PRIMATEALIGN: UNIFYING ZERO-SHOT CROSS-SPECIES VISUAL DECODING AND NEURAL TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent neural decoding advances have successfully translated brain activity into visual representations. Despite neuroscientific evidence for profound similarities in primate visual processing, computational decoding models remain single-species. To address this limitation, we propose PrimateAlign, the first unified framework for cross-species visual decoding. Our approach projects human electroencephalography (EEG) and macaque multi-unit activity (MUA) into a shared multimodal latent space. By combining global semantic matching with a novel composite loss formulation for dense spatial alignment, the framework explicitly preserves both high-level conceptual meaning and fine-grained visual details. This shared representational space creates a computational bridge between distinct biological signals. In doing so, it unlocks a fundamentally novel capability of bidirectional cross-species neural translation on unseen stimuli. Extensive experiments demonstrate that PrimateAlign achieves state-of-the-art image retrieval performance in both modalities, reaching 85.9% (+5.9%) on THINGS-EEG2 and 96.4% (+27.1%) on THINGS-TVSD.

1 INTRODUCTION

Neural decoding involves translating complex biological brain signals into interpretable representations, a process with profound implications for advancing brain-computer interfaces and understanding primate cognition (Lebedev & Nicolelis, 2006). In cognitive neuroscience, electroencephalography (EEG) provides scalable, non-invasive capture of macroscopic brain dynamics but suffers from low signal-to-noise ratio and significant inter-subject variability (Makeig et al., 2004; Niedermeyer & da Silva, 2005). In contrast, invasive multi-unit activity (MUA) recordings provide unparalleled spatial and temporal fidelity, but their invasiveness restricts application (Buzsáki, 2004; Maynard et al., 1997).

Visual stimulus retrieval from evoked brain activity has advanced in both invasive macaque recordings and non-invasive EEG benchmarks such as THINGS-EEG2 (Gifford et al., 2022), with state-of-the-art frameworks projecting brain signals into the latent space of contrastive vision-language models. However, a critical limitation persists across these paradigms, as they rely almost exclusively on global semantic embeddings. By bottlenecking the decoding process through a single semantic vector, these models lose access to localized visual features and fine-grained details that can support richer neural-visual alignment.

Beyond these representational bottlenecks, current computational models remain siloed by species, despite decades of neuroscientific evidence for conserved visual-stream geometries across humans and macaques (Orban et al., 2004; Tsao et al., 2008; Kriegeskorte et al., 2008). Foundational representation learning methods such as CEBRA (Schneider et al., 2023) align neural activity across subjects and recording modalities within a single species, yet extending these unified spaces across species to capture high-level visual semantics shared across the primate visual system remains largely unexplored.

To address these representational and cross-species limitations, we introduce PrimateAlign, a unified human EEG and macaque MUA decoding architecture with a species-agnostic visual latent space.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Its dense alignment and joint cross-species training establish a shared visual-semantic representation across the two recording modalities.

Our main contributions are as follows:

- We propose a unified decoding architecture that jointly models human EEG and macaque MUA in a shared, species-agnostic latent space, capturing mutual biological information and enabling a fundamentally new capability of bidirectional cross-species neural translation.
- We propose a novel composite loss formulation for dense spatial alignment. To circumvent the limitations of a single global bottleneck, this new objective integrates token-level maximum similarity matching (MaxSim) with semantic soft-target InfoNCE regularization, allowing the model to establish fine-grained correspondences between neural tokens and localized visual features.
- We achieve state-of-the-art zero-shot image retrieval, reaching 85.9% (+5.9%) accuracy on the THINGS-EEG2 benchmark and 96.4% (+27.1%) on THINGS-TVSD.

2 Related Work

Visual Decoding from Neural Signals. Visual decoding from brain activity has advanced rapidly with the release of large-scale datasets and vision-language models. Early work demonstrated the feasibility of recovering visual categories and synthesizing images from EEG (Spampinato et al., 2017; Palazzo et al., 2020; Kavasidis et al., 2017; Tirupattur et al., 2018). To avoid the temporal autocorrelation confounds present in early block-design experiments (Ahmed et al., 2021), the field has shifted toward rapid serial visual presentation (RSVP) paradigms (Grootswagers et al., 2019). Leveraging this shift, the THINGS-EEG2 benchmark (Gifford et al., 2022) then enabled systematic zero-shot image retrieval, with strong baselines built on temporal-spatial encoders (Lawhern et al., 2018; Schirrmeister et al., 2017; Song et al., 2022) and multimodal contrastive frameworks (Du et al., 2023; Song et al., 2024). Subsequent frameworks have improved retrieval accuracy by incorporating cognitive and visual priors (Zhang et al., 2026; Liu et al., 2026; Wu et al., 2025), enforcing cycle and semantic consistency (Wei et al., 2024; Chen et al., 2024), and unifying multimodal features (Zhang et al., 2025; Li et al., 2025a). Beyond static retrieval, EEG decoding has expanded to dynamic video perception (Liu et al., 2024), language alignment (Duan et al., 2023), and diffusion-based image reconstruction (Li et al., 2024; Bai et al., 2023), paralleling the rapid generative breakthroughs seen in MEG and fMRI (Benchetrit et al., 2024; Chen et al., 2023; Lin et al., 2022; Huo et al., 2024; Xia et al., 2024; Scotti et al., 2023; Takagi & Nishimoto, 2023). Meanwhile, self-supervised foundation models for brain signals pursue task-agnostic pretraining on large-scale EEG and intracranial corpora (Kostas et al., 2021; Jiang et al., 2024; Xiao et al., 2026; Yang et al., 2023; Zhang et al., 2023; Wang et al., 2025; 2024a; Jiang et al., 2025), while complementary self-supervised methods target cross-recording generalization through time-frequency consistency (Zhang et al., 2022) and topology-agnostic representations (Yi et al., 2023). In the invasive domain, MUA from macaque visual cortex offers high-resolution recordings of inferotemporal responses (Hung et al., 2005; Majaj et al., 2015; Yamins et al., 2014; Kar et al., 2019; Bao et al., 2020; Chang & Tsao, 2017; Bashivan et al., 2019; Higgins et al., 2021). These recordings further support generative image reconstruction (Dado et al., 2024). Spike-based foundation models have also emerged, pretraining transformers and sequential autoencoders on large-scale population recordings (Azabou et al., 2023; Ye et al., 2023), though none target cross-species visual alignment. The recent THINGS-TVSD dataset (Papale et al., 2025) offers a direct counterpart to THINGS-EEG2 on identical stimuli, supporting CLIP-based retrieval (Ciferri et al., 2025) and image reconstruction (Le et al., 2024; Fei et al., 2025). Despite this natural pairing, prior work has treated EEG and MUA decoding as separate problems.

Cross-Species Neural Homology. Comparative neuroscience has long established that humans and macaques share conserved representations across homologous regions of the ventral visual stream (Kriegeskorte et al., 2008; Orban et al., 2004; Tsao et al., 2008; Rajalingham et al., 2018; DiCarlo et al., 2012). Recent work on the THINGS stimulus set further demonstrates geometrically aligned object spaces between human fMRI and macaque MUA (van Bree & Hebart, 2026), and aligning vision models to macaque IT representations improves their behavioral and adversarial generalization (Dapello et al., 2022). More broadly, both goal-driven and self-supervised deep networks have emerged as state-of-the-art predictors of ventral-stream responses in macaques and humans

(Schrimpf et al., 2020; Zhuang et al., 2021; Konkle & Alvarez, 2022), justifying contrastive image-text embeddings as a principled shared space across species. Temporal dynamics in EEG have also been shown to track progressively from visual to linguistic representations (Rong et al., 2025), motivating time-resolved neural representations. The closest methodological precedent is CEBRA (Schneider et al., 2023), which uses contrastive learning to align neural activity across subjects and recording modalities within a single species. While this established the viability of intra-species neural embedding, our approach targets a fundamentally different challenge of aligning representations across species, bridging the evolutionary gap between non-invasive human EEG and invasive macaque MUA. To the best of our knowledge, PrimateAlign is the first to embed neural signals from different species in a shared representational space that supports both zero-shot cross-species visual retrieval and direct neural signal translation.

