Generativ AI and the brain: how context shapes visual perception

Balázs Meszéna

Wigner RCP (RMI), KTO (CSNL, KTT)

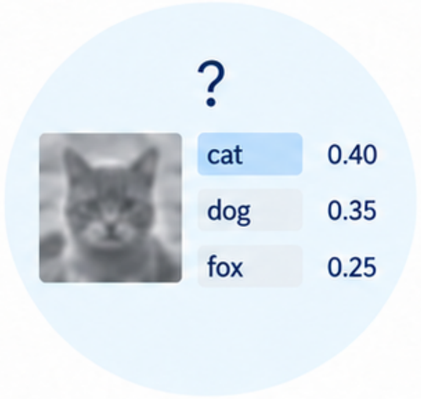
AIME, 2026.11.12



http://golab.wigner.mta.hu/

ICLR

Challenges in Machine Learning



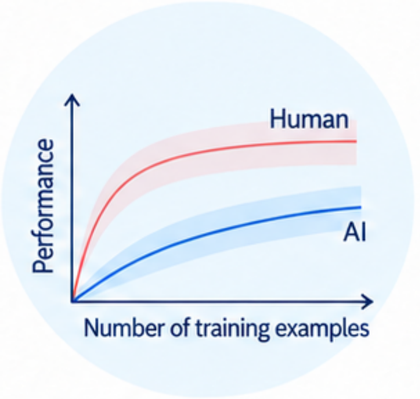
Uncertainty estimation

Can the system recognize when its prediction is uncertain?



Distribution shift

What happens when test conditions differ from training?



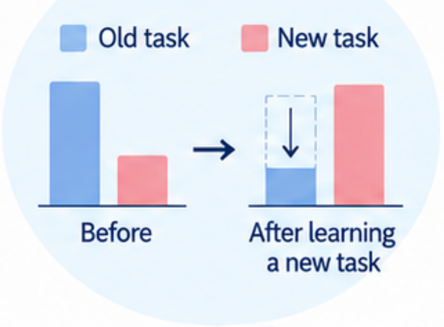
Rapid adaptation

How can a model adapt using little new data?



Robust generalization

How can a model remain reliable in new, noisy, or partly unfamiliar situations?

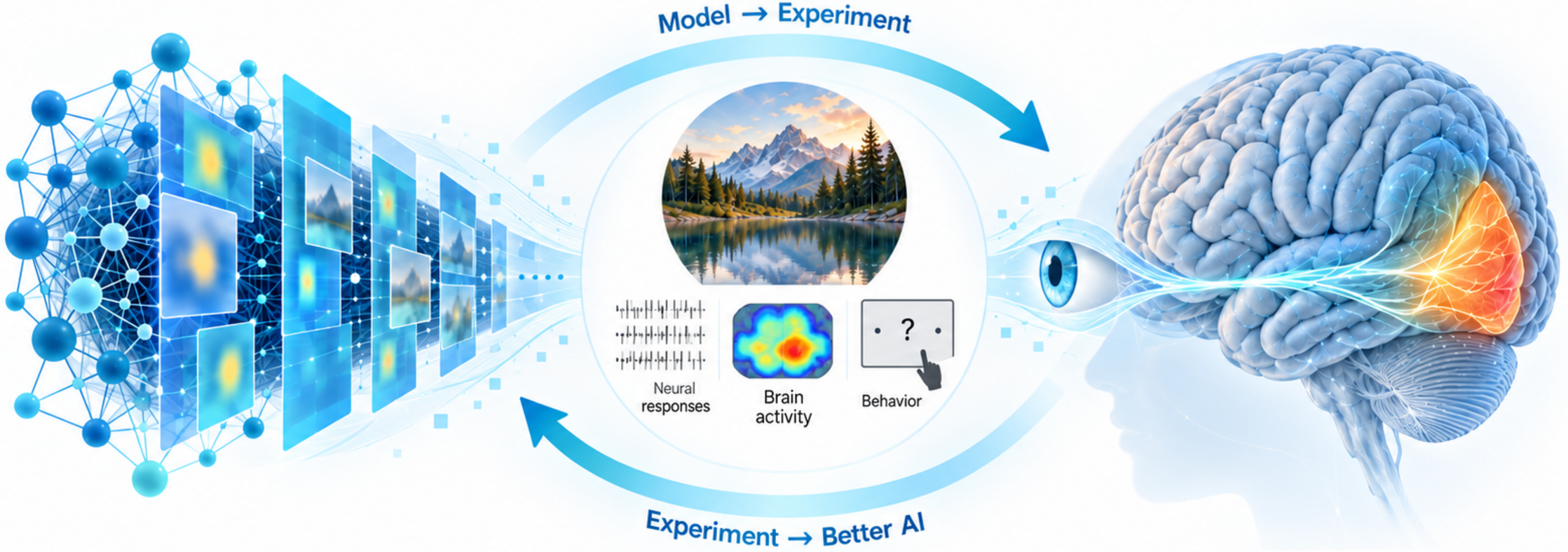


Catastrophic forgetting

How much previously learned knowledge is lost when learning a new task?

AI models to understand the brain

Modeling neural and behavioral phenomena — and learning from biology



AI → Neuroscience

Build normative and mechanistic models that explain neural responses and behavior.

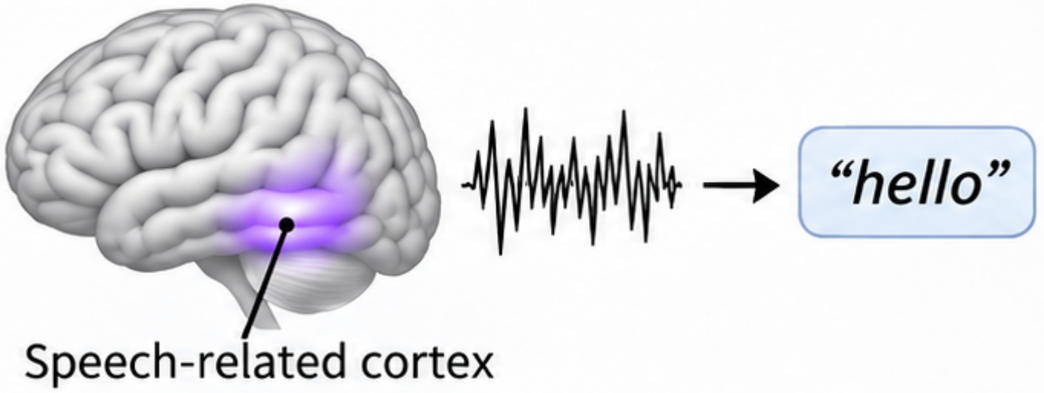
Neuroscience → AI

Biological principles can inspire more robust, flexible and data-efficient AI.

Why model sensory neural responses?

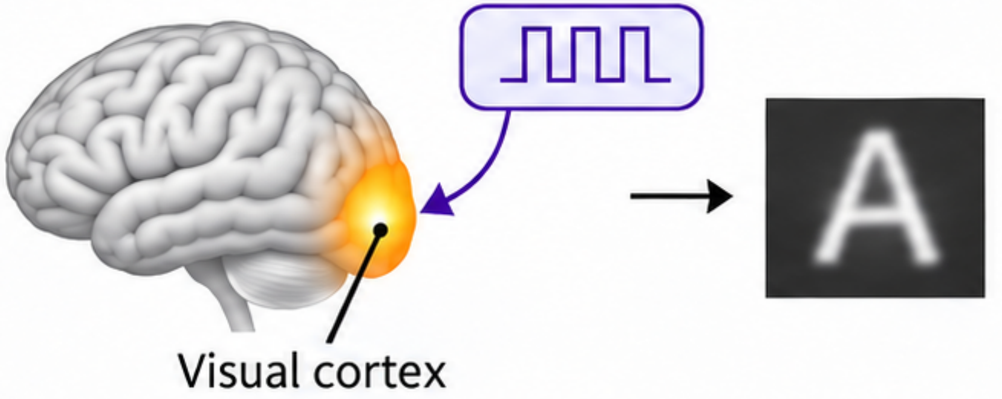
Linking brain activity and sensory experience enables novel applications.

Decoding speech from speech-related cortex



Potential impact: communication for people with speech impairments such as ALS, stroke, or locked-in syndrome.

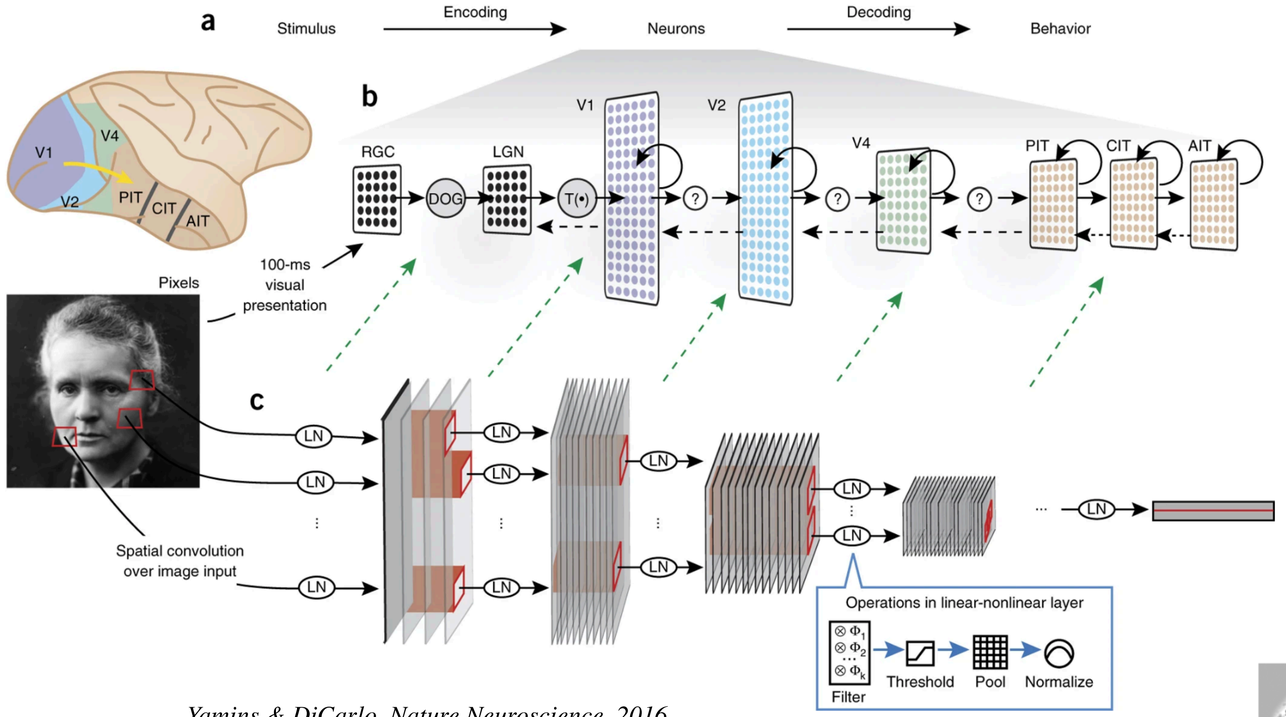
Restoring vision via visual cortex stimulation



Potential impact: partial vision restoration by targeted stimulation.

Deeper understanding of sensory processing and perception is needed.

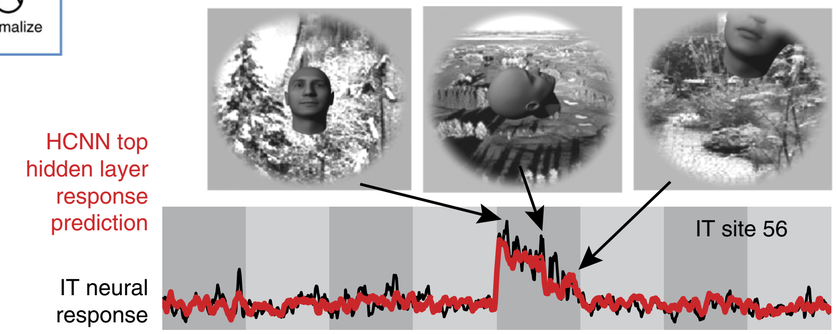
Predicting neuronal responses in the visual cortex using deep (discriminative) neural networks



Yamins & DiCarlo, Nature Neuroscience, 2016

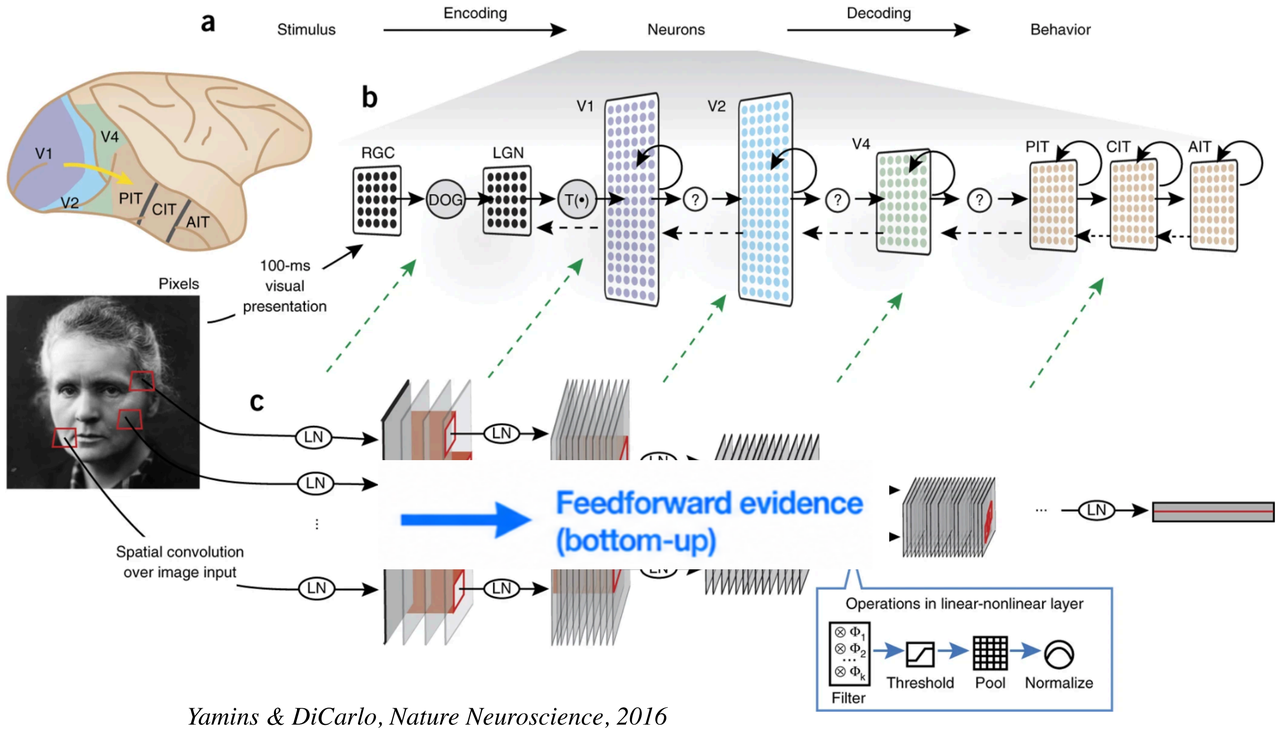
Biological neurons

Model CNN layers



Yamins et al (2016) Nat Neurosci

Predicting neuronal responses in the visual cortex using deep (discriminative) neural networks



