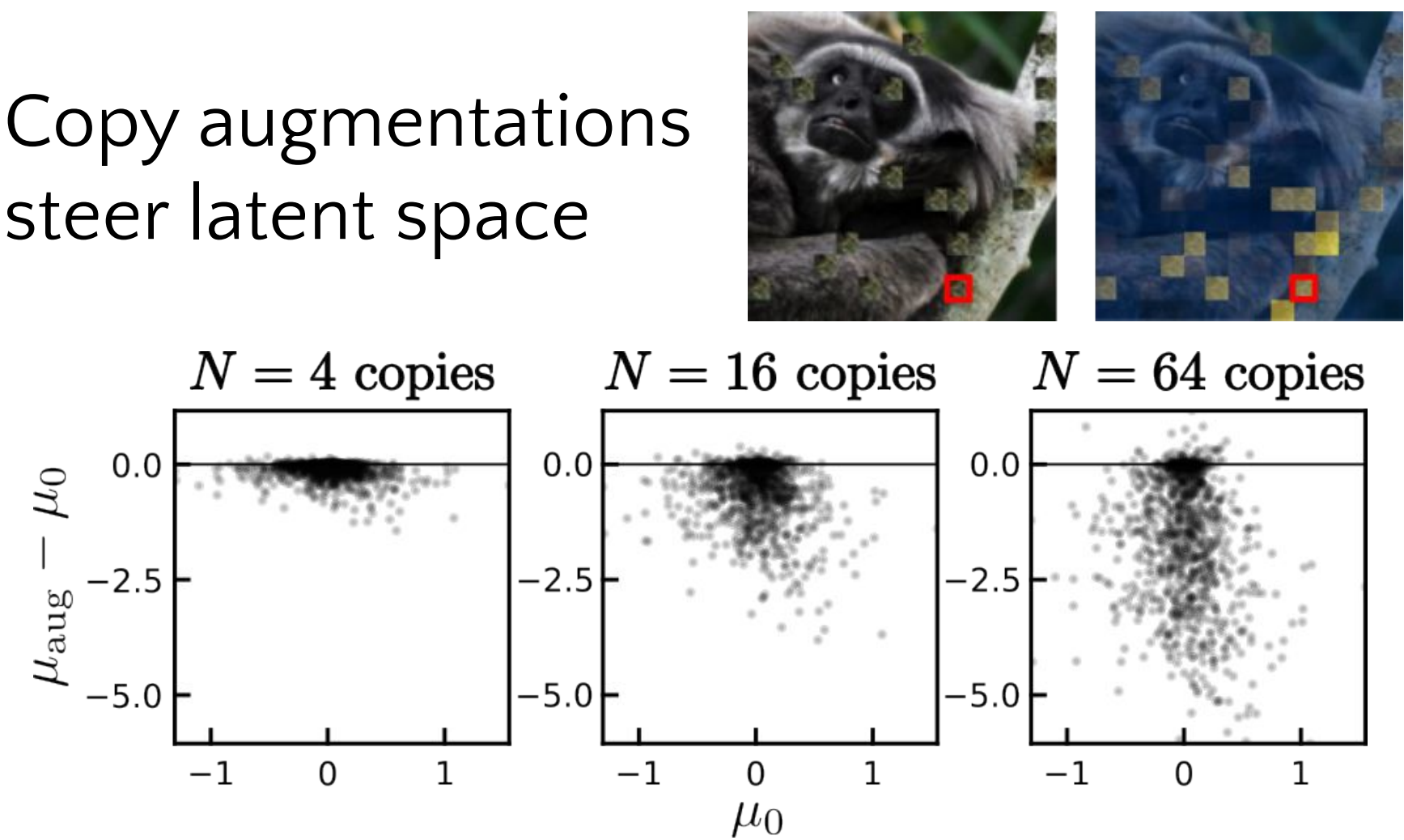
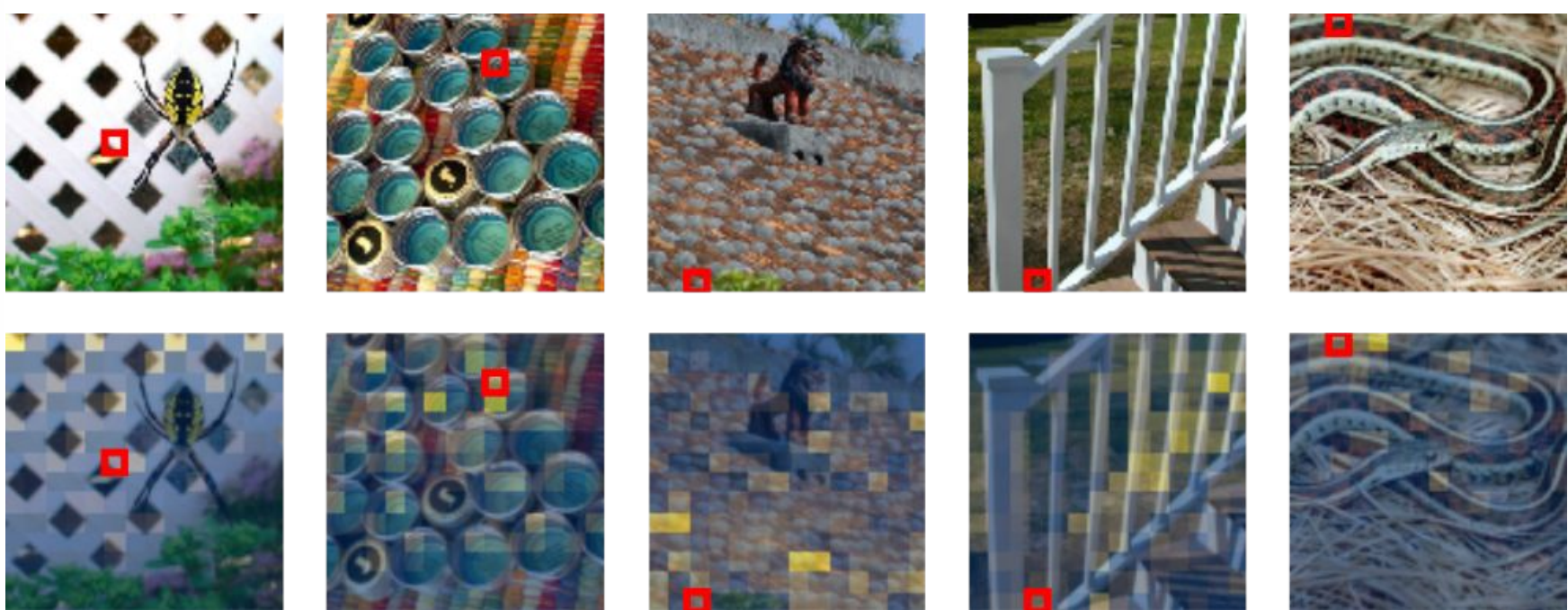
Foreground often receives majority of updates



- 2

Are the low info heads interpretable?

Repetition detection robust at high β



PAPER

Transformers as Communication Systems:
Controlling Information Flow with Bottlenecks

Mechanistic Interpretability Workshop, ICML 2026

Kieran A Murphy
New Jersey Institute of Technology