TEMPORAL CONTRASTIVE ALIGNMENT OF EEG AND NATURALISTIC VIDEO ONLY BEATS LINEAR DECODING ON IN-SAMPLE STIMULI

Yuanyuan Yao[1], *Simon Geirnaert*[1,2], *Alexander Bertrand*[1]

[1]KU Leuven, Department of Electrical Engineering (ESAT),
STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Belgium
[2]KU Leuven, Department of Neurosciences, Research Group ExpORL, Belgium

ABSTRACT

Objective. Neural tracking, the temporal alignment between (features of) a continuous stimulus and the elicited brain response, enables decoding attention from electroencephalography (EEG). Heavily studied for speech and music, it remains largely unexplored for naturalistic video, relying on linear canonical correlation analysis (CCA). We investigate whether a nonlinear temporal contrastive model can strengthen EEG-video alignment, and examine the physiological and practical basis of what it learns. *Approach.* We pair a lightweight EEG encoder with a projection of VideoMAE features, training both branches via a correlation-based InfoNCE objective. The model is benchmarked against CCA under three generalization regimes. To probe the basis and utility of the learned representations, we introduce a sham control with circularly shifted video features, a zero-shot transfer test to selective attention decoding on superimposed videos, and an evaluation against a group-averaged EEG reference. *Main results.* The model outperforms CCA on unseen subjects, matches CCA on held-out segments of seen videos, but falls behind on entirely novel videos. A sham control shows that this high accuracy on unseen subjects does not certify physically meaningful alignment: a model retrained on circularly shifted video features, with genuine temporal alignment destroyed, still reaches comparable accuracy on seen stimuli. Yet the coupling stays useful: on a zero-shot selective-attention decoding task with seen stimuli, the model still beats CCA, since decoding needs only a strong statistical coupling, regardless of whether it is physiologically meaningful. Finally, against a group-averaged EEG reference, genuine, generalizable coupling is nonetheless learned, only masked in single-subject EEG by measurement noise and subject-specific response differences. *Significance.* These results demonstrate that neural tracking models should be evaluated by holding out subjects and stimuli jointly. Progress toward open-set generalization hinges on data scale, stimulus diversity, and raising the single-trial signal-to-noise ratio.

***Index Terms*—** contrastive learning, neural tracking, EEG-video alignment

1. INTRODUCTION

During the processing of continuous sensory stimuli such as natural speech or dynamic visual scenes, cortical neural activity synchronizes with the fluctuations of the input, creating a temporal alignment between the stimulus and the elicited electrophysiological response [1, 2]. This phenomenon, commonly known as neural tracking, is modulated by attention, where attention acts as an amplifier leading to stronger temporal coupling [3, 4, 5]. Because of this differential modulation, it is possible to decode the specific target of a person's attention and their overall level of attentional engagement directly from their brain activity. This activity is typically recorded using electroencephalography (EEG) due to its high temporal resolution and non-invasiveness. A key focus in recent years has been on translating proof-of-concept attention decoding into real-world applications, such as neuro-steered hearing aids and attention monitoring systems for clinical and educational settings [2, 6].

However, such translation is non-trivial. Naturalistic stimuli, such as speech, music, or video, are rich and highly complex, posing significant challenges for finding relevant features that robustly correlate with the EEG response [7, 8]. Moreover, to achieve true ecological validity, real-world applications require single-trial decoding. In these scenarios, the model must extract reliable information from inherently noisy EEG signals without the luxury of averaging across multiple trials [6, 9]. Together, these constraints motivate investigating whether more expressive decoding models can form stronger temporal stimulus-EEG alignments and generalize well across diverse subjects and stimuli.

Extensive research in the speech domain has historically focused on linear methods, primarily linear stimulus reconstruction (LSR) and canonical correlation analysis (CCA).

This research is funded by the Research Foundation - Flanders (FWO) project No G081722N, junior postdoctoral fellowship fundamental research of the FWO (for S. Geirnaert, No. 1242524N), Internal Funds KU Leuven (projects IDN/23/006, C14/25/108, and C3/25/107), and the Flemish Government (AI Research Program).
All authors are also affiliated with Leuven.AI - KU Leuven institute for AI, Belgium.

LSR learns a linear mapping from EEG to the stimulus [7, 9, 10], whereas CCA finds linear projections of both modalities that are maximally correlated [11, 12]. Although data-efficient and interpretable, these models have limited capacity to capture nonlinear relationships between EEG and complex naturalistic stimuli. An early step was taken by de Taillez et al. [13], who extended the LSR framework with hidden layers, nonlinear activations, and dropout, representing a minimal, meaningful, but ultimately non-replicable departure from purely linear models [6]. Ciccarelli et al. [14] benchmarked this nonlinear LSR variant alongside a convolutional neural network (CNN)-based model, showing that the CNN significantly outperforms both LSR and the nonlinear extension, a finding that was also non-replicable [6]. Subsequent research explored a broader range of nonlinear architectures for EEG-stimulus alignment, including recurrent networks [15, 16], autoencoders [17, 18], and transformers [19, 20]. In parallel, stimulus representations evolved from hand-crafted features (e.g., speech envelopes, Mel spectrograms) [14, 15] to deep-learned semantic embeddings (e.g., Wav2Vec 2.0, HuBERT) [21, 22]. Orthogonal to the choice of feature, the way the two modalities are related has also changed: instead of regressing one modality onto the other, two-branch models map EEG and stimulus into a shared embedding space with a trainable encoder on each path, jointly optimized by a match-mismatch [23, 24] or, more recently, a contrastive [17, 20] objective, so that both representations are shaped by the alignment task itself. While these advances have improved decoding accuracy in certain conditions, their gains are fragile: higher-capacity models are more susceptible to exploiting dataset-specific shortcuts [25, 26, 27, 28], and performance often fails to generalize to new subjects, unseen stimuli, or different recording setups [6]. Whether these improvements reflect genuine advances in stimulus-EEG alignment, or merely better in-distribution fitting, therefore remains an open question.

In the visual domain, the exploration of EEG alignment with naturalistic video stimuli remains limited. Early work by Dmochowski et al. [8] tested a variety of low-level video features and identified two (temporal contrast and optical flow) that could be correlated with EEG using CCA. However, Yao et al. [29] later demonstrated that these significant correlations were primarily driven by shot cuts (sudden changes in scene or camera angle). These cuts introduce abrupt discontinuities that create artificial peaks in the extracted features while simultaneously eliciting strong visually evoked responses in the EEG. When tested on continuous, single-shot videos, these correlations became non-significant. To address this, the authors proposed object-based versions of the temporal contrast and optical flow features, successfully recovering significant correlations even in single-shot videos. To date, these remain the only features proven effective under such single-shot condition, and were subsequently utilized in a study applying CCA to decode attended videos in a superimposed-video paradigm [30]. It is worth noting that a parallel line of recent research focuses on classifying or reconstructing visual stimuli from EEG using deep generative models [31, 32, 33]. However, these approaches typically operate on static images or short, discrete video clips. They rely heavily on decoding overarching semantic information to perform classification and reconstruction, which fundamentally differs from the continuous, stimulus-EEG temporal alignment problem focused on in this paper.

In this work, inspired by the advances in auditory EEG-stimulus alignment, we investigate for the first time whether introducing nonlinearity with neural networks and using more advanced deep video representations can improve temporal alignment between EEG and naturalistic video stimuli, compared to the CCA baseline. Aware of the pitfalls of data leakage and overfitting, the model is trained and evaluated under three generalization regimes: cross-subject (leave-one-subject-out), within-stimulus (leave-video-segment-out), and novel-stimulus (leave-one-video-out). For each regime, we dive into the underlying causes of success or failure, and identify the key constraints that limit performance. The paper is organized as follows. Section 2 presents the datasets, the pre-trained video feature, the temporal contrastive model, and the evaluation protocol. Section 3 benchmarks the model against CCA under the three generalization regimes and on a downstream attention-decoding task. Section 4 examines what the model actually learns and what limits it, and Section 5 concludes.

2. METHODS

2.1. Data

In this study, we worked with two publicly available continuous Video-EEG datasets collected at KU Leuven [34, 35]. Dataset I is our primary dataset for all model training and the three generalization regimes, while Dataset II provides a separate, held-out selective-attention test. These evaluation setups are detailed in Section 2.4.3 and Section 2.4.4, respectively. In both datasets, participants watched naturalistic videos under free-viewing conditions, with no fixation constraint, to preserve ecological validity. The corresponding EEG data were acquired using a 64-channel BioSemi ActiveTwo system at a sampling rate of 2048 Hz. To record eye movements, four electrooculogram (EOG) electrodes were placed around the participants' eyes. The audio of all videos across both datasets was muted to prevent auditory-evoked neural responses from confounding the visual processing analysis.

In both datasets, the EEG was preprocessed following the same procedure as the studies that released them [29, 30], comprising standard steps such as bad channel interpolation, rereferencing to the average of all channels, high-pass filtering with a cutoff frequency of 0.5 Hz, removing powerline noise at 50 Hz, and eye-artifact removal by regressing the

EOG channels out of the EEG channels. For the present analysis, the preprocessed EEG was downsampled to 30 Hz for the CCA baseline, matching the video frame rate [29, 30], and to 128 Hz for the contrastive model, which learns its own temporal filters and downsamples internally to the video-feature rate (Section 2.3.2).

