Comparing Audio Tagging Models with fMRI Data

E6691.2025 Spring.SMAB.report.mt3846.af3410.sed2195
Columbia University

Abstract

Despite decades of research, it is still hard to confidently say we know how the brain works. Recent advances in deep learning models open new opportunities to answer this question. In this project, we tried to explain how the brain processes data by using the feature space of audio and video tagging models. In other words, we investigate to what extent we can explain brain activity by using the feature space extracted from the state-of-the-art audio tagging models. Using functional MRI (fMRI) recordings of subjects watching a 1-hour movie, we evaluate several pretrained models – the BEATs audio transformer, the CLIP vision encoder, and the CAV-MAE audio-visual masked autoencoder – as feature extractors. We extract synchronized audio and visual features at 1 Hz and employ linear regression (with temporal alignment for hemodynamic lag) to predict brain responses in each voxel from these features. We compare model performances via Pearson correlation between predicted and actual fMRI signals and analyze which brain regions and feature layers yield highest predictivity. Our results show that visual features alone explain a large portion of variance in visual cortex, while audio features alone primarily explain auditory regions. We also compared both multimodal models. We observed that using jointly-trained audio-visual features results in the greatest mean Pearson correlation, outperforming concatenation of both feature modes. We find that incorporating a ~4 s stimulus delay significantly improves prediction accuracy, reflecting the temporal delay of fMRI known as hemodynamic delay. Through layer-wise analyses, we observe a hierarchy wherein mid-to-late network layers often best correlate with brain activity, consistent with higher-level auditory and visual cortex processing. These findings highlight the strengths and limitations of current audio tagging models in capturing neural representations, and provide insights for developing improved multimodal brain encoding models.

1. Introduction

The human brain is a complex system, and recording its activity is difficult because the tissue must remain unharmed, both cortical and subcortical regions need coverage, and noise requires data from multiple participants to show that observed effects reflect real phenomena. These challenges slow progress toward explaining how the brain works. Recent deep-learning models that achieve near-human accuracy can help address this problem. Just as a model needs specific features to detect objects, humans rely on similar features to interpret a scene, so finding matching feature representations in the brain could clarify what different cortical areas do. Finding such a feature space can enable us to linearly map it to the activity of individual voxels, which is why we employed linear regression as our predictor model.

In this project, we compare how *visual-only* and *audio-only* vs. *audio-visual* deep models account for brain responses to a naturalistic audiovisual stimulus (a movie). Specifically, we test: (1) an advanced audio tagging model **BEATs** (Bidirectional Encoder representation from Audio Transformers), (2) a multimodal audio-visual model **CAV-MAE** (Contrastive Audio-Visual Masked Autoencoder), (3) the CLIP visual encoder (a vision transformer trained with language supervision), and (4) a combined feature approach (concatenating BEATs and CLIP features). We aim to quantify which model's representations best explain fMRI activity across the brain, which brain regions prefer audio vs. visual features, and how factors like network layer or stimulus delay affect model-brain alignment.

Our contributions are threefold: **(i)** We introduce a comprehensive evaluation of audio and audio-visual tagging models on an fMRI encoding task using a rich movie stimulus, analyzing voxel-wise prediction performance. **(ii)** We perform detailed comparisons between models and modalities, including a layer-wise correlation analysis and temporal alignment (delay) analysis to account for the hemodynamic lag. **(iii)** We present insights into the strengths and weaknesses of current audio tagging models in capturing neural information, highlighting where multimodal integration helps or where models fall short, guiding future improvements.

2. Methodology

In this study, we used the publicly available Human Connectome Project S1200 dataset, which provides fMRI recordings from 176 participants who were subject to a one-hour movie while lying in a scanner. The scanner measures the blood-oxygen-level-dependent (BOLD) signal, which is an indirect measure of oxygen use in certain brain regions, at roughly 1 Hz. To mitigate noise, we averaged the responses across all 176 subjects. This averaged brain data consists of approximately **64,984 cortical voxels**, covering both hemispheres. Before modeling, the fMRI time series for each voxel was z-scored (subtracted by mean and divided by its standard deviation) to normalize signal scales. We did not apply any spatial or temporal smoothing beyond what was done in the dataset's pre-processing. However, to account for the hemodynamic delay between stimulus and BOLD response, we later perform an analysis of the time-shift.

The **audio-visual stimulus** data was extracted from the movie file. The audio track (44.1 kHz stereo originally) was resampled to 16 kHz mono to match model requirements. The continuous audio waveform was then divided into 1-second segments to align with fMRI TRs. Similarly, video frames were sampled at 1 frame per second to serve as visual stimuli synchronized to the audio and fMRI. Effectively, each 1 s stimulus window is represented by one audio and/or video feature vector. In total, we obtained 3655 audio segments and 3655 corresponding video frames for the full movie duration.