Yamins & DiCarlo, Nature Neuroscience, 2016

Biological neurons

Model CNN layers

Only feedforward (bottom-up) signals

Perception as evidence-prior fusion

Bottom-up
(feedforward)
sensory evidence



Daylight, no occlusion, high SNR

Strong sensory evidence



Established

Perception as evidence-prior fusion

Bottom-up (feedforward) sensory evidence



Daylight, no occlusion, high SNR

Strong sensory evidence

Established



Top-down prior / prediction

Low light (or heavy occlusion, low SNR)

Weak sensory evidence

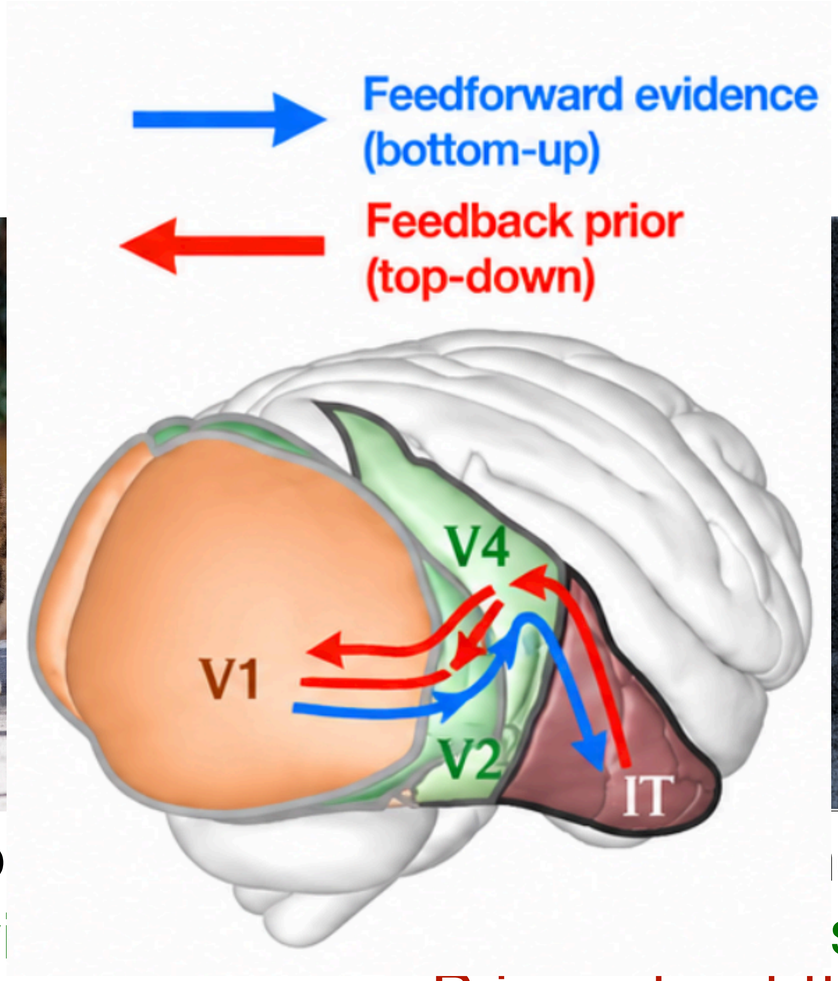
Priors about the world are crucial!

Struggling

Perception as evidence-prior fusion

Bottom-up
(feedforward)
sensory evidence







Top-down
prior /
prediction

Daylight, no occlusio

eavy occlusion, low SNR)

Strong sensory evi

sory evidence

Priors about the world are crucial!

Established

Struggling

Perception as inference in a latent generative model

Example latent causes z
(encoded in the brain)



IT: object hypothesis
"there is a tree"



V2: texture / category
"bark / grass texture"



V1: edges / orientations
"oriented edges"



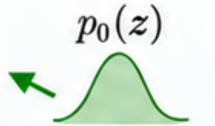
brain stores latent features + prior over them
(top-down)

Learned world model

Learned from natural images

z = **latent causes** (hierarchical factors)

$$z = [\, z_{IT}, z_{V2}, z_{V1} \,]$$

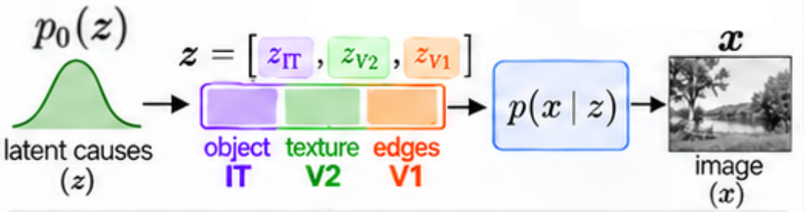


Generative model (from latent causes to images)

$$p_0(x) = \int p(x \mid z)\, p_0(z)\, dz$$

likelihood

prior over latent causes



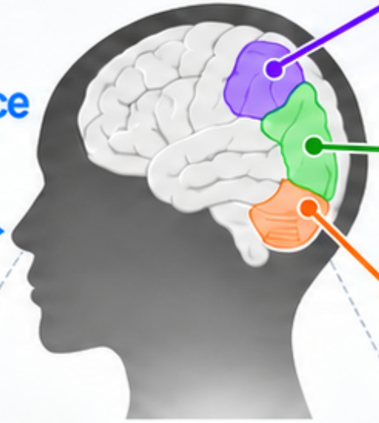
The brain represents multiple latent causes at different levels:
edges → textures → objects.

Perception as inference in a latent generative model

Observed scene



sensory evidence
(bottom-up)



Example latent causes z
(encoded in the brain)

IT: object hypothesis
"there is a tree"



V2: texture / category
"bark / grass texture"



V1: edges / orientations
"oriented edges"



brain stores latent features
+ prior over them
(top-down)

posterior / inferred latent state

$$q(z \mid x)$$

≈V1 neuronal response

V1:
edges / orientations



V2:
texture / category



IT:
object hypothesis



Learned world model

Learned from natural images

z = **latent causes** (hierarchical factors)

$$z = \left[z_{IT}, z_{V2}, z_{V1} \right]$$

$p_0(z)$

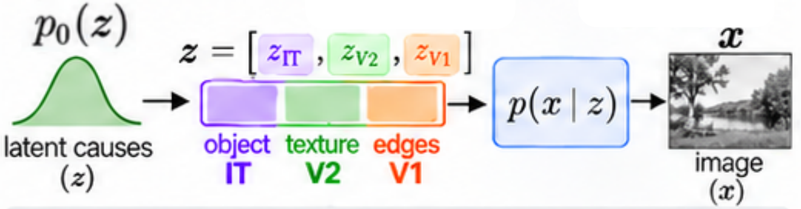
prior over latent causes

Generative model (from latent causes to images)

$$p_0(x) = \int p(x \mid z)\, p_0(z)\, dz$$

likelihood

prior over latent causes



The brain represents multiple latent causes at different levels:
edges → textures → objects.

Perception = fuse sensory evidence with prior knowledge

A Variational Autoencoder implements this inference framework

Natural images



Training
optimize neural-network weights

sensory evidence → inference

world model

prior
$p_0(z)$

x



image

Encoder
$q(z \mid x)$

z

latent variables
$z \approx$ V1 neurons

Decoder
$p(x \mid z)$

$\hat{x}$



reconstructed image

variational posterior
$q(z \mid x)$
≈ neuronal response (membrane potential)

Encoder = approximate inference
maps sensory input to latent variables

Decoder + prior = generative model
defines a model of natural images

Importance of uncertainty



- For calibrated bottom-up / top-down fusion we need proper uncertainty estimates

- VAEs are probabilistic models with uncertainties…

Importance of uncertainty

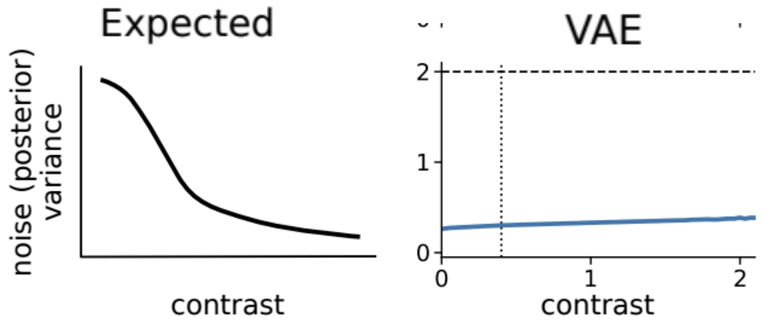


- For calibrated bottom-up / top-down fusion we need proper uncertainty estimates

- VAEs are probabilistic models with uncertainties… *In principle…*

Neuron model VAE





Importance of uncertainty

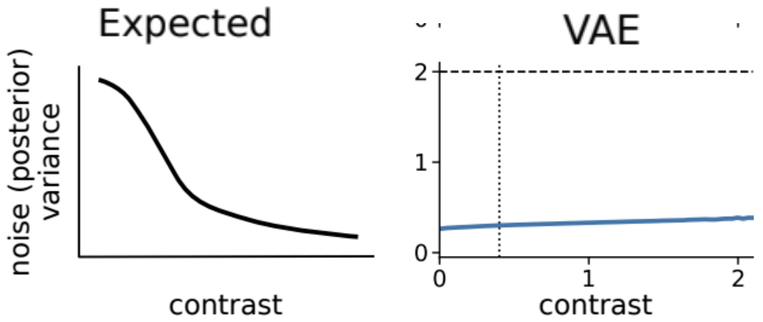


- For calibrated bottom-up / top-down fusion we need proper uncertainty estimates

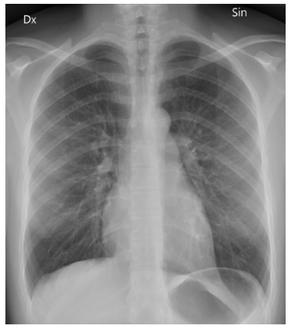
- VAEs are probabilistic models with uncertainties… *In principle…*

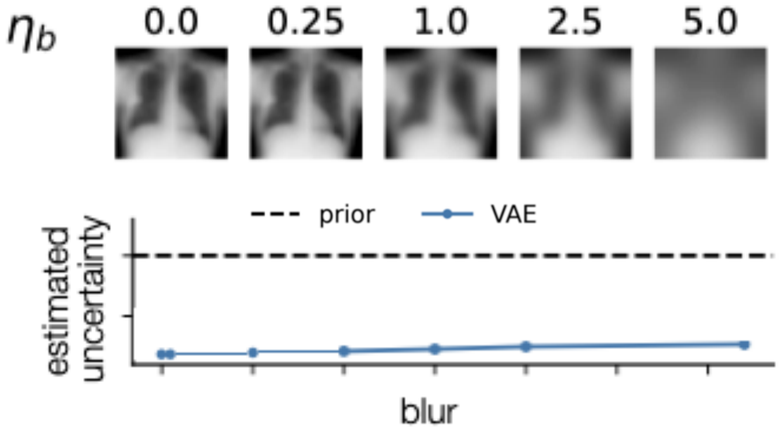
Neuron model VAE





medical image processing





Importance of uncertainty

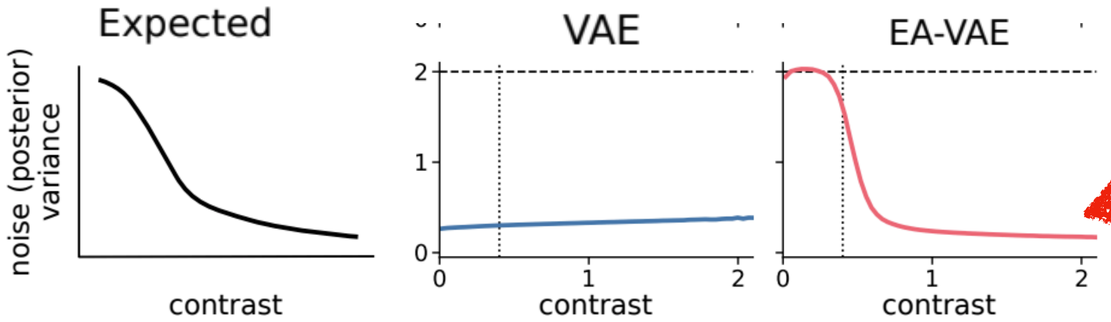


- For calibrated bottom-up / top-down fusion we need proper uncertainty estimates

- VAEs are probabilistic models with uncertainties… *In principle…*

Neuron model VAE



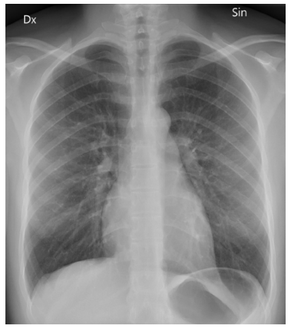


Scaling variable

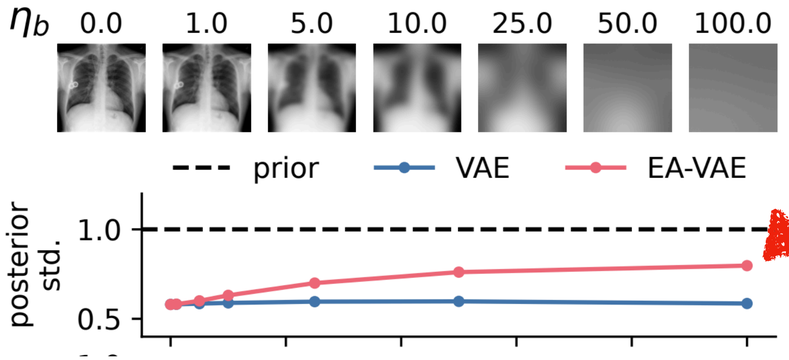
$$p_\varphi(\boldsymbol{x} \mid \boldsymbol{z}, s) = \mathcal{N}(\boldsymbol{x}; \exp(s) \cdot \boldsymbol{A}\boldsymbol{z}, \sigma^2 \boldsymbol{I})$$

Neuroscience inspired solution

medical image processing



Better OOD uncertainity representation in VAEs



IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)

With Domonkos Martos, Ferenc Csikor, Gergő Orbán et al.