2.1.1. Dataset I: Video-EEG encoding-decoding dataset

The first dataset [34] was designed to investigate the temporal correlations between natural, single-shot videos and EEG signals during passive viewing. It includes two distinct subsets based on the visual complexity of the stimuli:

- Single-Shot Subset: 19 participants passively viewed 13 continuous, single-shot videos (totaling approximately 63 minutes of data per participant). These videos depicted a single individual performing actions such as dancing, acrobatics, and mime.
- Complex Scene Subset: A subset of 9 participants viewed an additional 24-minute continuous clip from a television episode. This stimulus introduced more complex visual dynamics, including shot cuts, multiple moving objects, and complex interactions.

Since shot cuts introduce abrupt discontinuities that elicit strong event-related-potential-like responses in the EEG that may dominate the EEG-video coupling and overshadow the subtler neural tracking of the visual dynamics of interest [29], we focused our analysis on the single-shot subset (19 participants $\times$ 63 minutes of data).

2.1.2. Dataset II: Selective Visual Attention Decoding Dataset

The second dataset [35] was designed to evaluate EEG-based decoding of selective visual attention when participants were presented with superimposed video streams. This dataset also includes data from 19 participants (a separate cohort from Dataset I). The experimental paradigm used 14 single-shot videos (13 of which were inherited from Dataset I), combining them into 7 superimposed video pairs. Each trial was structured to guide and test sustained selective attention:

- Single-Object Phase (0-120s): Participants were presented with a single video stream and instructed to pay attention to the person in the video.
- Transition (120-122s): A two-second fade-in introduced a second, overlapping video stream.
- Superimposed-Object Phase (122s-End, 165s on average): Both videos were fully superimposed. Participants were instructed to maintain their visual attention solely on the original target from the first phase, actively ignoring the superimposed distractor.

Each video pair was viewed twice with the target and distractor roles reversed, leading to a total of 14 trials per participant. The dataset comprises 19 participants $\times$ 28 minutes of single-object viewing and 19 participants $\times$ 38 minutes of superimposed viewing.

2.2. Video Feature Extraction

Representation learning for video is an extensively studied topic in computer vision, with a wide variety of architectures proposed in the literature. For neural tracking, the key requirement is a representation that captures temporal dynamics rather than one optimized primarily for semantic classification. Previous work established that motion-based features (object-based optical flow) are effective for EEG-video alignment [29, 30], motivating a step toward richer spatiotemporal representations that go beyond a rough 'all-pixel' average motion feature. To this end, we chose VideoMAE [36], a masked autoencoder architecture that learns representations by reconstructing largely masked video patches. The masking applies both spatially and temporally, forcing the model to learn spatiotemporal features that capture the underlying dynamics of the video, making it a natural fit for temporal alignment tasks.

All videos were first standardized to a common format: resampled to 16 Hz and resized to 224×224 pixels. To preserve the original aspect ratio, frames were rescaled so that the longer side fit within 224 pixels and then symmetrically zero-padded to a square 224×224 canvas. These preprocessing choices match VideoMAE's original training configuration, which uses a 16-frame vision transformer base and an input resolution of 224×224. Each preprocessed video was partitioned into consecutive, non-overlapping 1-second windows of 16 frames. The model was loaded with a publicly released checkpoint on Hugging Face (MCG-NJU/videomae-base). Each input video segment yields 8 temporal token positions (16 frames / tubelet size 2) and $14 \times 14 = 196$ spatial patches per temporal position, for a total of 1568 tokens, each represented by a 768-dimensional embedding. The full hierarchy of representations of a 1-second video window includes the patch-embedding output together with the outputs of all 12 transformer blocks, giving 13 layers of hidden states: $(13, 8, 196, 768)$. The high dimensionality of this representation remains a challenge for downstream EEG alignment, so we applied a pooling step to reduce the dimensionality while preserving the temporal axis.

2.2.1. Attention-weighted spatial pooling

We pooled spatially, layer by layer: for each of the 13 layers, we collapsed its 196 spatial patches into a single feature vector per temporal position, leaving the layer axis untouched. Rather than uniform spatial averaging, we used an attention-weighted pooling scheme driven by the model's own attention, so that spatially salient regions contribute more to the

resulting representation. For a given layer, the multi-head attention tensor was averaged over heads, and for each token we computed the mean attention it receives across all query positions. These per-token saliencies (8 temporal $\times$ 196 spatial) were normalized with a softmax over the spatial dimension, which concentrates the pooling weights on the most attended regions. Figure 1 shows the attention maps of an example video for all 12 layers of the first window, next to the corresponding video frame, where the temporal axis is collapsed by averaging over the 8 positions (for display only). The attention maps reveal that early layers concentrate on the background and actively suppress the performer. This pattern inverts around layer 4, where the model begins to favor the performer and high-contrast structures, such as the light reflections on the floor. We observed that this progression is consistent across different time windows and distinct videos, though the onset of this inversion occasionally shifts one layer later.

Within each layer, the hidden states were then reduced to a weighted sum over the 196 spatial patches using these weights, yielding one 768-dimensional vector per temporal position and per layer. The patch-embedding layer, which has no associated attention map, was pooled using the attention weights of the first transformer block as the nearest available proxy. After pooling, each 1-second window was represented by a feature tensor of shape $(13, 8, 768)$, that is, layers $\times$ temporal positions $\times$ features. Concatenating the features across a T-second video, we obtained a feature tensor of shape $(T, 13, 8, 768)$.

2.2.2. Temporal differencing

Before using these features for EEG alignment, we asked whether they primarily reflect the video's temporal dynamics or a static, position-dependent component. We inspected the pooled features of each video via principal component analysis (PCA): for each layer, we flattened the temporal axis to a single dimension of size $8T$ (for a single video) and projected the 768-dimensional features onto their first two principal components. The resulting scatter plots revealed that the features of each video collapsed into eight tight clusters, consistent with the eight temporal positions within each 1-second window. Figure 2a (left) shows one such case. To further characterize how feature similarity depends on temporal separation, we computed a normalized autocorrelation function from the $8T \times 8T$ pairwise cosine-similarity matrix comparing each token of a video with every other token. The normalized autocorrelation function is the mean similarity as a function of temporal lag Δ, obtained by averaging the entries on the Δ-th off-diagonal of the similarity matrix. The function (Figure 2b, blue) shows a strong periodic pattern with period exactly one second: tokens at the same temporal position across different 1-second windows are highly similar. Moreover, this peak similarity shows only a marginal decrease as temporal distance increases. Both the clustering and the periodic autocorrelation were present in every transformer layer and in every video. This pattern is consistent with the influence of VideoMAE's positional embeddings: each window is encoded independently, but the token at a given temporal position carries the same fixed positional component in every window. This shared positional component, likely compounded by static background and semantic features, dominates the representation. Consequently, it drives a strong similarity between temporally aligned tokens, obscuring the genuine temporal variation we aim to study.

To remove this, we applied a first-order temporal difference across windows at each fixed temporal position, replacing each window's features by their difference from the preceding window:

$$\mathbf{f}'_m = \mathbf{f}_m - \mathbf{f}_{m-1} \quad \text{for } m = 2, \ldots, T, \tag{1}$$

where $\mathbf{f}_m$ is the feature tensor of the m-th window with shape $(13, 8, 768)$, and $\mathbf{f}'_m$ is the differenced feature. Because the position-dependent offset is constant across windows for a given temporal slot, the subtraction cancels this constant component while preserving the dynamic component of the signal[1]. Figure 2 confirms that differencing removes the positional component. In the PCA (Figure 2a, right), the eight clusters disappear, and the variance explained by the first two principal components drops from 51.0% to 15.1%. In the autocorrelation function (Figure 2b, orange), the periodic peaks vanish, and the similarity instead decays to around zero. The resulting differenced features, flattened to an 8 Hz time series, have shape $(8T, 13, 768)$.

2.3. Temporal Contrastive Model

To go beyond linear CCA while remaining tractable on a small dataset, we adopted a simple two-branch architecture trained with a contrastive objective. One branch processed T_w-second EEG windows; the other processed T_w-second video feature windows. Following CCA's use of temporal correlation as the similarity measure, both branches produced one scalar per temporal position, and similarity was computed as the Pearson correlation between the two output sequences. The overall architecture is illustrated in Figure 3.

2.3.1. Video branch

Since spatiotemporal representation learning was delegated to the pretrained VideoMAE, the video branch consists of a lightweight projection head.

[1] Because VideoMAE uses joint space-time attention, token i integrates information from the entire 1-second window, meaning the eight temporal slots do not strictly correspond to eight isolated 125 ms intervals. However, our differencing step does not assume this one-to-one correspondence. It solely requires the slot-specific nuisance component to be identical across windows, a condition guaranteed by the frozen encoder's repeated positional embeddings and empirically confirmed by Figure 2.

