BIT-LLM: Brain Instruction Tuned LLM with persistent cross-attention for fMRI-to-text decoding

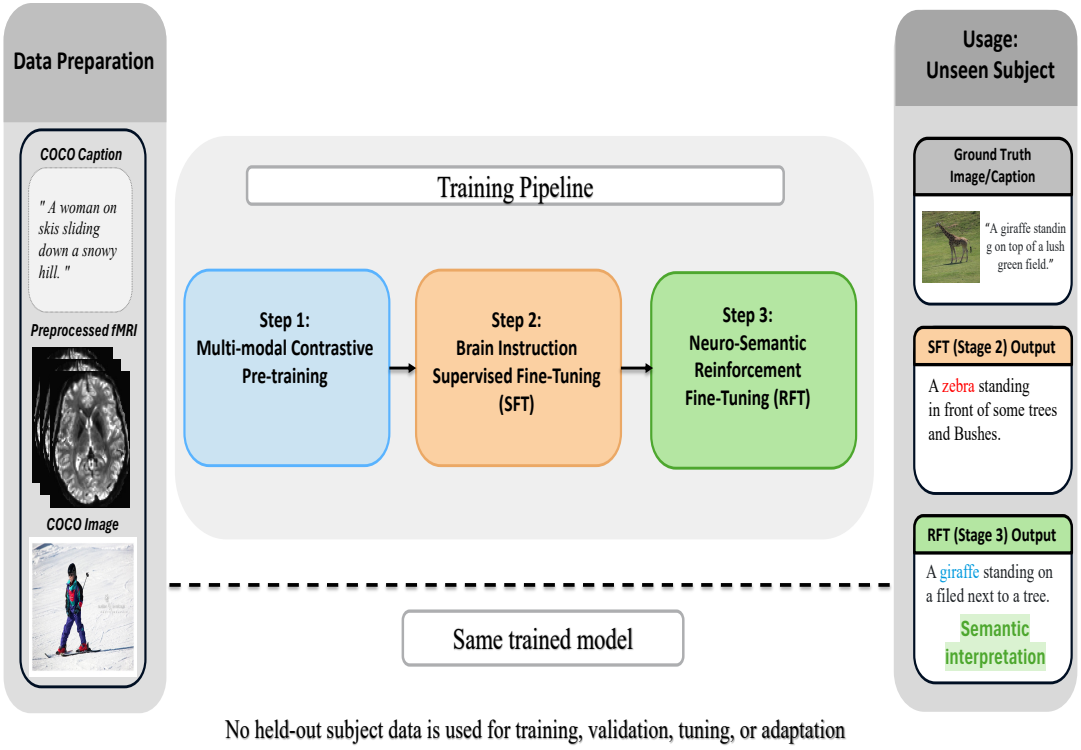
Sung Hwan Lee[1], Jihun Kim[1], Chae Lynn Kim[1] , Ji Yun Park[1],Jong-Hwan Lee[1]

[1]Dept. of Brain Engineering, Korea University

KOREA UNIVERSITY
Brain and Cognitive Engineering

Brain Signal Processing Laboratory
https://bspl.korea.ac.kr

Problem: language priors can dominate fMRI-to-text decoding



No held-out subject data is used for training, validation, tuning, or adaptation

Held-out protocol

Train	S1	S2	S3	S4	S5	S6	S7
Test	S8	← never used for training, validation, prompt tuning, or adaptation					

+ Test images: NSD (Natural Scenes Dataset) shared-image stimuli (also unseen)

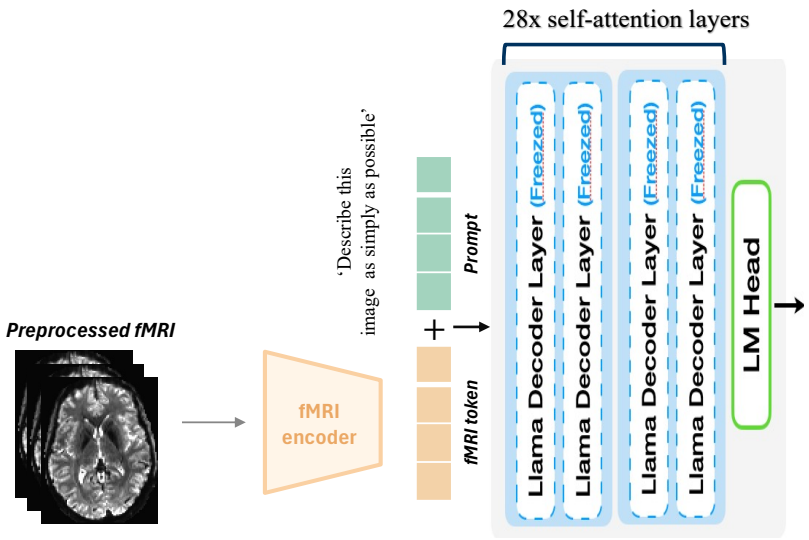
Goal: map fMRI responses to a caption of the viewed image.