Dense Alignment for Brain Decoding. Most neural decoding models rely on a single global representation such as the CLIP global pooled embedding, which discards spatial structure on the visual side. In contrast, vision-language research has shown that dense, token-level alignment improves cross-modal grounding. Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023) use cross-attention to distill spatial features into a fixed set of query tokens, FineCLIP (Jing et al., 2024) distills region-level CLIP features for fine-grained matching, and FILIP (Yao et al., 2021) matches local components through token-wise maximum similarity. In neuroimaging, structured visual latents have improved fMRI reconstruction (Takagi & Nishimoto, 2023), and semantic soft targets help preserve continuous representational geometry (Scotti et al., 2023). PrimateAlign builds on these lines of work by combining a query-based dense-token bottleneck with FILIP-style MaxSim alignment and MindEye-style soft semantic targets.

Generalization Across Subjects and Species. Establishing subject-invariant representations remains a primary challenge in neuroimaging, driving extensive domain adaptation efforts in both EEG (Nørskov et al., 2023; Wang et al., 2024c; Mellot et al., 2024; Li et al., 2025b) and fMRI (Wang et al., 2024b; Quan et al., 2024; Zhou et al., 2024). PrimateAlign extends this generalization axis from subjects to species by mapping disparate human EEG and macaque MUA signals into a shared latent space.

3 METHODS

PrimateAlign uses modality-specific encoders, shared visual-space projectors, and conditional decoders to align human EEG and macaque MUA, in order to establish a direct correspondence between the distinct neural signals and the visual features of their eliciting stimuli, while supporting bidirectional cross-species neural translation, which requires mapping a neural representation from a source species directly into the native, high-dimensional signal space of a target species. Human EEG first passes through subject-specific affine adapters to mitigate inter-subject variability; both EEG and MUA then feed into their respective encoders, extracting a global semantic vector for high-level conceptual alignment and a sequence of dense temporal tokens for fine-grained correspondence with localized visual features. The overall architecture is illustrated in Figure 1.

3.1 ARCHITECTURE

During training, the framework processes paired batches of human EEG signals $X_H \in \mathbb{R}^{C_H \times T_H}$ and macaque MUA signals $X_M \in \mathbb{R}^{C_M \times T_M}$, where C_s and T_s represent the number of recording channels and time points for modality $s \in \{H, M\}$. These signals are recorded in response to identical visual stimuli. To overcome physiological heterogeneity across human subjects, subject-specific affine transformations align distinct neural topologies (Dai et al., 2025; Scotti et al., 2024; Défossez et al., 2023). For a given human subject i_H, raw EEG undergoes a learned element-wise affine transformation, yielding the adapted signal $\tilde{X}_H \in \mathbb{R}^{C_H \times T_H}$:

$$\tilde{X}_H = X_H \odot W^{(i_H)} + B^{(i_H)} \tag{1}$$

where $W^{(i_H)}, B^{(i_H)} \in \mathbb{R}^{C_H \times T_H}$ are learnable subject-specific parameter matrices, and $\odot$ denotes the Hadamard product. The adapted human signal $\tilde{X}_H$ and the raw macaque signal X_M are then processed by modality-aware encoders E_H and E_M.

Macaque MUA Encoding. For macaque MUA, which provides high-fidelity localized spiking, the encoder E_M applies a spatial 1D convolution to fuse cross-channel information, followed by a

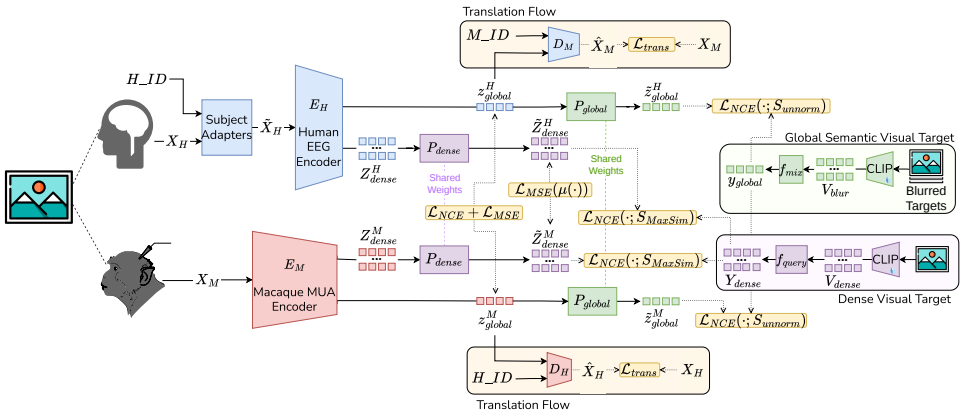

Figure 1: Overview of the PrimateAlign architecture. Modality-specific encoders (E_H, E_M) extract global vectors and dense temporal tokens from human EEG and macaque MUA. Both are projected into a shared space and aligned with hierarchical visual targets (y_{global}, Y_{dense}). Generative decoders (D_H, D_M) then perform direct cross-species translation from the aligned global representations.

series of bidirectional Mamba blocks (Gu & Dao, 2023; Zhu et al., 2024) for long-range temporal modeling. The resulting sequence is split into two representations:

$$(z^M_{global}, Z^M_{dense}) = E_M(X_M) \tag{2}$$

where $z^M_{global} \in \mathbb{R}^{D_g}$ is the global semantic vector obtained via attention pooling, and $Z^M_{dense} \in \mathbb{R}^{K_M \times D_d}$ are K_M dense temporal tokens produced by strided 1D convolution.

Human EEG Encoding. For human EEG, subject to volume conduction that reduces spatial resolution while preserving temporal dynamics, E_H adopts a dual-routing strategy. Following the spatial-temporal backbone of Liu et al. (2026), $\tilde{X}^H$ passes through an absolute spatial convolution and a temporal MLP to extract global semantics. Simultaneously, to capture fine-grained temporal patches for localized alignment, intermediate spatial features route through a bidirectional Mamba pathway, which has shown strong performance for modeling EEG signals (Gui et al., 2024), yielding:

$$(z^H_{global}, Z^H_{dense}) = E_H(\tilde{X}_H) \tag{3}$$

where $z^H_{global} \in \mathbb{R}^{D_g}$ is the global semantic vector, and $Z^H_{dense} \in \mathbb{R}^{K_H \times D_d}$ denotes the K_H dense temporal tokens.

Alignment Projections and Translation. The neural representations of both modalities ($s \in \{H, M\}$) are mapped into a shared D-dimensional visual space through modality-agnostic projectors. The global vectors pass through P_{global} (a residual MLP), while the dense tokens are mapped via P_{dense} (a standard MLP). Formally, these shared projections yield:

$$\tilde{z}^s_{global} = P_{global}(z^s_{global}), \quad \tilde{Z}^s_{dense} = P_{dense}(Z^s_{dense}) \tag{4}$$

where $\tilde{z}^s_{global} \in \mathbb{R}^D$ and $\tilde{Z}^s_{dense} \in \mathbb{R}^{K_s \times D}$. Furthermore, cross-species translation is enabled by convolutional decoders conditioned on target subject identities via learned embeddings. Using i_H and i_M to denote specific human and macaque subjects, the bidirectional signal reconstruction is formulated as:

$$\hat{X}_M = D_M(z^H_{global}, i_M), \quad \hat{X}_H = D_H(z^M_{global}, i_H) \tag{5}$$

where D_M and D_H decode the global semantic representation of the source modality into the predicted neural signal of the target modality.

