paper webpage

Generating metamers of human scene understanding

Ritik Raina[1]*, Abe Leite[1], Alexandros Graikos[1], Seoyoung Ahn[2], Dimitris Samaras[1], Gregory J. Zelinsky[1]

1 Stony Brook University 2 Berkeley UNIVERSITY OF CALIFORNIA

* correspondence to: `ritik.raina@stonybrook.edu`

ICLR 2026

Humans understand scenes by combining **low-resolution peripheral gist** with **high-resolution foveal information** from sparse viewing fixations.

How can we probe the properties of their internal scene representations?

Foundations of scene understanding

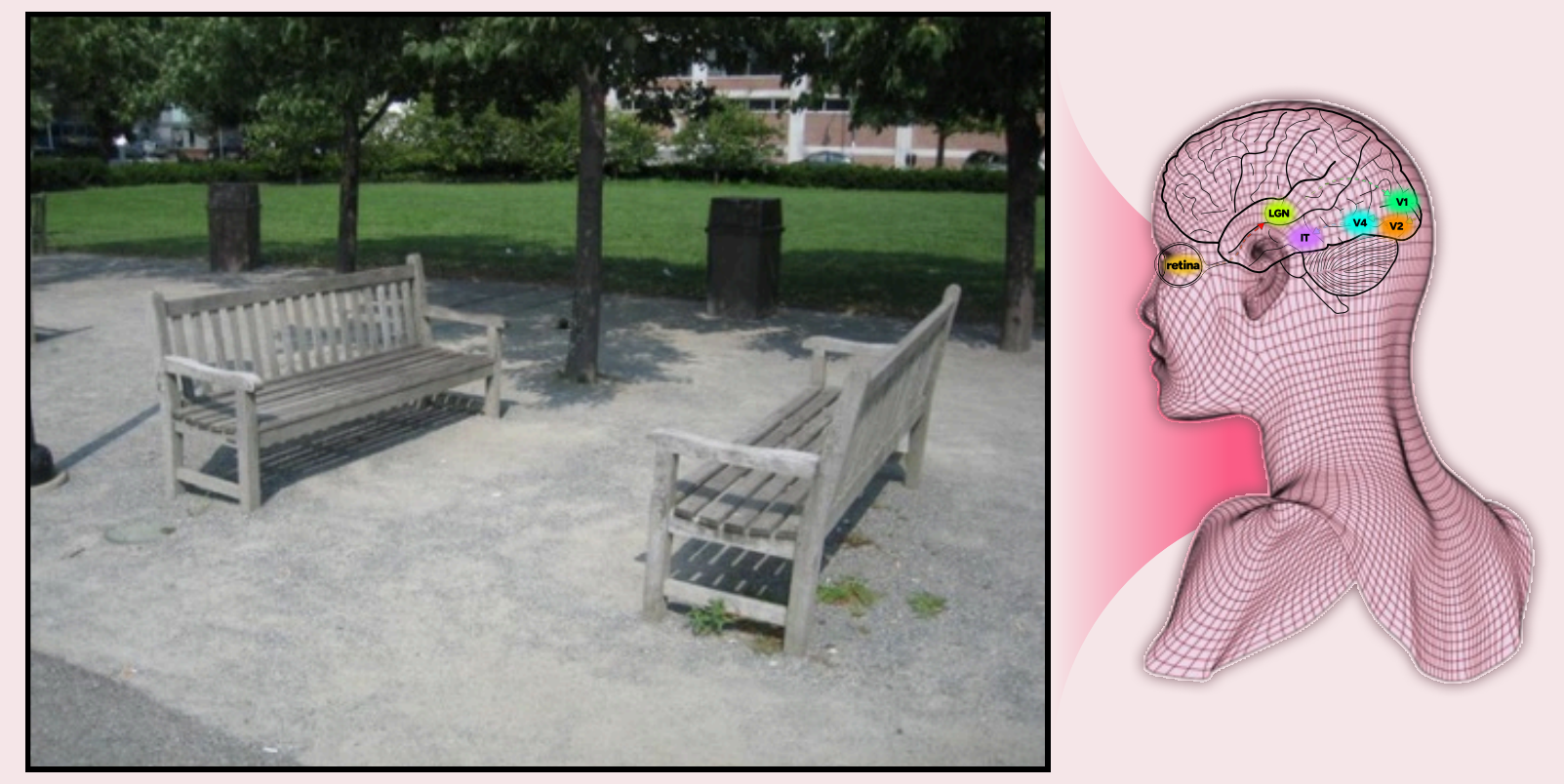
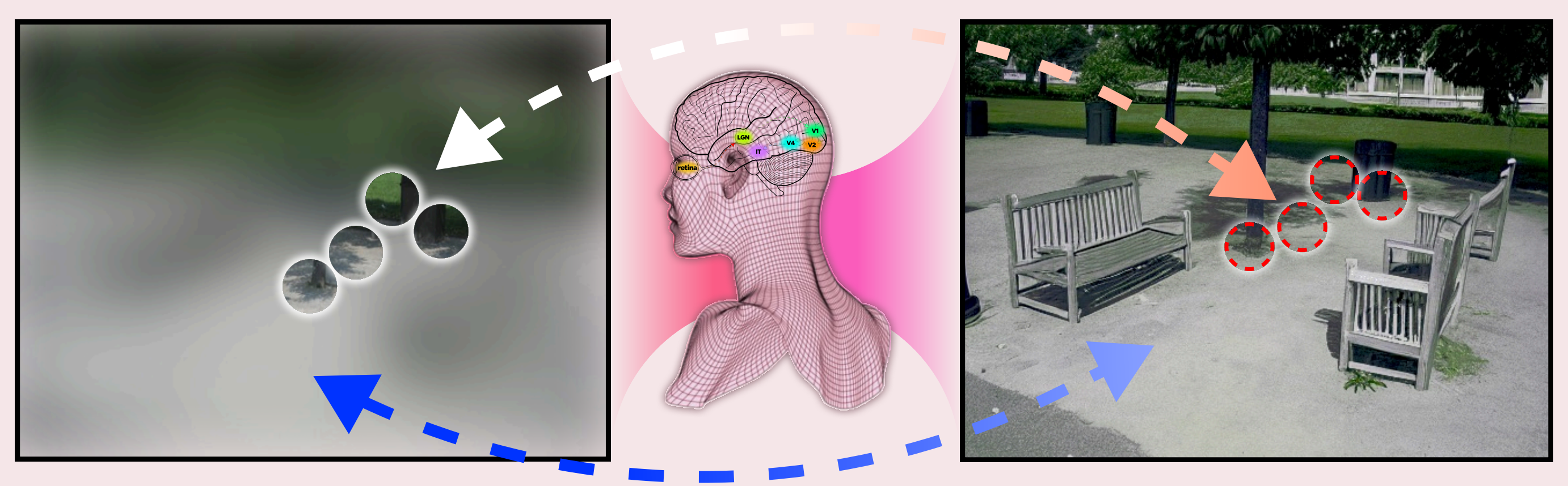