Importance of uncertainty

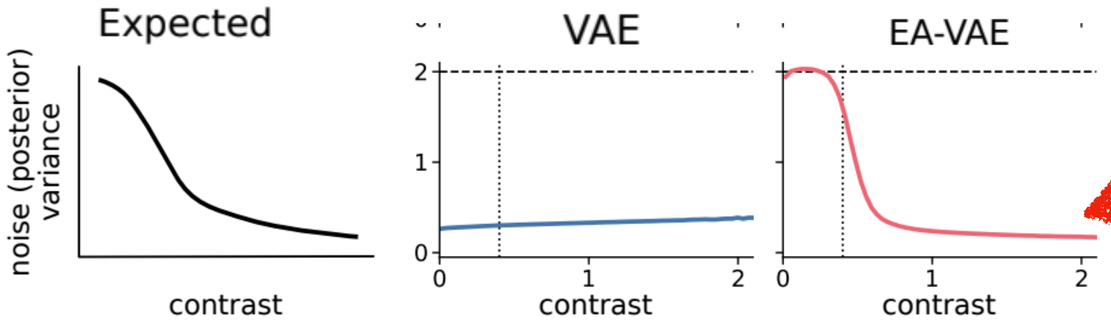


- For calibrated bottom-up / top-down fusion we need proper uncertainty estimates

- VAEs are probabilistic models with uncertainties… *In principle…*

Neuron model VAE



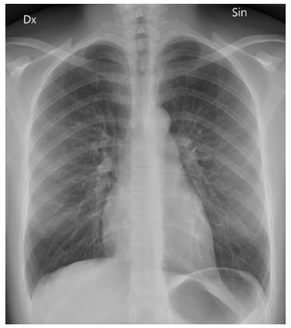


Scaling variable

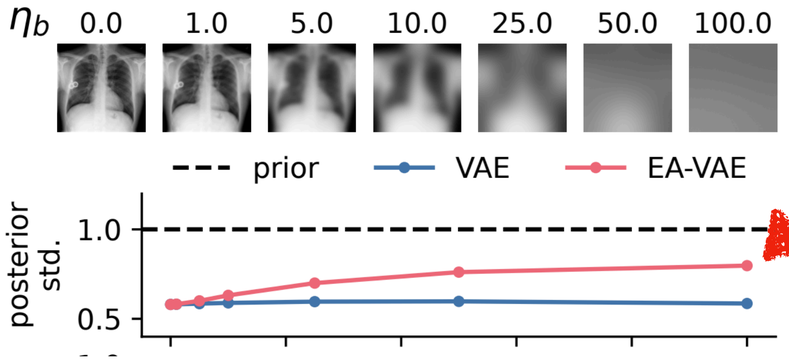
$$p_\varphi(\boldsymbol{x} \mid \boldsymbol{z}, s) = \mathcal{N}(\boldsymbol{x}; \exp(s) \cdot \boldsymbol{A}\boldsymbol{z}, \sigma^2 \boldsymbol{I})$$

Neuroscience inspired solution

medical image processing

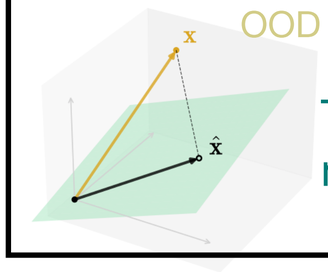


Better OOD uncertainity representation in VAEs



Half-analytical guarantees

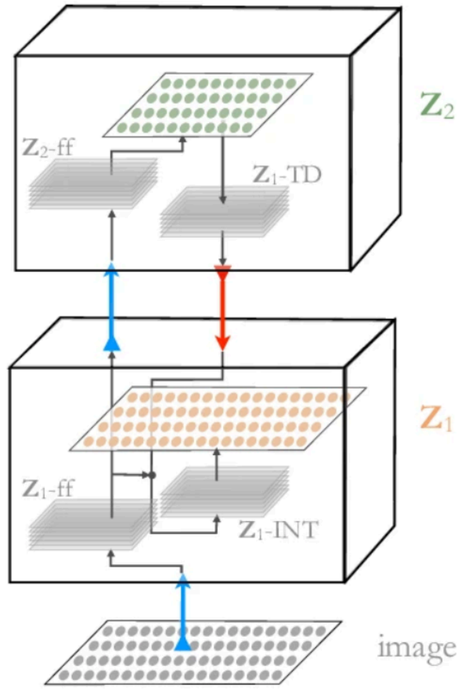
EAVAE



IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)

With Domonkos Martos, Ferenc Csikor, Gergő Orbán et al.

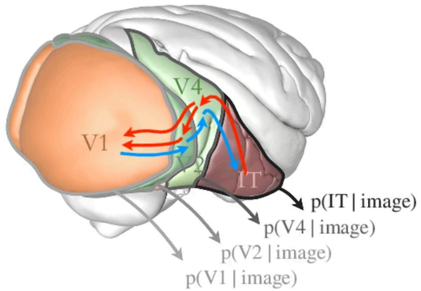
TDVAE: V1-V2 hierarchical system



Extension to V1-V2

TDVAE

Top-down signals from higher visual regions



NeurIPS SVRHM (2022)

Nature Communications (2025)

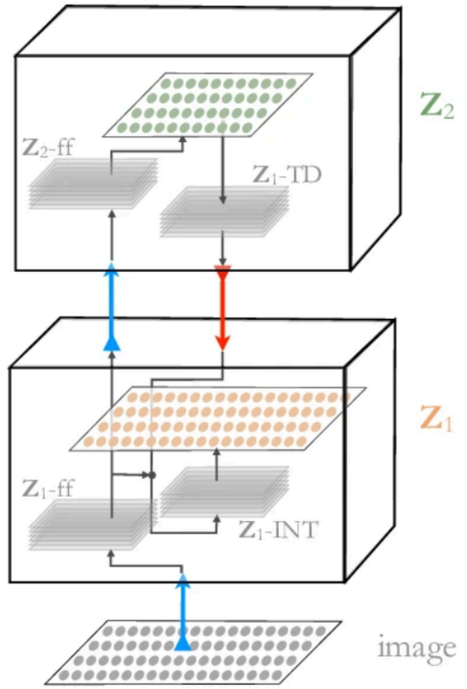
With Ferenc Csikor, Katalin Ócsai, Gergő Orbán

Top-down perceptual inference shaping the activity of early visual cortex

Ferenc Csikor, Balázs Meszéna, Katalin Ócsai & Gergő Orbán

Nature Communications **16**, Article number: 9998 (2025) | Cite this article

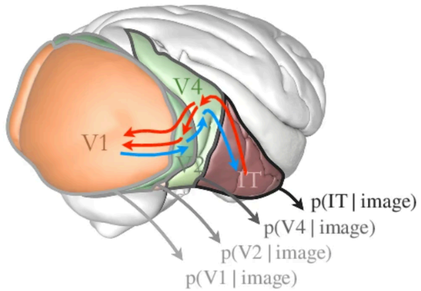
TDVAE: V1-V2 hierarchical system



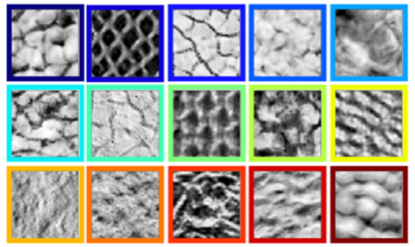
Extension to V1-V2

TDVAE

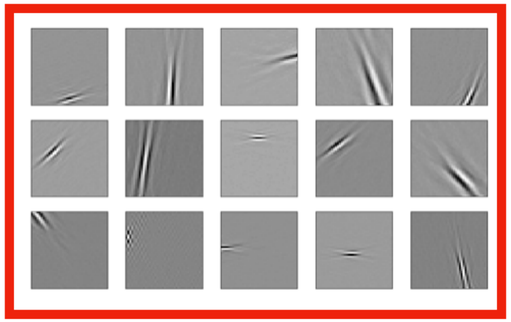
Top-down signals from higher visual regions



Z2: textures



Z1: Gábor filters (edges)



NeurIPS SVRHM (2022)

Nature Communications (2025)

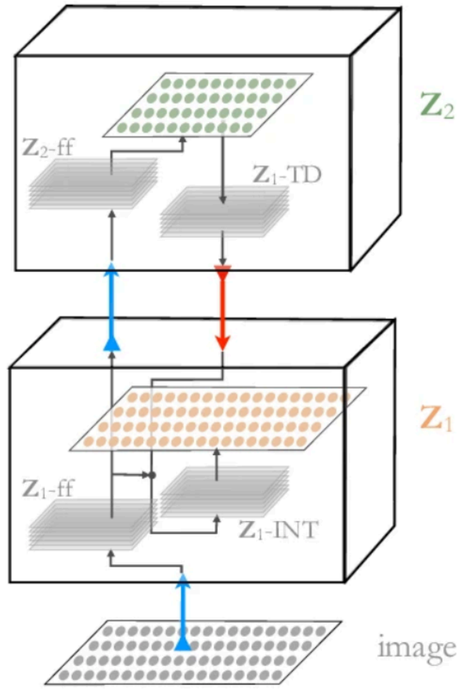
With Ferenc Csikor, Katalin Ócsai, Gergő Orbán

Top-down perceptual inference shaping the activity of early visual cortex

Ferenc Csikor, Balázs Meszéna, Katalin Ócsai & Gergő Orbán

Nature Communications **16**, Article number: 9998 (2025) | Cite this article

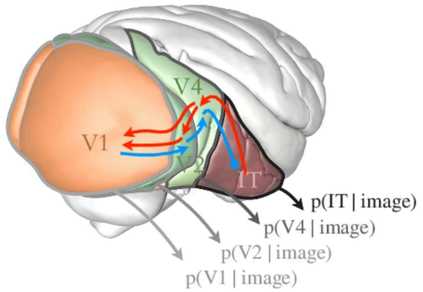
TDVAE: V1-V2 hierarchical system



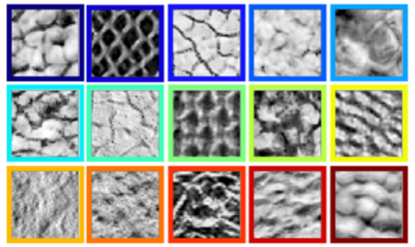
Extension to V1-V2

TDVAE

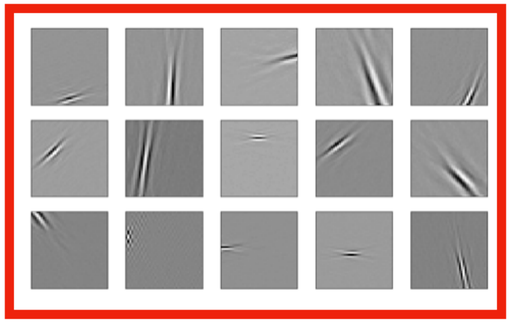
Top-down signals from higher visual regions



Z2: textures



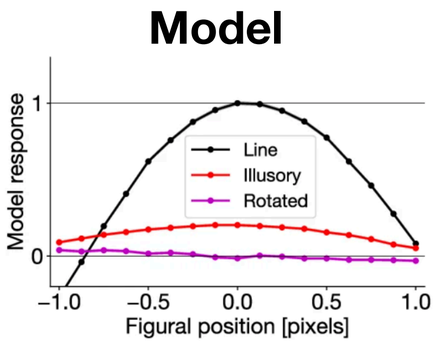
Z1: Gábor filters (edges)

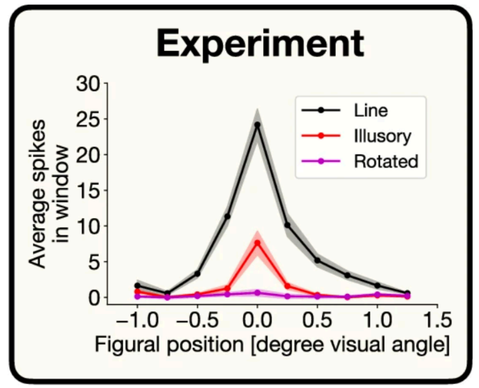


Illusory contour
V1 response









NeurIPS SVRHM (2022)

Nature Communications (2025)

With Ferenc Csikor, Katalin Ócsai, Gergő Orbán

Top-down perceptual inference shaping the activity of early visual cortex

Ferenc Csikor, Balázs Meszéna, Katalin Ócsai & Gergő Orbán

Nature Communications **16**, Article number: 9998 (2025) | Cite this article

Context shifts the prior



Garage / garden context

Prior favors hose



Grass / trail context

Prior favors snake

Same local evidence, different likely explanation

Joint work with:

Paper

Experiment

Keith T. Murray



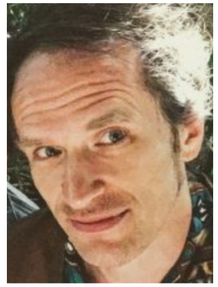
PRINCETON UNIVERSITY

Gergő Orbán



csnl

Márton A. Hajnal



HUN REN | wigner

Julien Corbo



O. Batuhan Erkat



Pierre-Olivier Polack



RUTGERS UNIVERSITY | NEWARK

Passive viewing paradigm

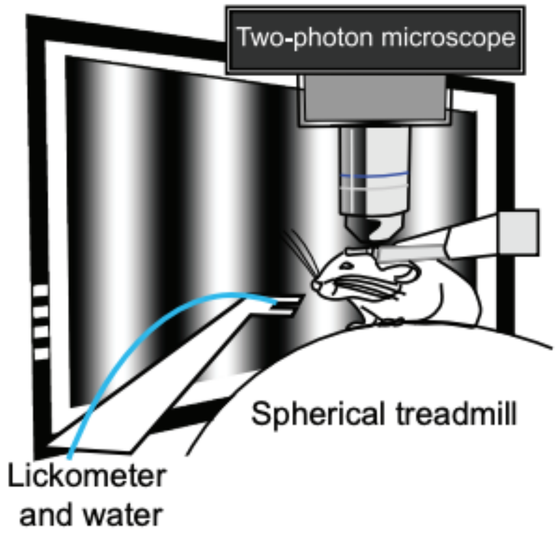


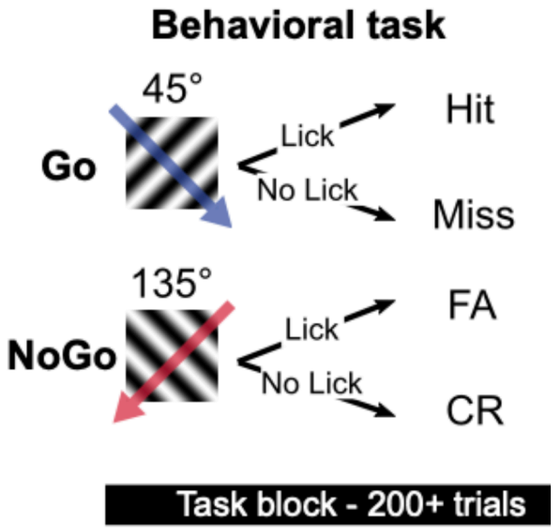
Recording during task context!