Visual Target Generation. Simultaneously, visual stimuli are processed into targets within this shared visual space. Following Liu et al. (2026), hierarchical blur features $V_{blur} \in \mathbb{R}^{L \times D_{blur}}$ pass through target mixer f_{mix}, which applies a learned per-feature softmax-weighted sum and MLP adapter to extract the global visual target. For high-dimensional dense visual patches $V_{dense} \in \mathbb{R}^{N_v \times D_v}$, a Perceiver-inspired query aggregator (Jaegle et al., 2021; Alayrac et al., 2022; Li et al.,

2023) uses learnable query embeddings $Q \in \mathbb{R}^{N_q \times D}$ to distill vital spatial information via Multi-Head Cross-Attention (MHCA), followed by a residual MLP, yielding:

$$y_{global} = f_{mix}(V_{blur}), \quad Y_{dense} = f_{query}(V_{dense}; Q) \tag{6}$$

where $y_{global} \in \mathbb{R}^D$ is the global visual target, and $Y_{dense} \in \mathbb{R}^{N_q \times D}$ denotes the N_q distilled dense visual tokens.

3.2 Loss Function

The framework is trained end-to-end to align multimodal representations and perform cross-species neural translation.

We optimize alignment and latent objectives with InfoNCE (Oord et al., 2018; Chen et al., 2020; Radford et al., 2021). Let $\mathcal{L}_{\text{NCE}}(X, Y; S)$ be a generalized temperature-scaled cross-entropy with similarity S between source representations X and target representations Y:

$$\mathcal{L}_{\text{NCE}}(X, Y; S) = -\frac{1}{B} \sum_{i,j=1}^{B} p_{i,j} \log \left(\frac{\exp(S(x_i, y_j)/\tau)}{\sum_{k=1}^{B} \exp(S(x_i, y_k)/\tau)} \right) \tag{7}$$

where τ is a fixed or learned temperature and $p_{i,j}$ the target probability distribution. We use three variants: standard InfoNCE ($p_{i,j}$ as a hard one-hot identity matrix and $S_{\cos}(x, y)$ as the L2-normalized dot product), magnitude-preserving InfoNCE (leaving the source representation unnormalized, $S_{\text{unnorm}}(x, y) = \frac{x \cdot y}{\|y\|}$, to preserve magnitude information) (Zhang et al., 2026), and InfoNCE with semantic soft targets ($\mathcal{L}^{\text{soft}}_{\text{NCE}}$), replacing hard one-hot labels $p_{i,j}$ with a softmax over pairwise cosine similarities in a pre-trained CLIP embedding space (Scotti et al., 2023).

Modality-to-Vision Alignment. To align the neural representations with the visual targets, we optimize across two spatial scales. For global semantics, we apply a symmetric InfoNCE loss between the projected brain vectors and the global visual target:

$$\mathcal{L}^s_{global} = \frac{1}{2}\Big(\mathcal{L}_{NCE}(\tilde{z}^s_{global}, y_{global}; S_{unnorm}) + \mathcal{L}_{NCE}(y_{global}, \tilde{z}^s_{global}; S^T_{unnorm})\Big) \tag{8}$$

where $s \in \{H, M\}$ and $S^T(X, Y) \triangleq S(Y, X)$, utilizing the magnitude-preserving metric S_{unnorm}. For fine-grained spatial alignment, we enforce local correspondence between the sequence of dense brain tokens and the dense visual tokens:

$$\mathcal{L}^s_{dense} = \frac{1}{2}\Big(\mathcal{L}^{soft}_{NCE}(\tilde{Z}^s_{dense}, Y_{dense}; S_{MaxSim}) + \mathcal{L}^{soft}_{NCE}(Y_{dense}, \tilde{Z}^s_{dense}; S^T_{MaxSim})\Big) \tag{9}$$

where $\mathcal{L}^{soft}_{NCE}$ uses token-wise maximum similarity S_{MaxSim}. Traditional contrastive learning pools features into a single global embedding, discarding token-level distinctions; we therefore adapt MaxSim (Yao et al., 2021) to align temporal sequences with localized visual features. For token sets $X = \{x_1, \ldots, x_N\}$ and $Y = \{y_1, \ldots, y_M\}$, similarity uses the best-matching target token for each source token:

$$S_{MaxSim}(X, Y) = \frac{1}{N} \sum_{i=1}^{N} \max_{1 \le j \le M} \left(\frac{x_i \cdot y_j}{|x_i||y_j|} \right) \tag{10}$$

This token-level similarity captures fine-grained local correspondences beyond a single global representation. In our cross-species setting, this mechanism is particularly valuable as it anchors generalized temporal neural motifs directly to localized visual patches.

Cross-Species Latent Constraints. We enforce direct biological-modality alignment at both spatial scales through a unified latent loss $\mathcal{L}_{latent}$ with three terms. First, we apply symmetric InfoNCE (standard cosine similarity S_{cos}) and Mean Squared Error (MSE) to unprojected global latents. Second, we align aggregate dense representations by computing MSE over the token-wise means of the projected dense latents:

$$\begin{aligned} \mathcal{L}_{latent} = \frac{1}{2}\Big(&\mathcal{L}_{NCE}(z^H_{global}, z^M_{global}; S_{cos}) + \mathcal{L}_{NCE}(z^M_{global}, z^H_{global}; S_{cos})\Big) \\ &+ \gamma_1 \text{MSE}(z^H_{global}, z^M_{global}) + \gamma_2 \text{MSE}(\mu(\tilde{Z}^H_{dense}), \mu(\tilde{Z}^M_{dense})) \end{aligned} \tag{11}$$

where $\mu(\cdot)$ averages over token dimension K_s, and γ_1 and γ_2 are empirically determined scaling weights.

Cross-Species Signal Translation. To ensure the learned global representations retain sufficient temporal dynamics for cross-species decoding, we compute a MSE translation loss between the generated signal $\hat{X}_s$ and the ground truth biological signal X_s:

$$\mathcal{L}^s_{trans} = \text{MSE}(\hat{X}_s, X_s) \tag{12}$$

Total Objective. The final loss function is defined by grouping the modality-specific components and the cross-species latent constraints:

$$\mathcal{L}_{total} = \sum_{s\in\{H,M\}} \left(\lambda^s_1 \mathcal{L}^s_{global} + \lambda^s_2 \mathcal{L}^s_{dense} + \lambda^s_3 \mathcal{L}^s_{trans}\right) + \lambda_4 \mathcal{L}_{latent} \tag{13}$$

where λ^s_1, λ^s_2, λ^s_3, and λ_4 are empirically determined scaling weights.

3.3 Inference

Zero-Shot Decoding. During evaluation, the trained model performs zero-shot image retrieval, cross-species biological retrieval, and cross-species translation. For visual decoding, given test biological signal X_s ($s \in \{H, M\}$) and candidate visual target $V = (V_{blur}, V_{dense})$, the model extracts projected neural features $\tilde{z}^s_{global}, \tilde{Z}^s_{dense}$ and computed visual counterparts y_{global}, Y_{dense}. The final image retrieval score is a multi-scale convex combination:

$$\text{Score}(X_s, V) = (1-\beta) S_{cos}(\tilde{z}^s_{global}, y_{global}) + \beta S_{MaxSim}(\tilde{Z}^s_{dense}, Y_{dense}) \tag{14}$$

where $\beta \in [0, 1]$ is an empirically determined hyperparameter balancing the global and dense spatial semantic landscapes.

For cross-species biological retrieval, the visual projection heads are bypassed to evaluate the alignment of the shared latent space directly. The cross-species retrieval score is calculated by computing the standard cosine similarity between the unprojected global vectors: $S_{cos}(z^H_{global}, z^M_{global})$.