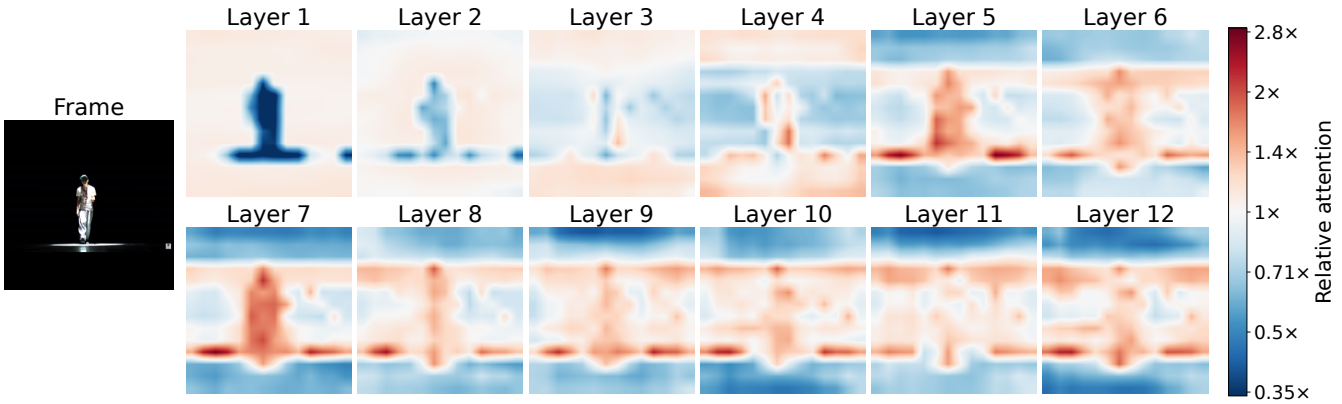
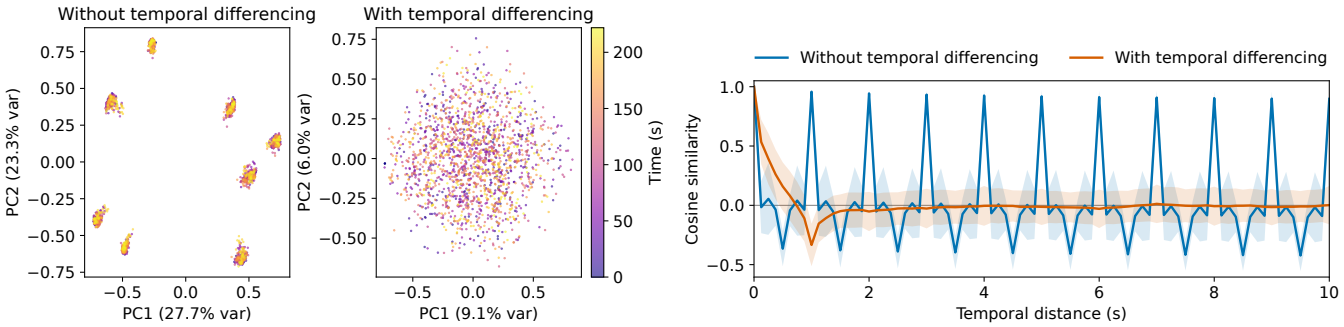


Fig. 1: Layer-wise received attention across all 12 layers of VideoMAE for the first 1-second window of an example video. Left: the middle frame of the window, shown for reference. Right: received attention per spatial patch for each layer, averaged over heads, query tokens, and the 8 temporal positions, and upsampled from the 14×14 patch grid to the frame resolution. Colors show attention relative to a uniform baseline ($1\times$) on a shared $\log_2$ scale: blue represents below-uniform attention, and red represents above-uniform attention.



(a) PCA scatter

(b) Normalized autocorrelation function

Fig. 2: VideoMAE features of a representative video without and with temporal differencing. (a) PCA of the 8-Hz layer-12 features: each point is one time step projected onto the first two principal components (variance explained in the axis labels), colored by its time in the video. Without differencing (left), the features form eight clusters corresponding to the eight temporal positions within the 1-second window; with differencing (right), the clustering structure disappears. (b) Mean cosine similarity between feature vectors versus their temporal separation. Without differencing (blue), the similarity peaks at every integer-second lag; with differencing (orange), the periodic peaks disappear and the similarity decays to around zero.

Denote the video features of a given window by $\mathbf{F}_w \in \mathbb{R}^{8T_w \times 13 \times 768}$. As observed in Figure 1, the model transitions from background to foreground focus around layer 4. We specifically target these middle layers (transformer layers 4–7), as they capture rich, object-level information before the network's features become excessively abstract or specialized in the deepest layers. We therefore applied a learnable aggregation over $\mathcal{L} \in \{4, 5, 6, 7\}$, the only step that mixes layers, to produce a single feature vector per temporal position:

$$\bar{\mathbf{F}}_w[t, :] = \sum_{\ell \in \mathcal{L}} a_\ell \mathbf{F}_w[t, \ell, :] \in \mathbb{R}^d, \quad t = 1, \ldots, 8T_w, \quad (2)$$

where a_ℓ are softmax-normalized learnable weights over the layer dimension, and $d = 768$ is the feature dimension. The aggregated features were then instance-normalized per window (each feature dimension zero-mean, unit-variance over the $8T_w$ timesteps) to account for scale variation in the pretrained features across videos, and passed through a multilayer perceptron (MLP) h applied at each timestep:

$$h(\cdot) = \text{Linear}_{d \to d_h} \to \text{GELU} \to \text{IN} \to \text{Linear}_{d_h \to 1}, \quad (3)$$

where d_h is the projection width, IN denotes instance normalization over time, and dropout is applied before the final linear layer. The output of the video branch was a scalar sequence of length $8T_w$, denoted $\mathbf{z}_w^{\text{vid}}$ with elements z_t^{vid} for $t = 1, \ldots, 8T_w$:

$$\mathbf{z}_w^{\text{vid}} = h\left(\text{IN}(\bar{\mathbf{F}}_w)\right). \quad (4)$$

The only trainable parameters of the video branch were the layer aggregation weights a_ℓ and the MLP weights. The VideoMAE encoder was kept frozen throughout training.

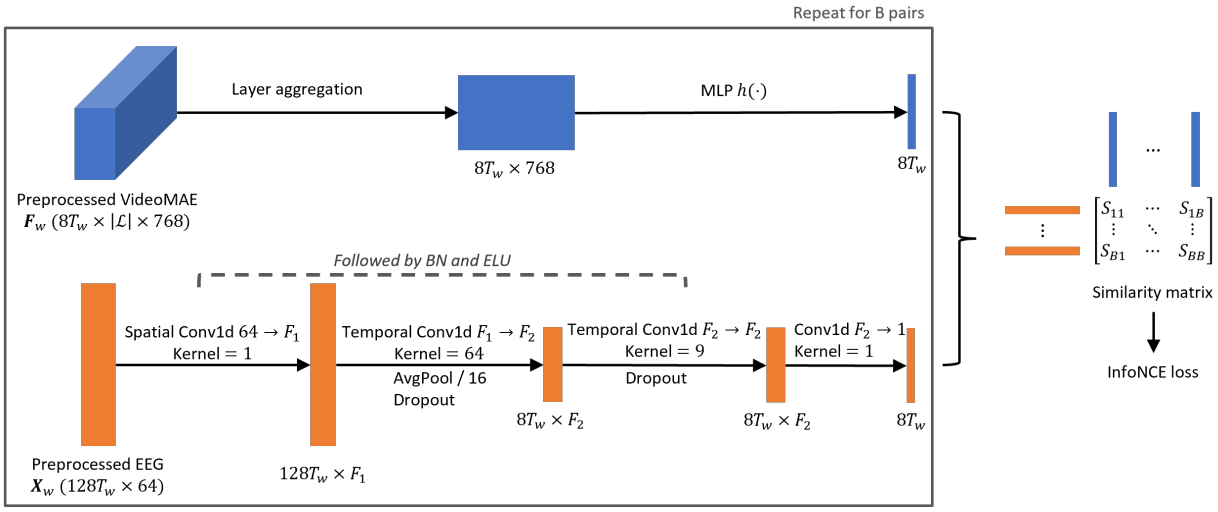


Fig. 3: Architecture of the temporal contrastive model. The EEG branch (orange) applies spatial filtering followed by temporal convolutions to produce a scalar sequence at 8 Hz. The video branch (blue) aggregates pretrained VideoMAE features over the middle transformer layers and projects them to a scalar sequence via a lightweight MLP. Both outputs are zero-meaned and L_2-normalized such that their inner products become equivalent to a Pearson correlation, which populates the similarity matrix. The model is trained with a symmetric InfoNCE loss.

2.3.2. *EEG branch*

The EEG branch accepted 128 Hz EEG windows $\mathbf{X}_w \in \mathbb{R}^{128T_w \times C}$, where $C = 64$ is the number of EEG channels. Inspired by EEGNet [37], a spatial convolution first collapsed the channel dimension, mixing the C electrodes at each time sample with F_1 learned spatial filters:

$$\underbrace{(128T_w, C)}_{\text{time}\times\text{channels}} \xrightarrow{\text{spatial Conv1d } C\to F_1,\ k=1} \underbrace{(128T_w, F_1)}_{\text{time}\times\text{spatial filt.}}. \quad (5)$$

With the channel axis collapsed, the remaining convolutions were temporal, operating along time. A 0.5-second kernel ($k = 64$) extracted temporal features, then average pooling with stride 16 downsampled the time axis from 128 Hz to 8 Hz, aligning EEG to the video time grid:

$$\underbrace{(128T_w, F_1)}_{\text{time}\times\text{spatial filt.}} \xrightarrow[\text{AvgPool } /16]{\text{temporal Conv1d } F_1\to F_2,\ k=64} \underbrace{(8T_w, F_2)}_{\text{time}\times\text{temporal filt.}}. \quad (6)$$

A second filter ($k = 9$) then operated on this 8 Hz grid to refine the temporal features, followed by a layer that mixed the F_2 filter responses into one scalar per timestep:

$$\underbrace{(8T_w, F_2)}_{\text{time}\times\text{temporal filt.}} \xrightarrow[\text{Conv1d } F_2\to 1,\ k=1]{\text{temporal Conv1d } F_2\to F_2,\ k=9} \underbrace{(8T_w)}_{\text{scalar/step}}. \quad (7)$$

All temporal convolutions used zero-padding of $\lfloor k/2 \rfloor$ to preserve the time axis, so the stride-16 average pooling mapped the 128 Hz sequence onto the same 8 Hz grid as the video features. Each convolution (except the final projection) was followed by batch normalization and ELU activation, with dropout after the pooling and after the second temporal filter. The EEG embedding output is denoted $\mathbf{z}_w^{\text{eeg}}$.

