DFT Tutorial – Part 3
Hybrid models

ICANN 2025

Raul Grieben, Minseok Kang and Stephan Sehring

Raul.Grieben@ini.rub.de

DFT vs. ML

- Traditional ML:
  - Focus on static input-output mappings
    - classification, regression, generation, …
- DFT:
  - Focus on neural dynamics that drive autonomous sequences
  - Addresses problems ML struggles with:
    - stability, continual adaptation, grounding, …
  - Useful for:
    - Robotics & embodied AI
    - Continuous control/online learning tasks
    - Adaptive systems that must stay stable under perturbation

	Typical Deep Learning XAI	**DFT Models**
Interpretation level	Post-hoc (e.g., Grad-CAM, SHAP)	Intrinsic, at all time steps
Representation	Latent, often abstract	Explicit, continuous feature maps
Dynamics	Implicit (via layers)	Explicit dynamical system equations
Transparency	Low (millions of params)	High (few interpretable params)
Causality	Hard to infer	Built-in via dynamical equations
Focus	Explaining predictions	Explaining neural processes underlying predictions

DFT: Interpretability

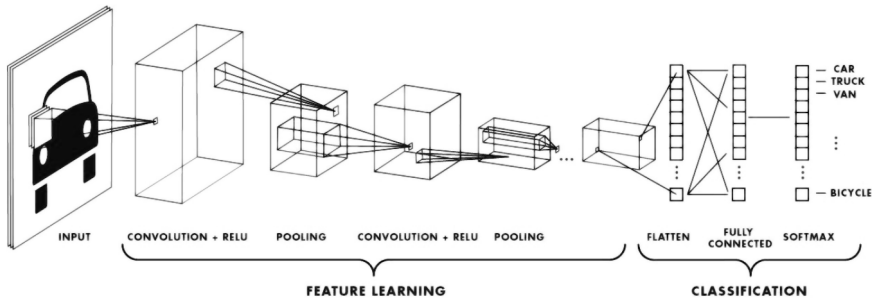
- DFT models are inherently interpretable because:
  - **Internal state = Activation**: Explicitly shows what and where the system is attending, what is in memory, what selection decision was made, …
  - **Attractors = Decisions or Memory States**: Easy to visualize as peaks
  - **Instabilities = Events**: Detectable as transitions in activation patterns
- The entire computation is a **transparent dynamical system**, not a black box

DFT: Explainability

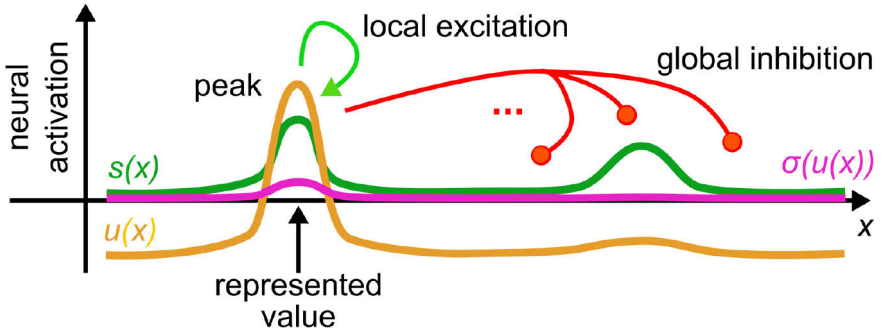
- Mechanistic Transparency:
  - Every prediction or decision is the result of explicit equations governing neural field dynamics
  - Peaks in neural fields directly represent what the system attends to, remembers, or selects
- Process-Level Explanation:
  - Explains how outcomes arise, not just what the prediction is
  - State trajectories show step-by-step causal transitions
- Explicit Architecture:
  - Kernels and connectivity patterns map to neural principles
  - Feature dimensions are explicit