Translation. Finally, cross-species signal translation uses the conditional convolutional decoders to predict a target subject's neural response to the same visual stimulus via $\hat{X}^M = D_M(z^H_{global}, i_M)$ for a target macaque and $\hat{X}^H = D_H(z^M_{global}, i_H)$ for a target human, where i_s denotes the target subject identity.

4 Experiments and Results

Datasets and Preprocessing. We evaluate PrimateAlign on two complementary datasets that share an identical set of visual stimuli, namely the non-invasive human THINGS-EEG2 dataset (Gifford et al., 2022), consisting of scalp EEG recordings, and the invasive macaque THINGS Ventral-Stream Spiking Dataset (TVSD) (Papale et al., 2025), consisting of MUA recordings. We follow standard preprocessing procedures from prior work on these datasets, including averaging trial repetitions within the official test splits to improve the signal-to-noise ratio (SNR).

Although official test splits are designed to be disjoint, an audit found that 178 of the 200 human test images overlapped with the default macaque training partition. To prevent leakage, we removed those 178 images from macaque training and reserved them for cross-species testing. This creates a controlled SNR asymmetry, where human EEG benefits from multi-trial averaging, whereas the corresponding macaque MUA comes from unaveraged single-session recordings.

The final joint training set comprises 16,540 unique natural images; during training, each human sample is paired with a randomly selected macaque sample of the same stimulus.

Evaluation Metrics. We evaluate zero-shot image retrieval with standard 200-way human EEG (Gifford et al., 2022) and 100-way macaque MUA (Papale et al., 2025) protocols. Accuracy is Top-1/Top-5 in CLIP embedding space (Song et al., 2024), with human EEG evaluated under both within-subject and leave-one-subject-out (LOSO) settings. To assess direct cross-species alignment,

Table 1: Cross-species neural translation metrics on 178 unseen shared stimuli, averaged over all source–target subject pairs.

Direction	Method	MSE ↓	rRMSE ↓	Pearson ρ ↑
$H \to M$	PCA + Ridge	0.5317	0.7362	0.2994
	PCA Procrustes	0.5338	0.7376	0.2935
	PrimateAlign (Ours)	**0.1768**	**0.4264**	**0.3189**
$M \to H$	PCA + Ridge	0.0273	0.8554	0.2721
	PCA Procrustes	3.8115	10.3331	0.0237
	PrimateAlign (Ours)	**0.0205**	**0.7354**	**0.4189**

we introduce a novel $H \leftrightarrow M$ retrieval over all target-modality subjects, resulting in 1,780 candidates for macaque-to-human and 356 candidates for human-to-macaque. A hit occurs if any sample corresponding to the target visual stimulus ranks in Top-K. Since correct targets scale with subject pool size, chance remains 1/178 (0.56%) in both directions. We average accuracies per query subject.

For cross-species neural signal translation, we report Mean Squared Error (MSE, averaged across dimensions), relative Root Mean Squared Error (rRMSE, normalized by per-sample signal RMS), and Pearson correlation (Naselaris et al., 2011) (per channel over time). Translation is conditioned on target subject identity, and metrics are averaged over all same-stimulus source-target subject pairs per direction.

Baselines. For THINGS-EEG2 human retrieval, we compare with NICE (Song et al., 2024), ATM (Li et al., 2024), UBP (Wu et al., 2025), NeuroBridge (Zhang et al., 2026), and current state-of-the-art Liu et al. (2026). For THINGS-TVSD macaque retrieval, we use the sole published baseline of Ciferri et al. (2025). Since cross-species neural retrieval ($H \leftrightarrow M$) and continuous translation lack benchmarks, we establish two standard alignment baselines in a 256-dimensional PCA-reduced space, consisting of ridge regression and orthogonal Procrustes alignment adapted from cross-subject hyperalignment (Haxby et al., 2011).

Implementation Details. We encode visual stimuli with a pre-trained CLIP ResNet-50 (RN50) backbone (Radford et al., 2021), extracting dense spatial features (V_{dense}) from Layer 4. Models use Distributed Data Parallel (DDP) on two NVIDIA L40S GPUs with global batch size 512. PrimateAlign has 57.0M trainable parameters; on a single L40S, extracting global and dense representations from 256 neural signals takes 4.81 ms for human EEG and 5.65 ms for macaque MUA. We optimize with AdamW (weight decay 0.1, peak learning rate 0.001) for 100 epochs, with 10-epoch linear warmup and 10x decay at epoch 70. To prevent overfitting, we regularize the macaque MUA with additive noise alongside random channel and time masking. We apply 10% token dropout to the dense spatial tokens ($\tilde{Z}^s_{dense}$) to prevent representational collapse within the MaxSim objective. For human inter-subject (LOSO) evaluation, we apply a cross-subject Mixup by linearly interpolating EEG trials from distinct subjects viewing identical stimuli (applied with probability 0.5, using an interpolation weight in $[0, 0.5]$), which simulates inter-subject variance while leaving the visual target invariant. All PrimateAlign results are reported as means over five random seeds. The Supplementary Material details loss coefficients, temperature scales, and architectural dimensions.

4.1 Results

We primarily evaluate a shared PrimateAlign model trained across all subjects. For image retrieval, we also train per-subject and per-monkey variants for fair comparison with single-subject baselines. In inter-subject evaluation, the shared model omits subject-specific affine adapters.

We first assess continuous cross-species signal translation on the 178 shared stimuli (Table 1). The shared PrimateAlign model yields the best MSE, rRMSE, and Pearson correlation in both directions. While our model substantially reduces MSE in the $H \to M$ direction, Ridge regression remains competitive in the $M \to H$ direction, likely because predicting noisy EEG targets naturally encourages a conservative smoothed average. Turning to Pearson correlation, PrimateAlign achieves the highest score in the $H \to M$ direction. We hypothesize the baselines remain competitive here because PCA effectively maps scale-invariant macroscopic trends, while less accurately capturing

Table 2: Zero-shot Top-1/Top-5 retrieval accuracy (%) on THINGS-EEG2 (200-way).