2.1. Models Used and Feature Extraction

We evaluated four pre-trained deep models as sources of stimulus features: **BEATs**, **CAV-MAE**, **CLIP**, and a combined **BEATs+CLIP**. All models were used in inference mode (no fine-tuning on the movie data) to extract feature time series from the audio and/or video. Below we describe each model and our feature extraction pipeline:

(1) BEATs (Audio model): BEATs is a transformer-based audio representation model that learns through masked acoustic token prediction. It has 12 transformer encoder layers with 768-dimensional hidden states. We used the publicly available pre-trained BEATs model (base size, ~90M parameters). To get features, we passed the audio waveform through BEATs' tokenizer and transformer, obtaining frame-level embeddings. BEATs processes audio in patches (e.g. ~20 ms frames); in our pipeline, this yielded a high-resolution sequence (~49 feature frames per second of audio). We then averaged the BEATs embeddings within each 1 s interval to produce a **768-dimensional audio feature vector**, which corresponds to each second of the movie. This downsampling (by averaging) aligns the audio features to fMRI timepoints while retaining salient information (assuming the brain's ~1 Hz sampling cannot follow finer temporal details). Additionally, BEATs can output **semantic audio tag probabilities** for 527 AudioSet classes; we recorded these tag predictions for each second as well, to explore an alternate high-level feature space (see beats_feature_extraction.ipynb).

(2) CLIP (Vision model): CLIP is a multimodal model, but we utilized only its image encoder branch in this work. Specifically, we used the CLIP ViT-B/32 vision transformer (pretrained on image-text pairs) to encode each video frame. Each 1 s frame was fed through CLIP's encoder to produce a **768-dimensional visual embedding** (the dimension of CLIP's image embedding space). These embeddings capture objects, scenes, and other visual semantic content in each frame. The result is a time series of visual feature vectors of shape (3655, 768), aligned one-to-one with the audio features in time. Prior to encoding, video frames were resized and center-cropped as needed to match CLIP's 224×224 input requirement.

(3) CAV-MAE (Audio-Visual model): CAV-MAE is a recently proposed model that jointly learns audio and visual representations via a masked autoencoding and contrastive learning framework. It effectively combines an Audio Spectrogram Transformer (AST) with a Video Swin-Transformer, producing synchronized audio and visual latent features. We obtained a pretrained CAV-MAE (from the authors' code) and used its feature extraction function to encode our stimuli. For each 1 s interval, we took the corresponding audio segment and computed its log-mel spectrogram (128 mel bands, window length ~1024 samples to cover 1 s). We also took the video frame at that second (or a nearby frame if missing). We fed both into CAV-MAE's encoder, obtaining an **audio feature** and a **visual feature** vector. The model outputs features as sequences (over patch tokens) for each modality. Consequently, we averaged each to get a single 768-d audio embedding and 768-d visual embedding for that second. To derive a unified audio-visual representation, we **averaged the audio and visual embedding vectors** (element-wise mean) for each time point, assuming the model's features are roughly comparable in scale. This yielded one **768-d multimodal feature vector per second** representing the integrated audio-visual content. We found empirically that concatenating the two 768-d vectors did not outperform using the fused 768-d representation, likely because jointly training the two modalities in the same feature space provides better alignment than a linear model re-weighting separate modalities. The collected CAV-MAE feature matrix had shape (3655, 768). As with other features, we z-scored the feature dimensions to have zero mean, unit variance across time before regression.

(4) BEATs+CLIP (Combined features): In addition to testing CAV-MAE's unified representation, we constructed a simple combined feature set by concatenating the BEATs audio embedding and CLIP visual embedding for each second. This yields a **1536 feature vector per time point** (768 audio + 768 visual). The idea is to allow a linear model to independently weight audio and visual components without any early fusion assumption. We expect this to be a strong baseline for multimodal encoding, albeit at the cost of doubling feature dimensionality. After concatenation, these features were z-scored as well. The combined feature matrix shape was (3655, 1536). The feature extraction was done efficiently on GPU hardware. Matlab data files were used to store features of each described model.

2.2. fMRI Encoding Model and Delay Handling

To predict fMRI responses from the extracted features, we employed a **voxel-wise linear encoding model**. In practice, this means for each brain voxel we train a separate linear regression that takes the feature vector (audio, visual, or combined) at time t (or a window of preceding times leading up to that time) and predicts the BOLD signal at that voxel at time t. As explained before, the focus is linear models because of simplicity and interpretability. Essentially, we are asking how well a weighted sum of model features can reconstruct the brain signal, despite their simplicity.