Risk: prefix conditioning injects fMRI evidence **once** and fades, while the LLM caption prior persists across decoding.

Consequence: high BLEU / CIDEr can reflect fluent, stimulus-agnostic captions — not voxel-dependent evidence.

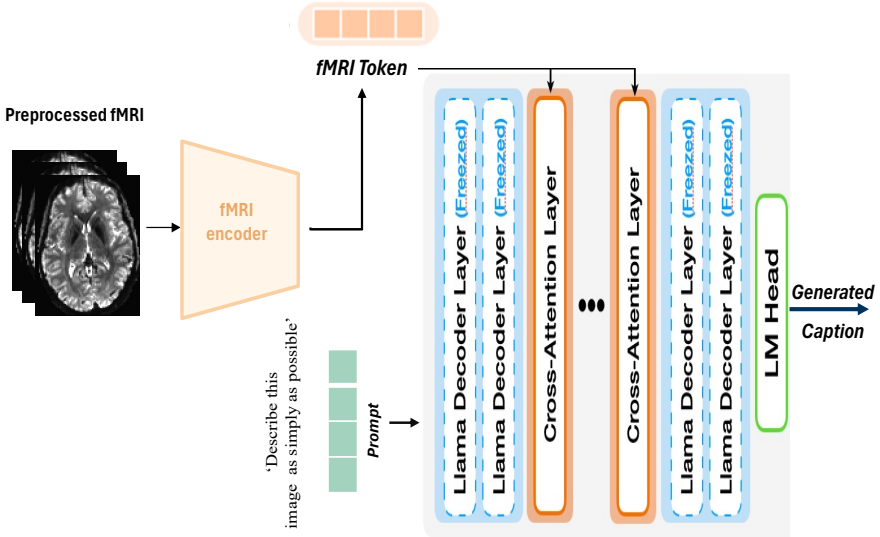
BIT-LLM: persistent neural access via interleaved cross-attention

Prefix conditioning (baseline)



fMRI injected once at the input, must survive implicitly through every layer.

BIT-LLM: persistent key–value memory



Text states repeatedly query brain KV at multiple depths during decoding.

Brain tokens are exposed as a **persistent key–value memory** queried repeatedly during decoding, rather than as a single prepended prefix.

fMRI encoder → 128 brain tokens **Projection** → LLM hidden space **LLM** Llama-3.2-3B-Instruct, largely frozen

Interface 4 cross-attention adapters after layers {3, 7, 11, 15}; Q = text states, K,V = projected fMRI.

A three-stage pipeline: align, connect, optimize

Stage 1

Multimodal Contrastive Pre-training

fMRI ↔ image / text
(CLIP-style space)

Goal: semantically aligned, value-sensitive brain encoder

Train: fMRI encoder + proj

Stage 2

Brain Instruction SFT

fMRI tokens → caption
(persistent cross-attention)