Method		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
Intra-Subject Evaluation												
NICE	Top-1	13.2	13.5	14.5	20.6	10.1	16.5	17.0	22.9	15.4	17.4	16.1
	Top-5	39.5	40.3	42.7	52.7	31.5	44.0	42.1	56.1	41.6	45.8	43.6
ATM	Top-1	25.6	22.0	25.0	31.4	12.9	21.3	30.5	38.8	34.4	29.1	27.1
	Top-5	60.4	54.5	62.4	60.9	43.0	51.1	61.5	72.0	51.5	63.5	58.1
UBP	Top-1	41.2	51.2	51.2	51.1	42.2	57.5	49.0	58.6	45.1	61.5	50.9
	Top-5	70.5	80.9	82.0	76.9	72.8	83.5	79.9	85.8	76.2	88.2	79.7
NeuroBridge	Top-1	50.0	63.2	61.6	61.0	54.8	69.7	62.7	71.2	64.0	73.6	63.2
	Top-5	77.6	90.6	91.1	90.0	85.0	92.9	88.8	95.1	91.0	97.1	89.9
Liu et al.	Top-1	81.9	81.7	78.3	76.9	71.4	84.3	78.2	84.5	78.7	84.1	80.0
	Top-5	96.8	96.8	96.0	97.2	93.7	98.7	96.7	98.1	96.8	98.1	96.9
PrimateAlign Per-Subject (Ours)	Top-1	85.4	84.1	82.4	83.9	75.5	87.7	**83.3**	88.9	83.1	**91.1**	84.5
	Top-5	97.8	97.6	97.1	**98.2**	94.7	98.5	**99.0**	98.7	**98.5**	98.7	97.9
PrimateAlign Shared (Ours)	Top-1	**88.4**	**87.0**	**84.0**	**85.0**	**81.0**	**88.2**	82.6	**89.6**	**85.6**	87.3	**85.9**
	Top-5	**98.5**	**99.2**	**99.0**	**98.2**	**97.0**	**99.3**	98.3	**99.4**	98.4	**99.0**	**98.6**
Inter-Subject Evaluation (LOSO)												
NICE	Top-1	7.6	5.9	6.0	6.3	4.4	5.6	5.6	6.3	5.7	8.4	6.2
	Top-5	22.8	20.5	22.3	20.7	18.3	22.2	19.7	22.0	17.6	28.3	21.4
ATM	Top-1	10.5	7.1	11.9	14.7	7.0	11.1	16.1	15.0	4.9	20.5	11.9
	Top-5	26.8	24.8	33.8	39.4	23.9	35.8	43.5	40.3	22.7	46.5	33.8
UBP	Top-1	11.5	15.5	9.8	13.0	8.8	11.7	10.2	12.2	15.5	16.0	12.4
	Top-5	29.7	40.0	27.0	32.3	33.8	31.0	23.8	32.2	40.5	43.5	33.4
NeuroBridge	Top-1	23.2	21.2	13.2	17.0	14.5	**25.0**	15.3	20.1	13.7	27.2	19.0
	Top-5	52.4	49.3	36.5	45.3	37.7	**55.0**	45.1	44.9	36.5	56.3	45.9
Liu et al.	Top-1	27.3	32.0	11.1	18.6	16.9	16.1	16.0	18.3	13.6	30.1	20.0
	Top-5	54.9	63.0	34.5	46.0	40.7	43.2	45.4	**49.0**	41.2	62.1	48.0
PrimateAlign (Ours)	Top-1	**29.1**	**43.9**	**15.7**	**25.5**	**28.3**	24.3	**24.8**	**24.2**	**26.2**	**35.7**	**27.8**
	Top-5	**55.6**	**71.9**	**39.6**	**54.3**	**55.0**	49.7	**49.7**	45.1	**55.7**	**66.7**	**54.6**

Table 3: Zero-shot Top-1/Top-5 retrieval accuracy (%) on THINGS-TVSD (100-way). averaged over both macaque subjects.

Method	Top-1	Top-5
Ciferri et al.	69.3	93.6
PrimateAlign Per-Monkey (Ours)	91.1	98.5
PrimateAlign Shared (Ours)	**96.4**	**99.7**

Table 4: Zero-shot cross-species retrieval accuracy (%) on the 178 shared stimuli.

Direction	Method	Top-1	Top-5
$H \rightarrow M$	PCA + Ridge Regression	3.8	11.2
	PCA Procrustes Alignment	3.3	11.2
	PrimateAlign (Ours)	**44.9**	**77.5**
$M \rightarrow H$	PCA + Ridge Regression	5.3	21.1
	PCA Procrustes Alignment	2.0	7.6
	PrimateAlign (Ours)	**53.1**	**77.8**

absolute signal values, as reflected by the higher MSE. However, in the $M \rightarrow H$ direction, PrimateAlign substantially outperforms all baselines on Pearson correlation. This suggests our model successfully reconstructs the dynamic temporal shape of the target EEG rather than defaulting to a flattened prediction.

For intra-modality image retrieval on THINGS-EEG2 (Table 2), PrimateAlign achieves state-of-the-art intra-subject performance, suggesting its dual-pathway architecture extracts richer semantics than single-vector approaches. This performance extends to the inter-subject (LOSO) evaluation, where PrimateAlign yields the highest average generalization and achieves top accuracy on most of the unseen subjects. On THINGS-TVSD (Table 3), PrimateAlign exceeds the published baseline with over 90% accuracy, suggesting the Mamba backbone captures long-sequence dynamics in binned MUA population signals.

Finally, in zero-shot cross-species retrieval (Table 4), PCA-based baselines perform near chance, whereas PrimateAlign reaches approximately 50% accuracy in both directions, showing that the species-agnostic latent space bridges non-invasive waves and invasive spikes.

Table 5: Subject-averaged PrimateAlign retrieval ablations (%). β denotes global/dense mixing.

	$H \rightarrow$Image			$M \rightarrow$Image			$H \rightarrow M$		$M \rightarrow H$	
Ablation	**T1**	**T5**	β	**T1**	**T5**	β	**T1**	**T5**	**T1**	**T5**
PrimateAlign (Full)	**85.9**	**98.6**	0.7	96.4	99.7	0.7	**44.9**	**77.5**	**53.1**	**77.8**
w/o Dense Pathway	79.9	97.2	0.0	88.7	99.2	0.0	–	–	–	–
w/ Hard Targets	75.5	95.7	0.7	95.0	99.9	0.7	–	–	–	–
	83.5	98.1	0.3	94.4	**100.0**	0.3	–	–	–	–
w/o Query Aggregator	81.4	97.6	0.7	94.5	99.6	0.7	–	–	–	–
w/ MeanSim	84.5	**98.6**	0.7	94.4	99.8	0.7	–	–	–	–
Human-Only Training	84.2	98.3	0.7	–	–	–	–	–	–	–
Monkey-Only Training	–	–	–	**98.1**	99.9	0.7	–	–	–	–
w/o H–M Objectives	–	–	–	–	–	–	17.4	45.2	29.5	61.8

4.2 Ablation Analysis

We ablate architectural components on THINGS-EEG2 and THINGS-TVSD (Table 5). The full architecture gives the strongest balanced performance among the joint-model variants. Because Top-5 accuracy remains high and comparatively less discriminative across most variants, we focus on Top-1 accuracy.

Removing the dense pathway disables all dense-related loss terms (λ_2^H, λ_2^M from Eq. 13, and γ_2 from Eq. 11) and sets $\beta = 0$ at inference, forcing the model to rely solely on global semantic vectors (z_{global}). This reduces Top-1 accuracy by 6.0% for human EEG and 7.7% for macaque MUA, demonstrating that the dense pathway contributes substantially beyond the global representation.

Within the dense pathway, we evaluate three design choices. At the fixed $\beta = 0.7$, replacing the soft-target objective ($\mathcal{L}_{\text{NCE}}^{\text{soft}}$) with hard-target InfoNCE ($\mathcal{L}_{\text{NCE}}$, using a one-hot identity matrix for $p_{i,j}$) reduces Top-1 accuracy by 10.4% for human EEG and 1.4% for macaque MUA, demonstrating that semantic soft targets are particularly important for EEG retrieval. Reducing β to 0.3 partially recovers human performance to 83.5%, while providing no corresponding Top-1 improvement for macaque MUA. This suggests that EEG retrieval with hard-target dense similarities benefits from lower weighting, whereas the soft-target formulation performs strongly under the common $\beta = 0.7$ setting. Substituting the max operation in S_{MaxSim} (Eq. 10) with mean pooling reduces Top-1 accuracy by 1.4% for human EEG and 2.0% for macaque MUA, indicating that best-patch matching retains more discriminative localized information than averaging similarities across all visual patches. Bypassing the cross-attention bottleneck in f_{query} (Eq. 6) and routing visual patches directly through the residual MLP reduces Top-1 accuracy by 4.5% for human EEG and 1.9% for macaque MUA, supporting the role of cross-attention in distilling task-relevant spatial representations.

We evaluate cross-species training from two complementary perspectives. Training each modality separately lowers human Top-1 by 1.7% but raises macaque Top-1 by 1.7% relative to joint training, suggesting that high-fidelity MUA regularizes noisier EEG representations. Within the joint model, removing H–M latent and translation objectives while retaining shared visual supervision reduces cross-species Top-1 from 44.9/53.1% to 17.4/29.5% ($H \rightarrow M$/$M \rightarrow H$), demonstrating substantial gains from direct coupling.

