Brain2VLM

MARS2 @ ECCV 2026

Hierarchical Alignment Between Cortical Representations and Vision–Language Latent Spaces

N. A. Adarsh Pritam[1] Jeba Shiney O[1] Sanyam Jain[2,3]

[1]Alliance School of Advanced Computing, Alliance University, India • [2]Dept. of Dentistry and Oral Health, Aarhus University, Denmark • [3]Dept. of Computer Science and Communication, Østfold University College, Norway

ECCV 2026 | MALMÖ | SEPT 8–12

AARHUS UNIVERSITY

Østfold University of Applied Sciences

CODE github.com

PREPRINT doi.org

PRESENTER Sanyam Jain

FIRST AUTHOR Adarsh Pritam

NOT ONE MAP

Linear for structure, strongly non-linear for semantics: △ **+0.48** on CLIP *c* against **+0.06** on diffusion *z*.

ONLY THE INTERFACE

No new generator, no new conditioning — and it still beats far heavier pipelines on every semantic metric.

BETTER LATENTS, NOT PIXELS

An expressive decoder drifts out of the generative prior: semantics improve, texture does not.

Brain-to-image decoders keep getting heavier *generators*. But what kind of map is the brain → latent interface itself?

1 THE PIPELINE

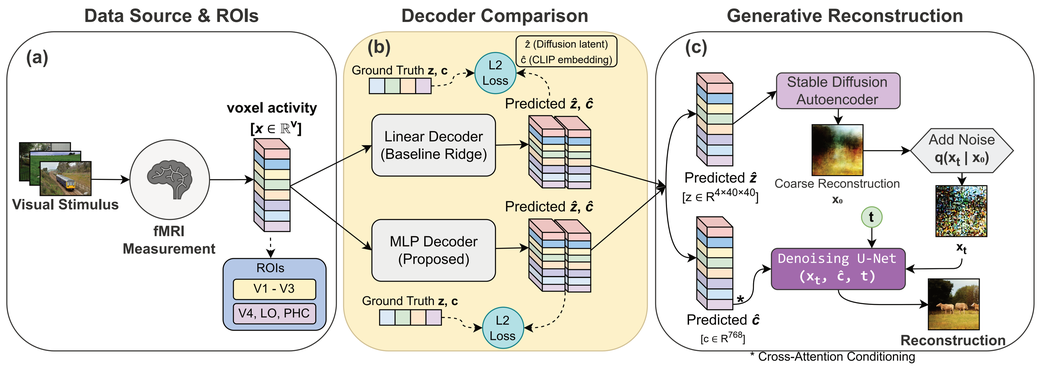

Fig. 1. **(a)** Voxels from early (V1–V3) and ventral (V4, LO, PHC) ROIs. **(b)** Ridge *vs.* residual MLP predict $\hat{z}$ and $\hat{c}$. **(c)** Frozen Stable Diffusion v1.4 decodes $\hat{z}$, re-noises it and denoises under $\hat{c}$. Nothing inside the generator is trained.

2 THE DECODER

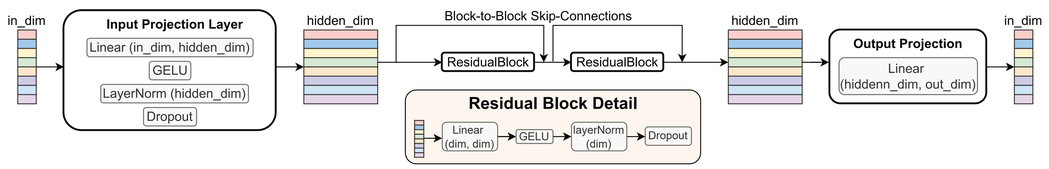

Fig. 2. Input projection, *N* residual blocks (Linear · GELU · LayerNorm · Dropout + skip), output projection — 2048 units, ℓ_2 loss, **identical hyper-parameters for all four subjects**.

3 RECONSTRUCTIONS

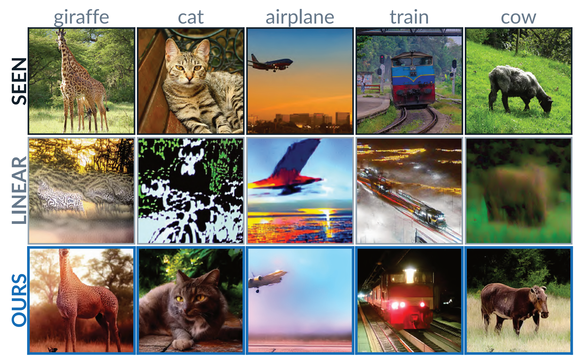

Fig. 3. subj01. Object identity is recovered far more often; texture and fine detail are *not*.

4 ABLATIONS

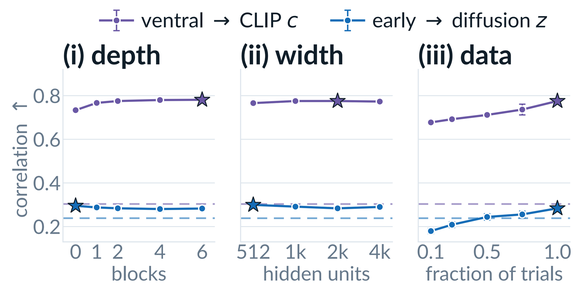

Fig. 4. Depth saturates after ∼2 blocks and width after 2048 — but accuracy still climbs with data.

5 DISTRIBUTIONS

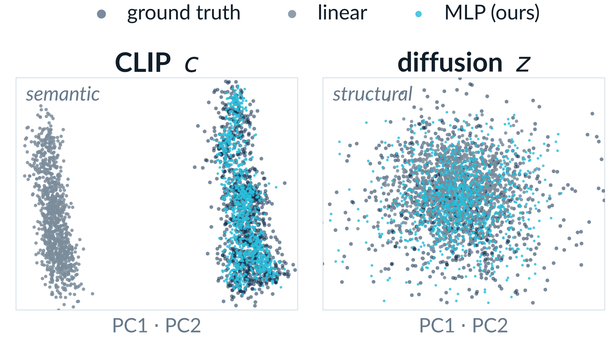

Fig. 5. MMD to the true latents: CLIP *c* **.359 → .042**, 8.5× closer; diffusion *z* .121 → .133, slightly worse.

6 CORTICAL MAP

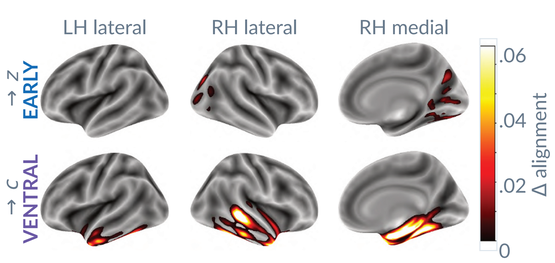

Fig. 6. △ alignment (non-linear − linear): sparse in early V1–V3, strong and coherent along the **ventral stream** V4·LO·PHC.

Decoder residual MLP, 2048 units, ℓ_2, identical hyper-parameters for all four subjects • **Data** NSD 7T, subj01/02/05/07, ∼10 000 MS COCO stimuli each, 982 shared test images • **Next** distribution-aware losses so the decoder stays on-manifold