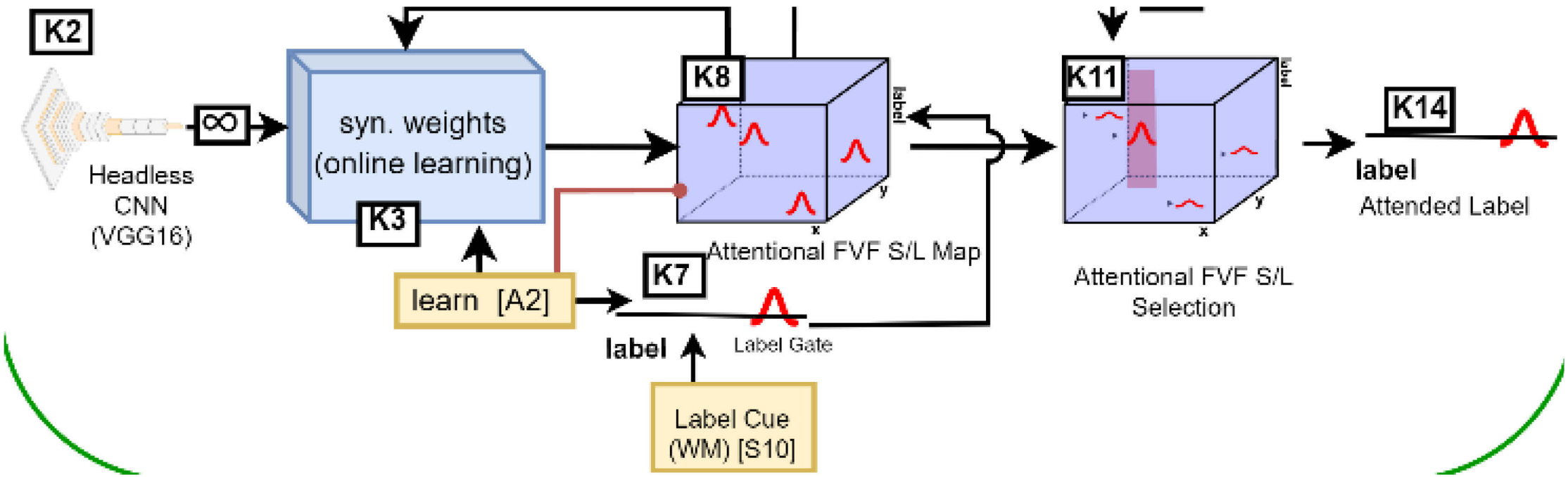
DFT: Causality

- Behavior emerges causally from explicit dynamical equations that define how every state evolves
- Every state change is the outcome of inputs and recurrent dynamics
- Each decision can be traced exactly to the fields, inputs, instabilities that caused it


CAR
TRUCK
VAN
BICYCLE
INPUT
CONVOLUTION + RELU
POOLING
CONVOLUTION + RELU
POOLING
FLATTEN
FULLY CONNECTED
SOFTMAX
FEATURE LEARNING
CLASSIFICATION

interface


neural activation
local excitation
peak
global inhibition
s(x)
σ(u(x))
u(x)
x
represented value

