

RESEARCH ARTICLE

Computational Neuroscience

Deep learning reveals cross-modal neural representations of auditory and visual mental imagery in MEG

 **Alina Schüller, Constantin Jehn, Marietta Stegmaier, Jasmin Riegel, and Tobias Reichenbach**

Department Artificial Intelligence in Biomedical Engineering (AIBE), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany

Abstract

Mental imagery provides a unique window into the brain's ability to internally simulate sensory experiences, offering valuable insights for both cognitive neuroscience and brain-computer interface (BCI) research. This study examined the neural representations of imagined auditory and visual stimuli using magnetoencephalography (MEG) and assessed the ability of machine learning models to decode these mental processes. MEG data were recorded from 18 right-handed participants during auditory and visual imagery tasks and source reconstructed within modality-specific cortical regions of interest. We compared a convolutional neural network (CNN) and a linear logistic regression model within a subject-specific classification framework. Both approaches achieved above-chance decoding accuracies, with the CNN outperforming the linear model in both tasks, yielding a mean decoding accuracy of >70% for the visual imagery task. Notably, the CNN achieved significant decoding performance even when trained on nontask-relevant cortical regions, indicating that imagined stimuli are represented in distributed and partially overlapping neural networks across modalities. This cross-modal decoding capability highlights the potential of deep learning models to capture complex, multimodal neural patterns and suggests that future brain-computer interfaces could benefit from integrating auditory and visual information. These findings advance our understanding of cross-modal mental imagery and point toward more flexible and personalized approaches in BCI design.

NEW & NOTEWORTHY By comparing linear and deep classifiers, this work shows that convolutional networks capture rich, cross-modal neural representations of auditory and visual mental imagery in magnetoencephalography (MEG). Significant decoding from nontask-relevant regions indicates distributed cortical engagement, highlighting deep learning's potential for robust, modality-independent brain-computer interfaces.

imagined speech; magnetoencephalography; memory; neural decoding; visual imagery

INTRODUCTION

Neurological disorders such as amyotrophic lateral sclerosis (ALS) or severe brain injury can lead to locked-in syndrome, a condition in which individuals retain full cognitive function but lose voluntary muscle control. For these patients, brain-computer interfaces (BCIs) offer a means of communication and interaction independent of motor output (1, 2). BCIs translate neural activity into control signals for external devices such as communication aids or prosthetic systems, thereby improving autonomy and quality of life.

Conventional BCIs often rely on motor imagery or stimulus-driven paradigms that require sustained attention and

extensive user training (3, 4). Additionally, many BCIs require users to spell words letter by letter or imagine movements for every command, leading to high cognitive effort. Most systems also suffer from a low information transfer rate (ITR), making them inconvenient for regular use (4). As a more natural and intuitive alternative, decoding inner or imagined speech, the silent internal rehearsal of words or phrases, has gained increasing interest (5, 6). Neural activity during inner speech engages brain regions involved in overt speech perception and production, particularly in temporal and frontal cortices (6, 7), making it a promising signal source for a more seamless and user-friendly communication system.



Correspondence: A. Schüller (alina.schueller@fau.de).

Submitted 14 November 2025 / Revised 29 May 2026 / Accepted 30 June 2026



In parallel, visual imagery, the internal generation of mental images, has also been explored as a complementary modality for BCI control (8). The human brain allocates extensive cortical resources to vision and can extract complex scene information rapidly (e.g., within 150 ms) (9), making visual imagery a fast and robust channel for representing perceptual content. This efficiency and capacity for detailed visual representation motivate the use of visual imagery as a promising modality for BCI control. Importantly, recent neuroimaging studies have shown that auditory and visual imagery are not fully independent processes: visual areas can become active during inner speech, and auditory or linguistic regions may be recruited when processing visual scenes (7, 10). This functional overlap suggests that mental imagery involves distributed and cross-modal neural representations rather than strictly segregated modality-specific networks.

Cross-modal representations, whereby information from one sensory modality is decodable from brain regions not traditionally associated with that modality, are now well established in multisensory and cognitive neuroscience research. It is broadly recognized that unisensory information is not confined to preferred sensory cortices or higher multisensory association areas, but can be found throughout the cortex (11). Consistent with this view, several neuroimaging studies have demonstrated cross-modal decoding of both perceived and imagined stimuli: Vetter et al. (12) showed that both presented and imagined auditory information can be decoded from early visual cortex using functional MRI (fMRI). More recently, Pollicina et al. (13) decoded semantic sound categories from visual cortex, and Gu et al. (14) demonstrated cross-modal representations in both early visual and auditory cortices using multi-voxel pattern analysis. Using electroencephalography (EEG), Iamshchinina et al. (15) resolved the time course of visual and auditory object categorization, highlighting that MEG and EEG offer unique advantages over fMRI through their high temporal resolution. A key strength of the present MEG study compared to prior fMRI work is precisely this temporal resolution, which in principle allows the time course of imagery-related neural reactivation to be tracked. Supporting the equivalence of auditory and visual imagery as BCI channels, Lee et al. (16) reported comparable imagery intensities between auditory and visual tasks, indicating that both modalities may serve as equally viable channels for neural communication.