2.3.3. *Contrastive loss*

Both branch outputs were zero-meaned and L_2-normalized over time, so their inner product is equivalent to the Pearson correlation. For a batch of B matched pairs $\{(\mathbf{z}_i^{\text{eeg}}, \mathbf{z}_i^{\text{vid}})\}_{i=1}^{B}$, define the similarity matrix

$$S_{ij} = \frac{\langle \mathbf{z}_i^{\text{eeg}}, \mathbf{z}_j^{\text{vid}} \rangle}{\tau} = \frac{\rho_{ij}}{\tau}, \quad (8)$$

with temperature τ. The matched pairs lie on the diagonal, and we minimized the symmetric InfoNCE loss:

$$\text{Loss} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{e^{S_{ii}}}{\sum_j e^{S_{ij}}} + \log \frac{e^{S_{ii}}}{\sum_j e^{S_{ji}}} \right]. \quad (9)$$

This trains the matched EEG-video pair (on the diagonal) to score above all mismatched pairs in the batch, directly optimizing the match-mismatch discrimination used for evaluation (Section 2.4.3).

2.4. Training and Evaluation

2.4.1. *Batch construction and window augmentation*

We sampled T_w-second windows from the training set to form a batch of B matched EEG-video pairs. We did not assign a fixed start position for each video, as this would create a discrete set of (video, window-index) identities. A model could then memorize a unique code for each such unit without learning any genuine temporal structure, a form of table lookup that would fail on unseen stimuli. Instead, the B windows were drawn sequentially on-the-fly, with a running record of already-chosen start times per video. A candidate start t_0 was

valid for a given video only if $|t_0 - t_0^*| \geq \Delta_t$ for every previously chosen start t_0^* on that video, where $\Delta_t = \lceil 0.2T_w \rceil$. This capped the overlap between any two same-video windows in a batch, keeping all $B(B-1)$ off-diagonal similarities as unambiguous negatives. Start position was drawn uniformly within the available non-conflicting intervals of the chosen video. At each step the video was sampled proportional to the number of its remaining valid integer start positions, so videos with no room left dropped out automatically.

To encourage the model to learn representations that generalize across timescales, we applied a window augmentation strategy. For each batch, we randomly sampled a window length T_w from a uniform distribution over $\{5, 10, 15, 20, 25, 30\}$ seconds. For the chosen (video, start, T_w), a subject was drawn uniformly from the training pool and the EEG window was sliced and z-scored per channel. The video features were sliced from the pre-extracted tensor at the same integer second boundaries.

We will investigate three regimes, each with their own train, test, and validation split (see Section 2.4.3). In validation, we used fixed $T_w = 30$-second windows extracted with a 5-second stride, to avoid the randomness of start positions affecting the evaluation. In test, we used the same 5-second stride but additionally evaluated the model on multiple window lengths to assess its robustness to timescale variation.

2.4.2. Training

Parameters were optimized using Adam, with learning rate set under a cosine-annealing schedule. The optimization hyperparameters were tuned once per evaluation regime and reserved validation set (Section 2.4.3), not per test fold, using Optuna, choosing the best trial with the lowest validation loss out of 50 trials with epoch-level median pruning [38]. The search space is summarized in Table 1. Each trial was trained for up to 75 epochs of 60 batches each, with early stopping on the validation loss (patience of 10 epochs).

2.4.3. Evaluation under three generalization regimes

We separated three different generalization questions:

1. Leave-one-subject-out (LOSO): Let $\mathcal{S}$ be the subject set ($|\mathcal{S}| = N_s$). Subject s_0 is reserved as the validation set. Each of the remaining $N_s - 1$ subjects is held out in turn as the test subject s_i, while the other $N_s - 2$ subjects are used for training. All videos appear in both splits. This scheme tests cross-subject generalization.

2. Leave-video-segment-out (LVSO): Partition each video into $N_f = 8$ contiguous, non-overlapping segments. Fold f_0 of each video is reserved as the validation set, while each of the remaining $N_f - 1$ folds is held out in turn as the test fold f_i. All subjects and all videos appear in both splits. This regime tests within-manifold generalization, i.e., to unseen parts of seen videos.

3. Leave-one-video-out (LOVO): Let $\mathcal{V}$ be the video set ($|\mathcal{V}| = N_v$). Video v_0 is reserved as the validation set. Each of the remaining $N_v - 1$ videos is held out in turn as the test video v_i, while the other $N_v - 2$ videos are used for training. All subjects appear in both splits. This regime tests off-manifold generalization, i.e., to unseen videos.

The choice of the reserved subject s_0, fold f_0, and video v_0 affects the hyperparameter tuning and early stopping. Therefore, per regime we repeated the training and evaluation with three different reserved choices, which we will refer to as 'hyperparameter configurations'. Results for each test subject, fold, or video were averaged over the three hyperparameter configurations in which it was tested. All three regimes operate exclusively on Dataset I.

To evaluate the model, we used pairwise match-mismatch (MM) accuracy, a standard metric in neural tracking [12]. Given a video window and two EEG windows, one of which is the true match and the other an imposter, the model needs to identify which EEG window corresponds to the video. A dual formulation is to ask, given an EEG window and two video windows, which video is the true match. The model gives a similarity score for each pair, as defined in Equation 8. The pairwise accuracy is the fraction of comparisons in which the matched similarity strictly exceeds the mismatched one. Formally:

$$\text{acc} = \frac{\sum_i \sum_{j \neq i} [I(S_{ii} > S_{ij}) + I(S_{ii} > S_{ji})]}{2B(B-1)}, \qquad (10)$$

where $I(\cdot)$ is the indicator function. All windows in an evaluation batch, the true match and the imposters alike, were drawn from the held-out set, so no training data enters the comparison. The batch size was set to $B = 8$ in validation and test, since LVSO's and LOVO's held-out sets are too short to accommodate larger batches. To give a robust estimate of the model's performance, we repeated the evaluation 5 times with different random seeds and reported the average accuracy.

The chance performance was estimated with a permutation test using the trained decoders. The EEG was circularly shifted relative to the stimulus features per video, so that the alignment between the two modalities was destroyed while preserving the temporal structure of each. The same evaluation procedure was applied to the permuted data. After 1000 permutations, the significance level is the 95th percentile of the null accuracies across folds and subjects, computed separately for each decoder and window length.

2.4.4. Downstream selective attention decoding

The three regimes above probe generalization within Dataset I. To test whether the learned alignment also transfers to a qualitatively different task on a separate dataset, we additionally evaluated the model on the selective visual attention paradigm

Table 1: Hyperparameter search space. Learning rate, weight decay, and temperature were sampled log-uniformly over the stated interval; the remaining parameters were sampled from the listed discrete sets. Tuning used Optuna (50 trials per study), minimizing the validation loss on a reserved held-out fold.

Hyperparameter	Role	Range / values
Learning rate	Adam step size	$[1 \times 10^{-4},\ 5 \times 10^{-3}]$, log-uniform sampling
Weight decay	Adam L_2 regularization	$[1 \times 10^{-4},\ 1 \times 10^{-1}]$, log-uniform sampling
Temperature τ	InfoNCE loss temperature	$[0.05,\ 0.5]$, log-uniform sampling
F_1	EEG encoder spatial filters	$\{4, 8, 16\}$
D	EEG encoder depth multiplier, $F_2 = F_1 \times D$	$\{1, 2\}$
EEG dropout	Dropout in the EEG encoder	$\{0.3, 0.4, 0.5, 0.6, 0.7\}$
Projection hidden dim d_h	Width of the video projection MLP	$\{32, 64, 128\}$
Video dropout	Dropout in the video projection head	$\{0.3, 0.4, 0.5, 0.6, 0.7\}$
Batch size	Contrastive batch size	$\{8, 16, 32, 64, 128\}$

of Dataset II, without updating any model parameters on that dataset. For this evaluation, a dedicated model was trained for match-mismatch discrimination on all 19 subjects of Dataset I, with hyperparameters retuned for this setup, and then applied directly to attention decoding on Dataset II. The single-object phase of Dataset II served as the validation set for hyperparameter selection and early stopping. The superimposed-object phase was never seen during training or model selection and constitutes the true evaluation set.

During the superimposed-object phase, two video streams are presented simultaneously and fully overlaid. For a T_w-second EEG window $\mathbf{X}_w$, the model computes a similarity score against each competing stream:

$$s_k = \langle \mathbf{z}_w^{\text{eeg}},\, \mathbf{z}_w^{\text{vid},(k)} \rangle, \quad k \in \{1, 2\}, \tag{11}$$

where $\mathbf{z}_w^{\text{vid},(k)}$ is the normalized video branch output for stream k. The model predicts the attended stream as $\hat{k} = \arg\max_k\, s_k$. Attention decoding accuracy is then the fraction of windows for which $\hat{k}$ matches the ground-truth attended stream.

Of the 14 videos in Dataset II, 13 were inherited from Dataset I. We excluded the one pair containing the single new video, leaving six pairs formed from twelve videos, all of which were present in the training set. Because each pair was viewed twice with the target and distractor roles reversed, every one of these videos appears equally often as the attended and as the unattended stream. The two competing streams in a trial are therefore equally represented in training, and the model holds templates of equal provenance for both. This is important for the interpretation in Section 4.2: above-chance decoding cannot arise from the model scoring a better-learned video above a less familiar one, because neither stream is privileged. The evaluation is thus close to the LOSO regime, except that the task is now attention decoding, which tests whether the learned alignment is modulated by attention and decodable. The significance level was again estimated with a permutation test, this time by circularly permuting the EEG relative to both video streams.