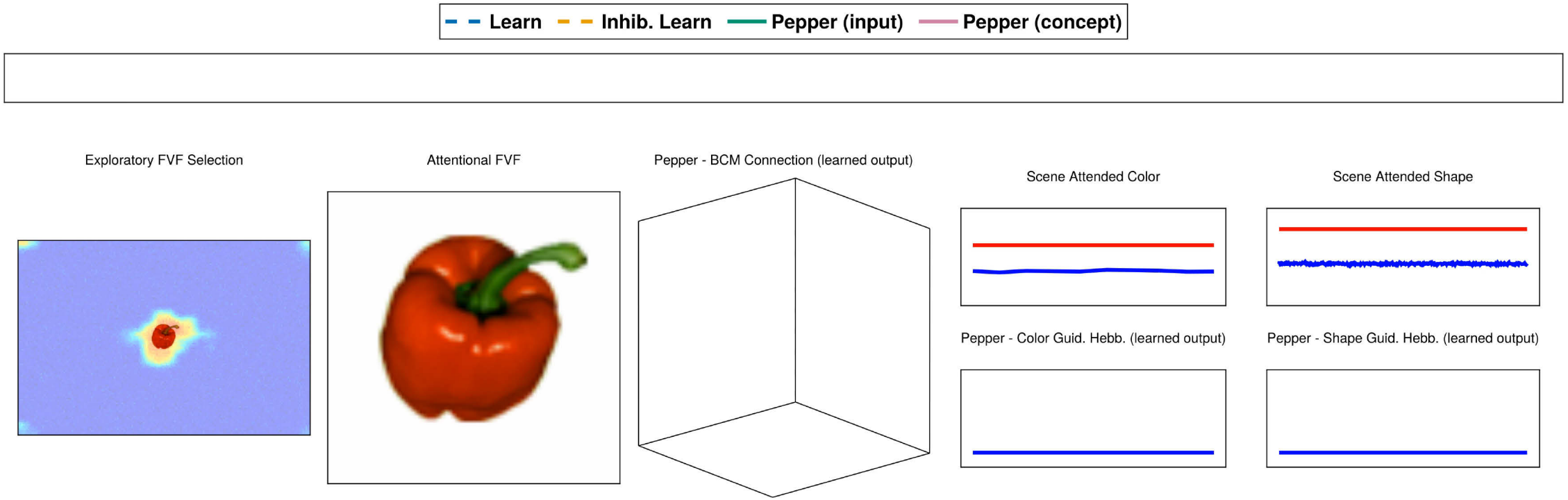


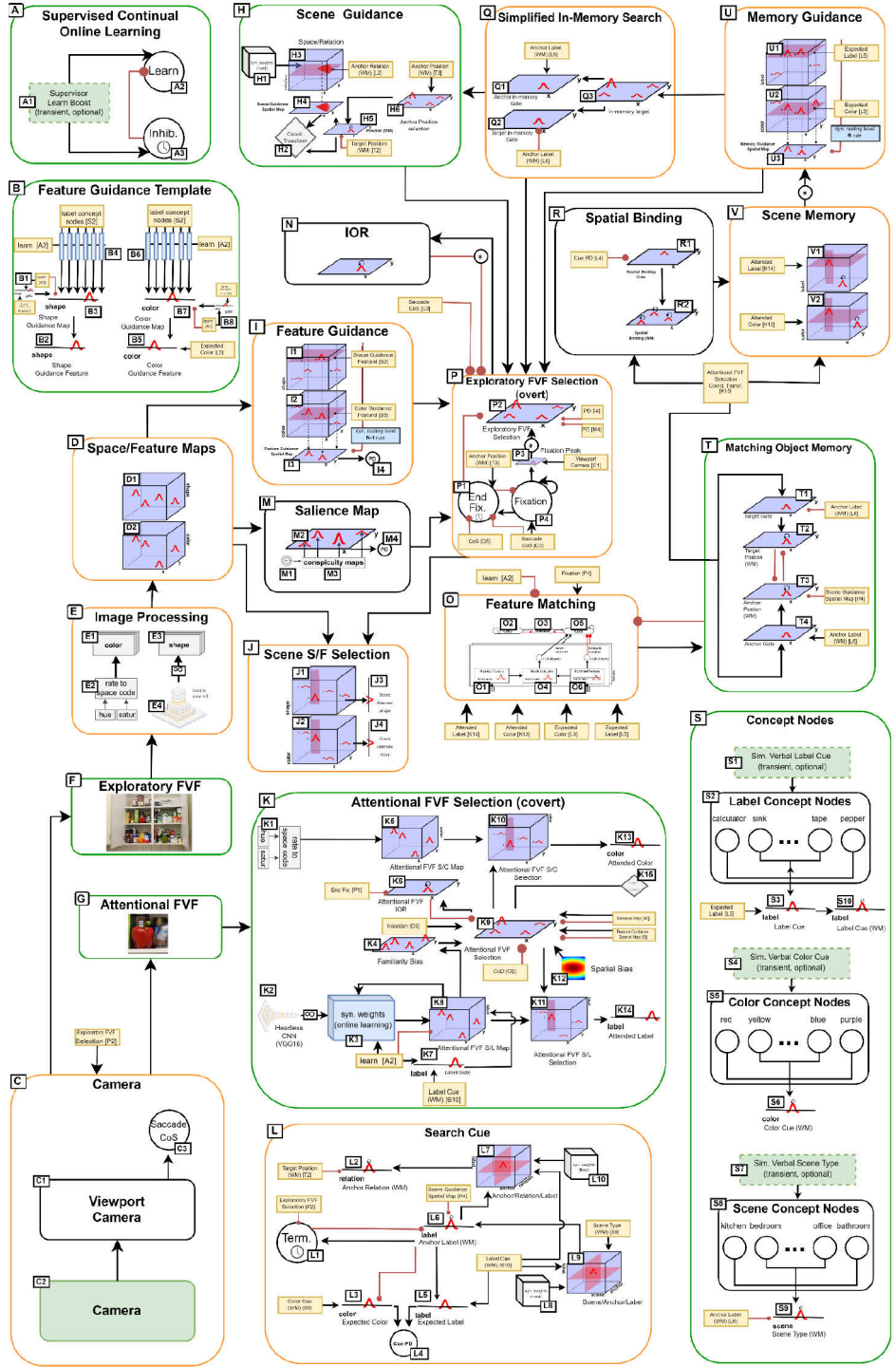
$$\tau_w \dot{w}_{m_\mathrm{f},u_\mathrm{fsml}}(\boldsymbol{x},t) = \eta\ \sigma(u_\mathrm{learn})\ y\ (y-\Theta)\ \frac{m_\mathrm{f}(x_1,x_2,t)}{\Theta}$$