Several studies have explored the decoding of neural signals and the application of machine learning techniques in brain-computer interfaces. Marion et al. (17) explored EEG responses to imagined music and demonstrated that imagined neural signals can be predicted accurately through regression analyses. The model outputs were robust enough to allow for precise identification of the imagined musical piece. Similarly, Kosmyna et al. (8) also analyzed the decoding of electroencephalographic signals during imagery tasks, yet focusing on visual instead of auditory imagination. Given that EEG-based BCIs cannot only differentiate between visual observation and visual imagery tasks but also distinguish between different observed stimuli, the visual modality shows a promising potential to expand BCI control strategies (7, 8, 16).

Decoding such distributed activity patterns requires recording methods with high temporal and spatial precision. While EEG has traditionally dominated BCI research due to its portability and applicability in real-world settings (17–19), its limited spatial resolution constrains the identification of distributed neural sources. Magnetoencephalography (MEG) offers a powerful alternative, providing millisecond temporal accuracy and precise spatial localization (20). These properties make MEG particularly suited for capturing the complex, time-frequency dynamics of imagined speech and visual imagery (10). Findings by Dash et al. (1), for instance, demonstrated the feasibility of decoding both imagined and spoken phrases directly from noninvasive MEG signals using machine learning approaches. Although MEG necessitates a complex measurement system and a magnetically shielded room, making it infeasible for real-world BCI applications, it can hence aid in guiding BCI research by identifying suitable neural response patterns.

In the present study, we recorded MEG data while participants performed auditory and visual imagery tasks. To assess the decoding of imagined content, we trained two complementary classifiers: a linear logistic regression model and a convolutional neural network. The convolutional neural network (CNN) was chosen for its ability to extract multi-scale spatial-temporal patterns and its demonstrated success in imagined speech decoding (1). Our primary goal was to evaluate how well each model decodes auditory and visual imagery from MEG data and to determine whether decoding generalizes beyond task-relevant cortical regions, thereby testing for cross-modal representations. Additionally, it is a well-known fact that people have diverse cognitive preferences for learning, utilizing distinct sensory modalities to optimize information acquisition and comprehension (21). Hence, we wondered whether preferred learning styles and memory capacities can be quantified using the decoding accuracies of the neural signals. We explored whether decoding accuracy relates to individual cognitive styles, as assessed by the visual, auditory, read-write, and kinesthetic (VARK) questionnaire (21) and a modality-specific memorization test. Together, these analyses aim to advance our understanding of distributed neural processing during mental imagery and to inform future BCI designs that leverage multimodal and individualized decoding strategies.

MATERIALS AND METHODS

Data Acquisition

A total of 18 healthy, right-handed individuals (9 male and 9 female; 19–31 yr) without hearing impairments took part in the study. Participants with visual impairments performed the task with corrected-to-normal vision. The study was granted ethical permission by the ethics board of the University Hospital Erlangen (Registration 22-361-S). The experiment lasted 1 h and consisted of auditory and visual stimuli.

Auditory stimuli comprised three German abstract nouns: *Ehre* (honor), *Absicht* (intention), and *Idee* (idea). They were carefully chosen for their similar length (all consisting of 2 syllables), distinct initial vowels, and low potential to evoke visual associations. The three words were synthesized using

an artificial intelligence (AI)-based text-to-speech system (Google Text-to-Speech, female German voice). Visual stimuli consisted of three color-distinct images: a yellow star, a red heart, and a green-brown tree. Both visual and auditory stimuli were presented visually on a screen positioned at face height at a viewing distance of ~100 cm. Importantly, the auditory stimulus words (Ehre, Absicht, and Idee) were presented as written text on the screen, and participants were instructed to imagine pronouncing the word after it was presented. No auditory stimulation was used during the experiment. The block structure, illustrated in Fig. 1, was inspired by earlier imagery research (1, 8, 19) and consisted of auditory as well as visual blocks.

In one trial of an auditory block, the participants first entered a 1-s concentration phase to focus on the upcoming sound, followed by a 1.5-s stimulus presentation. They then engaged in 2.5 s of auditory imagery, mentally rehearsing the word they had read. This was followed by 1 s of relaxation and 2 s of rest before the next trial. Each of the 3 words was presented 15 times per block, yielding 45 trials per block and resulting in 225 auditory trials in total per participant after completing 5 auditory blocks. The presentation order of the words was randomized in each block.