2.5. CCA baseline

We compared the contrastive model against a CCA baseline from previous studies [29, 30]. In those studies, object-based optical-flow features were extracted from 30-Hz videos, the EEG was downsampled to 30 Hz, both modalities were hankelized to incorporate temporal information (15 lags spanning -467 to 0 ms for the video feature and 3 lags spanning -33 to 33 ms for each EEG channel), and were correlated using CCA. The sum of the first two canonical correlation coefficients served as the similarity score for the MM task. We followed the same procedure and parameter settings, with a single deviation: those studies fit a separate CCA per subject, whereas we fit one CCA on the EEG and video features of the subjects in the training split, concatenated along the time axis. To keep the results directly comparable, the CCA baseline was trained and tested under exactly the same three regimes as the contrastive model (Section 2.4.3); the reserved validation set was excluded from fitting but served no further purpose, as CCA requires no hyperparameter tuning or early stopping. It was scored with the same pairwise accuracy metric, batch samplers, and seeds.

3. RESULTS

3.1. Contrastive Model Outperforms CCA Under LOSO Without Exploiting Video Identity

Under LOSO, the contrastive model outperforms the CCA baseline across all tested window lengths. Figure 4 shows the MM accuracies of both methods, averaged over the three hyperparameter configurations with different reserved validation subjects. The mean accuracy of the contrastive model versus CCA is 65.0% versus 58.6%, 70.8% versus 61.9%, and 74.6% versus 64.5% for 10-, 20-, and 30-second windows respectively. Across the individual hyperparameter configurations, the mean accuracy gain of the contrastive model over CCA stays within $5.6\% \sim 6.6\%$, $8.0\% \sim 9.2\%$, and $9.1\% \sim 10.5\%$ for 10-, 20-, and 30-second windows respec-

tively. The narrowness of these ranges shows that the advantage is insensitive to which subject is held out as validation set for hyperparameter tuning. The improvements are statistically significant under a Wilcoxon signed-rank test across the 19 test subjects, with accuracies averaged over hyperparameter configurations ($p < 0.001$ for all window lengths after Holm-Bonferroni correction).

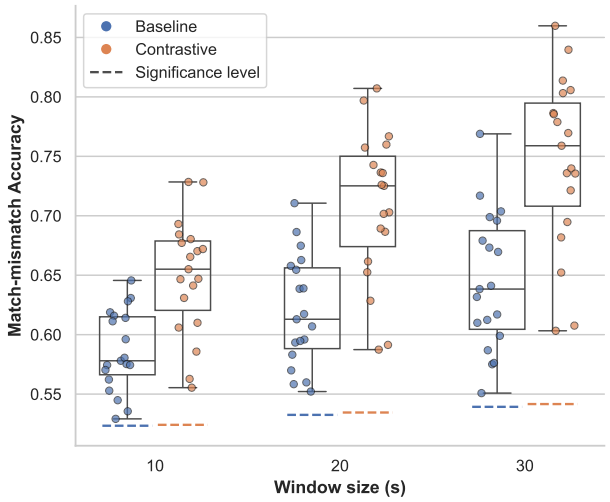


Fig. 4: MM accuracy under LOSO for 10-, 20-, and 30-second windows: contrastive model (orange) vs. CCA baseline (blue). Each dot is one test subject averaged over hyperparameter configurations with different reserved validation subjects; boxes show the median and interquartile range (IQR), and whiskers extend to the most extreme data point within $1.5\times$IQR of the box edges. Dashed lines mark the significance level of each method.

Because all videos appear in both the training and test sets under LOSO, a natural concern is that the contrastive model achieves its advantage by exploiting video identity together with video-dependent EEG cues rather than genuine temporal alignment. In the standard evaluation, batches are formed without any constraint on which video each window comes from, so an imposter is usually (though not always) from a different video than the true match. A model could then separate matched from mismatched pairs using coarse video-level appearance cues in the EEG instead of moment-to-moment coupling. To rule this out, we performed a within-video evaluation in which all samples in a batch are drawn from the same video, so every imposter shares the same video as the true match and video identity carries no discriminative information. Figure 5 compares the within-video and standard evaluations, averaged over hyperparameter configurations. The accuracies barely change across subjects and window lengths, indicating that the model does not rely on shortcuts such as video identity, and this also holds within each hyperparameter configuration individually.

The within-video control rules out the coarsest shortcut, video identity, but not every non-temporal route to high LOSO accuracy. In Section 4.1 we probe a subtler, synthesized correspondence that can inflate LOSO performance without physically meaningful EEG-video coupling.

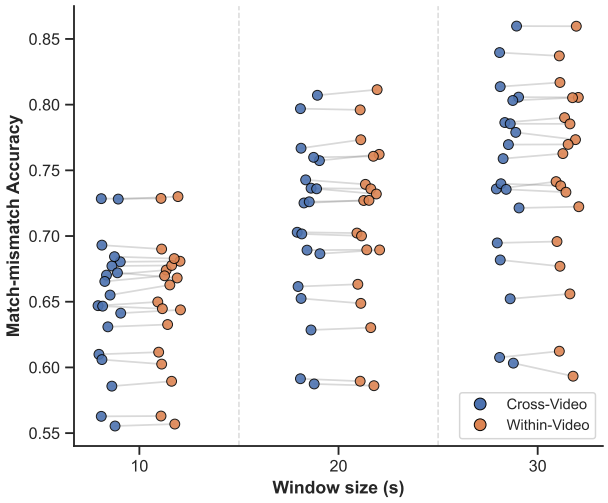


Fig. 5: MM accuracy of the contrastive model under LOSO: standard cross-video evaluation (blue) vs. within-video evaluation (orange), where every imposter in a batch comes from the same video as the true match. Each connected pair of dots is one test subject, averaged over hyperparameter configurations with different reserved validation subjects.

3.2. The Advantage Disappears Under LVSO

Under LVSO, the contrastive model's advantage over CCA disappears. This regime holds out unseen temporal segments of videos that are otherwise present in training, so it probes on-manifold (interpolation) generalization, in contrast to the novel-video extrapolation tested under LOVO (Section 3.3). Figure 6 shows the MM accuracies of both methods, averaged over hyperparameter configurations with different reserved validation folds. The mean accuracy of the contrastive model versus CCA is 58.0% versus 59.7%, 62.0% versus 62.9%, and 65.9% versus 64.7% for 10-, 20-, and 30-second windows respectively. None of these differences is statistically significant under a Wilcoxon signed-rank test across the 8 test folds after Holm-Bonferroni correction ($p_{\text{adj}} = 0.052, r = -0.94$ for 10 s; $p_{\text{adj}} = 0.53, r = -0.39$ for 20 s; $p_{\text{adj}} = 0.53, r = 0.44$ for 30 s, where p_{adj} is the Holm-Bonferroni-corrected p-value and r is the rank-biserial correlation effect size), although at 10 s there is a trend in favor of CCA.

The differences are also small in magnitude. The contrastive model is more sensitive to the choice of validation fold: averaged over folds, the standard deviation (SD) of its accuracy across hyperparameter configurations is 1.2%, 2.1%, and 2.7% for 10-, 20-, and 30-second windows, compared with at most 0.4% for CCA, which involves no hyperparameter tuning or early stopping and varies only because

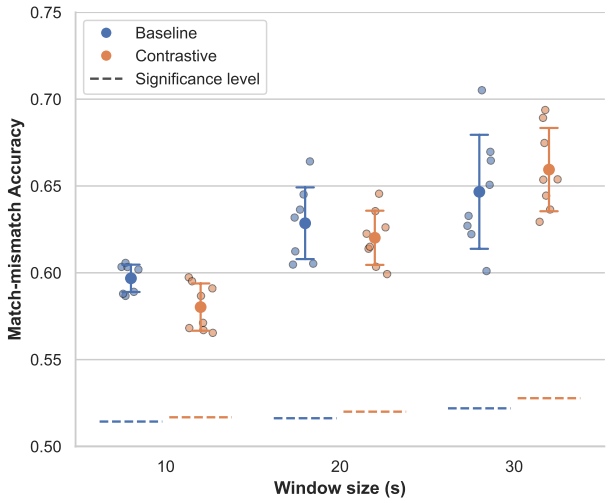


Fig. 6: MM accuracy under LVSO for 10-, 20-, and 30-second windows: contrastive model (orange) vs. CCA baseline (blue). Each dot is one test fold averaged over hyperparameter configurations with different reserved validation folds; the large dot is the mean over folds and the error bars show ± 1 SD. Dashed lines mark the significance level of each method.

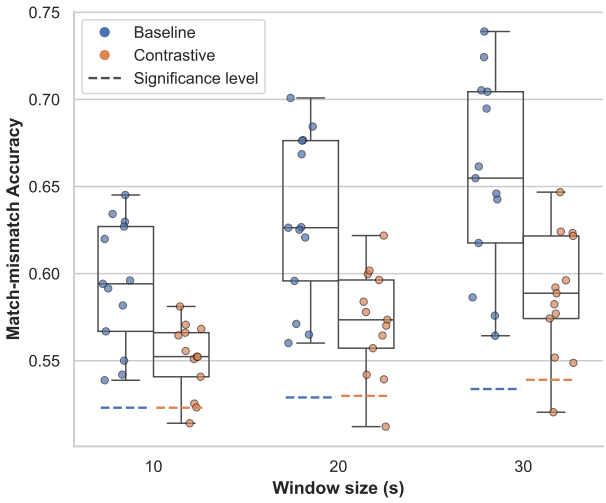


Fig. 7: MM accuracy under LOVO for 10-, 20-, and 30-second windows: contrastive model (orange) vs. CCA baseline (blue). Each dot is one test video averaged over hyperparameter configurations with different reserved validation videos; boxes show the median and IQR, and whiskers extend to the most extreme data point within $1.5\times$IQR of the box edges. Dashed lines mark the significance level of each method.

the reserved fold is excluded from its training data. At 20 and 30 s, the difference between the two methods is smaller than this validation-induced variability, and at 10 s it is comparable. We therefore interpret the two methods as performing on par on unseen video segments, with no systematic advantage for either method.

