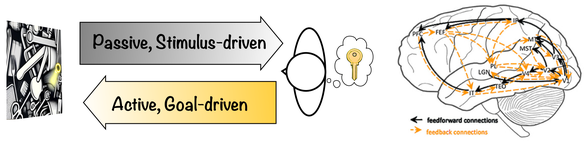
Cognitive Steering in Deep Neural Networks via Long-Range Modulatory **Feedback Connections**

Talia Konkle & George Alvarez
Harvard University

Kempner INSTITUTE
For the Study of Natural & Artificial Intelligence at Harvard University

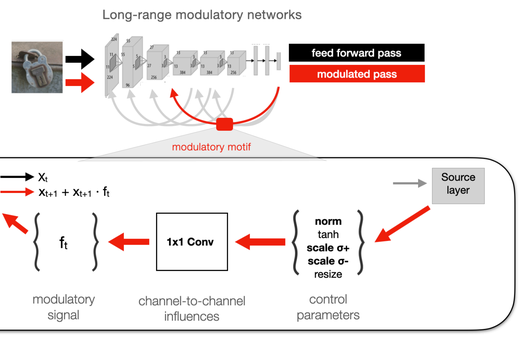
The world is cluttered…

Computer vision systems process images in a **passive** way. But humans can actively look at the world with a **goal** in mind (where are my keys again?) —- which **flexibly** adjusts our visual system to enhance detection.



Long-Range Modulatory Feedback Pathways

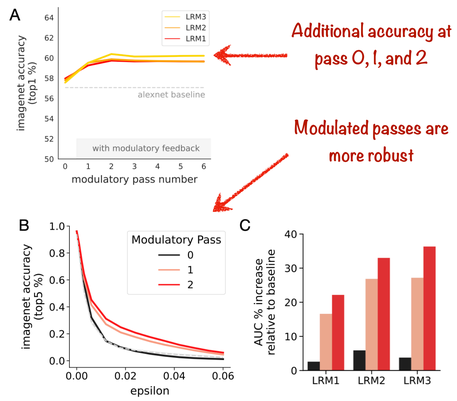
We designed long-range modulatory (LRM) feedback pathways, based on empirical findings in neuroscience and visual cognition, which can add on to any standard computer vision model. These allow later layers to influence processing in earlier layers, via learned channel-to-channel modulations.



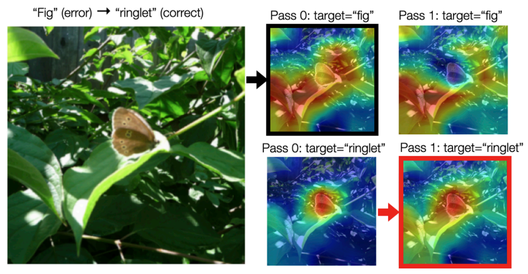
- **Architectures**: LRM1, LRM2, LRM3 with increasingly more feedback pathways on an Alexnet base encoder
- **Modulatory Motif**: Learnable influences between source layer channels and destination layer channels; learnable +/- gain
- **Dynamics**: Consecutive feed forward passes. A standard feed-forward pass (t=0); activations are remixed through learned pathways, and modulate the next forward pass (t=1). Multiplicative effect on destination layer: $x_t + (x_t \cdot f_{t-1})$
- **Loss**: Standard supervised cross-entropy loss computed separately for pass 0 and pass 1, then averaged.
- **Dataset:** ImageNet
- **Code**: Any model can be outfitted with LRM pathways. Link with QR code:

Inference time: Default Operation

With default feedback dynamics, LRM-enhanced models naturally have both improved **ImageNet recognition accuracy** and increased **adversarial robustness** compared to baseline models



The learned feedback pathways naturally help **re-route misclassified images** towards more accurate representations.



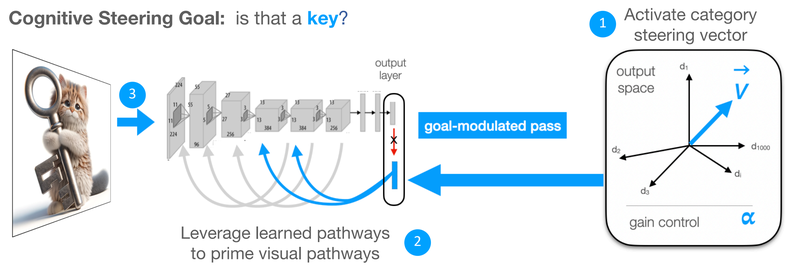
And the features learned in LRM models have **higher brain alignment** (assessed through Brain Score platform).

Brain Area	Baseline Alexnet	LRM3 (pass 0)	LRM3 (pass1)	Δ (from baseline)	rank change
IT	r = 0.358	r = 0.393	r = 0.400	**+0.042**	#145 → **#35**
V4	r = 0.443	r = 0.454	r = 0.467	**+0.024**	#153 → **#97**
V2	r = 0.353	r = 0.341	r = 0.333	-0.020	#13 → #48
V1	r = 0.507	r = 0.492	r = 0.531	**+0.024**	#68 → **#32**

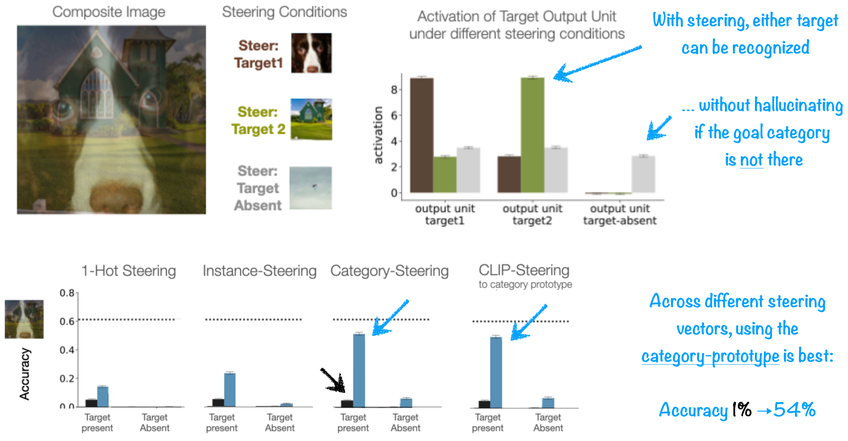
Table 1: Brain-Score results for the Baseline Alexnet model and LRM3 model. The r-value indicates the average single-unit neuron predictivity scores, reported for different visual areas along the ventral visual stream hierarchy.

"Cognitive Steering": Goal-directed encoding

The **output layer** of these models can now function as an **interpretable 'cognitive steering' interface**: Target goals are specified as vectors in the 1000-d output space, where the learned back-projections modulate earlier layers, enhancing target-relevant features present in the input.



Key Result: With steering, LRM-enhanced models can accurately recognize either of two categories present in a composite image, where matched baseline models fail dramatically. And, our multiplicative feedback motif prevents rampant hallucinations of the target, keeping false-alarms at negligible rates.



Take homes…

Long-range modulatory feedback pathways allow for different goal states to make flexible use of fixed visual circuitry, **enabling dynamic goal-based routing of incoming visual information**

These architectural pathways offer new possibilities for integrative systems (e.g. **multi-modal vision-language alignment**; RL agents **with goal-directed visual encoding**) to enable communication between visual and cognitive components of AI models