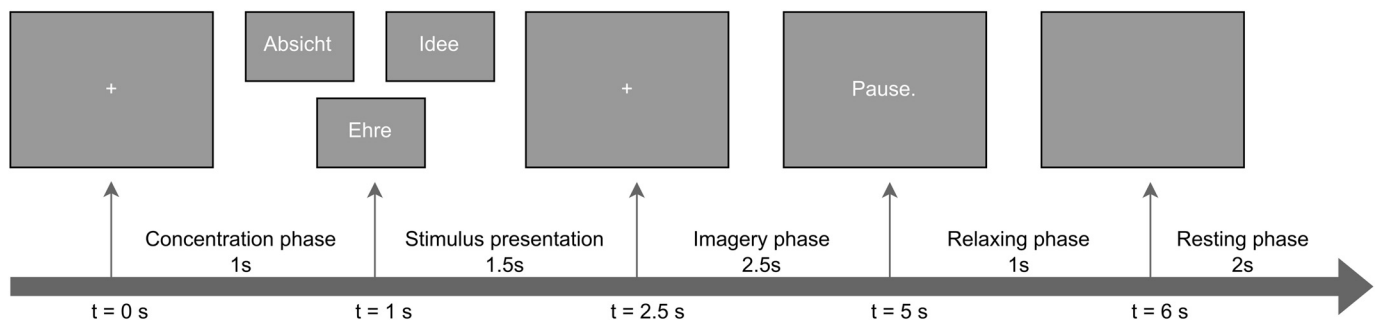
Each trial in the visual block began with a 1-s concentration phase as well, after which the image was displayed for 4 s. Participants then spent 4 s visualizing the image as vividly as possible while avoiding verbal associations with the corresponding object name (star, heart, and tree). A 5-s

relaxation period followed before the next trial. Each image was shown 6 times per block, resulting in 18 trials per block and 90 visual trials in total per participant after completing the 5 visual blocks. The presentation order of the images was randomized in each block.

Each session consisted of five auditory and five visual blocks that were presented alternately. For half of the participants, the session began with an auditory block followed by a visual block and so on; the remaining participants started with a visual block. Short breaks were provided between blocks.

MEG recordings were obtained with a four-dimensional (4-D) neuroimaging system (248 magnetometers) at the University Hospital Erlangen, following the technical specifications conducted within an earlier study by Schüller et al. (22). Three broken channels were excluded, yielding 245 channels for analysis. Data were recorded at a sampling frequency of 1,017.25 Hz with participants lying supine, eyes open, and an analog high-pass filter at 1 Hz was applied online. A Polhemus-integrated digitizer recorded five anatomical landmarks. Environmental noise was suppressed using the manufacturer's calibrated linear weighting algorithm on 23 reference sensors. A stimulation computer interfaced with an external USB audio device providing five analog outputs; two fed an audio amplifier, and one was additionally routed to an MEG analog input channel. To mark imagery onset and offset, inaudible 300-Hz, 0.01-s trigger tones were generated and recorded via an analog audio

One trial in the auditory block



One trial in the visual block

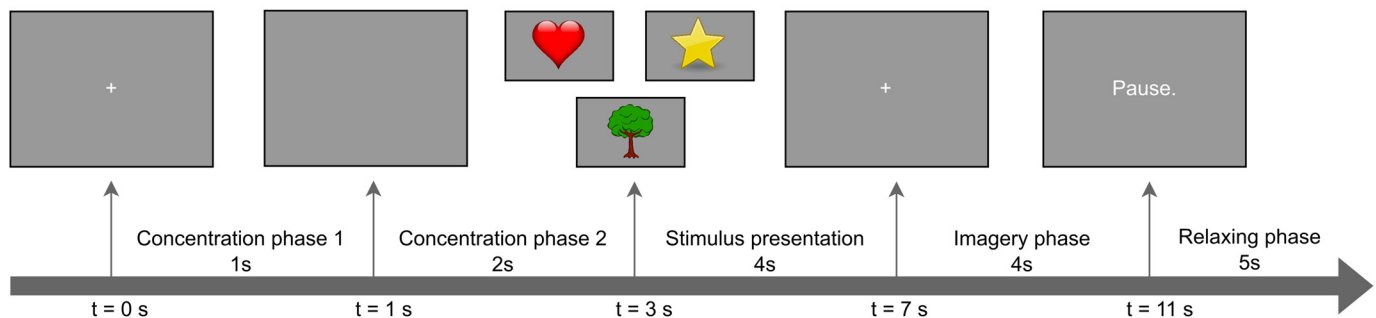


Figure 1. Trial workflow for auditory and visual imagery tasks. Each trial in the auditory block began with a 1-s concentration phase, followed by a 1.5-s word presentation, 2.5-s auditory imagery, 1-s relaxation, and 2-s rest. Each trial in the visual block began with a 1-s concentration phase, followed by a 4-s image presentation, 4-s visual imagery, and 5-s relaxation period. Auditory stimuli consisted of the 3 German abstract nouns Ehre (honor), Absicht (intention), and Idee (idea). The visual stimuli comprised 3 color-distinct images (yellow star, red heart, green-brown tree).

reference channel, enabling temporal alignment of the imaged intervals with MEG signals. The participants' heads were recorded with a camera and monitored outside the MEG chamber to control for head and facial movements. Electrooculogram (EOG) activity was recorded to monitor facial and lip movements, and no interference signals were observed.

Finally, participants underwent magnetic resonance imaging (MRI) to facilitate individual source reconstruction based on each participant's anatomical template. T1-weighted MR images were acquired using a Siemens 3 T Magnetom Cima.X scanner.