Goal: learn the brain–LLM interface

Train: projector + x-attn adapters
Frozen: encoder + Llama backbone

Stage 3

Neuro-Semantic RFT

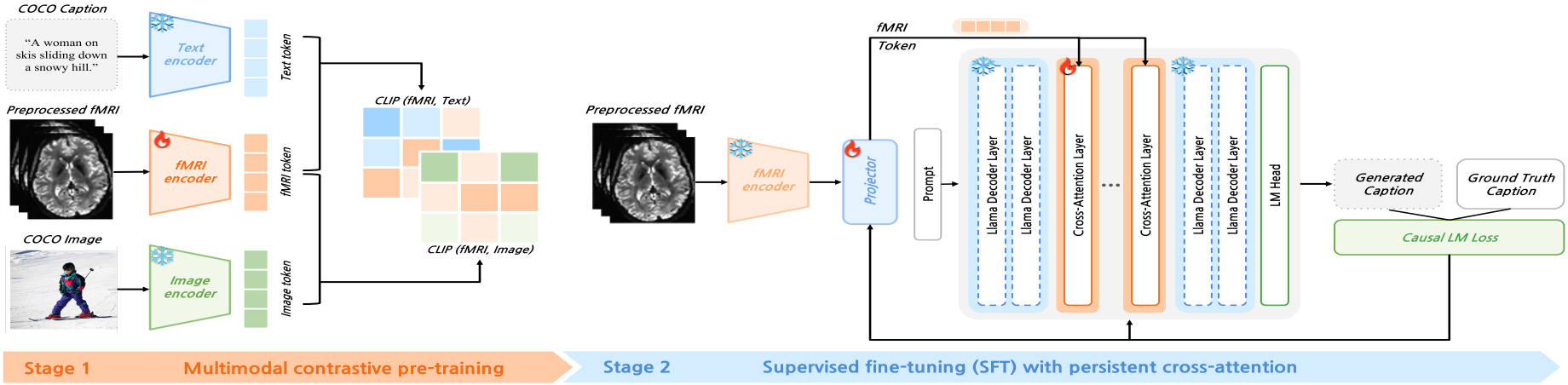
GRPO with caption rewards
(CIDEr + BERTScore − IP)

Goal: optimize captions at the sequence level

Train: LoRA adapters

Align brain representations → connect them to a frozen LLM → optimize caption quality.

Training Methods: Stage 1/2 Training Pipeline



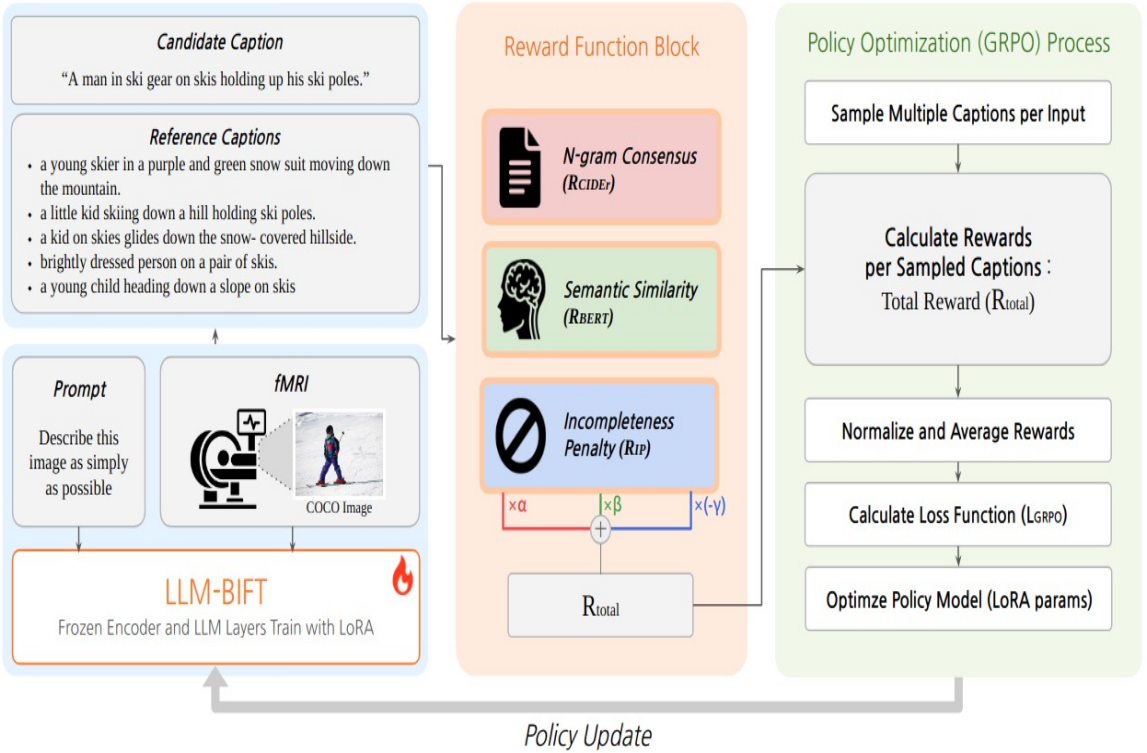
Stage 1: Multi-Modal Contrastive Pre-training (Alignment Stage)

- Objective: To learn a **subject-agnostic brain embedding space** by aligning fMRI representations with image and text embeddings in a shared semantic space.
- Employs a CLIP-style symmetric InfoNCE objective to map brain signals, images, and text into a common contrastive space.