Image $\mathcal{I}$

- Within 100ms, humans extract scene gist with remarkable accuracy (Potter, 1975).
- Recognition of spatial properties form a spatial envelope (Greene & Oliva, 2009).
- Processing of object co-occurrence gives rapid scene context (Li et al., 2007).

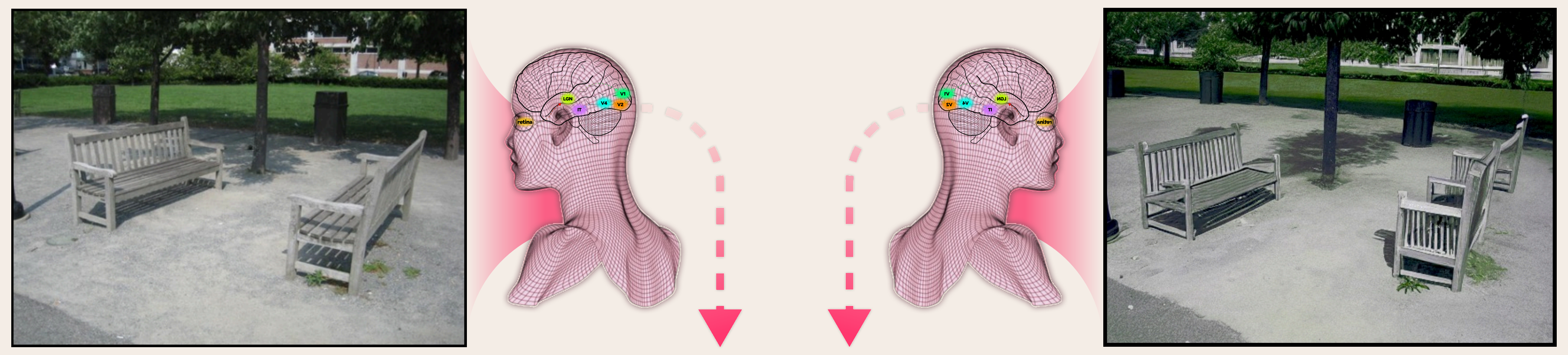


Fixated objects retained

$\mathcal{I}_{\text{foveated}}$ Peripheral objects estimated $\mathcal{I}_{\text{represented}}$

Where we look shapes our understanding of a scene.