Learning Preferences and Memorization Tasks

Apart from MEG acquisition, on a different day, participants completed visual and auditory memorization tasks. In the visual task, a 6×6 grid of white squares was presented on a computer screen. In each trial, seven randomly chosen squares were sequentially highlighted in red, after which participants were asked to reproduce the sequence by clicking them in the correct order. A recalled item was scored as correct only if both its identity and its serial position matched the presented sequence. In the auditory task, subjects heard a sequence of 7 syllables randomly drawn from a set of 35, presented via earphones by an AI-generated female voice (Google Text-to-Speech, female German voice). Participants reproduced the sequence by selecting the correct syllables from a displayed grid; again, both identity and serial position were required for a response to be scored as correct. Both tasks were performed five times resulting in five trials. Performance was quantified through the auditory memorization score $MT_a = \frac{n_{corr,a}}{35}$ for the auditory memorization task and the visual memorization score $MT_v = \frac{n_{corr,v}}{35}$ for the visual memorization task, where $n_{corr,a}$ and $n_{corr,v}$ are the number of correctly recalled syllables and squares, respectively.

All stimulus presentations described above, including the stimulus sequence during the MEG session and the memorization tasks, were programmed in PsychoPy, an open-source Python package for designing and running experiments in behavioral and cognitive neuroscience (23).

In addition to the memorization tasks, participants completed the VARK questionnaire, developed by (21), which assesses individual learning preferences through 16 questions, with each offering four response options corresponding to four modalities: visual (V), auditory (A), read/write (R), and kinesthetic (K). The resulting four VARK scores were calculated for each participant as follows: $VARK_i = \frac{n_i}{16}$, where n_i denotes the number of selected response options and $i = a, v, r, \text{ or } k$ denotes auditory, visual, read/write, or kinesthetic, respectively. Table 1 depicts the scores of the memorization tasks and VARK questionnaire, averaged across subjects.

Source Reconstruction

The acquired MR scans were processed using FreeSurfer (24), and an aparc + aseg volume parcellation was applied to the brain volumes. To localize the neural signals spatially, the raw MEG data were source-reconstructed using a linearly constrained minimum variance (LCMV) beamformer onto regions of interest (ROIs) defined from each participant's individual MRI template. The LCMV beamformer was chosen because it operates in the time domain and preserves the full temporal structure of source signals, which was required for the subsequent single-trial continuous wavelet transform decomposition. A frequency-specific approach such as dynamic imaging of coherent sources (DICS), while well suited for spectral power analyses, does not readily yield single-trial time-domain estimates compatible with our classification pipeline. Two separate cortical ROIs, one auditory and one visual, were created to target modality-specific processing. ROIs were defined using the aparc + aseg parcellation in each participant's native MRI space; the source-reconstructed data were subsequently morphed to the fsaverage brain template. The auditory ROI comprised 202 source points, and the visual ROI contained 107 source points.

Preprocessing

MEG data preprocessing was carried out in MNE-Python (25, 26). A 50-Hz notch filter was applied to remove power line noise, and the data were resampled to 1,000 Hz using MNE-Python's built-in resample function, which applies a zero-phase FIR low-pass filter at the Nyquist frequency of the target sampling rate (500 Hz) before decimation as anti-aliasing. Given the minimal difference between the original (1,017.25 Hz) and target (1,000 Hz) sampling rates, the effect of resampling on the neural signal is negligible.

Individual imagery trials were segmented from the continuous MEG signal using the trigger timestamps recorded in the reference channel. For the main classification analyses reported here, the full imagery phase was used: 0–2,500 ms for the auditory task (2,500 temporal samples) and 0–4,000 ms for the visual task (4,000 temporal samples). All epochs were time-locked to imagery onset, not stimulus onset, ensuring that stimulus-evoked neural activity does not directly contribute to the classification results. In a subsequent, independent analysis, we assessed the robustness across different time windows within the imagery phase by additionally training separate models on the first half (0–1,250 ms for the auditory task; 0–2,000 ms for the visual task) and second half (1,250–2,500 ms for the auditory task; 2,000–4,000 ms for the visual task) of the imagery phase.

Data Augmentation

To increase the size and variability of the MEG data set while preserving class labels, we applied a simple augmentation

Table 1. Mean scores of the memorization tasks and VARK questionnaire

	MT_a	MT_v	$VARK_a$	$VARK_v$	$VARK_r$	$VARK_k$
Mean score	0.29 ± 0.15	0.59 ± 0.14	0.45 ± 0.18	0.54 ± 0.17	0.36 ± 0.14	0.47 ± 0.21

Data are means $\pm$ standard error. The subject-averaged auditory and visual memorization task scores (MT_a and MT_v) and the 4 visual, auditory, read-write, and kinesthetic (VARK) modality scores ($VARK_a$, $VARK_v$, $VARK_k$, and $VARK_r$) are presented.

through temporal shifts to the source-reconstructed trials (auditory: $225 \times 202 \times 2,500$; visual: $90 \times 107 \times 4,000$; trials $\times$ source points $\times$ time points). For each trial, we created two additional, time-shifted copies by shifting the signal forward by 100 ms and 200 ms, respectively. This procedure was inspired by the study of Dash et al. (1).