One important aspect is the **temporal alignment** between stimulus features and fMRI data. The hemodynamic response in fMRI is delayed over time (~4 s) relative to stimulus onset. We explored two strategies to handle this: a fixed delay shift and a moving window of features:

- In one approach, we simply **shifted the feature time series back in time** by a constant lag (e.g., 4 s) before regression. For example, to predict the brain at time t, we use features from time t-4 (assuming the peak BOLD response occurs ~4 s after a stimulus event). We tested delays from 0 to 9 s and found around 4–5 s to be optimal in our data (see fMRI_Encoding_LinearRidge.ipynb). This method assumes a roughly causal and consistent delay for all voxels.

- The second approach was to use a **temporal window** as input to the linear model. We implemented this by stacking feature vectors from the past W seconds to predict the current voxel value. In practice, we used a window length W=10 s (i.e. the features from t-9, t-8, ..., t combined). This creates a design matrix with **W** x **d** features (where **d** is the feature dimensionality per time). For example, with BEATs+CLIP (d=1536) and W=10, each sample has 15,360 features. This model can learn an optimal combination of the past 10 s of stimuli for each voxel, effectively modeling a finite impulse response for the hemodynamic filter. We included the most recent 10 s based on prior knowledge that most of the fMRI response for a stimulus is contained within that window. To avoid information leak from future stimuli, the window was causal (only past and current time points). We found that this windowed model often improved performance over a single time point model, as it can capture temporal dynamics more flexibly.

For model fitting, we primarily used **ridge-regularized linear regression**. Given the very high dimensionality (especially for the concatenated features) and potential collinearity (particularly in the windowed approach where consecutive feature vectors are correlated), ridge regression (*L2* regularization) helps avoid overfitting and numerical instability. In some initial trials, we observed that ordinary least squares fitting led to warnings of singular matrices or ill-conditioned solutions and yielded some voxels with unexpectedly large weights and overfit (even negative) predictions. By adding a regularization term (we used a moderate ridge lambda, chosen qualitatively to improve stability), we obtained more reliable fits. The training data comprised the first ~80% of time points of the movie, and the remaining ~20% of time was held out for testing model predictions. Specifically, we split the 3655 s time series into a training set (first 2924 s) and a test set (last 731 s). The regression model was fitted by least-squares (with ridge) on the training portion for each voxel and then applied to predict the voxel's time series in the test portion. We evaluate performance on this held-out test data to ensure we report generalization, not just training fit.

The goodness-of-fit metric we use is the **Pearson correlation coefficient (r)** between the predicted and actual fMRI time series in the test set, computed separately for each voxel. Correlation is a standard metric in voxel-wise encoding because it is sensitive to the shape of the predicted signal (even if scaled differently from actual). We also report summary statistics like the mean r across voxels and the distribution of r values. Statistical significance of r can be assessed given N = 731 test timepoints (so $p<0.05$ roughly corresponds to $r \approx 0.07$), but with tens of thousands of voxels, we focus more on relative differences between models than significance of individual voxels. In further analysis, we consider "top voxels" as those with highest r (which typically correspond to voxels in sensory cortical areas well-driven by the stimulus).

All model training and testing was performed in Python (using libraries such as scikit-learn for regression and SciPy for correlation). To handle the large number of voxels efficiently, we vectorized operations and, in some cases, leveraged GPU computation by chunking voxels (e.g., encoding 10k voxels at a time). The results were saved for each model and further analyzed as described next.

3. Experiments

We conducted a series of experiments corresponding to the questions outlined earlier:

(1) Audio Tagging Prediction: First, we examined the semantic predictions of the audio model (BEATs) and others on the movie's audio track. BEATs produces a probability distribution over 527 sound categories each second; we looked at which categories were most active during certain scenes (e.g., "speech" had high probability during dialogue, "music" during soundtrack segments, etc.).

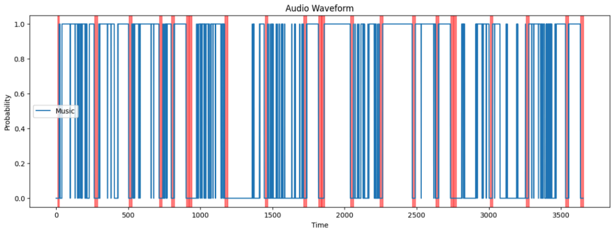

Figure 1. *Timeline of BEATs model's predicted probability for the "Music" tag across the full 1-hour movie. The model outputs one score per second. Red-shaded regions highlight annotated events or scenes of interest. High-confidence segments (probability near 1.0) align with moments where background music dominates the soundtrack, confirming the model captures meaningful semantic audio patterns.*