3.3. The Advantage Reverses Under LOVO

Under LOVO, which requires the model to generalize to entirely unseen videos (an off-manifold, extrapolation setting), the contrastive model's advantage reverses: it consistently underperforms the CCA baseline. Figure 7 shows the MM accuracies of both methods, averaged over hyperparameter configurations with different reserved validation videos. The mean accuracy of the contrastive model versus CCA is 55.1% versus 59.4%, 57.2% versus 63.1%, and 58.8% versus 65.5% for 10-, 20-, and 30-second windows respectively. The gap in favor of CCA is statistically significant under a one-sided Wilcoxon signed-rank test across the 13 test videos ($p_{\text{adj}} = 0.002$ for all window lengths after Holm-Bonferroni correction). The reversal is consistent, but it does not imply that the model has learned no genuine EEG-video coupling. As we show in Section 4.3, the contrastive model still learns coupling that generalizes to unseen videos; what limits its single-trial accuracy on novel stimuli is individual noise[2], not an absence of learned alignment.

3.4. The Advantage Transfers to Downstream Attention Decoding on Seen Videos

Finally, we asked whether the learned alignment transfers to a qualitatively different, application-relevant task: decoding which of two simultaneously presented videos a subject attends to. As described in Section 2.4.4, the model was trained on Dataset I, validated on the single-object phase of Dataset II, and tested on the superimposed-object phase of Dataset II, without any retraining. Dataset II adds one video to those in Dataset I, and we excluded the video pair containing that new video from the evaluation, so all test videos were seen during training and this evaluation is close to the LOSO regime. It is nonetheless harder than MM: both video streams are simultaneously encoded in the EEG, and the model must identify the attended one, which is only possible if the learned EEG-video alignment is modulated by attention. The contrastive model outperforms CCA at all window lengths (Figure 8): $p_{\text{adj}} = 0.014, r = 0.64$ for 10 s, $p_{\text{adj}} = 0.014, r = 0.62$ for 20 s, and $p_{\text{adj}} = 0.002, r = 0.85$ for 30 s.[3] The significance level is higher than in the MM task, because the attended versus unattended comparison is one-to-one, whereas in MM each matched pair is compared against multiple imposters. The larger effective sample size in MM concentrates the null accuracies near chance, lowering

[2]Throughout, *individual noise* denotes everything in a subject's EEG that is not shared across subjects: both measurement noise and individual differences in how subjects respond to the video.

[3]The CCA results are not directly comparable to the attention decoding results reported in [30] on the same superimposed-video paradigm, where the decoders were subject-specific (a separate decoder fit per subject). Also, the significance testing applied here is more rigorous than in [30], as their approach artificially introduced independent segments when estimating the null distribution, resulting in an anti-conservative estimate.

the significance threshold.

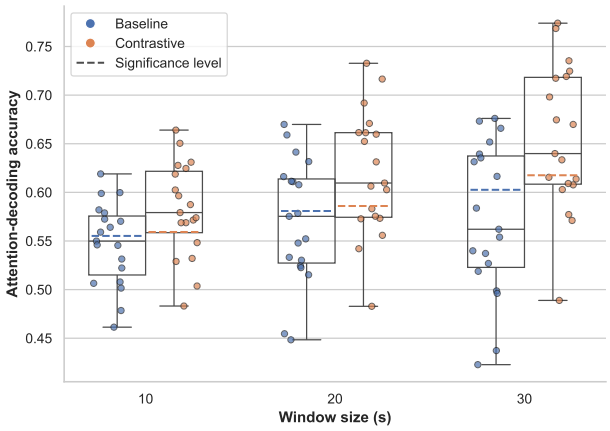
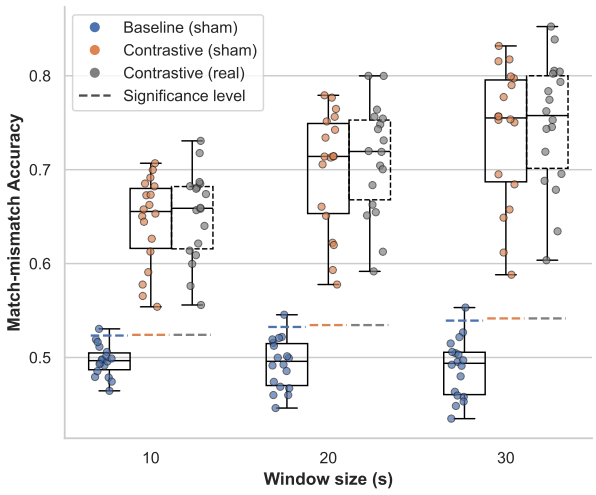


Fig. 8: Attention-decoding accuracy on the superimposed-object phase of Dataset II for 10-, 20-, and 30-second windows: contrastive model (orange) vs. CCA baseline (blue). Models are trained on Dataset I and validated on the single-object phase of Dataset II, without retraining. The one pair containing the video absent from Dataset I is excluded, leaving six pairs formed from twelve videos, all seen during training and each serving equally often as the attended and the unattended stream. Each dot is one subject ($n = 19$); boxes show the median and IQR, and whiskers extend to the most extreme data point within $1.5\times$IQR of the box edges. Dashed lines mark the significance level of each method.

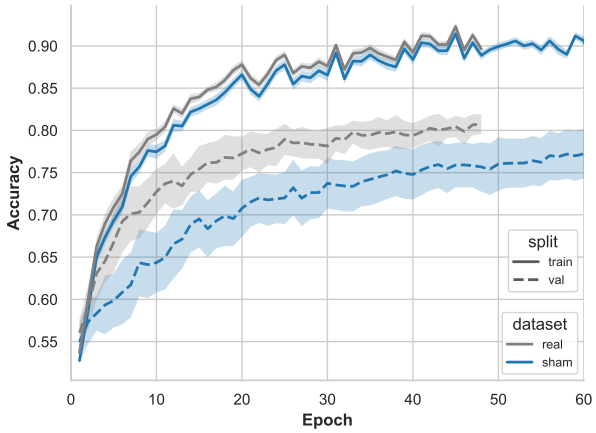
4. DISCUSSION

4.1. High Accuracy Under LOSO Does Not Guarantee Physically Meaningful EEG-Video Alignment

In Section 3.1 we showed that the contrastive model outperforms CCA under LOSO without using video identity as a shortcut, and the window augmentation of Section 2.4.1 further discourages the model from memorizing a discrete code for each (video, window-index) pair. To probe whether a more subtle form of shortcut exists, we constructed a sham dataset in which the genuine temporal alignment between EEG and video is destroyed while all other structure is preserved: we applied a per-video deterministic circular shift ($\geq$ 60 s), the same one for all subjects, to the video features and left the EEG unchanged. The model was retrained on this sham dataset and evaluated under LOSO, reusing the hyperparameter configuration selected on real data with Subject 1 as the reserved validation subject. Architecture, learning rate, batch size, and schedule were therefore identical across the two conditions, which differ only in whether genuine EEG-video alignment is present in the training data. Surprisingly, the sham model reaches accuracy close to that of the corresponding real-data Subject 1 run (Figure 9a), whereas the CCA baseline, given the same sham data, collapses to chance. The contrastive model can therefore reach high LOSO accuracy without genuine temporal alignment in the training pairs, through a route that is unavailable to CCA.



(a) Sham vs. real accuracy



(b) Training and validation curves

Fig. 9: (a) MM accuracy under LOSO for 10-, 20-, and 30-second windows: contrastive model trained on real data (grey, dashed box) versus contrastive model (orange) and CCA baseline (blue) trained on sham data (per-video circularly-shifted video features). Each dot is one test subject; boxes show the median and IQR, whiskers extend to the most extreme data point within $1.5\times$IQR of the box edges, and colored dashed lines mark the significance level of each method. (b) Training (solid) and validation (dashed) accuracy versus epoch for the contrastive model trained on real (grey) and sham (blue) data, averaged over leave-one-subject-out folds; shading is ± 1 SD across folds.

To see how this is possible, it helps to view the training set as a whole rather than as a collection of windows. Even though circular shift breaks the temporal alignment between EEG and video, the model may eventually build a complex, physically arbitrary mapping from the video-feature sequences to the corresponding EEG traces over the entire training set (i.e., across all subjects), which is then applied to each

window. This is what window augmentation cannot resolve, because it only prevents the model from memorizing a fixed set of window slots, not from overfitting or memorizing a function that maps video content to EEG content. This function is unlikely to extrapolate to unseen videos, but it can interpolate to windows within the training videos, and, through the subject-general EEG response component, to EEGs of unseen subjects viewing the same videos. This mirrors the pattern across the three regimes: any stimulus-specific correspondence, whether genuine but confined to the seen video content or synthesized through a physically arbitrary mapping, is available wherever the model only needs to interpolate between different window sizes and window start-times within seen videos, inflating LOSO (Section 3.1) and matching CCA under LVSO (Section 3.2), and it vanishes precisely where the model must extrapolate to novel videos (LOVO, Section 3.3).

The sham experiment exemplifies a synthesized correspondence that can inflate LOSO accuracy without physically meaningful EEG-video coupling, but it does not necessarily imply that the model trained on real data also relies on such a nonsensical correspondence. Deep networks have the capacity to fit pure noise [39]; yet when genuine structure is present, they prioritize learning simple, generalizable patterns before resorting to memorization [40, 41]. Here the physically meaningful coupling is plausibly simpler than the synthesized one. We therefore expect the model trained on real data to exploit the physically meaningful coupling whenever it is learnable.