Shifting was implemented at the sample level (sampling rate: 1,000 Hz) by removing the first n samples (corresponding to the desired delay) and padding zeros at the end of each trial. The original, 100-ms-shifted, and 200-ms-shifted trials were then concatenated, tripling the number of trials (auditory: 675; visual: 270). The class labels were duplicated accordingly so that each augmented trial retained the correct stimulus label.

Such temporal jittering also accounts for intertrial variability in participants' reaction times and imagery onset, thereby compensating for small temporal misalignments in the data while effectively increasing the data set size. To prevent data leakage, every original trial and its augmented copies (+100-ms and +200-ms time shifts) were assigned to the same data split.

Feature Extraction

For each MEG trial (auditory and visual), we derived time-frequency representations using a continuous wavelet transform (CWT). We employed the Python library `scipy.signal.cwt` together with the Morlet wavelet `morlet2` to compute time-frequency representations of each MEG trial. The CWT convolves the signal with scaled and shifted Morlet wavelets, yielding complex coefficients whose magnitude reflects signal power across frequencies and time.

Single-channel signals were decomposed with complex Morlet wavelets spanning 1–120 Hz on a logarithmically spaced grid of 200 center frequencies (sampling rate 1,000 Hz). The wavelet parameter was set to $w = 6$, balancing temporal and spectral resolution. The absolute values of the complex coefficients were retained to yield amplitude scalograms of shape (200 frequencies $\times$ T time points), where $T = 2,500$ for the auditory task and $T = 4,000$ for the visual task. For the first and second half imagery phase analyses, $T = 1,250$ for the auditory task and $T = 2,000$ for the visual task.

This procedure was applied independently to each source point signal, resulting in a set of channel-wise two-dimensional scalograms per trial. Inspired by the approach of Dash et al. (1), the individual channel scalograms were spatially arranged into a single composite image representing the time-frequency activity across all sources in the ROI. For the auditory ROI (202 source points arranged in a 15×14 grid), the composite image had dimensions $(15 \times 200) \times (14 \times T)$ pixels before resizing, i.e., $3,000 \times 35,000$ for the auditory task ($T = 2,500$ or $T = 1,250$). For the visual ROI (107 source points in an 11×10 grid), the pre-resize dimensions were $(11 \times 200) \times (10 \times T) = 2,200 \times 40,000$ for the visual task ($T = 4,000$ or $T = 2,000$). The amplitude matrices were visualized using the matplotlib `viridis` colormap, which maps low amplitude values to dark violet/blue and high-amplitude values to yellow/green tones. These composite images were saved as $1,000 \times 1,000 \times 3$ arrays by bilinear interpolation, where the three channels corresponded to the RGB-encoded red, green, and blue components of the color-

mapped amplitude values. This RGB encoding increased the dimensionality and complexity of the input, allowing the deep neural network to exploit richer feature hierarchies during training. We note that collapsing source location, time, and frequency into a single composite image sacrifices direct interpretability: while gradient-based attribution methods (e.g., GradCAM) were applied to the trained models, the resulting activation maps did not yield neuroscientifically interpretable patterns due to the nonphysical spatial layout of the composite images. Developing more interpretable representations, for instance, applying separate classifiers to predefined frequency bands, remains an important direction for future work.

Model Variants

Data were analyzed using a within-subject design, with separate models trained for each participant. Both ROIs described above were used to examine task-specific neural activity, resulting in two model groups. The first group used source-reconstructed data from the task-relevant cortical region (e.g., visual cortex for the visual task), ensuring that model inputs matched the functional demands of the condition. The second group used data from the nontask-relevant cortex (e.g., visual cortex for the auditory task). This setup enabled us to assess the specificity of neural representations and to reveal potential cross-modal activity.

Classification Frameworks

To evaluate both simple and complex decision boundaries, we trained two complementary classifiers on the wavelet-scalogram features: a regularized linear model and a deep convolutional neural network. For the linear model, the scalogram images ($1,000 \times 1,000 \times 3$) were converted to single-channel grayscale, as the linear classifier requires one-dimensional feature vectors rather than multi-channel image data. The grayscale images were then resized to 256×256 pixels to reduce computational load and z-scored per image. This simple linear architecture served as a baseline, assessing whether class information could be separated by a linear decision boundary in feature space. The model will, in the following, be referred to as LogReg. We performed a 70/15/15 split of data for training, validation, and testing, respectively. To prevent data leakage, every original trial and its augmented copies (+100-ms and +200-ms time shifts) were assigned to the same set.

To capture more complex, nonlinear relationships in the time-frequency representations, we used AlexNet (27). AlexNet comprises stacked convolutional layers with nonlinear activations and pooling, followed by fully connected layers. This hierarchical architecture learns progressively more abstract spatial features, from simple edges and frequency contours to complex joint time-frequency motifs, directly from the input images. Due to the MEG scalograms being effectively two-dimensional time-frequency images, AlexNet is particularly well suited to extract multiscale spatial structure and detect discriminative oscillatory patterns across sources and timescales. Importantly, Dash et al. (1) previously demonstrated that AlexNet yields strong performance for imagined speech decoding, providing additional justification

for its use in the present study. The model will, in the following, be referred to as CNN.