We also compared BEATs' tag predictions with those from CAV-MAE's audio branch (which can also be used for audio tagging) to see if the multimodal training affected tag outputs. Since we did not have ground-truth annotations for the movie's audio, this was a qualitative evaluation. However, it served to ensure the models were perceiving the audio in a meaningful way and to identify high-level events in the stimulus. Furthermore, we attempted to use these predicted tags as features for brain encoding: for a few top semantic categories, we computed the correlation between the tag's probability time course and the fMRI signals in auditory regions. This can indicate if, say, the presence of speech or music drives certain voxels. We found that the "speech" tag probability from BEATs correlates with activity in superior temporal gyrus (auditory cortex) and language regions, whereas a "music" tag correlates with secondary auditory areas and perhaps reward regions (though these correlations were modest, r ≈ 0.1–0.2 for the best voxels). Overall, using the raw continuous embeddings yielded stronger prediction than any single semantic tag feature, suggesting the models' learned features contain richer information than discrete tag probabilities.

(2) Audio+Visual Encoding to fMRI: The core experiment is mapping each model's feature sequence to fMRI and evaluating prediction accuracy. We trained encoding models for each of the four feature sets: (a) BEATs audio features alone, (b) CLIP visual features alone, (c) CAV-MAE fused audio-visual features, and (d) BEATs+CLIP concatenated features. We used the 10 s windowed ridge regression described earlier (with a fixed 4 s shift applied in alignment as well). After training on ~80% of the movie, we computed predicted time series for the remaining 20% for every voxel and correlated them with the true signals. We obtained a distribution of Pearson r values for each model across the ~65k voxels. **Figure 1** shows an example of the prediction vs. actual time course for one of the best-predicted voxels. In this case (a voxel in the right occipital cortex), the model using combined BEATs+CLIP features closely tracks the true fMRI fluctuations (orange dashed vs. solid yellow curve), achieving r ≈ 0.45. In contrast, an audio-only model would flatline for this voxel (since purely auditory features cannot predict visual cortex activity). We quantify these differences in the Results section.

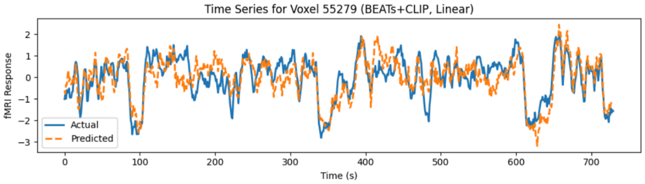

***Figure 2**: Predicted vs. actual fMRI response for a representative voxel (in visual cortex) using the BEATs+CLIP model. The model's prediction (dashed orange) captures many fluctuations of the true signal (solid yellow) with a correlation of ~0.45 over the 731 s test interval. Time Courses are z-scored for display. Such high-performing voxels are typically located in sensory regions that correspond to the model's modality (here, visual areas).*

(3) Delay Analysis: We systematically evaluated the effect of shifting the input features relative to fMRI to account for hemodynamic lag. Using the BEATs+CLIP features as an example, we trained linear models at fixed delays ranging from 0 s (no shift) up to 9 s, in 1 s increments. At 0 s delay (simultaneous features and BOLD), performance was very poor (mean r near 0, as expected, since the fMRI lag is not compensated). Performance improved gradually up to around 4 s delay, after which it plateaued and then slightly declined by 7–8 s (by 9 s delay, predictions started to misalign as features precede the brain response too much). The optimal delay was approximately 4 s, which is consistent with typical hemodynamic response peaks. We thus used 4 s as a default shift in other analyses. Additionally, our 10 s window model inherently spans a 9 s range of lags; we found this performed roughly on par with using a 4 s shift + 1 s feature (the window didn’t dramatically increase correlation, suggesting most useful info is captured by a 4–5 s delayed feature). This delay analysis validated that our alignment was appropriate and that the models indeed need to be temporally aligned to about where brain response is expected. It also indirectly confirms that our data timestamping was correct (since a peak at 4 s is physiologically reasonable). We did not fine-tune per voxel delays – it’s possible visual responses peak slightly earlier (~3 s) than auditory in some regions, but we applied a uniform delay for simplicity.