Scene metamerism

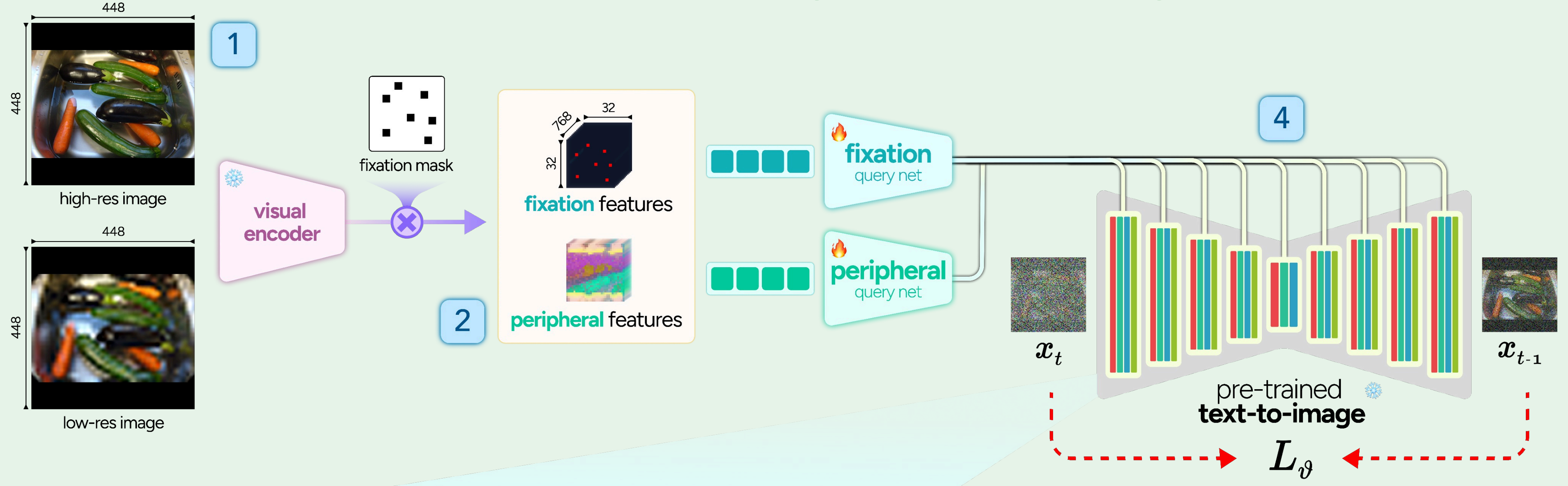


$\mathcal{I}_{\text{original}}$ **representation**$_{\text{original}}$ ⟷ **representation**$_{\text{plausible}}$ $\mathcal{I}_{\text{plausible}}$

physically different, understanding identical

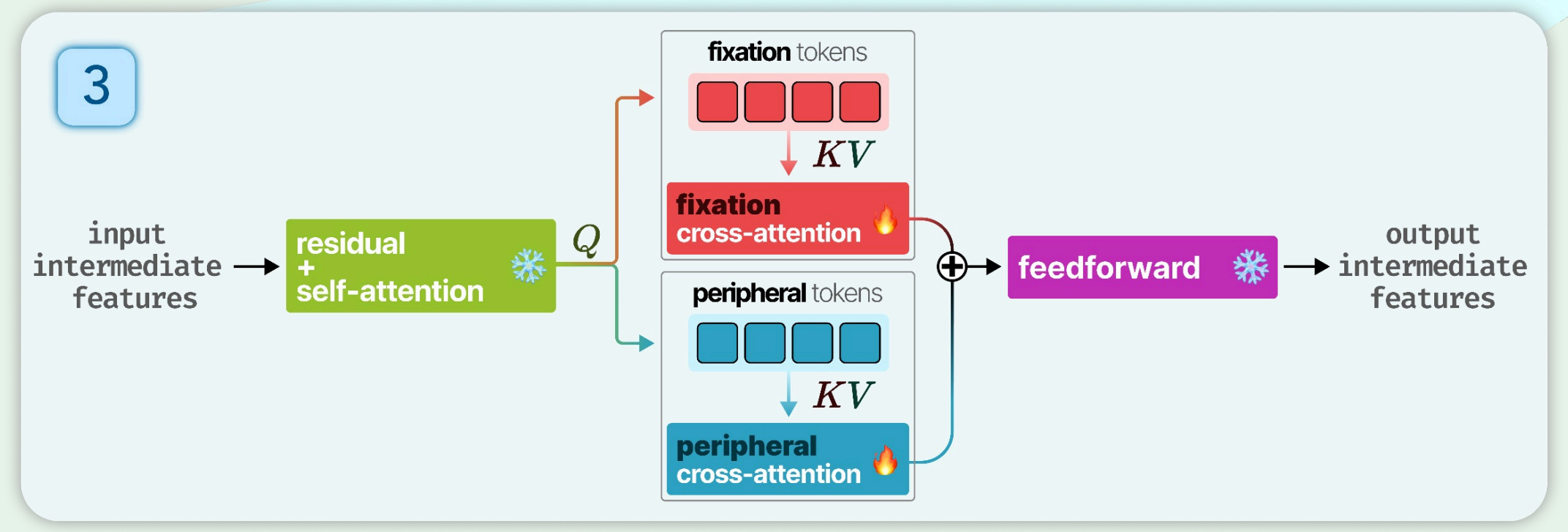
Can **metamerism** reveal the **features** underlying our **scene representations**?

MetamerGen — conditioning for metamer generation



448
high-res image
low-res image
1
visual encoder
fixation mask
32
768
fixation features
peripheral features
2
fixation query net
peripheral query net
4
x_t
x_{t-1}
pre-trained text-to-image
L_ϑ