Stage 2: Brain Instruction Fine-Tuning (SFT/BIFT Stage)

- Objective: To transform the encoder-decoder pair into an end-to-end fMRI-to-text generator that produces natural language captions.
- Optimizes a causal language modeling objective conditioned on brain tokens through interleaved cross-attention layers.

Stage 3: Optimize generated captions with Neuro-Semantic RFT



Motivation- SFT optimizes token-level likelihood,

but caption quality is evaluated at the sequence level.

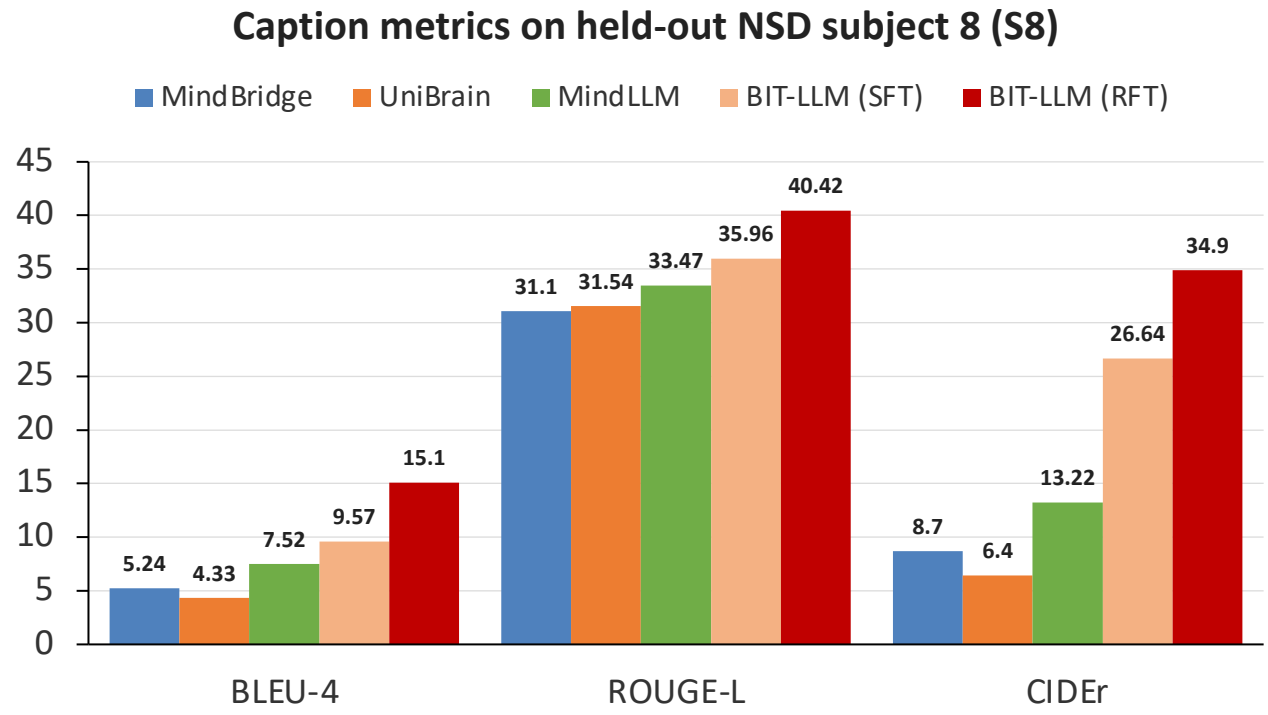
Reward - $R_{Total} = \alpha R_{CIDEr} + \beta R_{Bert} + \gamma R_{IP}$

Procedure:

1. Start from the Stage-2 checkpoint.
2. Sample G captions per fMRI input.
3. Score each caption with a composite reward.
4. Compute group-relative advantages.
5. Update LoRA parameters with GRPO.

(*Evaluation: greedy decoding — temp 0, top-p 1*)

BIT-LLM performance for Held-out subject



Brain alignment lifts CIDEr 4×; reward refinement adds another 30%.