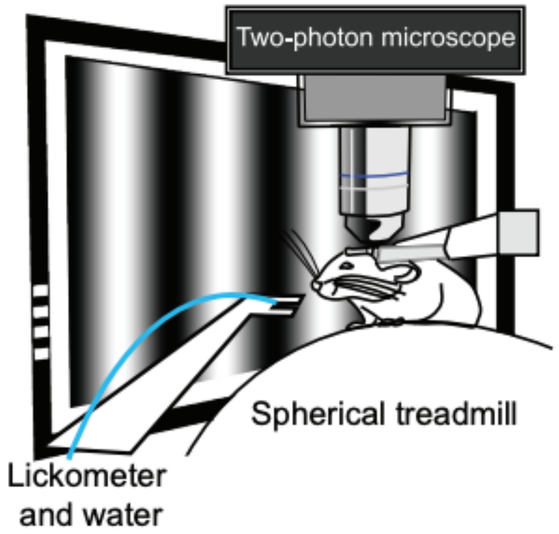
ICLR International Conference for Learning Representations, 2026

Experimental setup

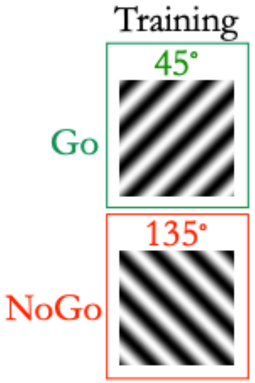




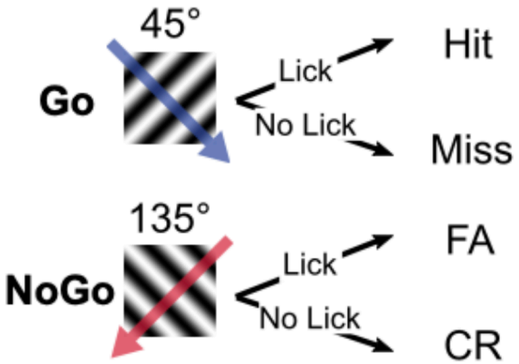
Experimental setup



9-18 days

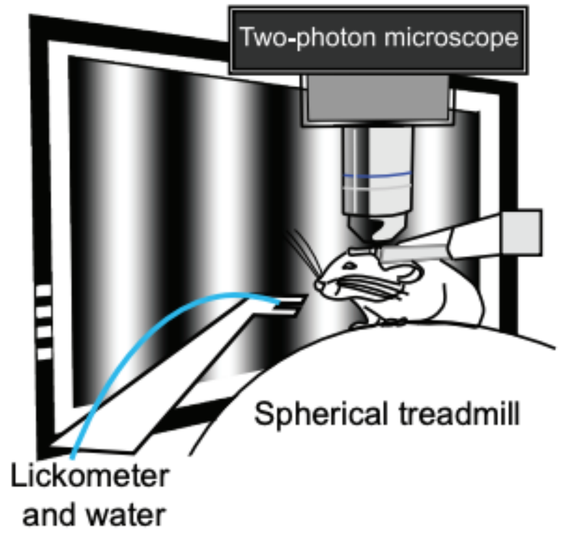


Behavioral task

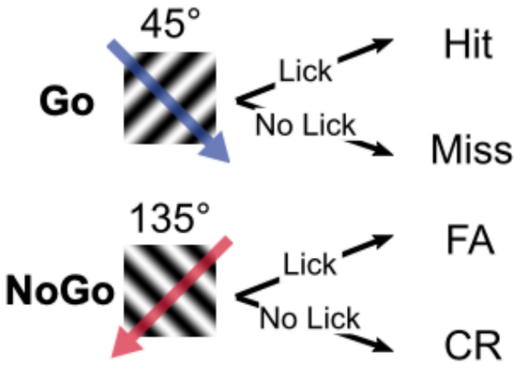


Task block - 200+ trials

Experimental setup

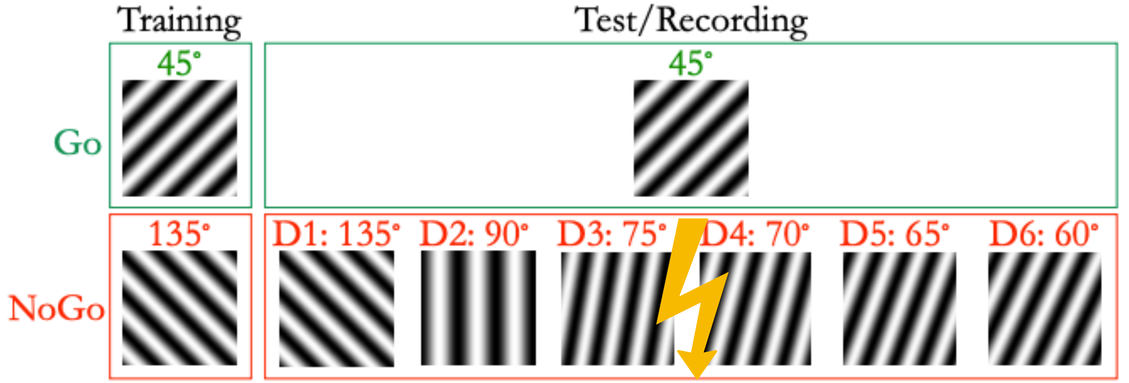


Behavioral task

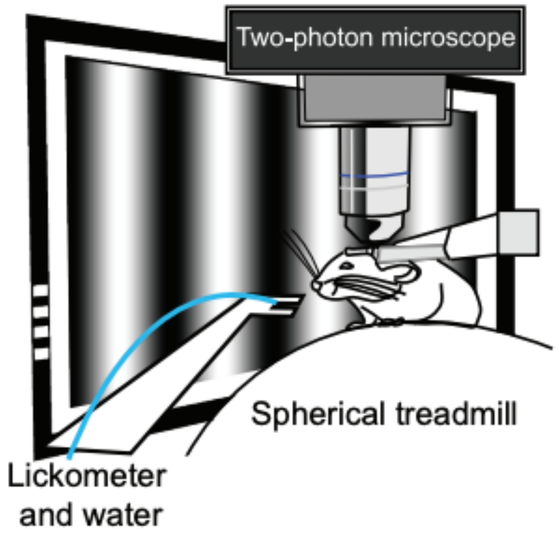


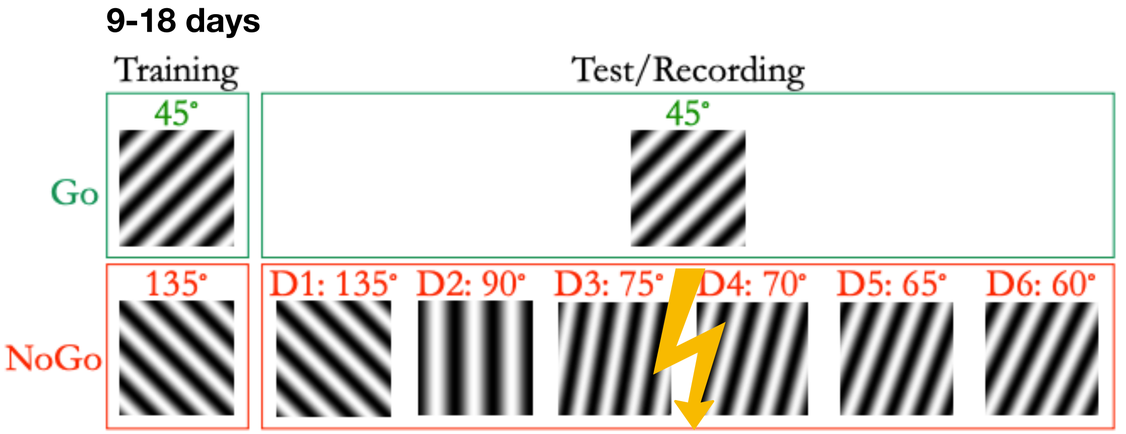
Task block - 200+ trials

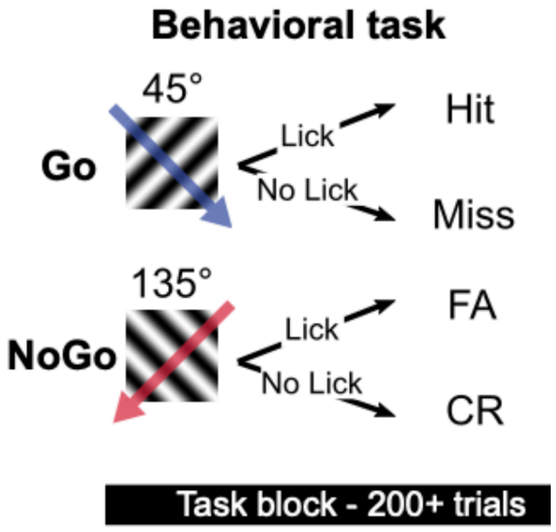
9-18 days

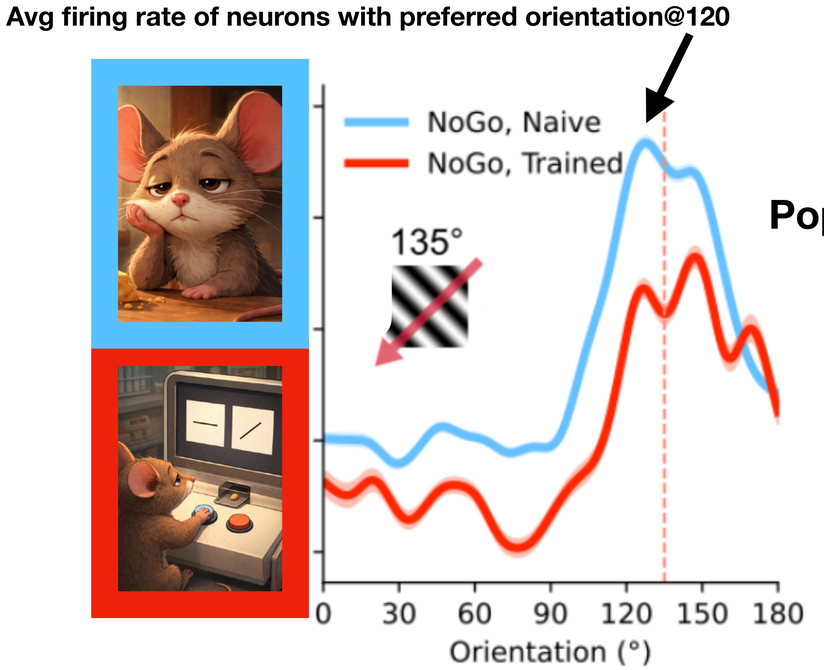


Experimental setup





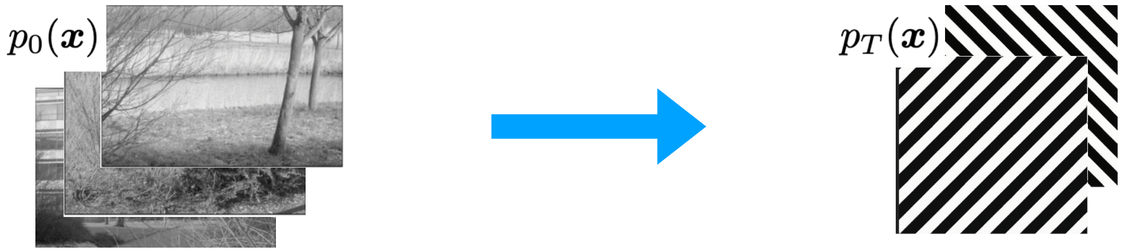


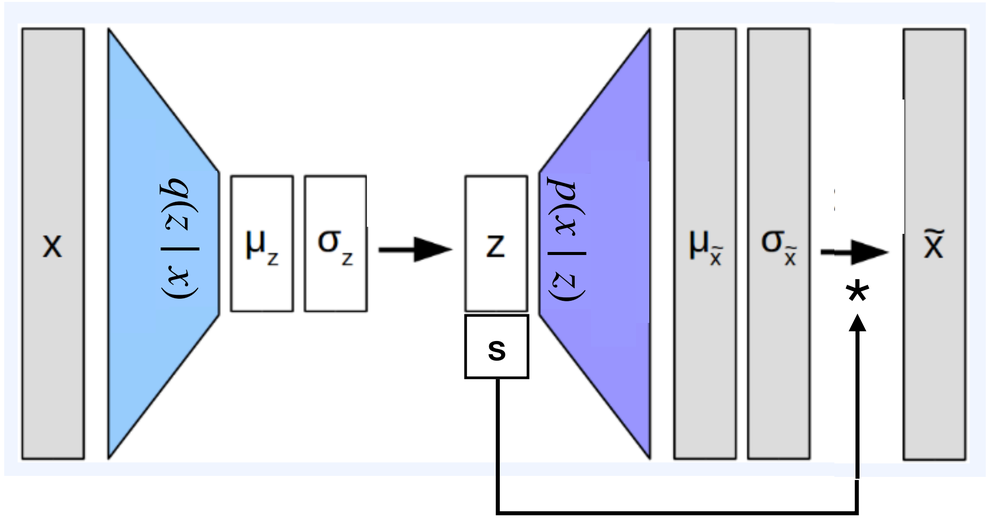


Population activity profiles

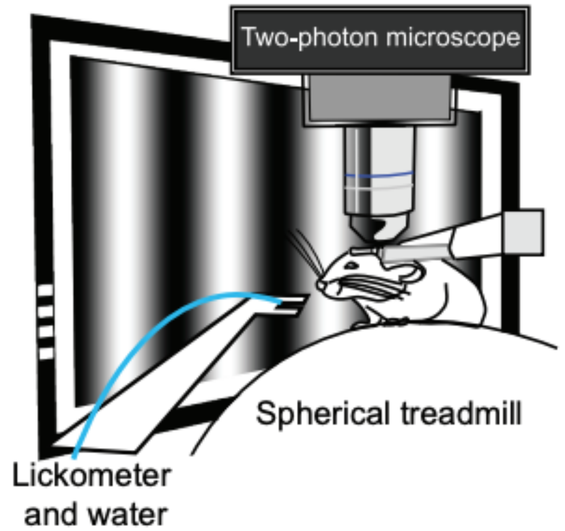
Goals (AI + neuroscience)