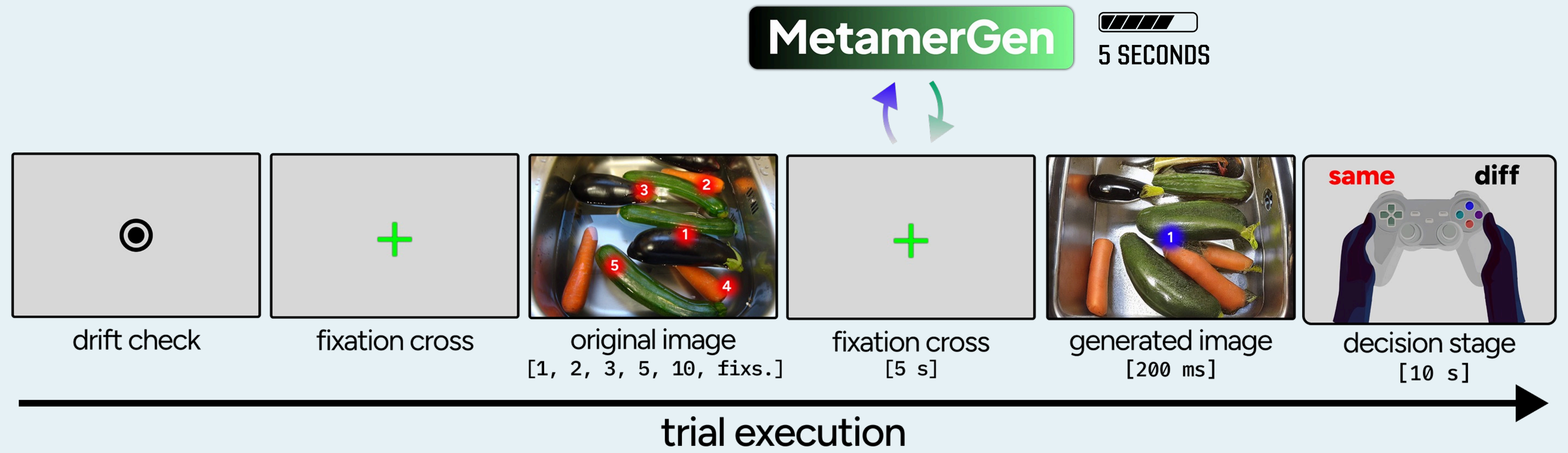
MetamerGen UNet block



3
input intermediate features
residual + self-attention
Q
fixation tokens
KV
fixation cross-attention
peripheral tokens
KV
peripheral cross-attention
feedforward
output intermediate features

1. Simulate periphery with blur and fixations by masking
2. Encode both visual streams using DINOv2
3. IP-Adapter to condition a T2I model with fixations
4. Synthesize a potential metamer image using both sources

Human-in-the-Loop metamerism experiment



MetamerGen 5 SECONDS

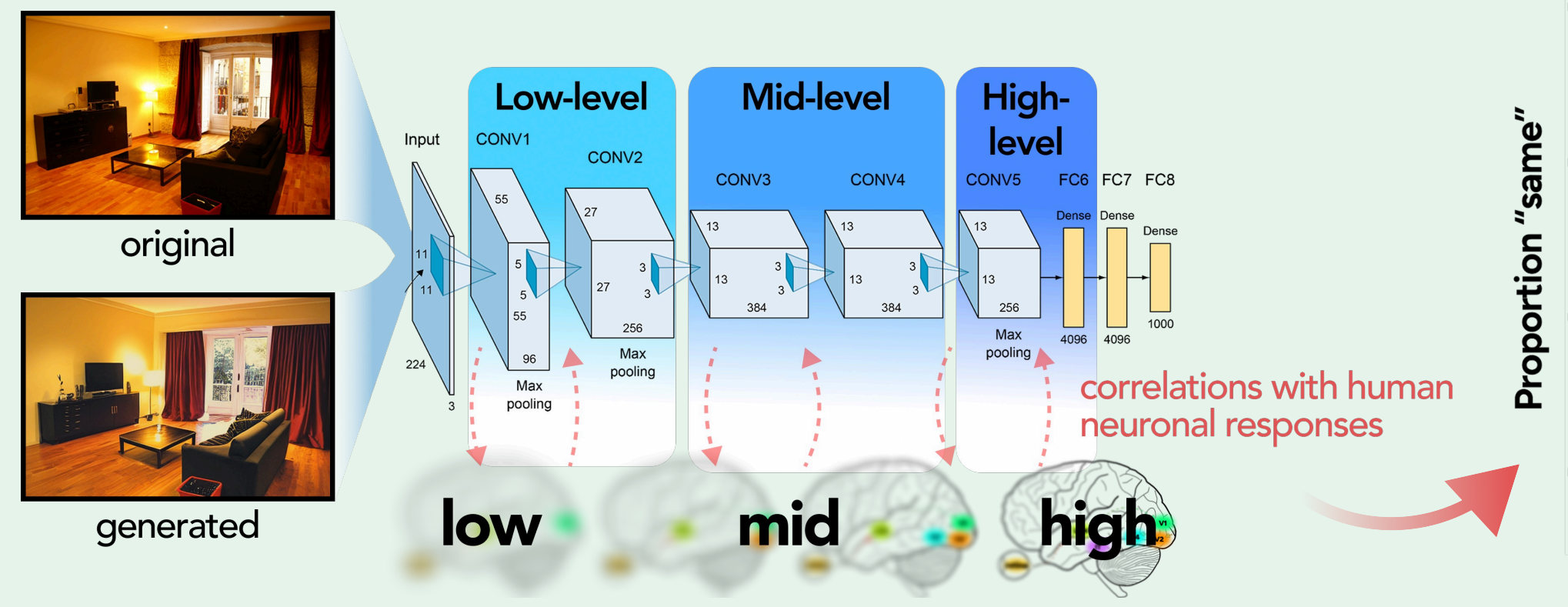
drift check | fixation cross | original image `[1, 2, 3, 5, 10, fixs.]` | fixation cross `[5 s]` | generated image `[200 ms]` | decision stage `[10 s]`

trial execution

- Behavioral same-different experiment **(N = 45)** where participants judged whether a generated image depicted the same scene as a briefly presented original.
- Generated images were conditioned on either the viewer's **own fixations** or **randomly sampled fixations** from other viewers on the same scene.
- **300 natural scenes** sampled from **Visual Genome** (Krishna et al., 2017).