For input to the CNN, the scalograms were, as for the LogReg model, resized to 256×256 pixels and normalized. The network was initialized with ImageNet weights and fine tuned employing the Adam optimizer using a learning rate of 0.001 and a batch size of 64. We applied the same grouping strategy to the train, validation, and test splits as for the linear classifier, guaranteeing that all augmented versions of a given trial remained within a single split. Training proceeded for 30 epochs.

All networks were implemented using the PyTorch framework in Python 3.12.2. Models were trained on an NVIDIA RTX A5000 GPU with 24 GB RAM and CUDA 12.6.

Statistical Evaluation

To determine whether model accuracies were above chance and to directly compare models, we performed one-sample and paired-sample *t* tests, with decoding accuracy as the unit of observation aggregated at the participant level. For each participant and condition, the mean decoding accuracy was tested against the theoretical chance level of 33% (three-class problem) using a one-sample *t* test. To assess whether one classifier outperformed another, we used a paired-sample *t* test comparing their accuracies across participants. All tests were performed in the Python library *scipy* (28), and significance was assessed at $P < 0.05$.

For the behavioral correlation analyses, linear mixed-effects models (LMEs) were fitted separately for each model-type-task-ROI combination, with decoding accuracy as the dependent variable and the behavioral measures (MT_a , MT_v , $VARK_a$, $VARK_v$, $VARK_k$, and $VARK_r$) as fixed-effect predictors. Participant identity was included as a random intercept to account for within-subject correlations. By considering all combinations of classifier type (CNN or logistic regression), task (auditory and visual), and ROI (task relevant or nontask relevant), we obtained eight model-type-task-ROI combinations. *P* values were corrected for multiple comparisons using Benjamini-Hochberg false discovery rate (FDR) correction applied jointly across all six behavioral predictors within each combination and across the eight combinations, yielding a total correction for 48 comparisons. The behavioral analyses are exploratory in nature, and results should be interpreted accordingly.

RESULTS

General Model Performances and Regional Specificity

The mean accuracies of the model for the auditory and visual decoding tasks are visualized in Fig. 2. The LogReg models, as well as the CNN approach, were trained and evaluated once per participant for the whole imagery phase. We trained one network per subject using functionally relevant data and a second model using data from the nontask-relevant ROI. The two investigated ROIs are shown in Fig. 2A. This resulted in task-relevant models, namely for the auditory task on auditory cortex (for both LogReg and CNN) and for the visual task on visual cortex (for both LogReg and CNN). We also computed nontask-relevant models, namely for the auditory task on visual cortex (for both LogReg and

CNN), as well as for the visual task on auditory cortex (for both LogReg and CNN).

Figure 2B demonstrates the classification accuracies of the CNN models and the LogReg models for the different conditions. Significance annotations indicate whether performance significantly exceeds chance. In Fig. 2B, the classification accuracies of the CNN models and the LogReg models are compared to the chance level (33%, three-class classification). Figure 2B, left, depicts the data from the auditory task, and in Fig 2B, right, the data from the visual task are presented. The performances of the LogReg models are shown in purple, and the CNN accuracies are shown in orange. The darker purple and orange depict accuracies of the models that were trained on the task-relevant ROI, and the lighter purple and orange are the accuracies of the models that were trained on the respective nontask-relevant ROI. All models trained on data from the functionally relevant ROI (dark purple and dark orange), on average, significantly exceeded the chance level (33%). This was the case for both the auditory and the visual tasks (Fig. 2B). For the CNN models, this difference was statistically significant with $P < 0.001$ for both tasks. The LogReg models also outperformed the chance level with $P < 0.05$ (auditory task) and $P < 0.01$ (visual task). Regarding the nontask-relevant ROIs, both the CNN and the LogReg model accuracies significantly exceeded chance level for the auditory task (CNN: $P < 0.001$; LogReg: $P < 0.01$), indicating that auditory imagery representations are decodable from both auditory and visual cortex. For the visual task, the CNN significantly outperformed chance level in both the task-relevant and nontask-relevant ROI ($P < 0.001$), whereas the LogReg model, on average, did not exceed chance level in the nontask-relevant auditory ROI.

Comparing the models trained on the task-relevant ROI with those trained on the nontask-relevant ROI, significant differences were found for both models in the visual task (LogReg: $P < 0.05$; CNN: $P < 0.001$, both Bonferroni corrected), where the task-relevant models outperformed the nontask-relevant models. No significant differences between ROIs were found for either model in the auditory task.

LogReg versus CNN: Performance Evaluation

To compare the CNN-based approach with the linear approach, we contrasted model accuracies of LogReg against the CNN accuracies, concentrating on the task-relevant ROI. As shown in Fig. 2B, in the auditory task, the CNN significantly outperformed LogReg ($P < 0.01$, Bonferroni corrected). In the visual task, the CNN likewise significantly outperformed LogReg ($P < 0.001$, Bonferroni corrected), reflecting the CNN's superior ability to exploit the richer and more stable neural representations present in the visual imagery condition.