- Develop efficient way of context (prior) switching in latent generative models





- Study experimental signature of context switching in the brain (visual cortex)



Task adaptation in generative models

$p_0(\boldsymbol{x})$



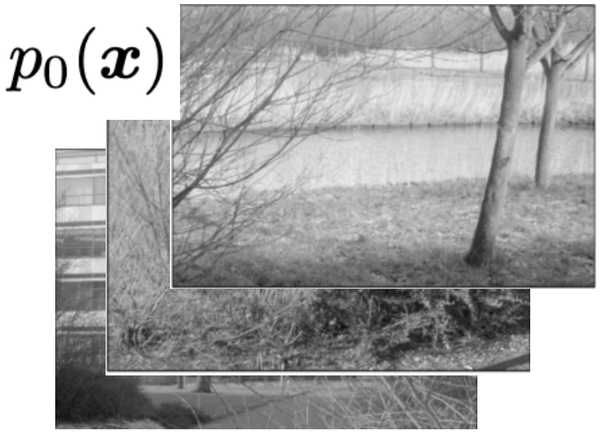
Task adaptation in generative models

$p_0(\boldsymbol{x})$



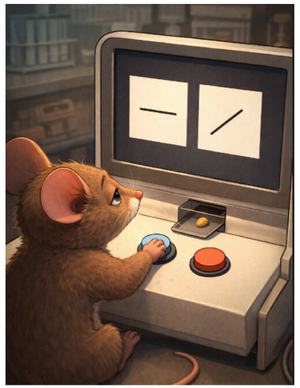


Task adaptation in generative models





How do we incorporate task adaptation?



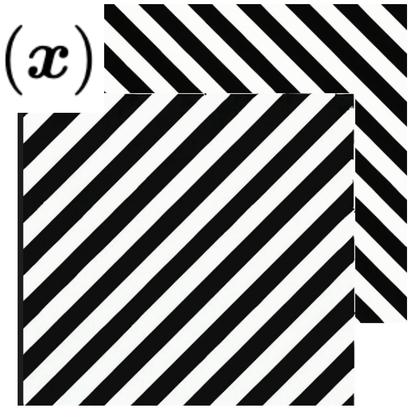
Task adaptation in generative models

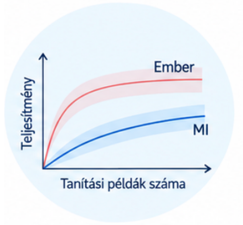
$p_0(\boldsymbol{x})$

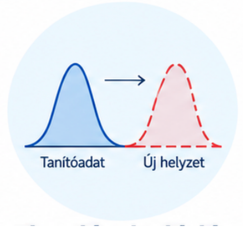


Change in stimulus statistics

$p_T(\boldsymbol{x})$









How do we incorporate task adaptation?



Task adaptation in generative models

Brute force: train new VAE for the new image set



Task adaptation in generative models

Brute force: train new VAE for the new image set



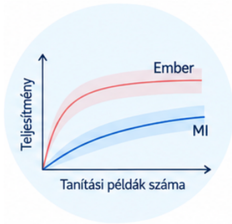
Not enough data



Task adaptation in generative models

Brute force: train new VAE for the new image set





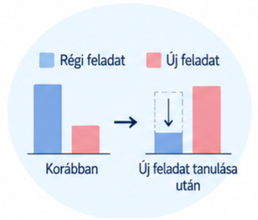
Not enough data



Tune away important representation
Catastrophic forgetting
Inefficient switching between contexts







Task adaptation in generative models

Brute force: train new VAE for the new image set



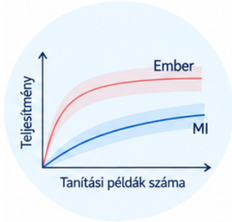
Not enough data

Tune away important representation
Catastrophic forgetting
Inefficient switching between contexts

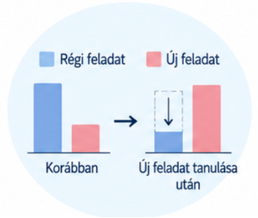
Instead: computational efficient, biologically plausible solution

Retain likelihood

$$p_T(\boldsymbol{x}) = \int \boldsymbol{p}(\boldsymbol{x} \mid \boldsymbol{z}) p_T(\boldsymbol{z})\, d\boldsymbol{z}$$



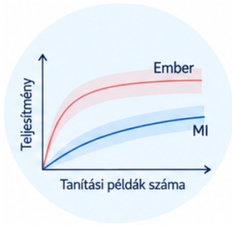




Task adaptation in generative models

Brute force: train new VAE for the new image set





Not enough data



Tune away important representation
Catastrophic forgetting
Inefficient switching between contexts

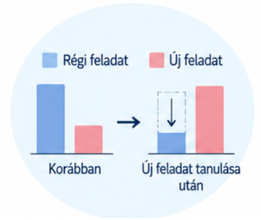




Instead: computational efficient, biologically plausible solution

Retain likelihood

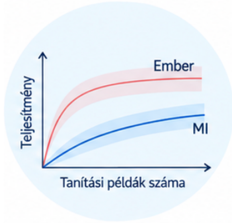
$$p_T(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{z}) p_T(\boldsymbol{z}) d\boldsymbol{z}$$



Task adaptation in generative models

Brute force: train new VAE for the new image set



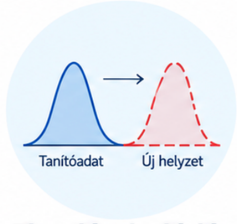


Not enough data

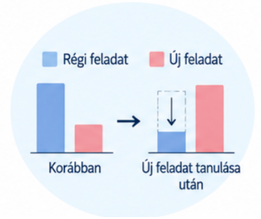


Tune away important representation
Catastrophic forgetting
Inefficient switching between contexts





Instead: computational efficient, biologically plausible solution



Retain likelihood

$$p_T(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{z}) p_T(\boldsymbol{z}) \, d\boldsymbol{z}$$



$$q_{\mathrm{T}}(\boldsymbol{z}|\boldsymbol{x}) \propto \frac{q(\boldsymbol{z}|\boldsymbol{x})\, p_{\mathrm{T}}(\boldsymbol{z})}{p_0(\boldsymbol{z})}$$

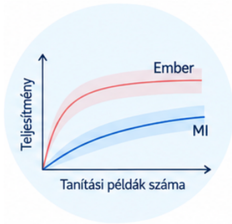
Inference in task



Task adaptation in generative models

Brute force: train new VAE for the new image set



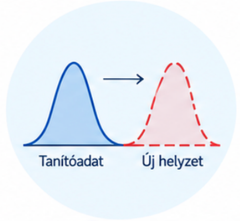


Not enough data

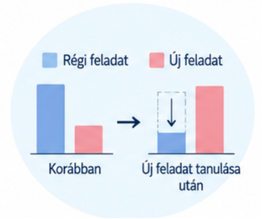


Tune away important representation
Catastrophic forgetting
Inefficient switching between contexts





Instead: computational efficient, biologically plausible solution



Retain likelihood

$$p_T(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{z}) p_T(\boldsymbol{z}) d\boldsymbol{z}$$



$$q_{\mathrm{T}}(\boldsymbol{z}|\boldsymbol{x}) \propto \frac{q(\boldsymbol{z}|\boldsymbol{x}) p_{\mathrm{T}}(\boldsymbol{z})}{p_0(\boldsymbol{z})}$$

Inference in task



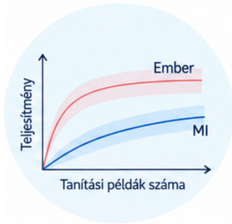
Context-free inference



Task adaptation in generative models

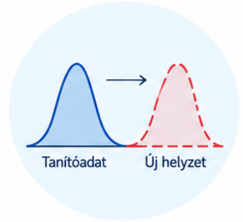
Brute force: train new VAE for the new image set



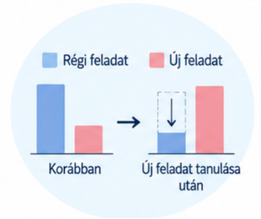


Not enough data

Tune away important representation
Catastrophic forgetting
Inefficient switching between contexts



Instead: computational efficient, biologically plausible solution



Retain likelihood

$$p_T(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{z}) p_T(\boldsymbol{z}) d\boldsymbol{z}$$

Cheap optimization

$$q_{\mathrm{T}}(\boldsymbol{z}|\boldsymbol{x}) \propto \frac{q(\boldsymbol{z}|\boldsymbol{x})\, p_{\mathrm{T}}(\boldsymbol{z})}{p_0(\boldsymbol{z})}$$

Simplified loss

$$L' = \sum_{\boldsymbol{x} \in X_T} \log \int d\boldsymbol{z}\, \frac{q(\boldsymbol{z}|\boldsymbol{x}) p_T(\boldsymbol{z})}{p_0(\boldsymbol{z})}$$