5 Conclusion

In this work, we introduced PrimateAlign, which unifies human EEG and macaque MUA for visual decoding and cross-species translation, achieving state-of-the-art intra-modality and strong cross-species retrieval. Ablations demonstrate the benefits of dense alignment and explicit H–M coupling, establishing a foundation for multi-species neural alignment.

AI USE STATEMENT

In this work, we used generative AI tools to assist with method implementation through software code generation based on our explicit specifications, provide feedback on research methodology and diagnostic experiments, and support the interpretation of experimental results. We additionally used generative AI tools for brainstorming, summarizing existing literature, and editing the manuscript to improve readability.

All AI-assisted work was critically reviewed and independently validated by the authors. AI-generated code was manually reviewed and tested for correctness, and literature summaries were verified against the original publications. The core research idea, task formulation, and architectural design were developed by the authors, who made all final methodological decisions and interpretations. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, the Experiments and Results section describes the datasets, preprocessing, evaluation protocols, baselines, and training setup, while Appendix A reports the architectural specifications and hyperparameter configurations. The supplementary material contains the anonymized source code and instructions for downloading and preprocessing the datasets and reproducing the training and evaluation procedures.

REFERENCES

Hamad Ahmed, Ronnie B Wilbur, Hari M Bharadwaj, and Jeffrey Mark Siskind. Object classification from randomized eeg trials. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3844–3853. IEEE, 2021.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva Dyer. A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, 36:44937–44956, 2023.

Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*, 2023.

Pinglei Bao, Liang She, Mason McGill, and Doris Y Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, 2020.

Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. In *International Conference on Learning Representations*, volume 2024, pp. 7846–7858, 2024.

György Buzsáki. Large-scale recording of neuronal ensembles. *Nature neuroscience*, 7(5):446–451, 2004.

Le Chang and Doris Y Tsao. The code for facial identity in the primate brain. *Cell*, 169(6):1013–1028, 2017.

Hongzhou Chen, Lianghua He, Yihang Liu, Longzhen Yang, Shaohua Shang, and MengChu Zhou. Visual neural decoding via improved visual-eeg semantic consistency. *arXiv preprint arXiv:2408.06788*, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22710–22720. IEEE, 2023.

Matteo Ciferri, Matteo Ferrante, and Nicola Toschi. Simple temporal attention beats complex decoders for neural-to-visual mapping from primate spiking data. In *NeurIPS 2025 Workshop on Foundation Models for the Brain and Body*, 2025.

Thirza Dado, Paolo Papale, Antonio Lozano, Lynn Le, Feng Wang, Marcel van Gerven, Pieter Roelfsema, Yağmur Güçlütürk, and Umut Güçlü. Brain2gan: Feature-disentangled neural encoding and decoding of visual perception in the primate brain. *PLoS computational biology*, 20(5): e1012058, 2024.

Yuqin Dai, Zhouheng Yao, Chunfeng Song, Qihao Zheng, Weijian Mai, Kunyu Peng, Shuai Lu, Wanli Ouyang, Jian Yang, and Jiamin Wu. Mindaligner: Explicit brain functional alignment for cross-subject visual decoding from limited fmri data. *arXiv preprint arXiv:2502.05034*, 2025.

Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Geary, Michael Ferguson, David D Cox, and James J DiCarlo. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. *BioRxiv*, pp. 2022–07, 2022.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5 (10):1097–1107, 2023.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. *arXiv preprint arXiv:2309.14030*, 2023.

Teng Fei, Srinivas Ravishankar, Hoko Nakada, Abhinav Uppal, Ian Jackson, Garrison W Cottrell, Ryusuke Hayashi, and Virginia R de Sa. Monkey perceptogram: Reconstructing visual representation and presumptive neural preference from monkey multi-electrode arrays. *arXiv preprint arXiv:2510.07576*, 2025.

Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Tijl Grootswagers, Amanda K Robinson, and Thomas A Carlson. The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188:668–679, 2019.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Yiyu Gui, MingZhi Chen, Yuqi Su, Guibo Luo, and Yuchao Yang. Eegmamba: Bidirectional state space model with mixture of experts for eeg multi-task classification. *arXiv preprint arXiv:2407.20254*, 2024.

James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1):6456, 2021.

Chou Hung, Gabriel Kreiman, Tomaso Poggio, and James Dicarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310:863–866, 11 2005. doi: 10.1126/science.1117593.

Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision*, pp. 56–73. Springer, 2024.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. In *International Conference on Learning Representations*, volume 2025, pp. 55436–55457, 2025.

Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *Advances in Neural Information Processing Systems*, 37:27896–27918, 2024.

Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B Issa, and James J DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, 22(6):974–983, 2019.

Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1809–1817, 2017.

Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Lynn Le, Paolo Papale, Katja Seeliger, Antonio Lozano, Thirza Dado, Feng Wang, Pieter Roelfsema, Marcel van Gerven, Yağmur Güçlütürk, and Umut Güçlü. Monkeysee: Space-time-resolved reconstructions of natural images from macaque multi-unit activity. *Advances in Neural Information Processing Systems*, 37:93826–93848, 2024.

Mikhail A Lebedev and Miguel AL Nicolelis. Brain–machine interfaces: past, present and future. *TRENDS in Neurosciences*, 29(9):536–546, 2006.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 102822–102864. Curran Associates, Inc., 2024. doi: 10.52202/079017-3266. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ba5f1233efa77787ff9ec015877dbd1f-Paper-Conference.pdf.

Dongyang Li, Haoyang Qin, Mingyang Wu, Chen Wei, and Quanying Liu. Brainflora: Uncovering brain concept representation via multimodal neural embeddings. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5577–5586, 2025a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Shanglin Li, Motoaki Kawanabe, and Reinmar Kobler. Spdim: Source-free unsupervised conditional and label shift adaptation in eeg. In *International Conference on Learning Representations*, volume 2025, pp. 101907–101931, 2025b.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35:29624–29636, 2022.

Wenchao Liu, Hongwei Li, Zhouyang Xu, Lin Ma, and Haifeng Li. Leveraging visual blur perception characteristics for eeg decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17580–17588, 2026.

Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. Eeg2video: Towards decoding dynamic visual perception from eeg signals. *Advances in Neural Information Processing Systems*, 37:72245–72273, 2024.

Najib J Majaj, Ha Hong, Ethan A Solomon, and James J DiCarlo. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *The Journal of Neuroscience*, 35(39):13402–13418, 2015.

Scott Makeig, Stefan Debener, Julie Onton, and Arnaud Delorme. Mining event-related brain dynamics. *Trends in cognitive sciences*, 8(5):204–210, 2004.

Edwin M Maynard, Craig T Nordhausen, and Richard A Normann. The utah intracortical electrode array: a recording structure for potential brain-computer interfaces. *Electroencephalography and clinical neurophysiology*, 102(3):228–239, 1997.

Apolline Mellot, Antoine Collas, Sylvain Chevallier, Alexandre Gramfort, and Denis A Engemann. Geodesic optimization for predictive shift adaptation on eeg data. *Advances in neural information processing systems*, 37:32828–32855, 2024.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Ernst Niedermeyer and FH Lopes da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

Anders Nørskov, Alexander Neergaard Zahid, and Morten Mørup. Cslp-ae: A contrastive split-latent permutation autoencoder framework for zero-shot electroencephalography signal conversion. *Advances in Neural Information Processing Systems*, 36:13179–13199, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Guy A Orban, David Van Essen, and Wim Vanduffel. Comparative mapping of higher visual areas in monkeys and humans. *Trends in cognitive sciences*, 8(7):315–324, 2004.

Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3833–3849, 2020.

Paolo Papale, Feng Wang, Matthew W Self, and Pieter R Roelfsema. An extensive dataset of spiking activity to reveal the syntax of the ventral stream. *Neuron*, 113(4):539–553, 2025.

Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 233–243. IEEE, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *The Journal of neuroscience*, 38 (33):7255–7269, 2018.