The learning dynamics are consistent with this hypothesis. Figure 9b plots mean training and validation accuracy across LOSO folds against epoch. The training curves are nearly superimposed: both conditions plateau around 90% (reaching 85% at 17 versus 19 epochs)[4]. The validation curves diverge: the real model reaches 70% in 8 versus 19 epochs for the sham model, and 75% in 15 versus 40 epochs, and its checkpoint is selected 10 epochs earlier on average.

During training, the goal is to reduce the loss as much as possible, and the model will exploit whatever structure is available. Since it does not necessarily require genuine video-to-EEG coupling, the training curves are remarkably similar when training on real versus sham data. When transferring to an unseen subject, only the found patterns of the subject-independent, stimulus-locked EEG component carries over, which is what the validation curves track. This divergence suggests that the real model may exploit the simpler, physically meaningful coupling early in training, whereas the sham model must rely on a more complex, artificial transformation of the sham videos to these general EEG patterns, which takes longer to develop [40, 41].

[4]The epoch-to-epoch fluctuations of the two training curves are highly correlated because the same random seed, shared across conditions and folds, determines the window length sampled at the start of each epoch, and window length dominates per-epoch training accuracy.

In summary, the sham experiment shows that high accuracy under LOSO does not by itself certify that the model has learned physically meaningful EEG-video alignment: a reproducible but synthesized correspondence fit to the seen stimuli is sufficient to explain it. At the same time, the faster emergence of cross-subject transfer in the real model suggests that genuine coupling is present and is exploited when available, which will be further demonstrated in Section 4.3 with actual held-out video data. Part of what is learned may nonetheless be stimulus-specific, either genuine but confined to particular video content, or synthesized and built up in the later stages of training.

4.2. Physically Meaningful Coupling Is Not Always Required: Attention Decoding as Template Matching

High LOSO accuracy may (at least partially) rest on a synthesized correspondence rather than (or in addition to) physically meaningful video-to-EEG coupling. What does this imply for the practical utility of the model? For applications that require generalization to unseen stimuli, the synthesized correspondence is insufficient, as shown in the LOVO setting. But many applications do not require generalization to unseen stimuli, only to unseen subjects. For example, a lecture video is fixed and known at training time, and the model is used to monitor attention of new students watching the same video. In this case, the model can operate as a template matcher: it learns a reproducible template for each video and matches incoming EEG against these templates. Whether the templates encode physically meaningful coupling and/or a synthesized correspondence to the EEG responses is irrelevant, as long as they are reproducible across subjects.

In the downstream attention decoding task of Section 2.4.4, every evaluated video was seen during training, so the task is close to the LOSO regime, and each video serves equally often as the attended and as the unattended stream, so neither competing template is better learned than the other. Attention amplifies EEG responses to the attended stream, and therefore leads to a stronger match with the corresponding video template, regardless of whether the template reflects physically meaningful coupling or a synthesized correspondence. In other words, this is possible because the discriminating information lies on the EEG side rather than in the fidelity of the video mapping. Therefore, the advantage of the contrastive model over CCA on this task demonstrates practical utility and attention modulation, but, consistent with Section 4.1, it does not by itself certify that the model has learned how to extract features from any video that are predictive for the EEG response.

4.3. Generalizable Video-EEG Coupling Is Learned but Masked by Individual Noise

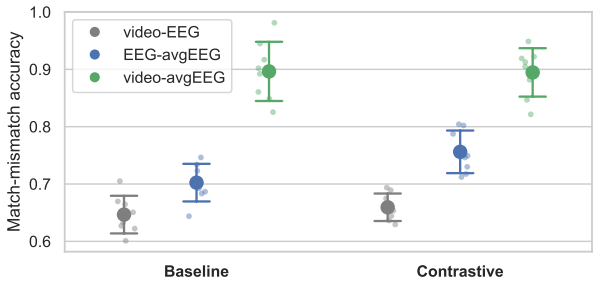
The LVSO tie and the LOVO reversal between CCA and the contrastive learning model raise a question: does the con-

trastive model fail to learn generalizable EEG-video coupling, or does it learn a generalizable coupling that is simply masked by individual noise in single-subject EEG? To separate these, we constructed a cleaner, group-averaged EEG reference, which suppresses individual noise because averaging cancels what is not shared across subjects. The regimes themselves are untouched: the model is fitted exactly as under LVSO or LOVO, so the held-out video segment or video is unseen during training for every subject, and the group-averaged reference enters only at evaluation time. When evaluating subject n on a held-out segment or video, the group-averaged reference is the average of the EEG embeddings of the other $N_s - 1$ subjects computed on that same held-out segment or video, time-aligned with the trial being scored.

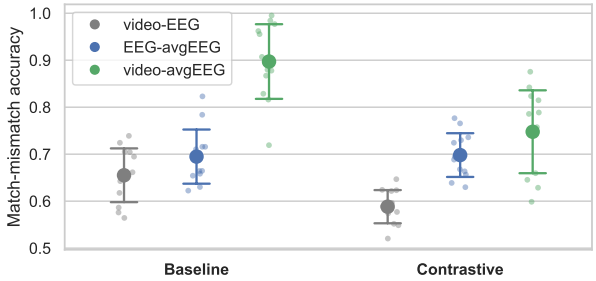
With this reference, we defined two tasks analogous to the standard video-EEG MM task: the video-avgEEG task matches the video embedding against the group-averaged reference EEG embedding, and the EEG-avgEEG task matches subject n's individual EEG embedding against the group-averaged reference (with subject n removed). Because individual noise is largely averaged out of the reference, video-avgEEG estimates how well the learned alignment generalizes once individual noise is removed, and comparing it against EEG-avgEEG isolates whether the limitation lies on the video side or the individual-EEG side.

The video-avgEEG accuracy is much higher than the standard (single-subject) video-EEG accuracy in all cases (Figure 10). Even under LOVO, where the model underperforms CCA on individual EEG, video-avgEEG reaches a mean accuracy around 75% on unseen videos. Genuine, generalizable EEG-video coupling that transfers to unseen videos is therefore learned even in the regime where the model appears to fail; it is simply masked by individual noise in the EEG embedding. Moreover, video-avgEEG on held-out videos or video segments is higher than EEG-avgEEG: the video embedding matches the group EEG better than an individual subject's EEG matches it. Both comparisons point to the same conclusion, that individual noise, rather than an absence of learned alignment, is what limits the contrastive model's decodability under LVSO and LOVO.

This does not conflict with the strong LOSO performance of Section 3.1. Individual noise limits every regime, LOSO included. What differs across regimes is not the noise, which is identical since the same subjects and EEG are used throughout, but the strength of the signal relative to that noise. For seen videos (LOSO, and partly LVSO), the model can additionally exploit stimulus-specific correspondence, whether genuine but confined to the seen content or synthesized (Section 4.1). For unseen videos (LOVO), the model can only leverage generalizable video features (omitting video-specific features), which is weaker in single-subject EEG. LOSO therefore sits highest not because it escapes individual noise, but because it brings more signal to bear against the same noise floor.



(a) Under LVSO



(b) Under LOVO

Fig. 10: MM accuracy of three tasks, video-EEG (grey), EEG-avgEEG (blue), and video-avgEEG (green), for the CCA baseline and the contrastive model, under (a) LVSO and (b) LOVO. The large dot is the mean and the error bars show ± 1 SD; each faint dot is one video-segment fold in (a) ($n = 8$) and one video in (b) ($n = 13$).

4.4. Group-Level Alignment Is Easy; Single-Subject Decoding Is Limited by Signal-to-Noise Ratio

Prior efforts to strengthen EEG-video alignment have concentrated on finding better video representations, from hand-crafted features [8, 29] to the pretrained VideoMAE features used here, in the hope of explaining more of the EEG variance. The group-averaged reference allows us to examine this direction separately for the shared and the individual part of the EEG. For the shared part, the video-avgEEG accuracy already reaches around 90% in both LVSO and LOVO with the simple hand-crafted optical-flow feature of the linear CCA baseline, even higher than the contrastive model's video-avgEEG (around 75% under LOVO; Section 4.3). On 30-second windows, the MM task against the group-averaged reference thus already approaches its ceiling, leaving little room to observe further gains from better video representations on this task.

The coupling with the individual EEG, which is the quantity that matters at deployment, is considerably weaker. The gap between the video-EEG and video-avgEEG tasks reflects a low single-subject signal-to-noise ratio (SNR) (Section 4.3): the stimulus-locked signal is weak relative to individual noise in a single subject. This bottleneck can be addressed from two sides. On the signal side, a more expressive video representation could capture a stronger stimulus-locked response and thereby raise the single-subject SNR. On the noise side, meth-

ods that suppress individual noise in the EEG could achieve the same, a direction that has so far received far less attention than feature design.

4.5. Implications for Evaluation, Modeling, and Data

The contrastive model is positioned differently against the CCA baseline under the three regimes (Sections 3.1–3.3), and the sham experiment (Section 4.1) shows that a high LOSO accuracy can be deceptive, reflecting a synthesized correspondence rather than physically meaningful coupling. Any single regime therefore gives an incomplete, and potentially misleading, picture. This leads to what we regard as the most impactful implication of this work: models for neural tracking of natural stimuli should be evaluated under the strictest regime, holding out subjects and stimuli jointly (leave-subject-and-video-out), whenever the dataset permits it.