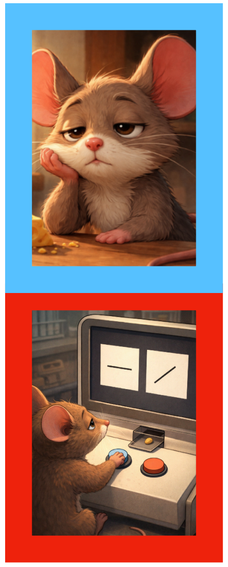
Inference in task



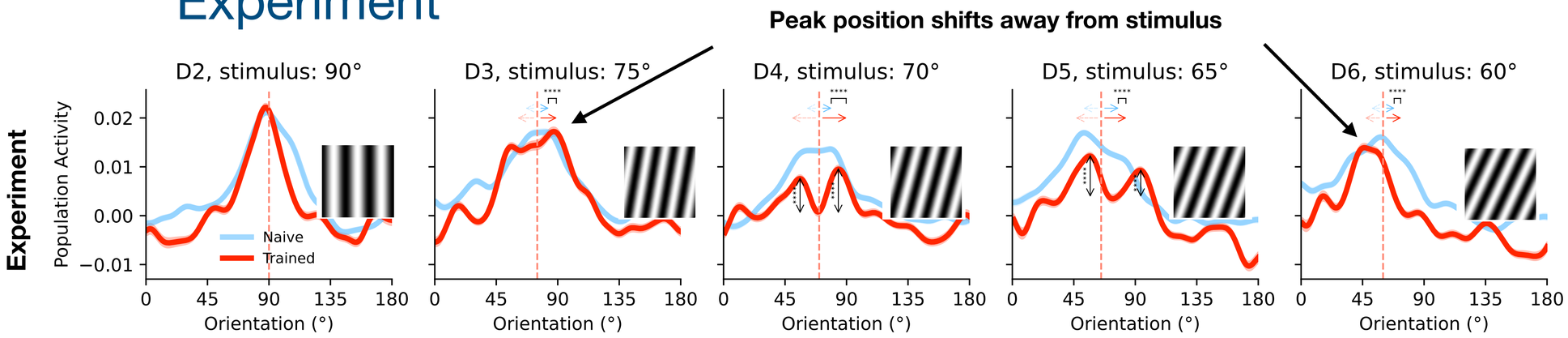
Context-free inference



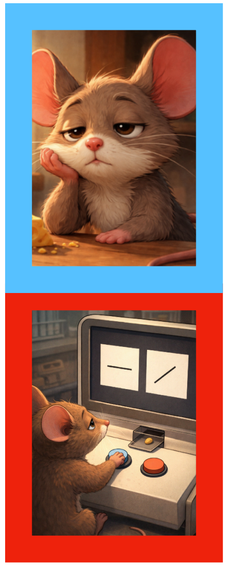
D2-6: task-stimulus mismatch



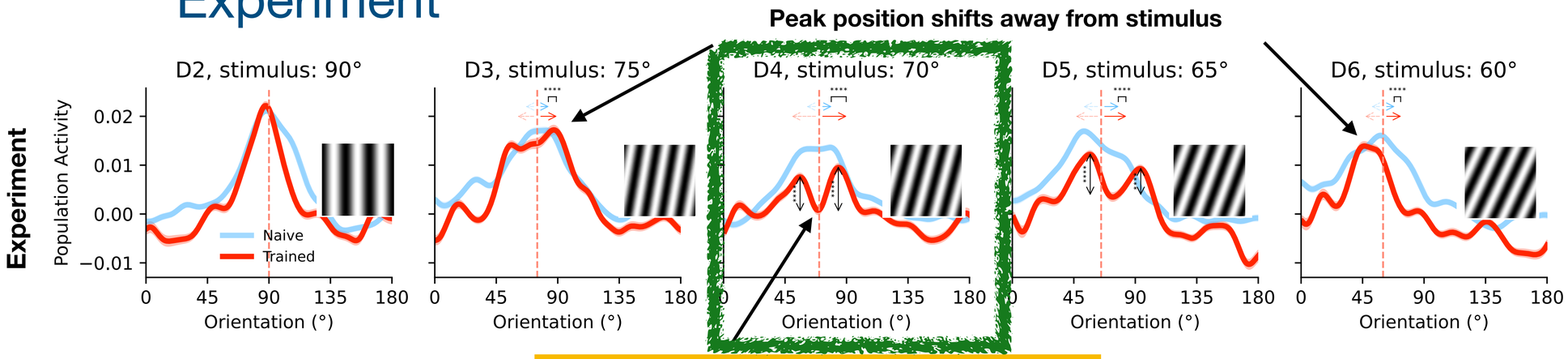
Experiment



D2-6: task-stimulus mismatch

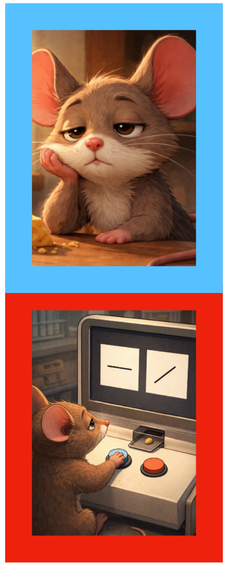


Experiment

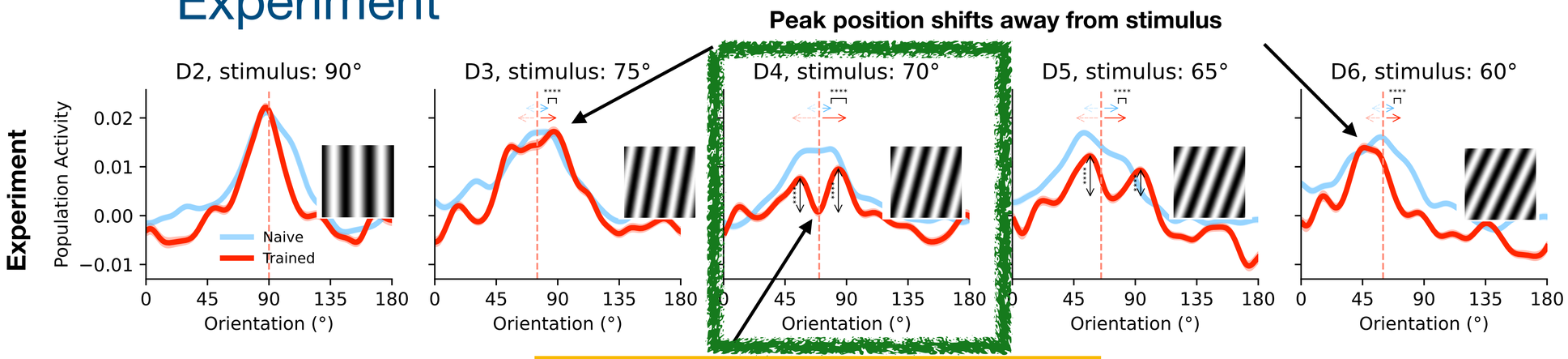


Bimodal response profile

D2-6: task-stimulus mismatch

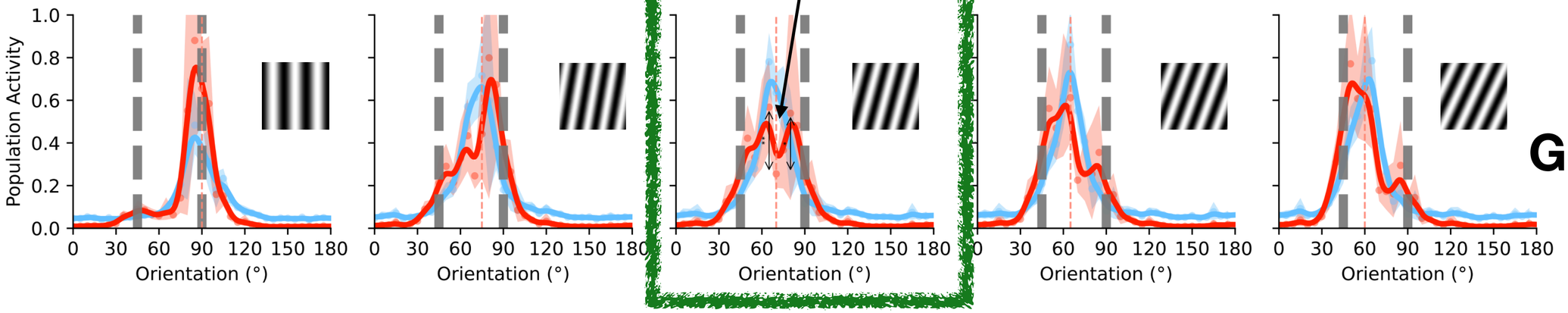


Experiment



Bimodal response profile

Model



Good match

Summary

1. **Fruitful connections between neuroscience and machine learning**

2. **Hierarchical Latent generative model: good model for the visual cortex**

3. **Efficient task adaptation by changing only the prior**
   Reuse the learned representation and encoder

4. **Prior-evidence mismatch produces structured response distortions**
   Including bimodality observed in V1 population activity

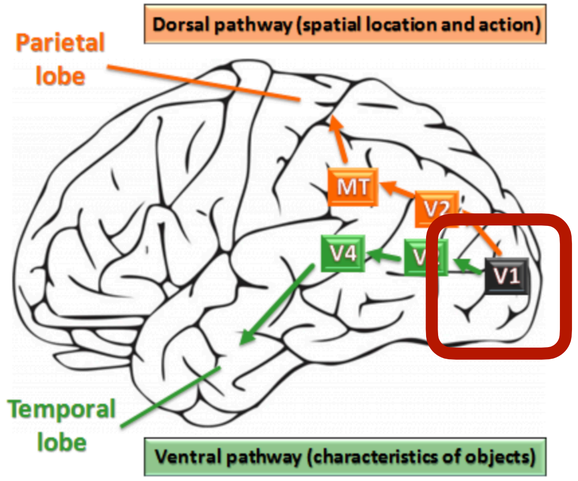
Funding

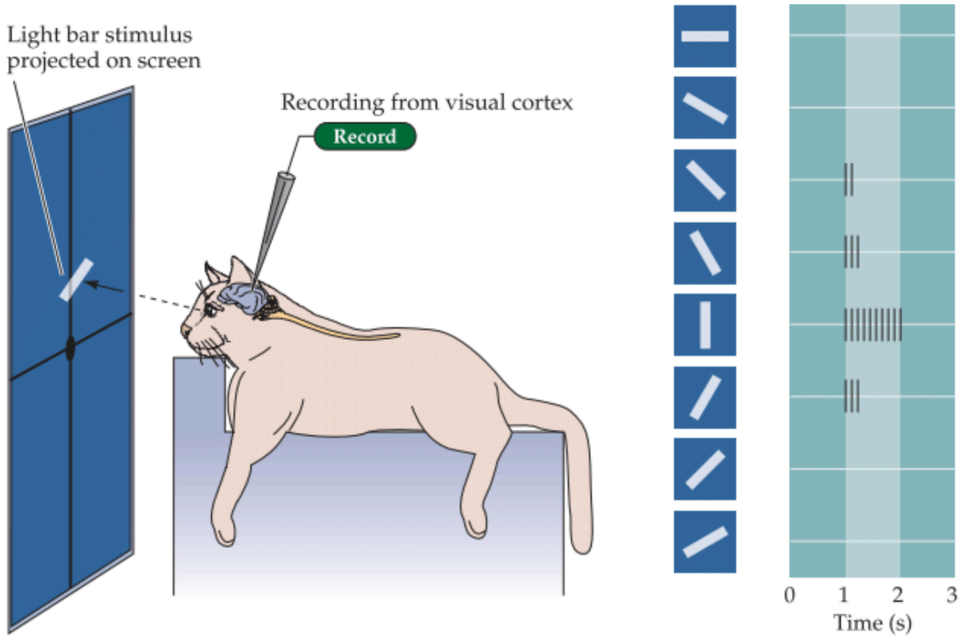
ARTIFICIAL INTELLIGENCE
National Laboratory

National
Research, Development
and Innovation Office

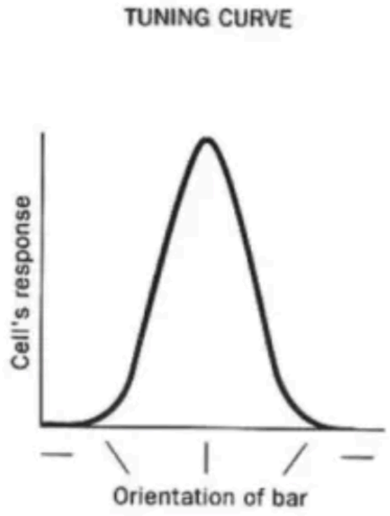
Backup slides

Orientation selectivity in V1



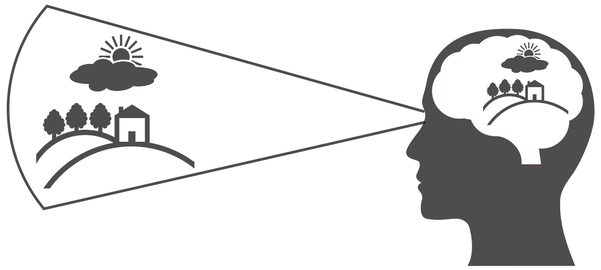


Orientation selectivity in V1

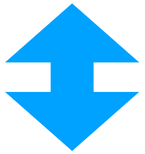


Hubel, D. H., & Wiesel, T. N. (1962)

Modeling vision with latent generative model

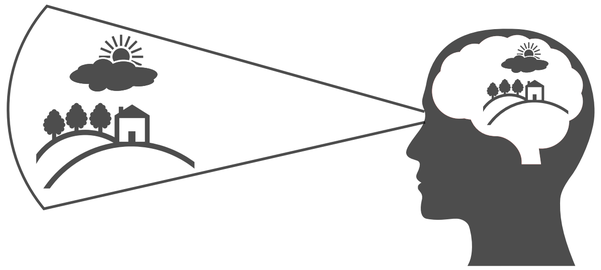


Visual cortex

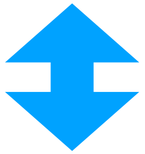


Generative model

Modeling vision with latent generative model

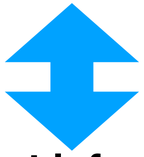


Visual cortex



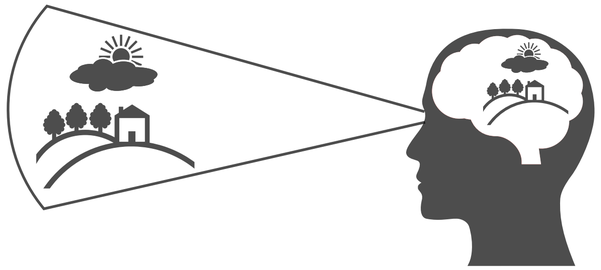
Generative model

Neuronal response



Latent inference

Modeling vision with latent generative model



Learning objective:
Learning regularities of natural images

$$p_0(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{z}) p_0(\boldsymbol{z})\, d\boldsymbol{z}$$

Likelihood Prior

Visual cortex ⇕ **Generative model**

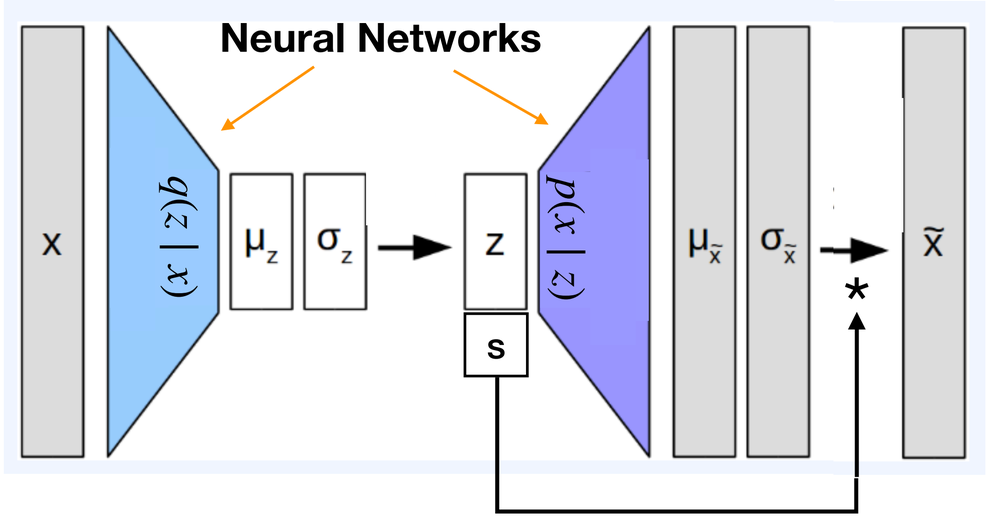
Neuronal response ⇕ **Latent inference**

natural images



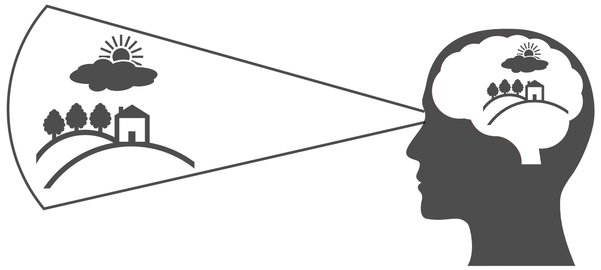
Training →

V1 VAE



Loss function (ELBO) $\mathcal{L}(\mathbf{x}, \varphi, \theta, \psi) = -\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\varphi(\boldsymbol{x}|\boldsymbol{z}, s)\right] + D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) + D_{KL}(q_\psi(s|\boldsymbol{x})||p(s))$

Modeling vision with latent generative model



Learning objective:
Learning regularities of natural images

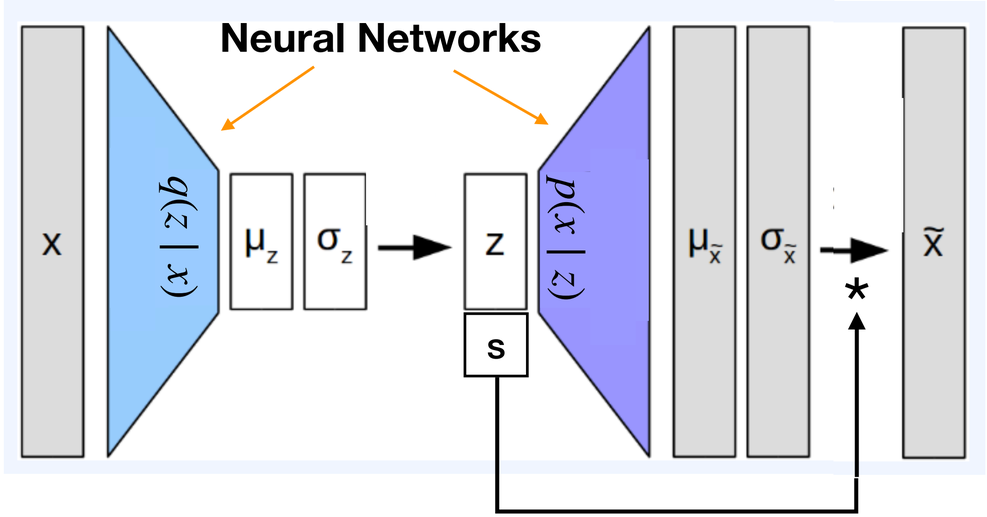
$$p_0(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{z}) p_0(\boldsymbol{z})\, d\boldsymbol{z}$$