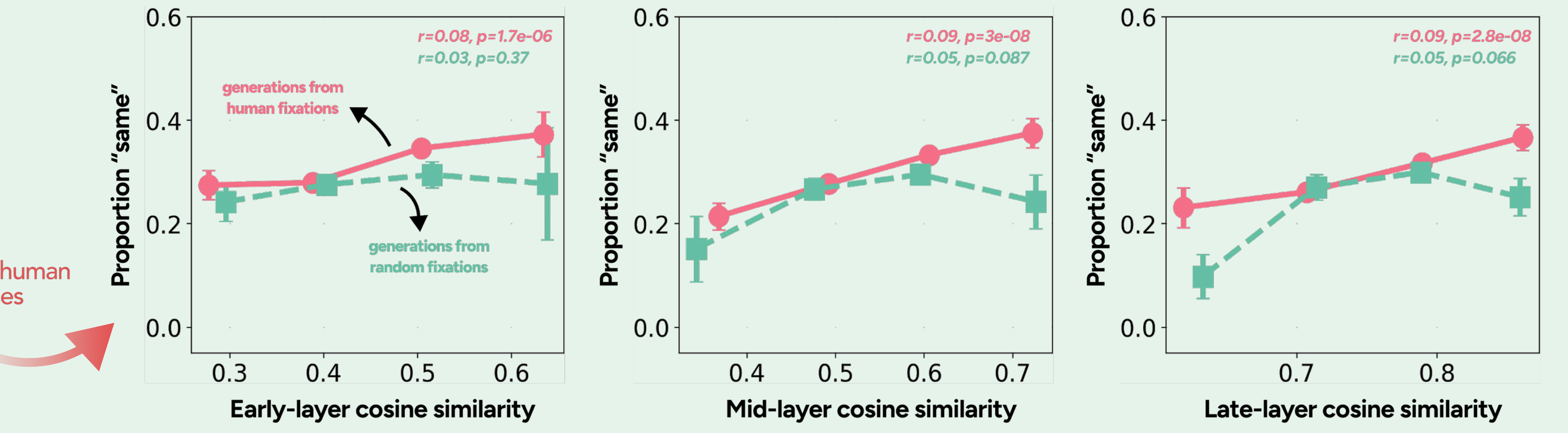
Probing the features of scene metamers

Neurally-grounded hierarchical feature similarity

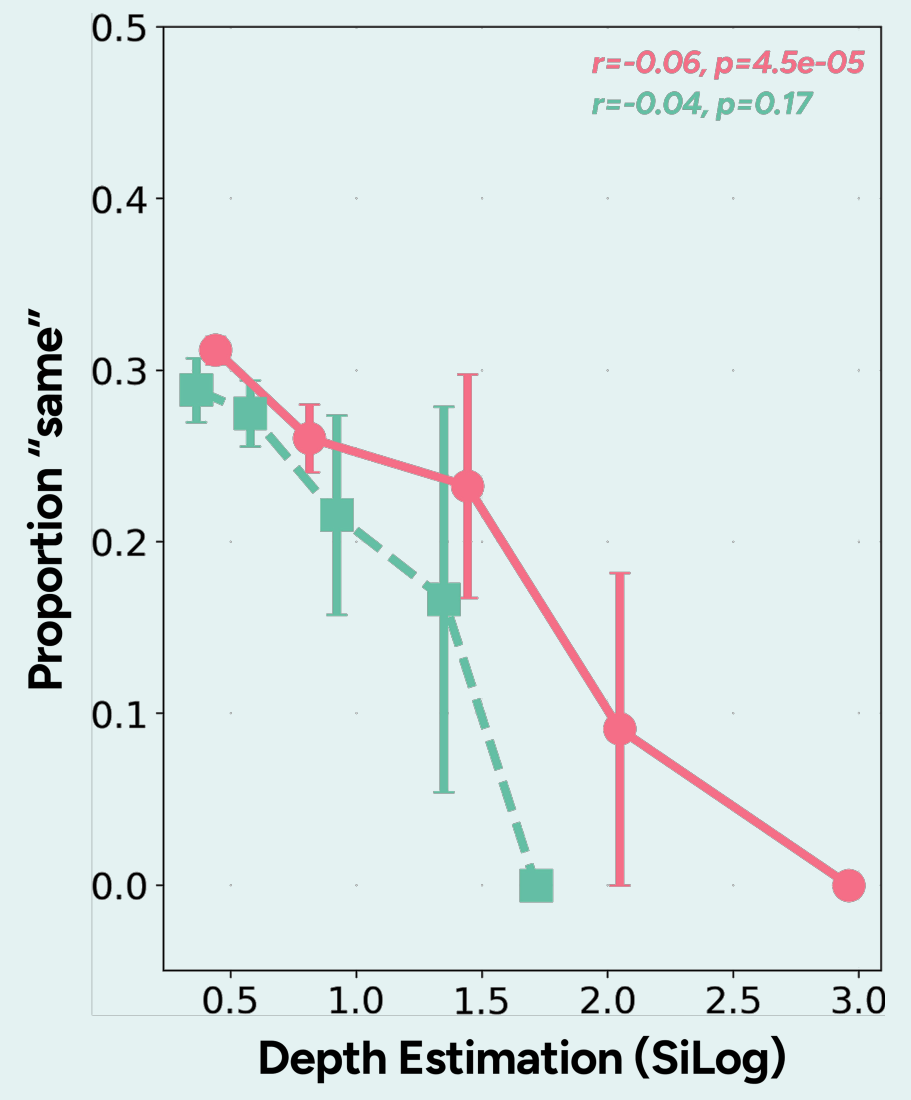


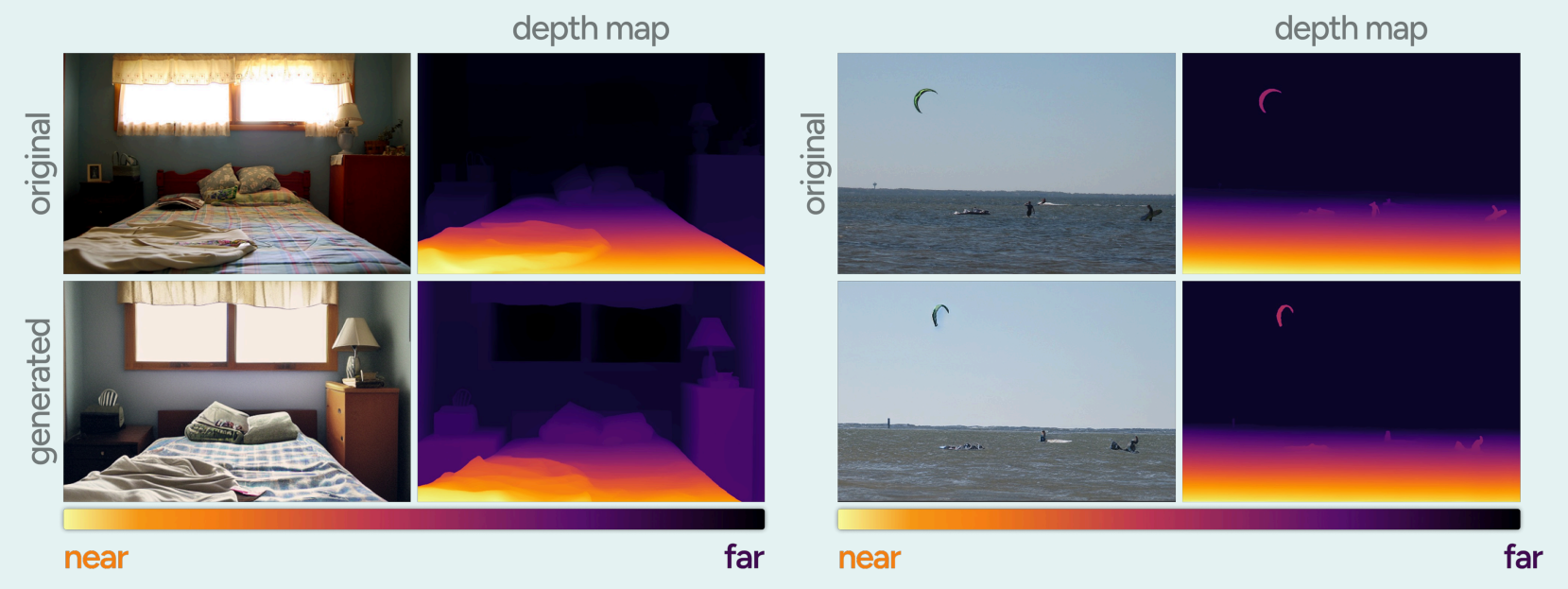
original
generated
Low-level