$$y = \sigma(u_\mathrm{fsml}(\boldsymbol{x},t))$$

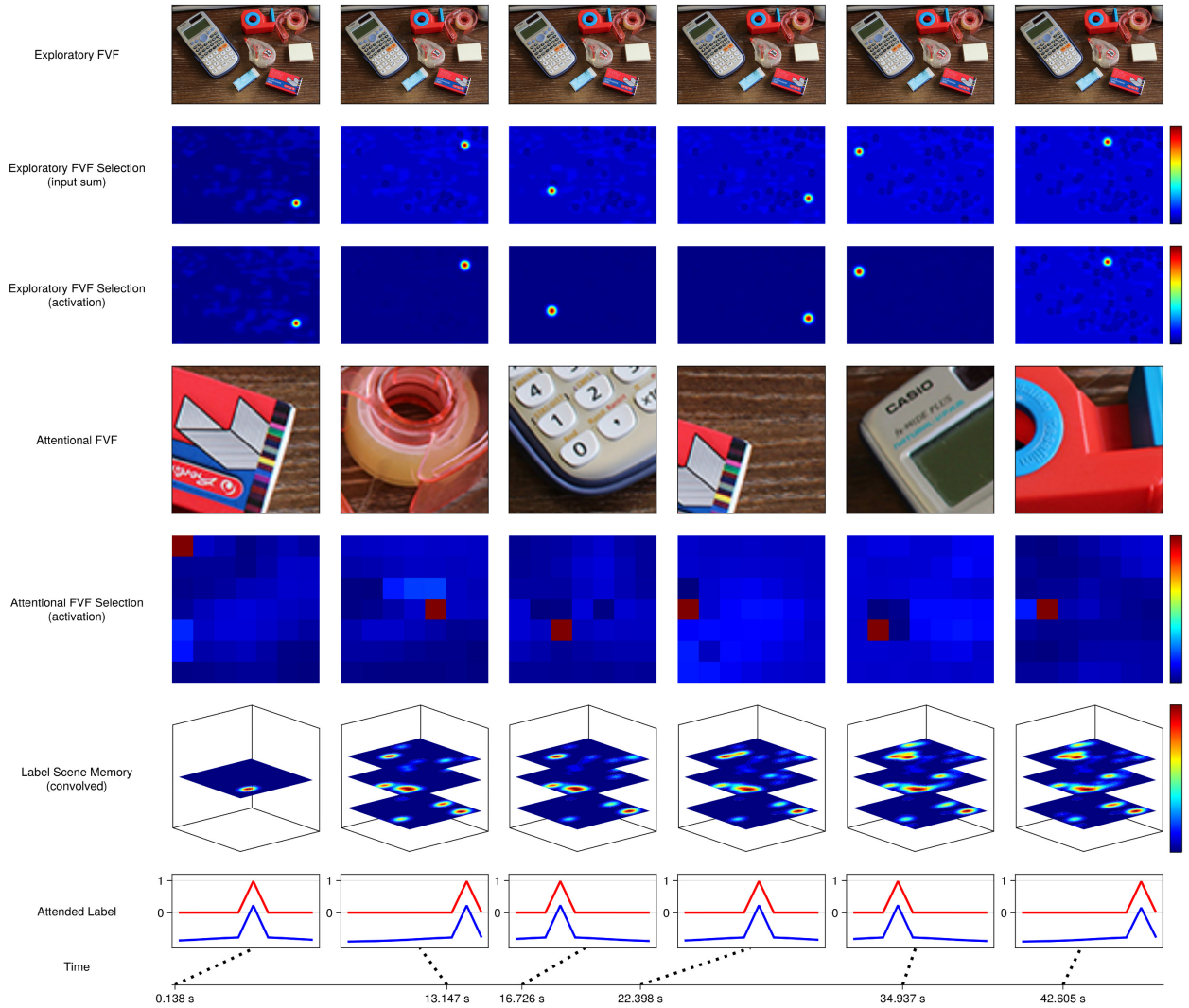
$$\tau_\Theta \dot{\Theta} = (y^2 - \Theta),$$

Learning

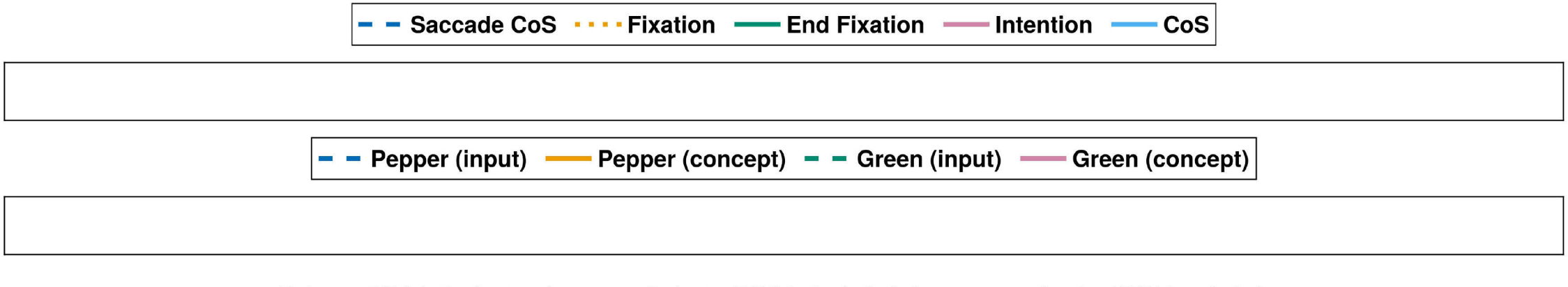

Learn
Inhib. Learn
Pepper (input)
Pepper (concept)
Exploratory FVF Selection
Attentional FVF
Pepper - BCM Connection (learned output)
Scene Attended Color
Scene Attended Shape
Pepper - Color Guid. Hebb. (learned output)
Pepper - Shape Guid. Hebb. (learned output)