Likelihood Prior

Visual cortex ⇕ **Generative model**

Neuronal response ⇕ **Latent inference**

V1 VAE



natural images



Training

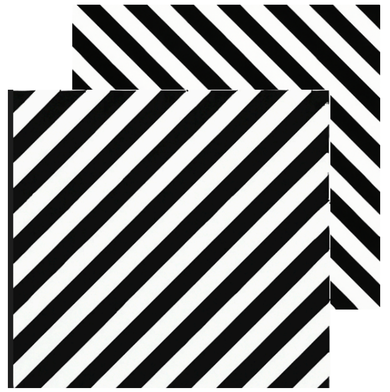
$q(z \mid x)$



Variational posterior ⇕ **Neuronal response (membrane potential)**

Loss function (ELBO) $\mathcal{L}(\mathbf{x}, \varphi, \theta, \psi) = -\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\varphi(\boldsymbol{x}|\boldsymbol{z}, s)\right] + D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) + D_{KL}(q_\psi(s|\boldsymbol{x})||p(s))$

Training of the task prior



Task dataset:
Gratings

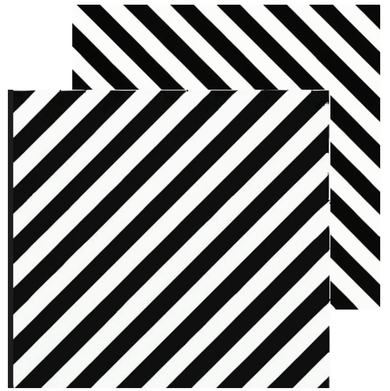
Training of the task prior

Prior $p_0(\boldsymbol{z})$ → $p_T(\boldsymbol{z}) = \mathrm{Laplace}(\boldsymbol{z}; \mathbf{0}, \underline{\sigma}_\mathrm{T}) = \prod_{i=1}^{N} \frac{1}{2\sigma_{T,i}} \exp\left(-\frac{|z_i|}{\sigma_{T,i}}\right)$

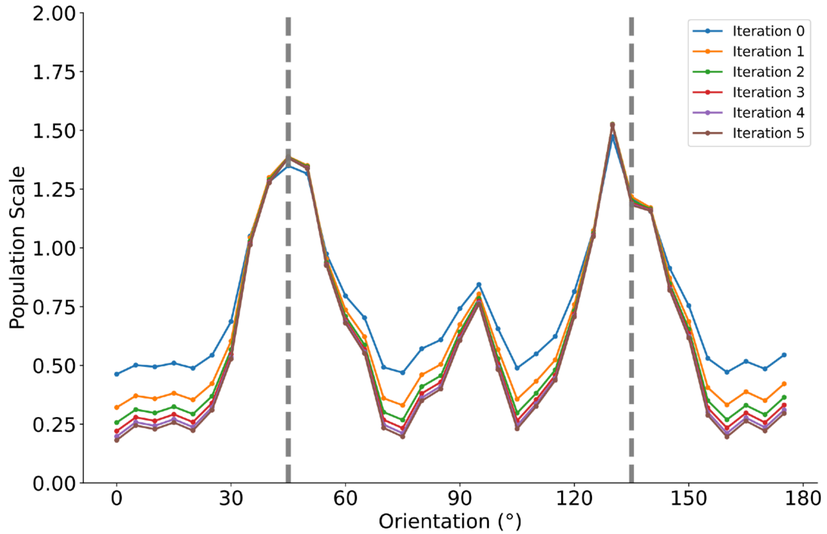
Self-consistency eq. for scales

$$\sigma_i = \frac{1}{n} \sum_{\boldsymbol{x} \in X_T} \mathbb{E}_{q_{T,\sigma}(z_i|\boldsymbol{x})} \left[|z_i|\right]$$

Task training for D1: 45° and 135° synthetic gratings with different phases



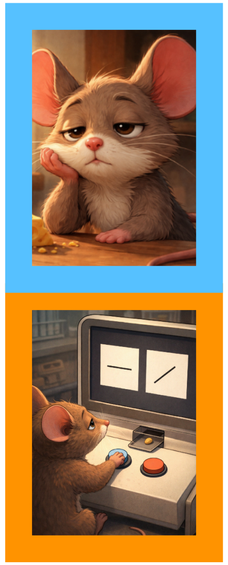
Task dataset: Gratings



Mean task-prior scale, grouped by preferred orientation

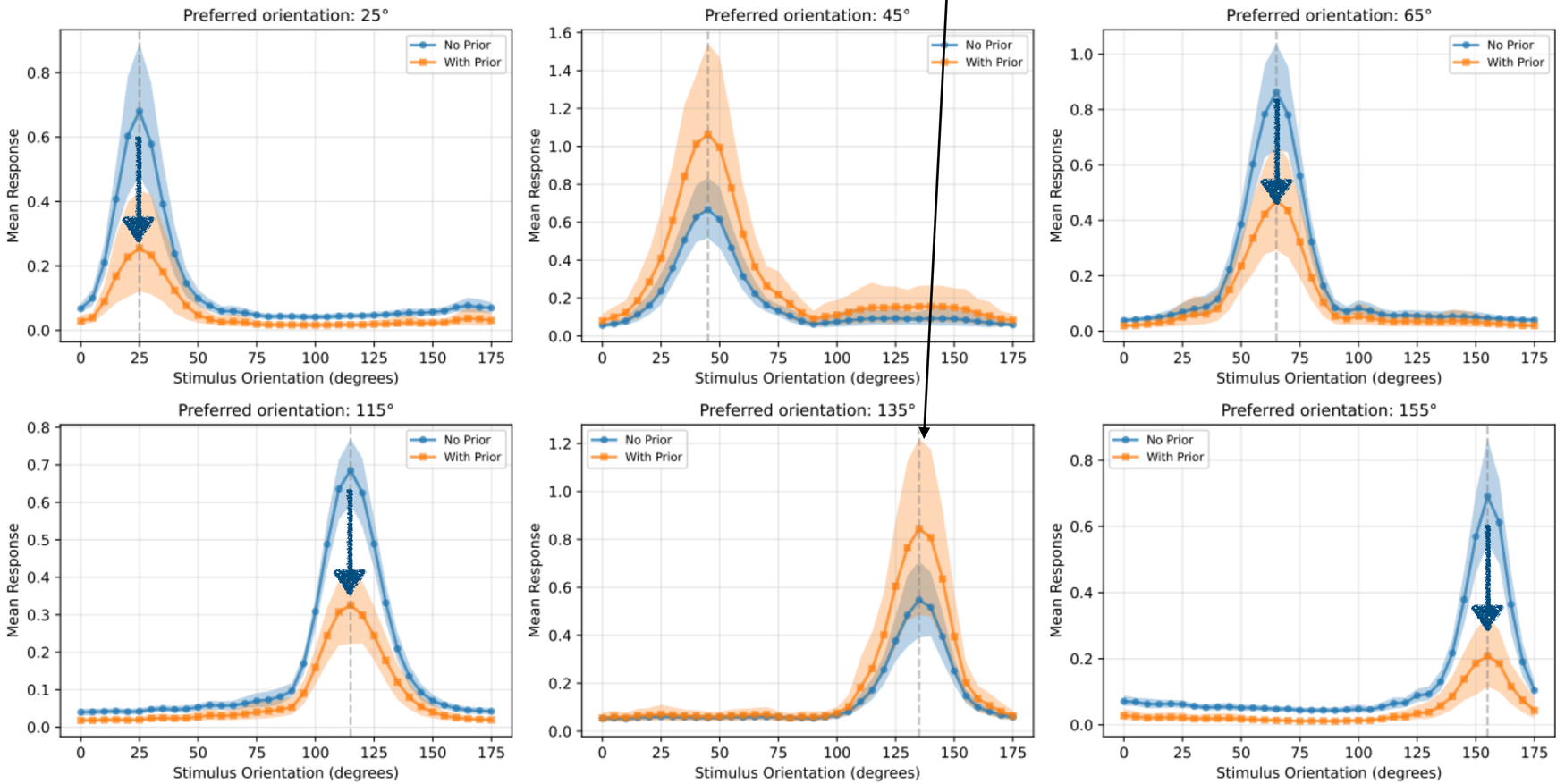
Gain modulation of tuning curves

Task prior modulates gain without preferred orientation



Prior orientations: (45, 135)

Left flanking orientations Aggregated Mean Tuning Curves by Orientation Bin **Right flanking orientations**



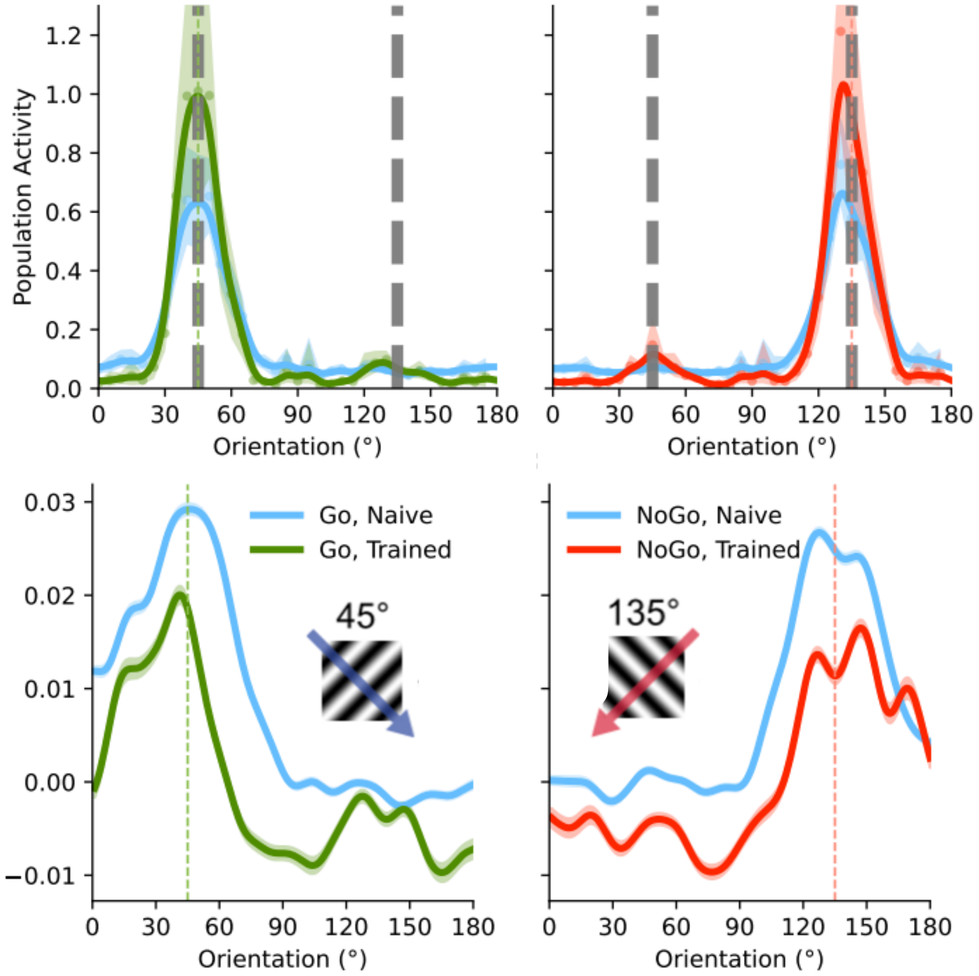
Tuning curves for individual neurons!

D1: task congruent stimuli

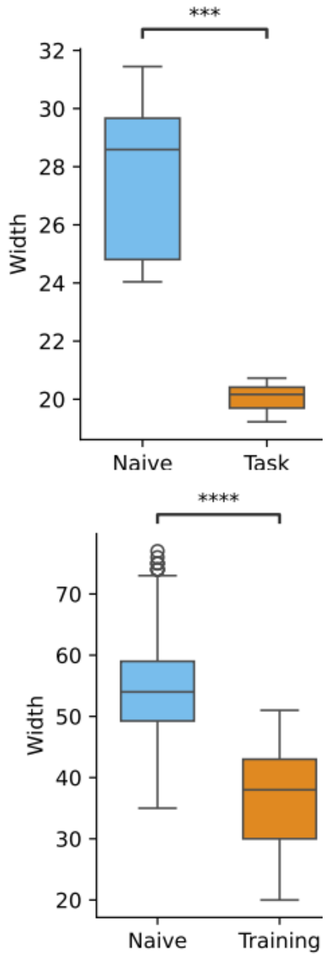
Population activity profiles: mean response of neurons grouped by preferred orientation