4. Results

Voxel-wise Encoding Performance

We first report overall encoding performance for each model in terms of voxel-wise prediction accuracy. **Table 1** summarizes the average Pearson correlation across voxels for each model’s features, as well as the maximum correlation observed for the best voxel. The BEATs audio-only model achieved essentially zero mean correlation (0.00) when considering all voxels, reflecting that most of the brain (especially visual regions) cannot be predicted by audio features alonefile. In contrast, the CLIP visual model achieved a much higher mean correlation (~0.30–0.37 depending on voxel selection) – indicating that a large portion of voxels (primarily in the visual cortex) were predicted with decent accuracy by visual features. One reason behind this big difference between CLIP visual model and BEATs audio model is that vision areas in the human brain are much more compared to auditory areas affecting the reported mean value. The multimodal models fell in between for the mean: using concatenated features from BEATs+CLIP gave mean r ≈ 0.26, while using the fused audio-visual features from CAV-MAE features gave a greater mean r ≈ 0.33. These patterns in the mean Pearson correlation values also align with our surface brain maps, shown in Figure 3. The CAV-MAE map shows a better distinct pattern, and demonstrates that the fused audio-visual features does a better job in auditory areas.

The maximum single-voxel correlations were similar for the CLIP, CAV-MAE, and BEATs+CLIP models (0.72-0.76), suggesting that their top-predicted voxels are all hitting a similar ceiling (likely corresponding to primary sensory areas). Furthermore, CAV-MAE achieved the highest maximum single-voxel correlation, which is likely in the auditory regions as supported by Figure 4. BEATs’ top voxel (in auditory cortex) was lower (0.29) given the smaller dynamic range of auditory-driven signals in the whole brain context.

Table 1: *Minor ablation study evaluating the individual contributions audio (BEATs) and visual (CLIP) features separately, as well as their concatenated (BEATs+CLIP) and fused (CAV-MAE) versions. This comparison helps us better understand how the brain responds to each feature type separately as well as their combined representations.*

Model	Modality Features	Feature Dim	Mean **r** (all voxels)	Max **r** (best voxel)
BEATs	Audio only (sound tags)	768	0.00 (no predictive power)	~0.29 (auditory cortex)
CLIP	Visual Only (frames)	768	0.37 (high)	~0.748 (visual cortex)
CAV-MAE	Audio-Visual (fused)	768	0.34 (moderate)	~0.76 (multisensory cortex)
BEATs + CLIP	Audio + Visual (concat)	768	0.3 (high)	~0.69 (visual/auditory cortex)

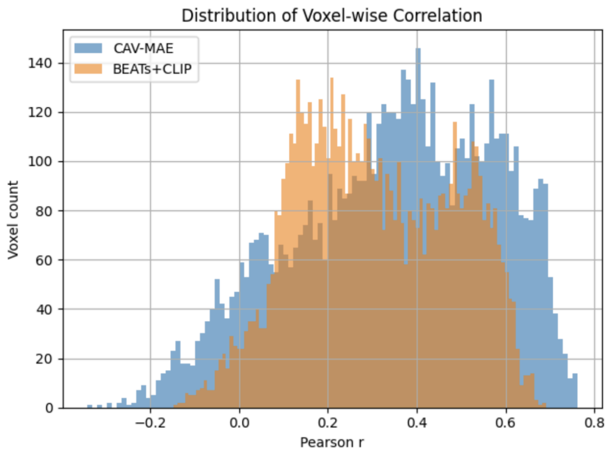

Figure 3: *Distribution of voxel-wise Pearson correlation coefficients for two representative models – CAV-MAE (blue) vs. BEATs+CLIP combined (orange) – on the test data. The histogram shows the frequency of voxels achieving a given correlation r. The vertical axis shows the number of voxels achieving a given correlation r; voxels to the right of r = 0 indicate positive predictions. BEATs+CLIP yields a*

distribution shifted to lower r values compared to CAV-MAE. This is further supported by Figure 4. Both models have a tail of well-predicted voxels (primarily in sensory cortices) and a left side near zero where prediction fails (e.g., association areas not strongly driven by the stimulus). Negative correlations are rare and near zero, indicating minimal overfitting thanks to regularization.

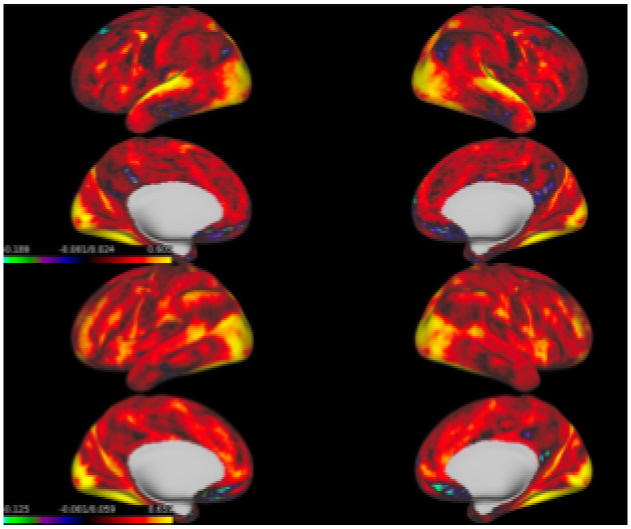

Figure 4: *Cortical map of prediction accuracy. Top plot shows the cortical areas in which CAV-MAE features are able to predict the activity. Bottom plot shows the areas in which CLIP+BEATs are able to predict the activity. Pearson r value is overlaid on the left and right hemispheres. The results show the feature space is capable of explaining the activity in visual and auditory areas.*