A
Supervised Continual Online Learning
Learn
Supervisor Learn Boost (transient, optional)
Inhib.
H
Scene Guidance
Space/Relation
Q
Simplified In-Memory Search
U
Memory Guidance
B
Feature Guidance Template
label concept nodes [S2]
learn [A2]
shape
Shape Guidance Map
Shape Guidance Feature
color
Color Guidance Map
Color Guidance Feature
Expected Color [L3]
N
IOR
R
Spatial Binding
V
Scene Memory
I
Feature Guidance
P
Exploratory FVF Selection (overt)
Saccade CoS [C3]
Attended FVF Selection Coord. Transl. [M10]
D
Space/Feature Maps
M
Salience Map
conspicuity maps
T
Matching Object Memory
E
Image Processing
color
shape
rate to space code
hue
saturation
J
Scene S/F Selection
O
Feature Matching
Attended Label [K10]
Attended Color [K13]
Expected Color [L3]
Expected Label [L7]
S
Concept Nodes
Sim. Verbal Label Cue (transient, optional)
Label Concept Nodes
calculator
sink
tape
peeper
Label Cue
Label Cue (WM)
Sim. Verbal Color Cue (transient, optional)
Color Concept Nodes
red
yellow
blue
purple
Color Cue (WM)
Sim. Verbal Scene Type (transient, optional)
Scene Concept Nodes
kitchen
bedroom
office
bathroom
Scene Type (WM)
F
Exploratory FVF
G
Attentional FVF
K
Attentional FVF Selection (covert)
Attentional FVF S/C Map
Attentional FVF S/C Selection
Attended Color
Attentional FVF IOR
Familiarity Bias
Attentional FVF Selection
Spatial Bias
syn. weights (online learning)
Attentional FVF S/L Map
Attentional FVF S/L Selection
Attended Label
learn [A2]
Label Cue (WM) [S10]
C
Camera
Saccade CoS
Viewport Camera
Camera
L
Search Cue
Target Position (WM) [T2]
relation
Anchor Relation (WM)
Anchor/Relation/Label
Term.
label
Anchor Label (WM)
Label Cue (WM) [S10]
Scene Type (WM) [S9]
Scene/Anchor/Label
color
Expected Color
label
Expected Label