Model

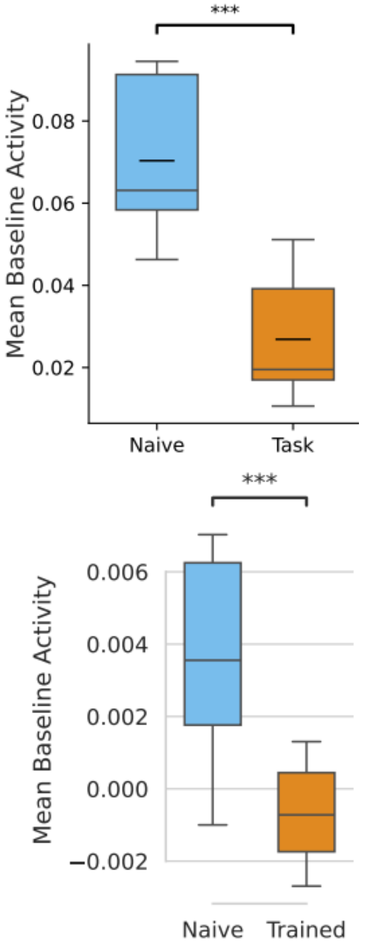
Experiment

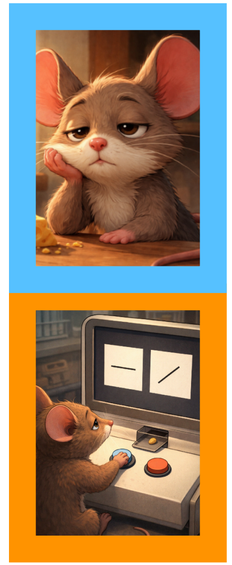


Sharpening of profiles

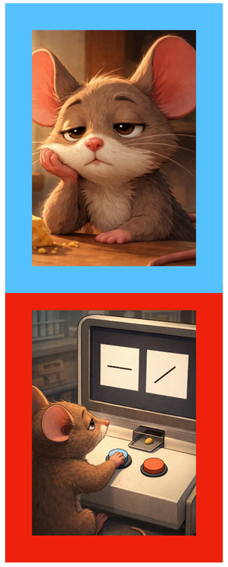


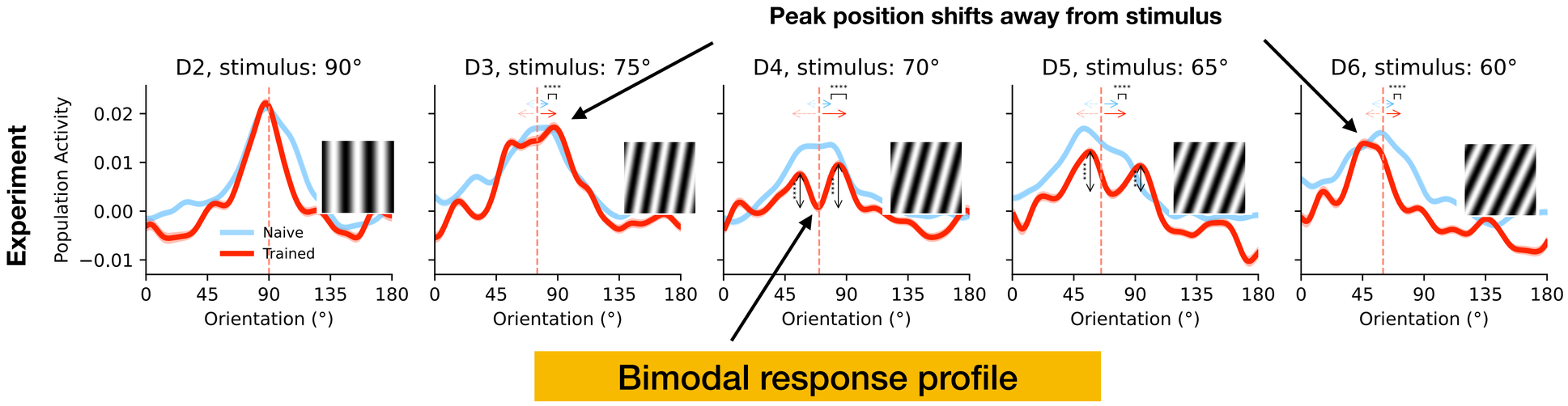
Baseline reduction



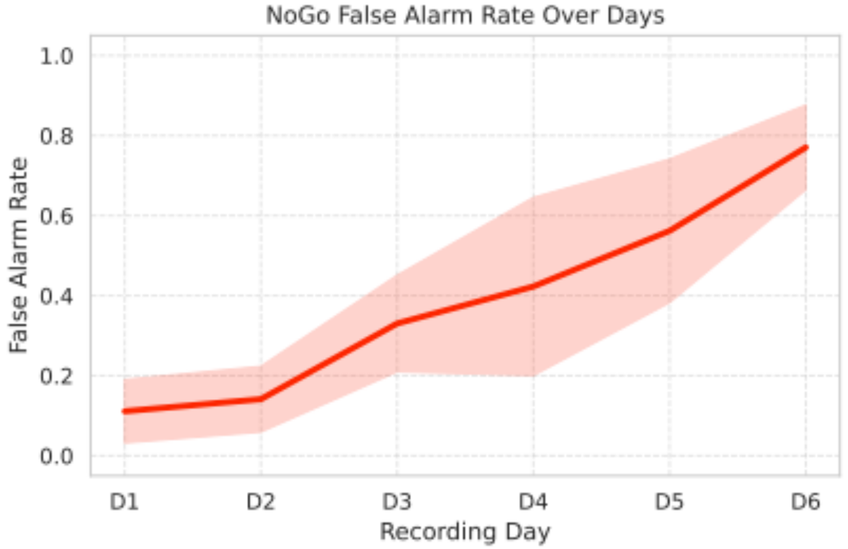


D2-6: task-stimulus mismatch





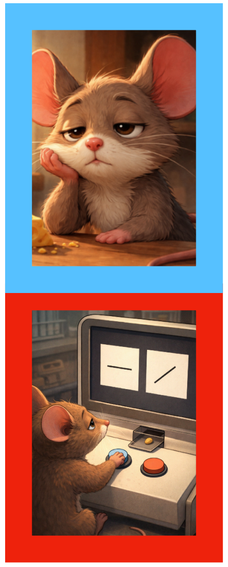
Bimodal response profile

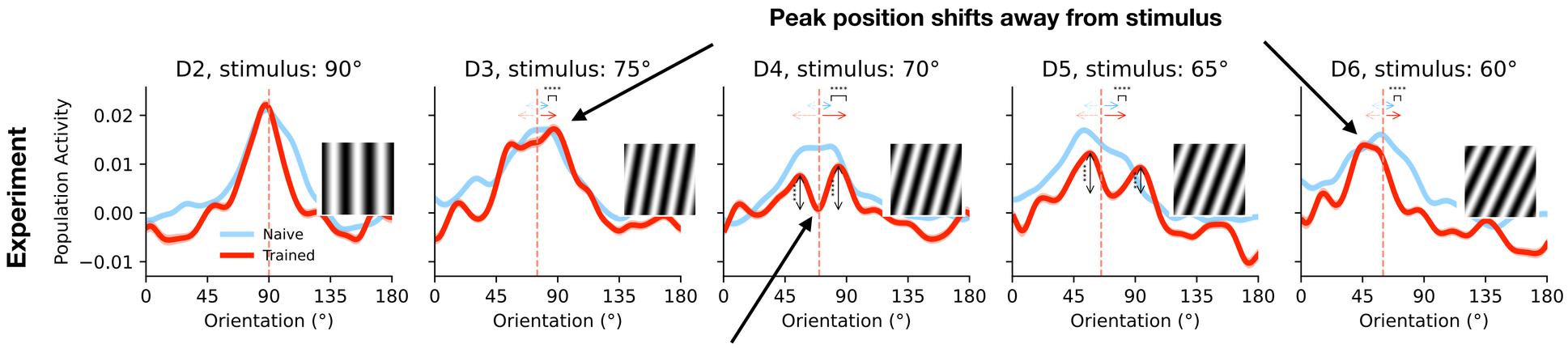


Three hypothesis

- **No adapter: retains (45, 135) prior angles across all sessions**
- **Eager adapter: updates to the current NoGO stimulus every day**
- **Partial adapter: updates on D2 to (45, 90), then retains this prior**

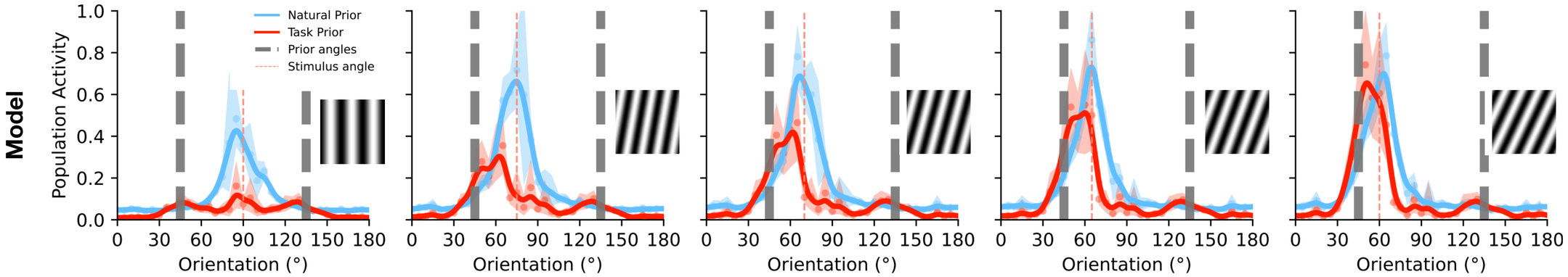
D2-6: task-stimulus mismatch





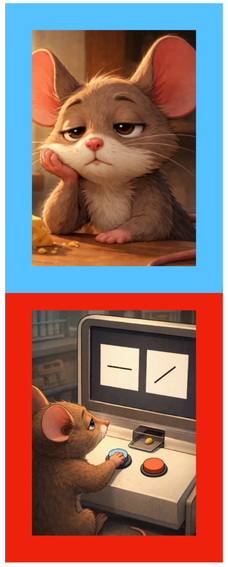
Bimodal response profile

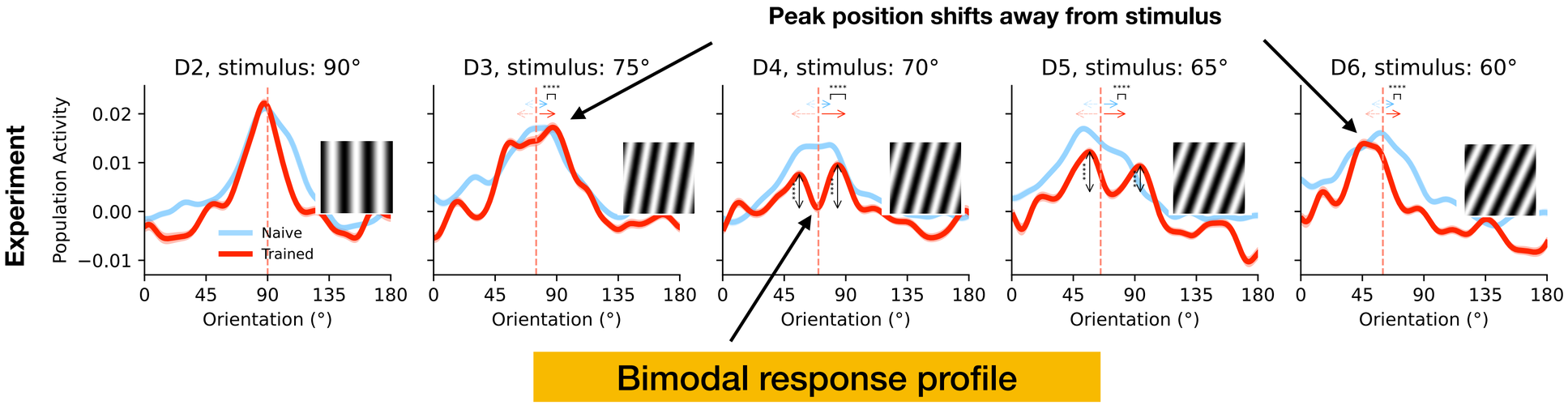
- No adapter: retains (45, 135) prior angles across all sessions



Poor match

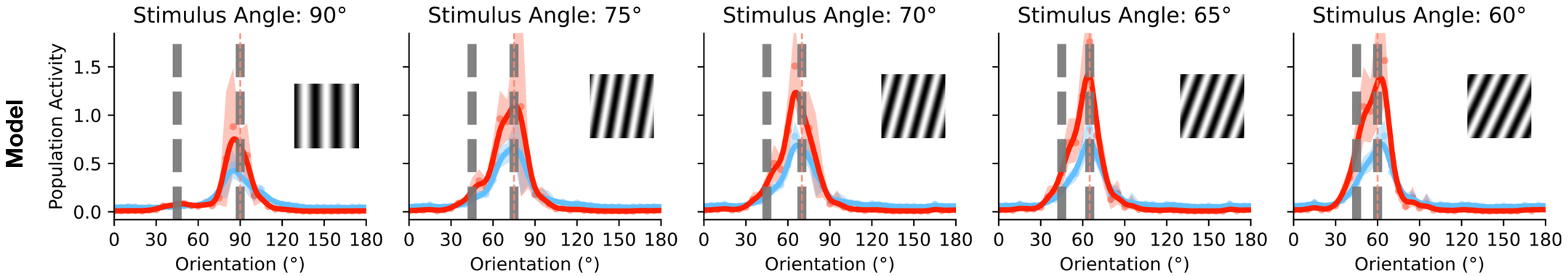
D2-6: task-stimulus mismatch





Bimodal response profile

- Eager adapter: updates to the current NoGO stimulus every day

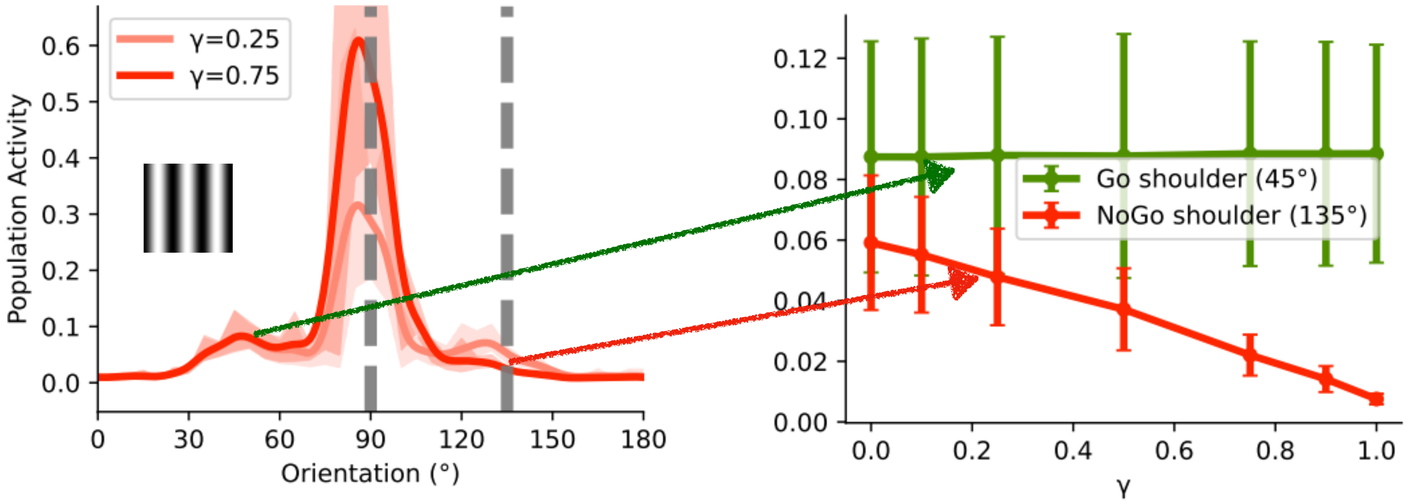


Poor match

D2 intraday adaptation

Model

Trained with NoGo gratings of 90(γ) and 135 (1-γ)



Experiment

Split D2 sessions into quintiles: Q1-5