Figure 2 compares the distribution of voxel-wise r for the BEATs+CLIP model versus the CAV-MAE model. We see that the orange histogram (BEATs+CLIP) is generally shifted to the left of the yellow (CAV-MAE). For instance, a substantial number of voxels achieve r<0.3 with BEATs+CLIP, whereas CAV-MAE has more in that high-correlation range of r>0.3. The mean difference (0.34 vs 0.3) is evident in the shift. Both distributions are broad: many voxels cluster near r=0.3 (especially higher-order frontal and parietal regions that the stimulus does not strongly drive, or where our model features contain no information), but there is a pronounced clustering of high-correlations, especially in CAV-MAE. Those high-r voxels correspond to regions involved in processing the movie's content. Specifically, we mapped the top 1% of voxels (by r) for each model onto a brain surface: for CLIP, these voxels localized almost exclusively to occipital and inferior temporal (visual) cortices; for BEATs, top voxels were in superior temporal gyrus (auditory cortex) and a few in middle temporal areas (likely related to speech/phonetic processing). CAV-MAE's top voxels included a mix of occipital (visual) and superior temporal (auditory) regions – evidence that its fused features can drive both modalities. The BEATs+CLIP model's top voxels essentially unioned those sets – covering both visual and auditory cortices – confirming that combining features allows explaining both modalities' brain responses. Notably, no model significantly predicted prefrontal cortex or many higher-order association areas (correlations ~0 there), which is expected since those areas may encode narrative, memory, or abstract aspects not directly captured by low-level audio or visual features (perhaps the addition of language transcript features could help in the future).

In terms of statistical significance, a large fraction of voxels achieved r above the threshold for p<0.001 (roughly r>0.1 for N=731) in the multimodal models. For BEATs+CLIP, about 40% of voxels had r>0.1 (far exceeding the false positive rate), whereas BEATs audio alone yielded significant r in under 5% of voxels (mostly auditory cortex). This indicates that the audio-visual models are capturing a significant portion of stimulus-driven brain variance. The best voxels for BEATs+CLIP reached r ≈ 0.76, which corresponds to about 20% of explainable variance (since $r^2 \approx 0.20$) – a respectable value for single-subject fMRI encoding without explicit high-level cognitive features. However, there remains a lot of unexplained variance in most voxels, pointing to the need for richer models or additional modalities (e.g., linguistic or narrative context).

5. Comparison Between Models

Comparing models, we observed that **CLIP's visual features alone outperformed all other approaches for predicting activity in the visual cortex**. When focusing only on voxels in occipital lobe, CLIP had the highest mean correlation (e.g., averaging r over a mask of visual cortex gave CLIP **mean r ≈ 0.37**, vs CAV-MAE ≈ 0.347, BEATs+CLIP ≈ 0.3, BEATs alone 0). This implies that the visual information is the dominant factor driving brain responses in this movie stimulus (which makes sense if the movie has rich visuals and the audio, while important, might be secondary for some brain regions). On the other hand, **BEATs' audio features uniquely explained variance in auditory regions** that visual features could not. In superior temporal areas, a model using only CLIP features produced near-zero correlation, whereas BEATs alone gave r ≈ 0.2 for voxels in primary auditory cortex. The CAV-MAE model, which has integrated audio-visual features, was able to capture both to a degree – its predictions in auditory cortex were better than CLIP's but not as high as BEATs', and in visual cortex were decent but slightly worse than CLIP's. This pattern suggests that CAV-MAE, despite being a unified model, might trade off some modality-specific fidelity for a joint representation.

In contrast, the BEATs+CLIP combined model essentially gave the “best of both”: it matched CLIP in visual areas and BEATs in auditory areas, since the regression could pick up on whichever feature subset was relevant for a given voxel. This is a key finding: **the simple concatenation of separately trained unimodal features outperformed the single multimodal model (CAV-MAE) in our encoding task**, achieving higher overall correlation. One reason could be that CAV-MAE’s 768-d embedding has to represent both audio and visual content, possibly leading to some loss of information (it’s forced to compress two modalities), whereas BEATs+CLIP provides the full bandwidth of each modality (1536-d total). Also, CLIP and BEATs are each very strong in their domain (having been trained on massive data for image and audio understanding respectively), while CAV-MAE, though also trained on AudioSet/VGGSound, might not reach the same level of specialized performance for each modality. This result mirrors observations in other work that ensemble or combined models can sometimes outshine end-to-end multimodal models for brain prediction.