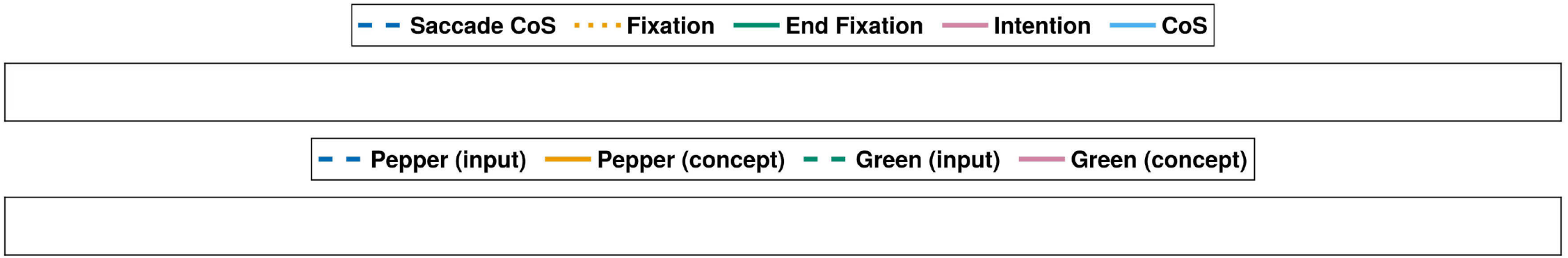
Exploration



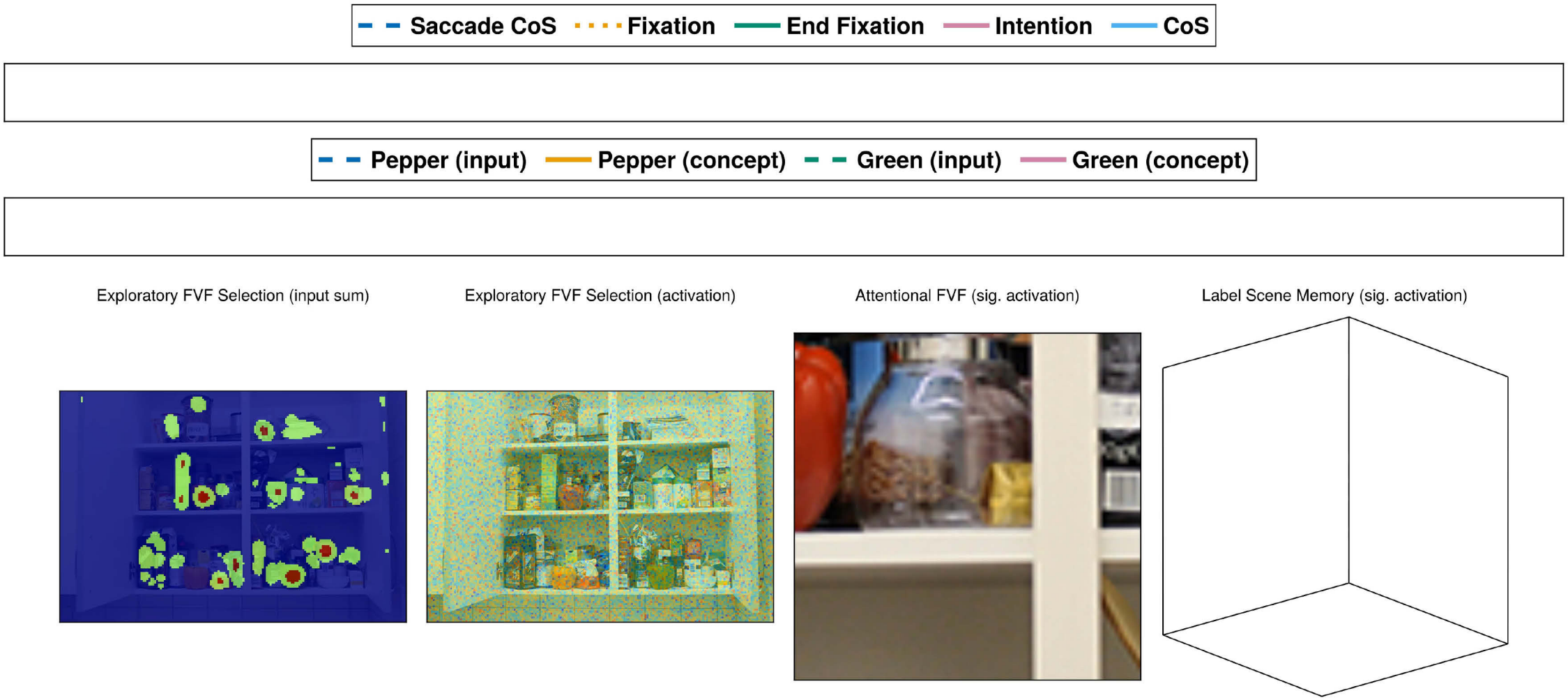
Categorical visual search



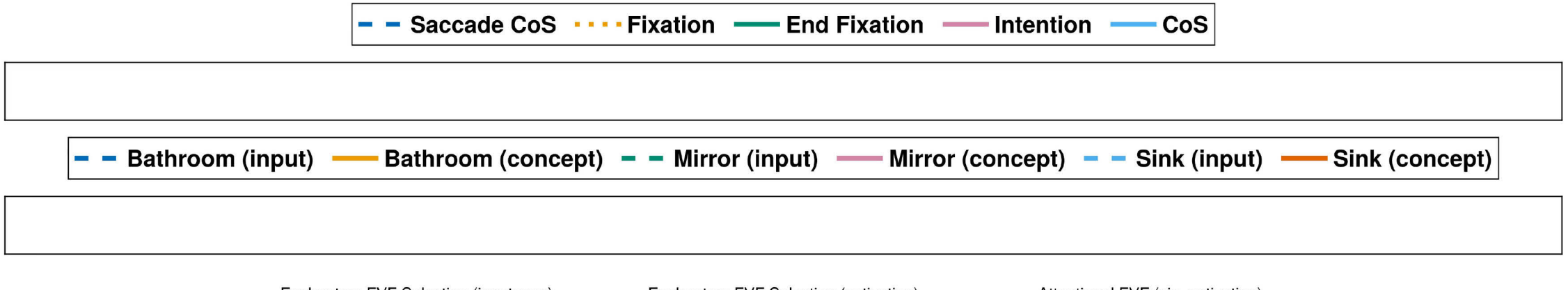
Combined categorical and feature search



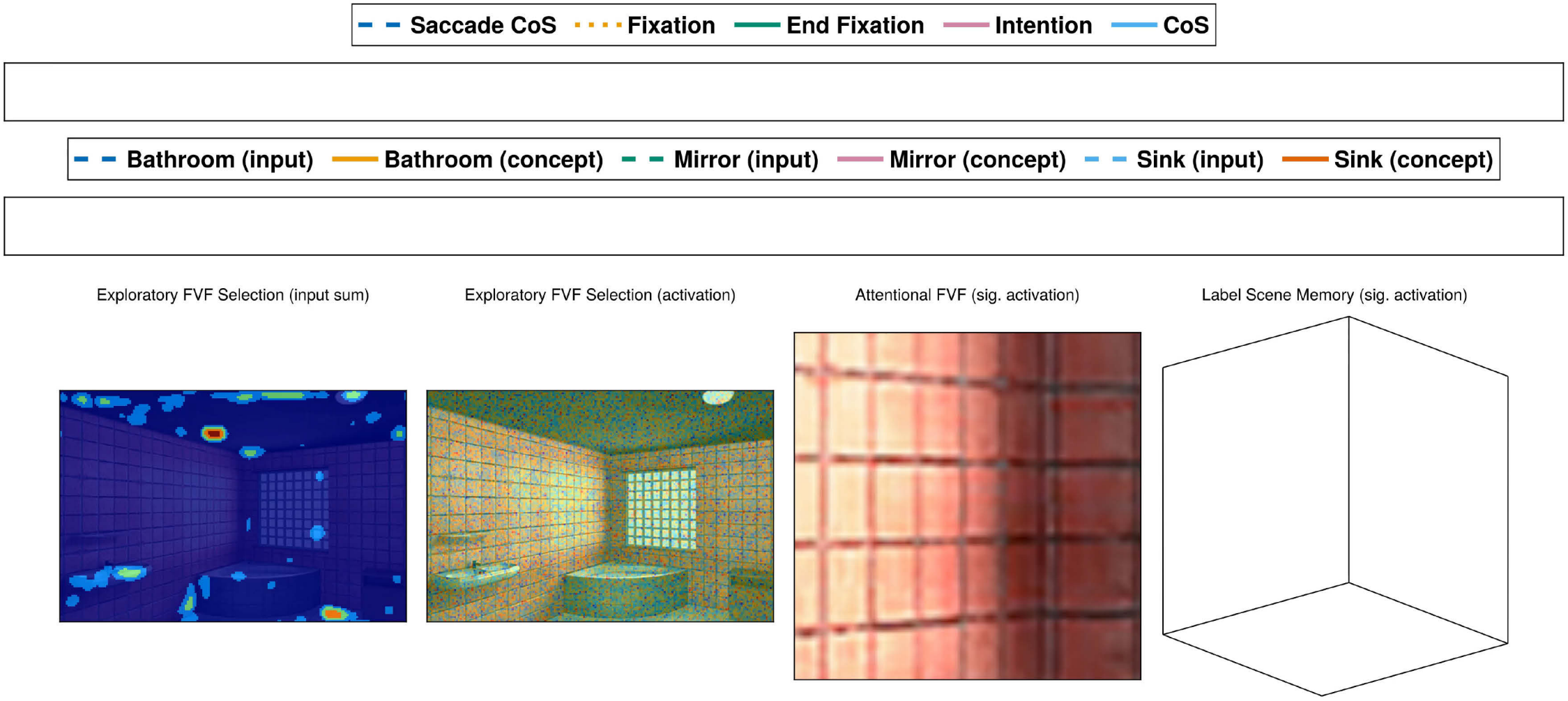
Memory guidance



Scene grammar



Scene grammar – Memory guidance



ROBOVERINE: A human-inspired neural robotic process model of active visual search and scene grammar in naturalistic environments

Raul Grieben[1], Stephan Sehring[1], Jan Tekülve[1], John P. Spencer[2] and Gregor Schöner[1]

***Abstract*— We present ROBOVERINE, a neural dynamic robotic active vision process model of selective visual attention and scene grammar in naturalistic environments. The model addresses significant challenges for cognitive robotic models of visual attention: combined bottom-up salience and top-down feature guidance, combined overt and covert attention, coordinate transformations, two forms of inhibition of return, finding objects outside of the camera frame, integrated space- and object-based analysis, minimally supervised few-shot continuous online learning for recognition and guidance templates, and autonomous switching between exploration and visual search. Furthermore, it incorporates a neural process account of scene grammar — prior knowledge about the relation between objects in the scene — to reduce the search space and increase search efficiency. The model also showcases the strength of bridging two frameworks: Deep Neural Networks for feature extractions and Dynamic Field Theory for cognitive operations.**

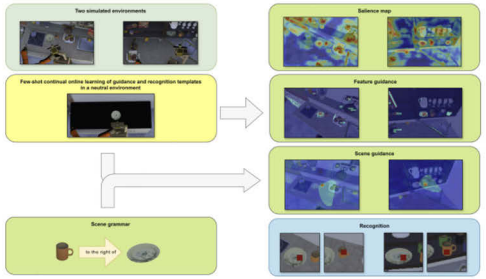


Fig. 1. A simplified overview of the problem (left) and the cognitive operations needed to solve it (right).