Model Performances during Auditory and Visual Tasks

When comparing the overall performance between data from the auditory versus visual imagery task, we observed superior decoding for the visual than for the auditory imagery condition (Fig. 3). While all task-relevant models significantly exceeded chance level, the CNN model for the visual task significantly outperformed the CNN model for the auditory task (Fig. 3, right, $P < 0.001$, two-sided paired

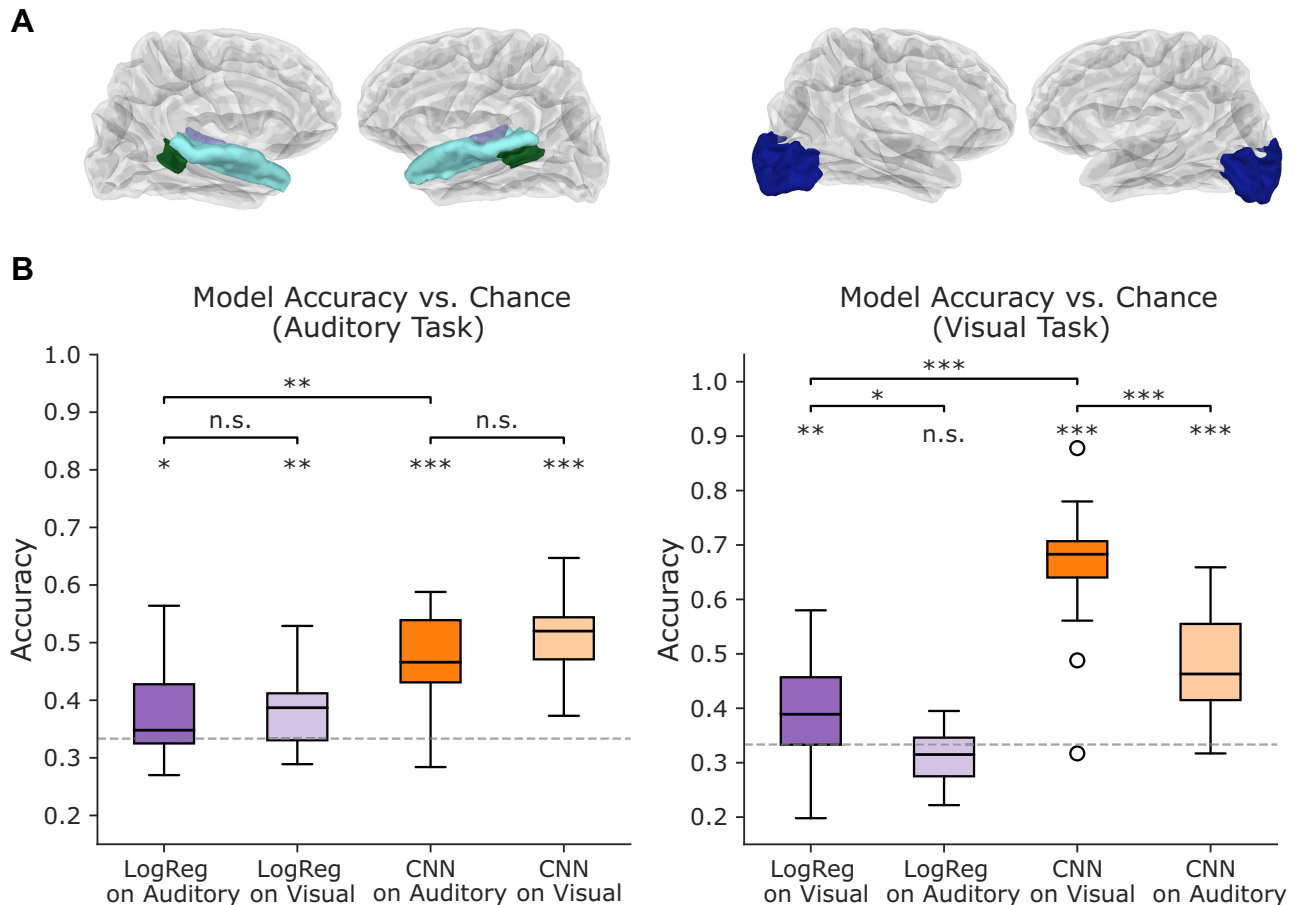


Figure 2. *A*: auditory and visual regions of interest (ROIs) consisting of the transverse temporal gyri (purple), the superior temporal gyri (turquoise), the banks of the superior temporal sulci (dark green) in the auditory ROI (*left*), and the occipital lobe (dark blue) in the visual ROI (*right*). *B*: classification accuracies of the logistic regression model (LogReg; purple) and AlexNet [convolutional neural network (CNN); orange] models for auditory (*left*) and visual (*right*) decoding tasks. Dark colors indicate models trained on the task-relevant ROI (auditory cortex for the auditory task; visual cortex for the visual task), and light colors indicate models trained on the nontask-relevant ROI. Dashed lines mark the three-class chance level (33%). * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$, one-sided t tests against chance, as well as two-sided paired t tests to compare model performances against each other (n.s., not significant).

t test). No significant task-related difference was observed for the LogReg models (Fig. 3, *left*).