Mid-level
High-level
Input
CONV1
CONV2
CONV3
CONV4
CONV5
FC6
FC7
FC8
Max pooling
Dense
low
mid
high

correlations with human neuronal responses



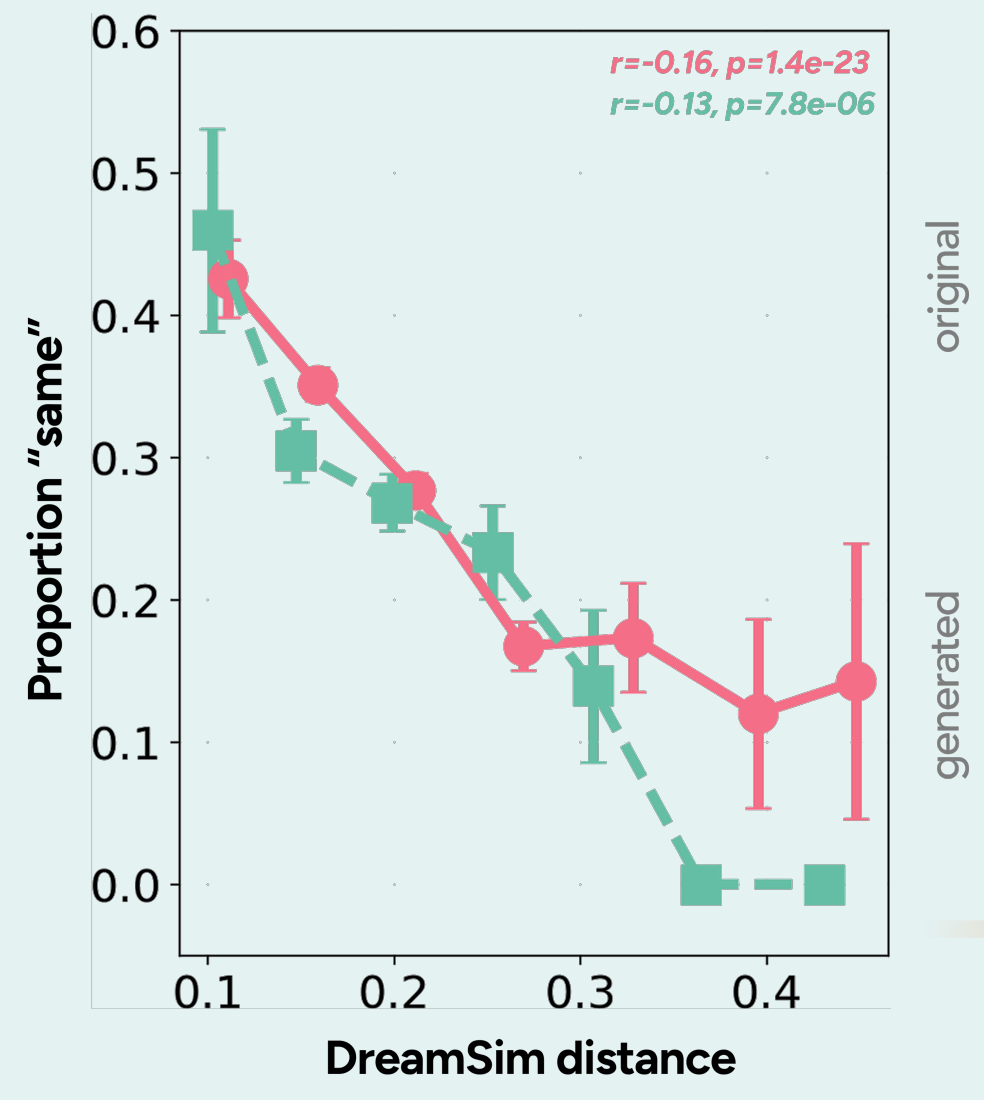
Mid & High-level visual feature similarity