Evaluation: same S1–7 → S8 held-out-subject protocol, unseen images, greedy deterministic decoding.

Key result.

BIT-LLM (RFT) tops every baseline on all three metrics

BLEU-4: 15.1

ROUGE-L: 40.42

CIDEr: 34.9

SFT → RFT (Δ from reward refinement)

BLEU-4 9.57 → **15.10 (+58%)**

ROUGE-L 35.96 → **40.42 (+12%)**

CIDEr 26.64 → **34.90 (+31%)**

Reward refinement aligns generations with semantic-similarity signals — biggest lift on BLEU-4 and CIDEr.

Cross-attention outperforms parameter-matched prefix conditioning

Interface	Trainable params	BLEU-4	ROUGE-L	METEOR	CIDEr
Prefix, proj-only	3.15M	9.45	34.99	15.12	24.80
Prefix, parameter-matched MLP	100.67M	7.69	33.59	14.58	23.98
Cross-attention, ours	**103.82M**	**9.57**	**35.96**	**15.76**	**26.64**

Parameter-matched prefix MLP does not close the gap.

Prefix with MLP:

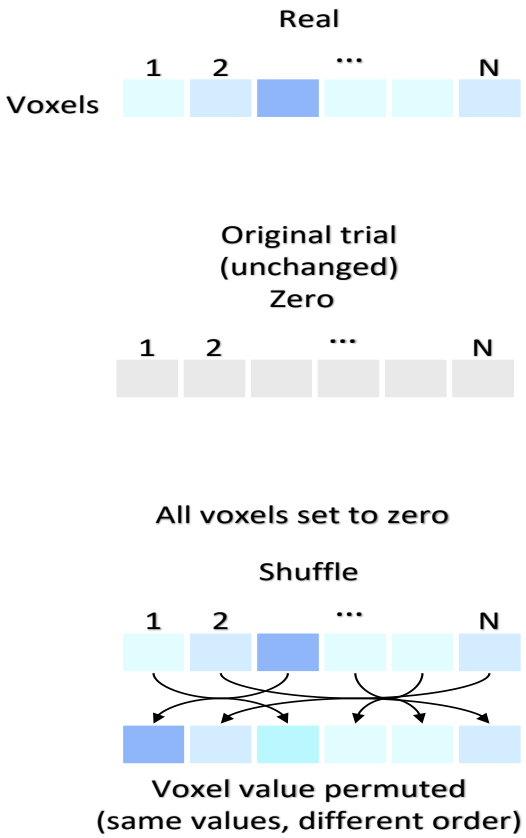
100.67M params, CIDEr 23.98

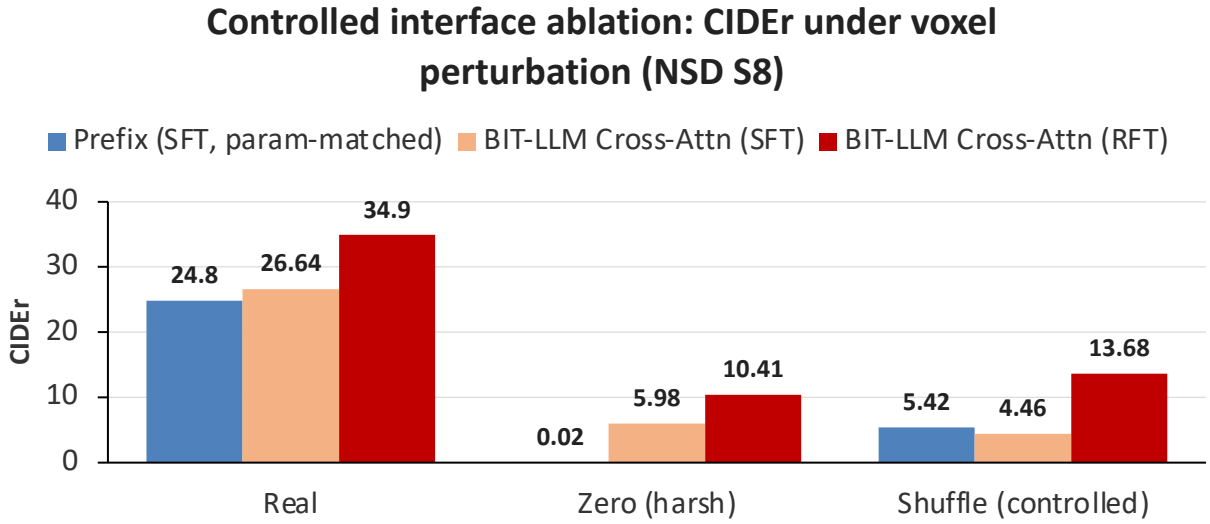
Cross-attention:

103.82M params, CIDEr 26.64

- The gain is not explained by interface capacity alone;
- persistent cross-attention provides a better conditioning path.