Boyan Rong, Alessandro Thomas Gifford, Emrah Düzel, and Radoslaw Martin Cichy. The time course of visuo-semantic representations in the human brain is captured by combining vision and language models. *arXiv preprint arXiv:2506.19497*, 2025.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. In *International conference on learning representations*, volume 2024, pp. 47648–47665, 2024.

Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4503–4511. IEEE, 2017.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14453–14463. IEEE, 2023.

Praveen Tirupattur, Yogesh Singh Rawat, Concetto Spampinato, and Mubarak Shah. Thoughtviz: Visualizing human thoughts using generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 950–958, 2018.

Doris Y Tsao, Sebastian Moeller, and Winrich A Freiwald. Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences*, 105(49):19514–19519, 2008.

Sander van Bree and Martin N Hebart. Shared and distinct object spaces in human and macaque inferotemporal cortex. *bioRxiv*, pp. 2026–05, 2026.

Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024a.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *International conference on learning representations*, volume 2025, pp. 75310–75346, 2025.

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11333–11342. IEEE, 2024b.

Yiming Wang, Bin Zhang, and Yujiao Tang. Dmmr: Cross-subject domain generalization for eeg-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 628–636, 2024c.

Yayun Wei, Lei Cao, Hao Li, and Yilin Dong. Mb2c: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8992–9000, 2024.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2246–2257. IEEE, 2025.

Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision*, pp. 242–259. Springer, 2024.

Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. Brainomni: A brain foundation model for unified eeg and meg signals. *Advances in Neural Information Processing Systems*, 38:41179–41212, 2026.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Joel Ye, Jennifer Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: multi-context pretraining for neural spiking activity. *Advances in Neural Information Processing Systems*, 36:80352–80374, 2023.

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36:53875–53891, 2023.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems*, 36:26304–26321, 2023.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14486–14493, 2025.

Wenjiang Zhang, Sifeng Wang, Yuwei Su, Xinyu Li, Chen Zhang, and Suyu Zhong. Neurobridge: Bio-inspired self-supervised eeg-to-image decoding via cognitive priors and bidirectional semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18028–18036, 2026.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022.

Qiongyi Zhou, Changde Du, Shengpei Wang, and Huiguang He. Clip-mused: Clip-guided multi-subject visual neural information semantic decoding. In *International Conference on Learning Representations*, volume 2024, pp. 52776–52792, 2024.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

A Experimental Settings and Hyperparameters

To ensure full reproducibility of our experiments, this appendix provides detailed architectural specifications and hyperparameter configurations for PrimateAlign. Table 6 summarizes the key architectural components and their parameter counts. We set the human EEG input dimensions to $C_H = 63$, $T_H = 250$, and macaque MUA to $C_M = 1024$, $T_M = 300$. The latent dimensions are $D_g = D = 1024$ and $D_d = 256$. The neural dense sequence lengths are set to $K_H = 25$ and $K_M = 30$. Global visual targets incorporate $L = 12$ hierarchical blur levels and 4 learnable input-independent default features ($D_{blur} = 1024$). For dense visual alignment, $N_v = 49$ spatial patches of dimension $D_v = 2048$ are distilled into $N_q = 16$ visual queries. Table 7 lists the optimization settings, loss weighting coefficients, contrastive temperatures, and inference blending weights.

Table 6: Architectural dimensions and parameter counts for PrimateAlign.

Component	**Input Dimension**	**Parameters**
Human Encoder (E_H)	$C_H \times T_H$	5.8M
Macaque Encoder (E_M)	$C_M \times T_M$	824K
Global Projector (P_{global})	D_g	2.1M
Dense Projector (P_{dense})	$K_s \times D_d$	329K
Subject Adapters	$C_H \times T_H$	315K
Target Mixer (f_{mix})	$L \times D_{blur}$	1.6M
Query Aggregator (f_{query})	$N_v \times D_v$	8.4M
Human Decoder (D_H)	D_g	17.0M
Macaque Decoder (D_M)	D_g	20.7M

Table 7: Comprehensive hyperparameter configurations and loss coefficients for PrimateAlign.

Category	Parameter / Symbol	Value
Optimization	Optimizer	AdamW
	Base Learning Rate (η)	0.001
	Weight Decay	0.1
	Warmup Epochs	10
	Total Epochs	100
	Global Batch Size	512
	Random Seeds	42, 2026, 1, 2, 3
Macro Objectives (Eq. 13)	Human Global Alignment (λ_1^H)	1.0
	Macaque Global Alignment (λ_1^M)	0.25
	Human Dense Alignment (λ_2^H)	0.4
	Macaque Dense Alignment (λ_2^M)	0.1
	Human Translation (λ_3^H)	0.125
	Macaque Translation (λ_3^M)	0.125
	Cross-Species Latent (λ_4)	1.0
Latent Constraints (Eq. 11)	Global Space MSE (γ_1)	0.5
	Dense Space Token-Mean MSE (γ_2)	0.5
Contrastive Temperatures	Human Global Temp.	Fixed (Temp. = 1.0, Softplus)
	Macaque Global Temp.	Learned (Init: $\ln(1/0.07)$, Softplus)
	Cross-Species Latent Temp.	Learned (Init: $\ln(1/0.07)$, Softplus)
	Human Dense Temp.	Fixed ($\tau = 0.005$)
	Macaque Dense Temp.	Fixed ($\tau = 0.005$)
Inference Blending (Eq. 14)	Mixing Weight (β)	0.7

B Variability Across Random Seeds

Table 8 reports standard deviations across the five random seeds specified above.

Table 8: Aggregate PrimateAlign results across five random seeds. Standard deviations are computed across seeds; retrieval values are percentages.

Evaluation	Metric	Mean $\pm$ Std.
Human Shared	Top-1	85.9 ± 0.329
	Top-5	98.6 ± 0.220
Human LOSO	Top-1	27.8 ± 0.655
	Top-5	54.6 ± 0.928
Macaque Shared	Top-1	96.4 ± 0.548
	Top-5	99.7 ± 0.274
$H \to M$ Retrieval	Top-1	44.9 ± 0.918
	Top-5	77.5 ± 1.137
$M \to H$ Retrieval	Top-1	53.1 ± 1.555
	Top-5	77.8 ± 1.948
$H \to M$ Translation	MSE	0.1768 ± 0.001
	rRMSE	0.4264 ± 0.001
	Pearson ρ	0.3189 ± 0.001
$M \to H$ Translation	MSE	0.0205 ± 0.000
	rRMSE	0.7354 ± 0.001
	Pearson ρ	0.4189 ± 0.001

C Extended Qualitative Visualizations

To provide deeper intuition into the learned representations, this section presents extended qualitative visualizations. Figure 2 displays a UMAP projection of the unified global latent space (z_{global}).