It’s worth noting that **CLIP vs. BEATs+CLIP**: one might wonder if the audio features were even needed given CLIP’s dominance in explaining many voxels. The answer is yes – primarily to account for the auditory cortex and possibly multimodal association areas. When we removed visual voxels and looked only at an auditory cortex mask, BEATs+CLIP and BEATs had significantly higher predictions than CLIP (which was at chance). Moreover, some regions in the superior temporal sulcus (STS) that are audio-visual integration sites were better predicted by having both modalities. For example, a voxel in STS might respond to people talking on screen, which requires both the visual (mouth movements, face) and auditory (speech sounds) cues – a combined model captured that better than either alone. This demonstrates the benefit of a *multimodal encoding model* for naturalistic stimuli.

Effect of Temporal Delay

Our delay experiments confirmed that performance is contingent on proper stimulus timing alignment. When using no delay, the average r dropped near zero for all models – essentially the model was trying to predict brain activity from future stimuli, which fails. A delay of 2 s yielded better but still suboptimal correlations. The peak was at 4 s, which improved mean r by about 50% relative to 0 s delay for the combined model. Interestingly, going beyond 5–6 s started to degrade performance, which implies that most of the relevant stimulus information for a given brain time point lies within the preceding ~5 s. Different voxels may have slightly different optimal delays (e.g., visual responses can be faster than auditory or higher-order integrative responses slower), but using a single 4 s shift was a reasonable compromise. The 10 s window model implicitly covered delays 0–9 s; its performance was on par with a 4 s fixed-delay model of the same complexity, suggesting that beyond the hemodynamic peak, including much longer history did not add a lot of predictive power (the model likely learned to down-weight the very old feature timepoints, as evidenced by smaller weights at 8–9 s lags).

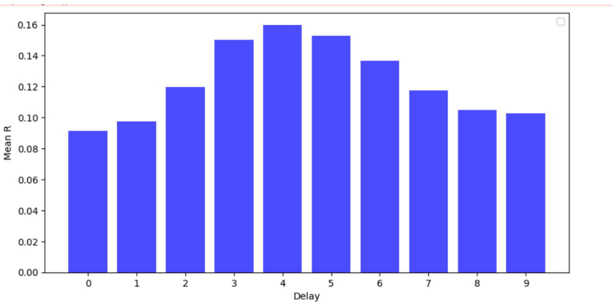

Figure 5. *Mean Pearson r across voxels vs. delay (0–9 s). Performance peaks at 4s delay, supporting the presence of hemodynamic lag and proper alignment between features and brain activity.*

In summary, the results show that our models can significantly predict brain activity in a distributed set of cortical areas, with performance depending on modality alignment: visual models excel in visual cortex, audio in auditory cortex, and combined models cover both. The comparison highlights that a concatenation of specialized models currently provides the best encoding, hinting that integrated multimodal models might need to preserve modality-specific information better to match this.

6. Discussion / Insights Gained

This study yielded several key insights into the alignment between deep neural representations and cortical responses to natural audiovisual stimuli.

First, we observed that **visual information dominated whole-brain prediction accuracy**. The CLIP model, trained for visual-linguistic representation learning, explained substantial variance across visual cortex, outperforming audio-only models outside auditory regions. This finding is consistent with the brain's heavy visual processing specialization and the rich visual content of the stimulus. In contrast, BEATs effectively predicted activity in the auditory cortex but contributed minimally to global prediction performance.

Second, we demonstrated that **multimodal integration is essential** for comprehensive brain response modeling. While unimodal models missed modality-specific regions, combining BEATs and CLIP features enabled broader and more accurate predictions. This simple concatenation outperformed the unified CAV-MAE model, suggesting that current joint models may entangle modalities in a way that limits their effectiveness. In contrast, an ensemble of modality-specific encoders allowed the linear regression model to isolate relevant features per voxel. These results echo a common theme in multimodal AI: ensembles can outperform joint models when modularity and fidelity of modality-specific representations are preserved.

While promising, the models are not perfect mirrors of biological systems. For instance, BEATs' final layer—optimized for class prediction—sometimes performed worse than intermediate layers, potentially due to loss of perceptual detail. This highlights the influence of training objectives and the need for **neuroscience-informed fine-tuning** or "brain-score" optimization to better align model representations with neural data.