This recommendation is not specific to video, as studied here, but applies equally to neural tracking of speech. In that domain, increasingly complex nonlinear models share the same capacity to synthesize a correspondence instead of capturing genuine neural coupling. Such models are commonly evaluated with only one axis held out, subjects or stimuli, but rarely with both held out simultaneously [14, 23, 24, 25, 42], which limits the confidence that can be placed in their reported gains. Our findings clearly demonstrate the risks of using a regime that holds out only one axis and leaves the other exploitable, despite the prevailing assumption that these are rigorous evaluation schemes that are not prone to shortcuts. When a dataset does not permit joint subject-and-stimulus holdout (e.g., due to small scale, as in our case), LOSO and LOVO should at least be reported together, since the gap between them exposes how much of the apparent performance depends on having seen the stimulus or the subjects during training.

To improve generalization under LVSO and LOVO, we evaluated several interventions: constraining output dimensionality via low-rank temporal bases, implementing a multi-component structure, applying a multi-scale correlation loss, augmenting EEG channels, and introducing an auxiliary EEG-denoising loss. None yielded meaningful improvements over the base model. This outcome suggests two non-mutually exclusive interpretations: either the architecture requires a fundamental redesign, or the dataset inherently limits the current model's capabilities in size and/or diversity. Both scenarios point to the same underlying requirement. If the dataset is the bottleneck, scaling it is directly necessary; conversely, if a more expressive architecture is required, it will demand more data to prevent overfitting. Ultimately, future progress is gated by the size and diversity of the available data.

This turns attention to how such data should be collected. The fact that the model ties CCA under LVSO but underperforms under LOVO is unsurprising: the latter demands extrapolation to unseen videos, whose distribution differs from the training stimuli, and datasets in this area reuse a small set of videos, leaving a deep model little basis to generalize to novel stimuli. Broadening the stimulus set is thus likely more valuable than recording more subjects on the same videos.

5. CONCLUSION

We present the first study of temporal contrastive alignment between EEG and naturalistic video. Our approach aligns EEG signals with pretrained VideoMAE features using a lightweight two-branch encoder trained via a correlation-based InfoNCE objective. We benchmarked this model against a CCA baseline across three generalization regimes and a downstream selective visual attention task. Results show that the contrastive model outperforms CCA on seen stimuli, both in cross-subject (LOSO) settings and downstream attention decoding. However, it performs only on par with CCA within the stimulus manifold (LVSO) and falls behind when forced to extrapolate to entirely novel videos (LOVO).

Two diagnostic analyses explain this pattern. A sham experiment shows that high accuracy on seen stimuli does not certify physically meaningful EEG-video alignment: a synthesized, physically arbitrary correspondence, fit over the training stimuli, is enough to reproduce it. Yet genuine, generalizable coupling is also learned. It reaches good accuracy against a group-averaged EEG reference even on unseen videos, and is simply masked in single-subject EEG by individual noise, i.e., measurement noise and individual differences in the response to the video. Building the coupling with this clean reference is already achieved by the linear CCA baseline with a simple motion feature, which points to the SNR of single-subject EEG as the primary bottleneck.

These findings carry two practical messages. For applications with a fixed, known stimulus set, genuine cross-stimulus coupling is not required, and template matching with stimulus-specific features over seen stimuli already suffices. The demand for genuine, generalizable alignment becomes binding only for open-set applications that must handle novel stimuli. For the field more broadly: neural tracking of natural stimuli, whether video or speech, should be evaluated with subjects and stimuli held out simultaneously, since single-axis regimes such as LOSO or LOVO, despite their rigorous reputation, can substantially overstate genuine alignment. Further progress is gated by the size and diversity of the data. Ideally, datasets should be broadened to include more stimuli, and methods should raise the single-trial SNR, both by capturing more stimulus-locked signal and by suppressing individual noise, the latter having received far less attention so far.

6. REFERENCES

[1] J. Obleser and C. Kayser, "Neural entrainment and attentional selection in the listening brain," *Trends in cognitive sciences*, vol. 23, no. 11, pp. 913–926, 2019.

[2] G. M. Di Liberto, J. Hjortkjaer, and N. Mesgarani, "Neural tracking: Closing the gap between neurophysiology and translational medicine," *Frontiers in Neuroscience*, vol. 16, p. 872600, 2022.

[3] S. A. Hillyard, E. K. Vogel, and S. J. Luck, "Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 353, no. 1373, p. 1257, 1998.

[4] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[5] J. J. Ki, S. P. Kelly, and L. C. Parra, "Attention strongly modulates reliability of neural responses to naturalistic narrative stimuli," *Journal of Neuroscience*, vol. 36, no. 10, pp. 3092–3101, 2016.

[6] S. Geirnaert *et al.*, "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, 2021.

[7] W. Biesmans *et al.*, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 25, no. 5, pp. 402–412, 2016.

[8] J. P. Dmochowski *et al.*, "Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity," *NeuroImage*, vol. 180, pp. 134–146, 2018.

[9] J. A. O'sullivan *et al.*, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[10] J. Vanthornhout *et al.*, "Speech intelligibility predicted from neural entrainment of the speech envelope," *Journal of the Association for Research in Otolaryngology*, vol. 19, no. 2, pp. 181–191, 2018.

[11] A. De Cheveigné *et al.*, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[12] ——, "Auditory stimulus-response modeling with a match-mismatch task," *Journal of Neural Engineering*, vol. 18, no. 4, p. 046040, 2021.

[13] T. De Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1234–1241, 2020.

[14] G. Ciccarelli *et al.*, "Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods," *Scientific reports*, vol. 9, no. 1, p. 11538, 2019.

[15] I. Kuruvila *et al.*, "Extracting the auditory attention in a dual-speaker scenario from EEG using a joint CNN-LSTM model," *Frontiers in Physiology*, vol. 12, p. 700655, 2021.

[16] Y. Lu *et al.*, "Auditory attention decoding from electroencephalography based on long short-term memory networks," *Biomedical Signal Processing and Control*, vol. 70, p. 102966, 2021.

[17] X. Chen *et al.*, "Auditory attention decoding with task-related multi-view contrastive learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6025–6033.

[18] L. Bollens, T. Francart, and H. Van Hamme, "Learning subject-invariant representations from speech-evoked EEG using variational autoencoders," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 1256–1260.

[19] Z. Xu *et al.*, "Decoding selective auditory attention with EEG using a transformer model," *Methods*, vol. 204, pp. 410–417, 2022.

[20] L. Bollens *et al.*, "Contrastive representation learning with transformers for robust auditory EEG decoding," *Scientific Reports*, vol. 15, no. 1, p. 28744, 2025.

[21] K. Thakkar, J. Hai, and M. Elhilali, "Investigating self-supervised deep representations for EEG-based auditory attention decoding," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1241–1245.

[22] H. Hu *et al.*, "The effect of speech representations on EEG-based auditory attention detection," *Pattern Recognition Letters*, vol. 203, pp. 146–151, 2026, [Online]. Available: https://doi.org/10.1016/j.patrec.2026.03.005.

[23] M. J. Monesi *et al.*, "An LSTM based architecture to relate speech stimulus to EEG," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 941–945.

[24] B. Accou *et al.*, "Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1175–1179.

[25] C. Puffay *et al.*, "Relating EEG to continuous speech using deep neural networks: a review," *Journal of Neural Engineering*, vol. 20, no. 4, p. 041003, 2023.

[26] I. Rotaru *et al.*, "What are we really decoding? unveiling biases in eeg-based decoding of the spatial focus of auditory attention," *Journal of Neural Engineering*, vol. 21, no. 1, p. 016017, 2024.

[27] G. Ivucic *et al.*, "The impact of cross-validation schemes for eeg-based auditory attention detection with deep neural networks," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2024, pp. 1–4.

[28] N. Heintz *et al.*, "Why performance metrics overpromise in auditory attention decoding: an information-theoretic reappraisal," *arXiv preprint arXiv:2608.14250*, 2026.

[29] Y. Yao *et al.*, "Identifying temporal correlations between natural single-shot videos and EEG signals," *Journal of Neural Engineering*, vol. 21, no. 1, p. 016018, 2024.

[30] ——, "EEG-based decoding of selective visual attention in superimposed videos," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 10, pp. 7248–7261, 2025.

[31] C. Du *et al.*, "Decoding visual neural representations by multimodal learning of brain-visual-linguistic features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 760–10 777, 2023.

[32] X.-H. Liu *et al.*, "EEG2Video: Towards decoding dynamic visual perception from EEG signals," *Advances in Neural Information Processing Systems*, vol. 37, pp. 72 245–72 273, 2024.

[33] S. Huang *et al.*, "Need: Cross-subject and cross-task generalization for video and image reconstruction from EEG signals," *Advances in Neural Information Processing Systems*, vol. 38, pp. 173 134–173 173, 2026.

[34] Y. Yao *et al.*, "Video-EEG encoding-decoding dataset KU Leuven," Zenodo, Jan. 2024, [Online]. Available: https://doi.org/10.5281/zenodo.10512414.

[35] ——, "Selective visual attention decoding dataset KU Leuven," Zenodo, Jun. 2025, [Online]. Available: https://doi.org/10.5281/zenodo.15665101.

[36] Z. Tong *et al.*, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10 078–10 093, 2022.

[37] V. J. Lawhern *et al.*, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.

[38] T. Akiba *et al.*, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[39] C. Zhang *et al.*, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[40] D. Arpit *et al.*, "A closer look at memorization in deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 233–242.

[41] D. Kalimeris *et al.*, "SGD on neural networks learns functions of increasing complexity," *Advances in neural information processing systems*, vol. 32, 2019.

[42] M. Thornton, D. Mandic, and T. Reichenbach, "Robust decoding of the speech envelope from EEG recordings through deep neural networks," *Journal of neural engineering*, vol. 19, no. 4, p. 046007, 2022.