I. INTRODUCTION

Most goal-oriented interactions with the environment entail a preceding visual search. Effective feature guidance [1] helps reduce the number of saccades needed to find the target object in a scene, and the combination of overt and covert attention shifts [2] allows us to scan complex scenes efficiently despite the visual system's limitations. Natural scenes tend to be cluttered but highly structured, and humans use their knowledge about the relation between objects in scenes - the scene grammar [3] - to reduce the search space. Importantly, humans are not limited to finding objects they already know. Cognitive robotics aims to develop autonomous agents with cognitive abilities similar to humans (see [4] for a recent overview of the state-of-the-art in human-inspired robotic vision). Begum and Karray [5]

3) Learn an object's visual features with minimal human supervision from different view angles.
4) Autonomous switching between exploration and visual search based on the task.
5) Learning while working, without needing a separate training phase (online learning).

Here, we present ROBOVERINE, a neural robotic process model that addresses these issues (Figure 1) building upon our previous work on human attention ([6], [7]). We show that it can control an autonomous agent in different simulated environments. Furthermore, we also included a neural process account of scene grammar (see [8] for a related approach). Interfacing the neural architecture based on Dynamic Field Theory (DFT; [9]) with a pre-trained headless convolutional neural network (CNN: VGG16: [10])

Grieben, R., Sehring, S., Tekülve, J., Spencer, J. P., & Schöner, G.

IROS 2024

A Neural Dynamic Model Autonomously Drives a Robot to Perform Structured Sequences of Action Intentions

Stephan Sehring (stephan.sehring@ini.rub.de)
Richard Koebe (richard.koebe@ini.rub.de)
Sophie Aerdker (sophie.aerdker@rub.de)
Gregor Schöner (gregor.schoner@ini.rub.de)

Institute for Neural Computation, Ruhr-Universität Bochum, 44780 Bochum, Germany

Abstract

We present a neural dynamic process model of an intentional agent that carries out compositionally structured action plans in a simulated robotic environment. The model is inspired by proposals for a shared neural and structural basis of language and action (Pastra & Aloimonos, 2012). Building on neural process accounts of intentionality we propose a neural representation of the conceptual structure of actions at a symbolic level. The conceptual structure binds actions to objects at which they are directed. In addition, it captures the compositional structure of action sequences in an action plan by representing sequential order between elementary actions. We show how such a neural system can steer motor behavior toward objects by forming neural attractor states that interface with lower-level motor representations, perceptual systems and scene working memory. Selection decisions in the conceptual structure enables the generation of action sequences that adheres to a memorized action plan.

Keywords: neural process model; dynamic field theory; action grammar; intention; action and language; autonomous robot

Introduction

Following instructions, or planning actions ourselves to reach goals often requires that we generate novel sequences of actions that we never before performed in exactly the same order or directed at precisely the same objects. The human faculty for intentional action comprises this remarkable ability to form a practically unlimited set of novel actions by flexibly recombining previously learned motor behaviors. Even rather global goals may thus be ultimately achieved by combining the limited set of movements available to the human human action. It would enable an agent to represent novel action plans that generalize beyond any specific instances it may have learned or stored earlier.

How could a neural system implement such a representational system and how could such an implementation drive intentional action? To address these questions we propose a neural process account of intentional action that enables an agent to autonomously direct action at objects in its environment (Tekülve & Schöner, 2019). Two key problems are addressed. First, we propose a neural representation of the conceptual structure of an action at a symbolic level which binds the action to the objects at which it is directed (see the top panel of Figure 1 for an illustration). This makes use of earlier work on neural binding through a shared "index" dimension (Sabinasz, Richter, & Schöner, 2023). We show how this neural implementation of a structured representation may guide the embodied realization of the intentional action directed at objects. Second, we show how a neural representation of the sequential order of elementary actions in a "dependency graph" may capture the compositional structure of actions described in syntax trees. We demonstrate how this representation may steer sequences of actions toward achieving goals.

As a proof of concept, we present a neural dynamic process model that controls a simulated robot arm in a table-top environment that carries out pick and place actions. The model represents action intentions as *action phrases*, that is, conceptual structures that bind action concepts to object concepts in

Sehring, S., Koebe, R., Aerdker, S., & Schöner, G.

CogSci 2024