Voxel perturbations show that RFT gains remain brain-dependent

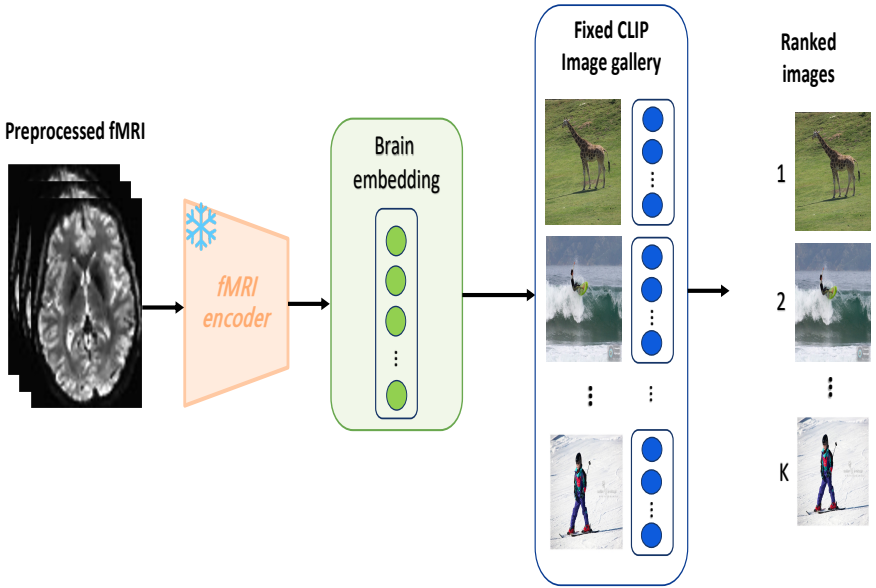




Same encoder, same SFT recipe — only the interface differs. Shuffle is the *distribution-controlled* diagnostic; Zero is an intentionally harsh stress test.

- **Prefix collapses** (CIDEr 0.02 under Zero) with *non-linguistic* generations — capacity-matched but conditioning fades.
- **Cross-attention degrades gracefully** — persistent neural access keeps outputs well-formed under perturbation.
- **RFT preserves voxel reliance** (Real 34.9 ≫ Zero 10.4, Shuffle 13.7) — gains don't erase neural grounding.

Encoder diagnostics confirm value-sensitive brain representations



Stage-1 contrastive pretraining makes the encoder sensitive to voxel values and preserves stimulus-specific semantic information.

Diagnostic 1

Real vs. all-zero embedding similarity

Encoder	Cosine similarity
MindLLM-base	0.9889 ± 0.0018
Ours, Stage-1 encoder	**−0.0080 ± 0.0345**

Near 1 = invariant to voxel values

Near 0 = value-dependent representation

Diagnostic 2

fMRI → image retrieval

Input	R@10 ↑	MedR ↓
Real	**0.4481**	**13**
Shuffle	0.0094	504
Zero	0.0102	500

Real fMRI preserves stimulus semantics;

Zero/Shuffle collapses to near-chance retrieval.

Takeaway: persistent access + staged training + grounding diagnostics

Main takeaways

1. Persistent neural access

Cross-attention keeps fMRI tokens queryable throughout autoregressive decoding.

2. Staged training

Contrastive pretraining grounds the encoder, SFT learns the interface, RFT refines generation.

3. Grounding diagnostics

Zero/shuffle perturbations and encoder retrieval show dependence on voxel values and spatial structure.

Limitations and next steps

1. Reward design

Current rewards still rely on caption metrics; stronger neuro-visual verifiers are needed.

2. Task scope

Captioning is not full language decoding; QA, dialogue, and multi-turn tasks remain open.

3. Generalization and ethics

Cross-dataset transfer, subject variability, privacy, and consent require careful treatment.

Thank You for Listening