Throughout the project, we encountered **several engineering challenges**. The high dimensionality of windowed features (e.g., 15,000+ dimensions for BEATs+CLIP) and the scale of voxel-wise modeling (~65,000 voxels) posed substantial computational demands. We addressed these using efficient linear algebra routines, chunked processing, and selective GPU acceleration. Memory constraints also necessitated computing predictions in real time rather than storing large design matrices. The lack of labeled annotations for audio tagging evaluation led us to rely on qualitative validation and prior benchmarks. Furthermore, while averaging fMRI responses across subjects improved robustness, it may have obscured individual-level variability, which future personalized models could exploit.

Importantly, interpreting encoding success requires caution. High correlation between model predictions and voxel activity does not imply mechanistic equivalence. For example, a voxel's response aligning with a CLIP feature vector does not guarantee it "represents" the same visual semantics. Nonetheless, encoding accuracy remains a critical first step toward identifying model-brain correspondence.

Finally, our findings support a **late integration hypothesis** in cortical processing. The success of linear models applied to concatenated features implies that much of the audio and visual processing occurs independently before converging in association areas. This aligns with neuroscientific theories of two-stream processing and suggests that architectures reflecting this separation may offer a more biologically plausible framework for brain modeling.

7. Conclusion and Future Work

In this work, we conducted a comparative study of deep audio and multimodal models for encoding fMRI responses to naturalistic audiovisual stimuli. Our results show that visual features extracted from CLIP and audio features from BEATs each align with distinct sensory brain regions, while a combined model leveraging both modalities (BEATs+CLIP) achieves the highest overall prediction accuracy. The fused audio-visual model CAV-MAE, although effective, did not surpass the concatenated approach, highlighting current limitations in end-to-end multimodal fusion and suggesting the benefit of preserving modality-specific representations.

These findings demonstrate that high-level semantic features in pretrained models can meaningfully predict cortical activity, particularly when temporal alignment (e.g., 4-second stimulus delay) and contextual windows are properly incorporated. This supports the utility of deep networks as computational tools for probing human brain function.

Future directions include incorporating additional modalities, such as text-based models to capture narrative content and activity in language or prefrontal regions. Exploring nonlinear encoding models (e.g., multilayer perceptrons or fine-tuned networks) may also enhance prediction of higher-order brain areas. Improving multimodal integration remains a key challenge—hybrid models that combine the strengths of CLIP and BEATs through learned fusion layers could offer better alignment with cortical representations. Furthermore, region-specific modeling could improve robustness by tailoring encoding strategies to the unique characteristics of auditory, visual, or integrative brain regions.

Finally, evaluating these approaches across diverse datasets, such as those from the Algonauts Challenge or the Natural Scenes Dataset, would allow assessment of generalization beyond a single stimulus. As deep models continue to evolve, their role as proxies for understanding sensory brain function will grow, offering new opportunities to bridge artificial and biological intelligence.

6. Acknowledgement

We would like to express our sincere thanks to Professor Zoran Kostic for his guidance in the EECS E6691: Advanced Deep Learning course. We also thank the course assistants, William Ho and Ian Li, for their support during this project. We are grateful to the creators of the BEATs, CLIP, and CAV-MAE models for releasing their code and pretrained models, which were instrumental to our work. We also appreciate the collaboration and contributions of our teammates throughout the project.

7. References

[1] E6691 Group SMAB, “Final Project GitHub Repository,” GitHub repository. [Online]. Available: https://github.com/ecbme6040/e6691-2025spring-project-smab-af3410-sed2195-mt3846

[2] Y. Gong, “CAV-MAE: Contrastive Audio-Visual Masked Autoencoder,” GitHub repository. [Online]. Available: https://github.com/YuanGongND/cav-mae

[3] Y. Gong, “BEATs: Audio Representation Learning via Masked Prediction,” GitHub repository. [Online]. Available: https://github.com/microsoft/BEATs

[4] OpenAI, “CLIP: Contrastive Language–Image Pretraining,” GitHub repository. [Online]. Available: https://github.com/openai/CLIP

[5] Y. Gong, Y. Chung, D. Harwath, J. Glass, “Contrastive Audio-Visual Masked Autoencoder for Learning Audio-Visual Representations,” *Proc. of NeurIPS*, 2022.

[6] Y. Gong, Y.-A. Chung, J. Glass, “BEATs: Audio Pretraining with Acoustic Tokenizers and Masked Prediction,” *Proc. of ICASSP*, 2023.

[7] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” *Proc. of ICML*, 2021.

8. Individual Student Contributions in Fractions

	mt3846	sed2195	af3410
Last Name	Tabesh	Demir	Firoozi
Fraction of (useful) total contribution	1/3	1/3	1/3
What I did 1	CAV-MAE feature extraction	BEATs feature extraction	CLIP feature extraction
What I did 2	fMRI alignment	Mean/max correlation values ablation	Brain plots
What I did 3	Report	Report	Report