depth map
original
generated
near
far

increasing depth estimation similarity

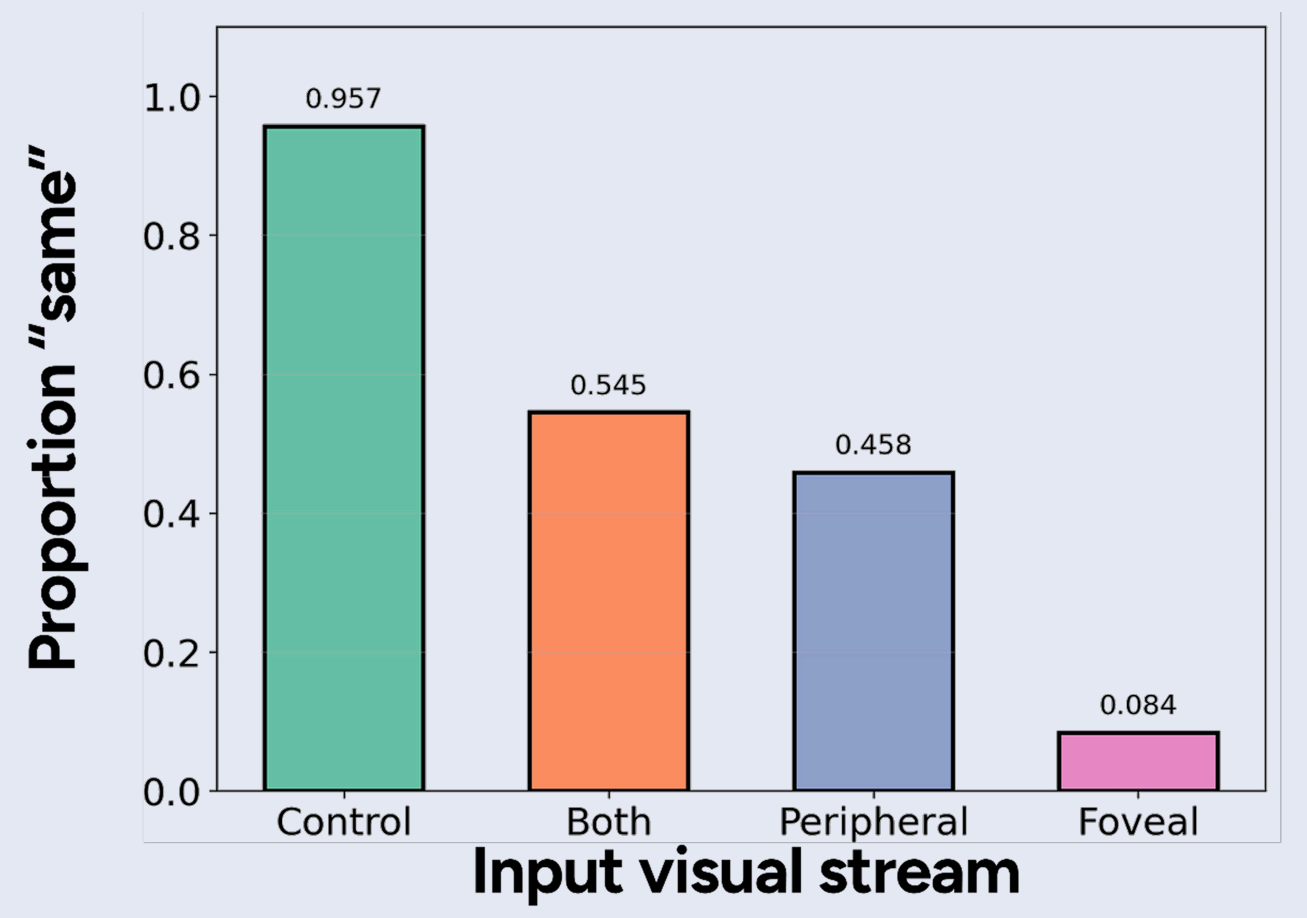


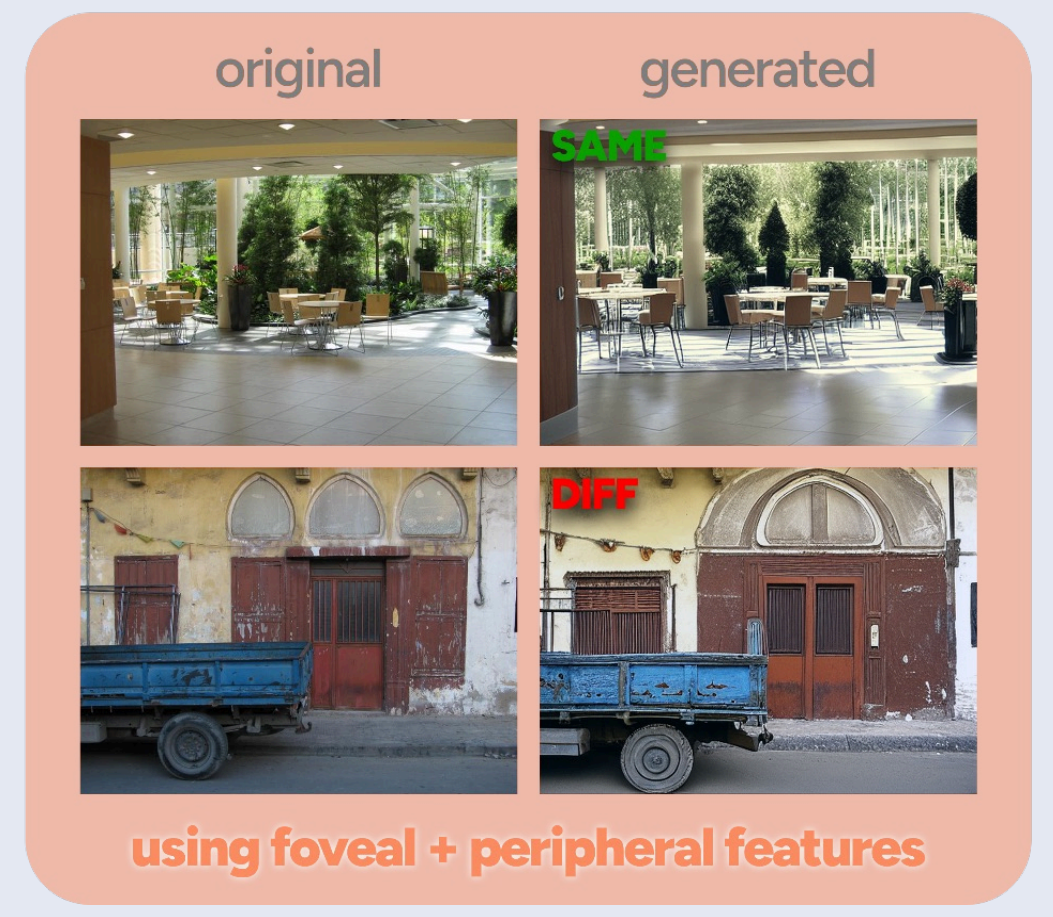
original
generated

increasing high-level visuo-semantic similarity

Both mid- and high-level similarities drive metameric judgments, with **semantic features showing the strongest effects.**

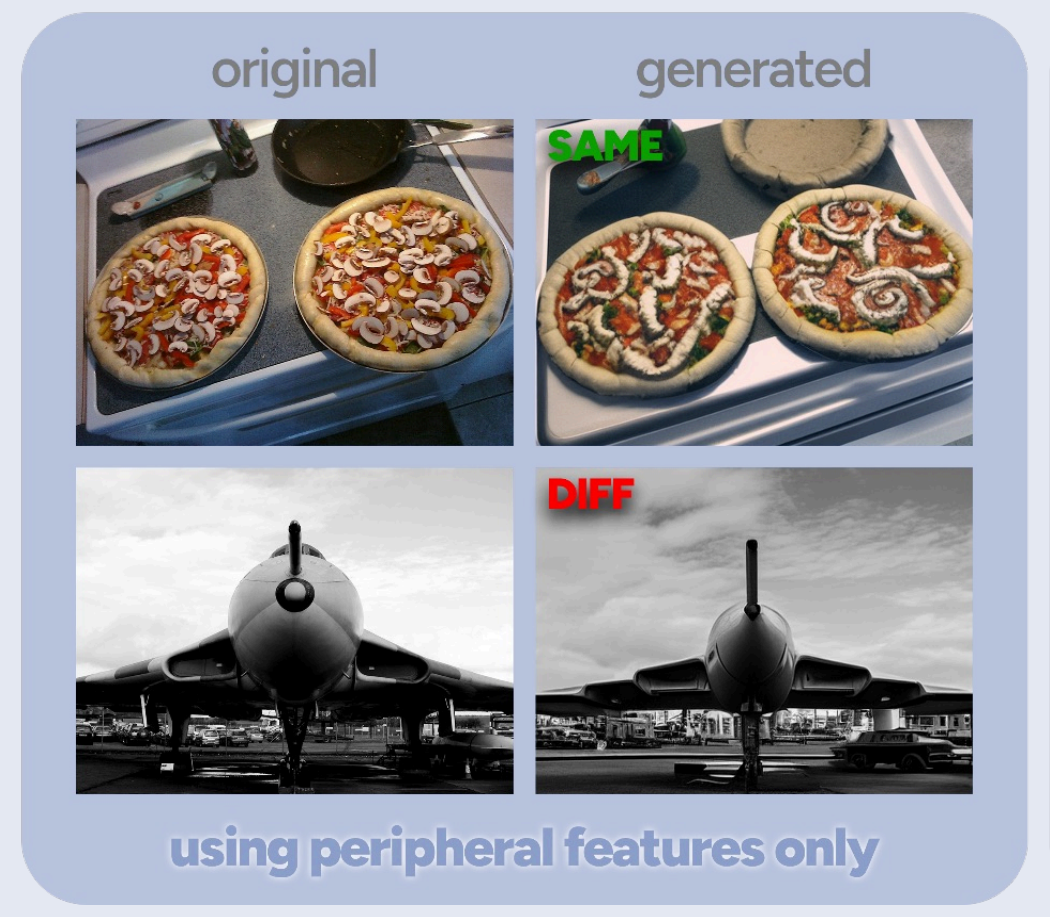
Foveal & peripheral contributions to metameric judgments





original
generated
SAME
DIFF

using foveal + peripheral features



original
generated
SAME
DIFF

using peripheral features only

original
generated
SAME
DIFF

using foveal features only

Scene understanding depends primarily on **peripheral scene structure**, with **foveal detail** providing an additional boost.