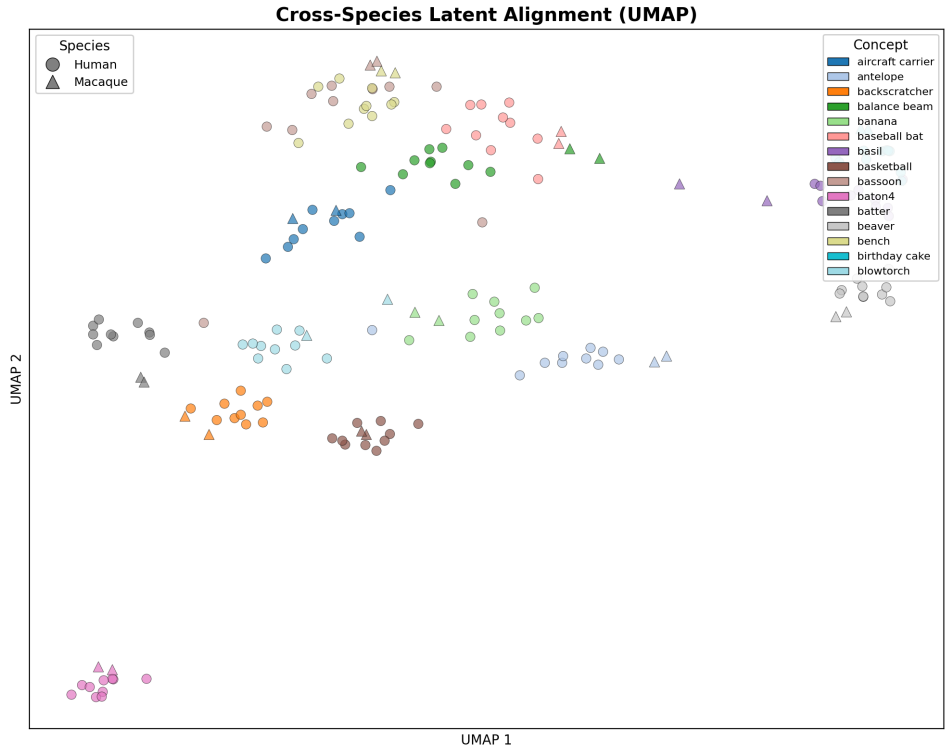

Figure 2: UMAP projection of z_{global} for the 15 most frequent concepts in the test set. Human EEG samples are denoted by circles, while macaque MUA samples are denoted by triangles. Colors indicate the specific stimulus concept.

For visual clarity, this projection plots the 15 most frequently occurring object concepts from the test set across all human and macaque subjects. Furthermore, Figures 3 through 6 provide top-5 zero-shot retrieval examples across the evaluated tasks. To ensure a representative evaluation, these specific queries were randomly sampled to showcase a realistic balance of successful matches alongside semantically informative failure modes.

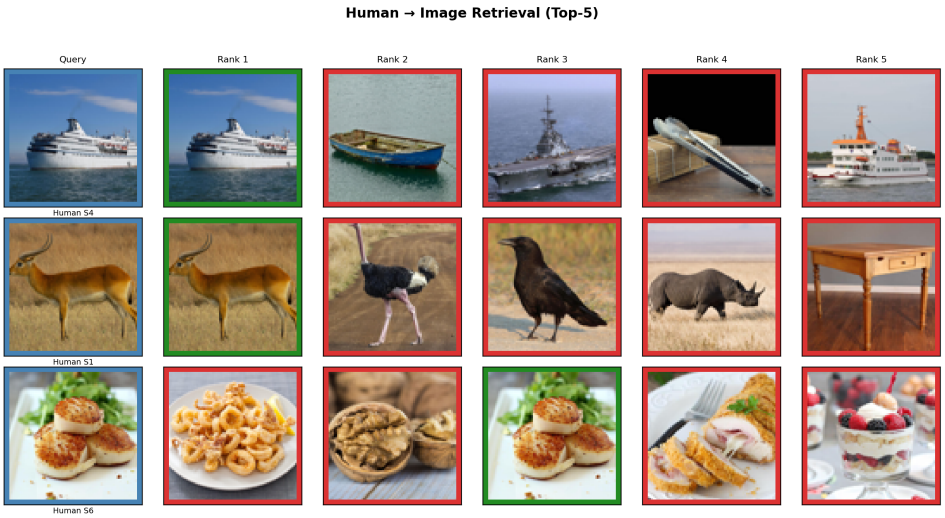

Figure 3: Qualitative examples of zero-shot image retrieval using human EEG queries. The query column displays the visual stimulus presented during the trial, with the corresponding human subject ID noted below. Green borders indicate exact ground-truth matches. Red borders indicate incorrect retrievals.

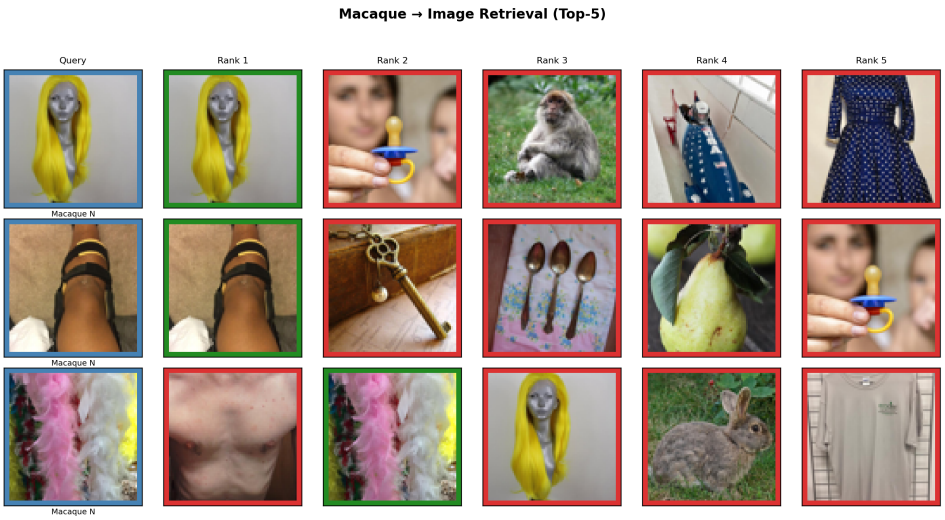

Figure 4: Qualitative examples of zero-shot image retrieval using macaque MUA queries. The query column displays the visual stimulus presented during the trial, with the corresponding macaque subject ID noted below. Green borders indicate exact ground-truth matches. Red borders indicate incorrect retrievals.



Figure 5: Qualitative examples of zero-shot cross-modal retrieval. The query column displays the visual stimulus presented to the human subject. The retrieved columns display the corresponding stimuli associated with the retrieved macaque MUA signals. The specific subject IDs are noted below each image. Green borders indicate exact ground-truth matches. Red borders indicate incorrect retrievals.

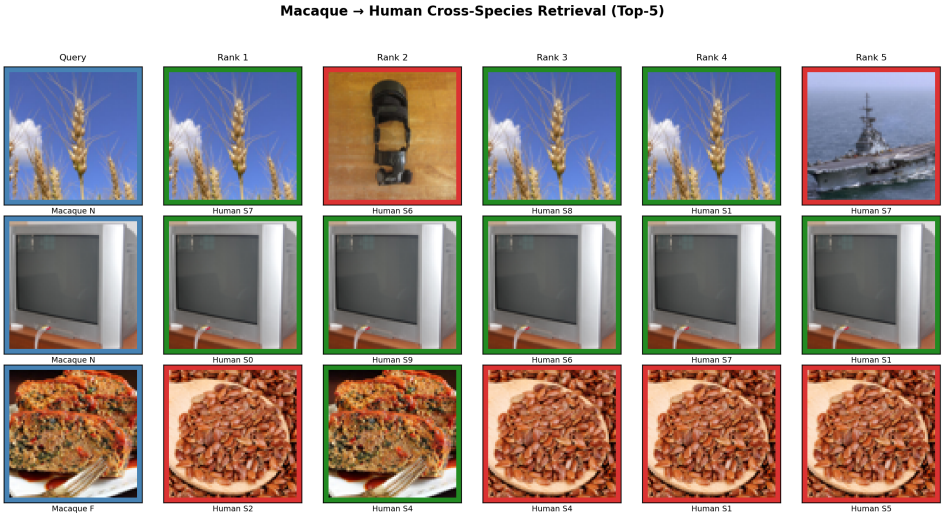

Figure 6: Qualitative examples of zero-shot cross-modal retrieval. The query column displays the visual stimulus presented to the macaque subject. The retrieved columns display the corresponding stimuli associated with the retrieved human EEG signals. The specific subject IDs are noted below each image. Green borders indicate exact ground-truth matches. Red borders indicate incorrect retrievals.