Decoding Accuracy across Imagery Time Segments

To characterize how decoding performance evolves across the imagery period and to assess whether results depend on the choice of time window, we trained models on three different segments. The first segment was the full imagery phase (auditory: 0–2,500 ms; visual: 0–4,000 ms), of which the results are already shown in Fig. 2*B*. Additionally, we trained models on the first half (0–1,250 ms and 0–2,000 ms, respectively), and on the second half (1,250–2,500 ms and 2,000–4,000 ms, respectively) of the imagery phase. Results are shown in Fig. 4, organized by classifier (LogReg: Fig. 4, *top*; CNN: Fig. 4, *bottom*), with task-relevant (Fig. 4, *right*, solid) and nontask-relevant (Fig. 4, *left*, transparent) ROIs shown side by side to facilitate direct comparison. For the auditory task, all three time segments yielded above-chance decoding for the CNN across both ROIs (Fig. 4, *bottom left*, $P < 0.001$), and for the LogReg in the task-relevant and nontask-relevant auditory ROIs (Fig. 4, *top left*, $P < 0.05$ and $P < 0.01$, respectively). For the visual task, the CNN significantly

exceeded chance level across all time segments and both ROIs (Fig. 4, *bottom right*, $P < 0.001$), while the LogReg exceeded chance only for the task-relevant visual ROI in the full and first-half segments (Fig. 4, *top right*, $P < 0.01$). Importantly, the second half of the visual imagery phase, which contains no stimulus-evoked activity, yielded above-chance CNN decoding in both task-relevant and nontask-relevant ROIs, confirming that imagery-related neural representations persist throughout the imagery period and are not exclusively driven by stimulus-evoked activity.

The models trained on the task-relevant cortex did significantly outperform the models trained on the nontask-relevant cortex for the visual task and when the full imagery period was considered only (Fig. 4, *right*). For all other models, we found no significant difference in performance when comparing the task-relevant versus nontask-relevant cortex.

Correlation between Memorization Task Performance, VARK Scores, and Model Accuracies

After evaluating the performance of the models in general, we investigated putative correlations between the decoding accuracies of the models trained on the full imagery phase

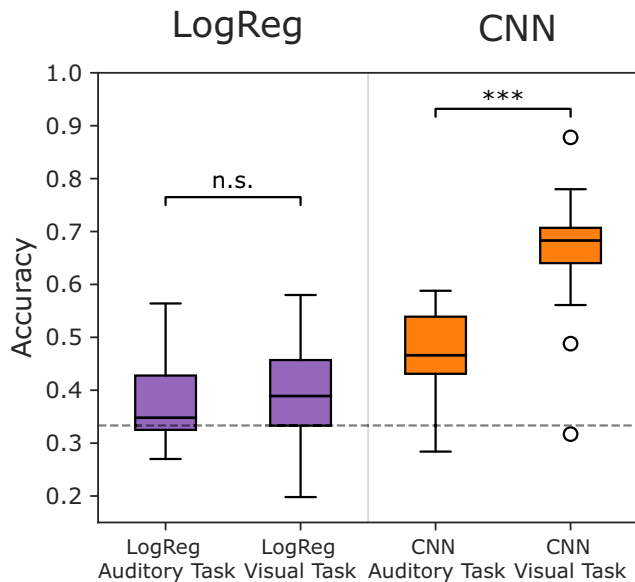


Figure 3. Comparison of decoding performances between auditory and visual imagination tasks. Mean classification accuracies of the logistic regression (LogReg; *left*, purple) and convolutional neural network (CNN; *right*, orange) models trained on task-relevant regions of interest (ROIs) are shown. All models significantly exceeded the three-class chance level of 33%. The CNN model achieved significantly higher accuracy in the visual task than in the auditory task (** $P < 0.001$, two-sided paired t test, Bonferroni corrected), whereas no significant task-related difference was observed for the LogReg models (n.s., not significant).

and the additional behavioral data, which indicated memory capacity and individual learning preferences collected from the subjects. We note that these analyses are exploratory in nature, given the small sample size ($n = 18$) and the absence of preregistered hypotheses; results should therefore be interpreted with caution. We computed linear mixed-effects regression models (LMEs), in which decoding accuracy served as the dependent variable, and the memorization task and VARK scores (MT_a , MT_v , $VARK_a$, $VARK_v$, $VARK_k$, and $VARK_r$) served as fixed-effect predictors, with participant as a random intercept. We performed this analysis for the decoding accuracies of the LogReg as well as the CNN model for both tasks, auditory and visual, on the task-relevant ROI and also on the nontask-relevant ROI.

Figure 5A depicts the results for the auditory (Fig. 5A, *left*) and visual (Fig. 5A, *right*) classification tasks with the data that was source-reconstructed in the task-relevant ROI (auditory task data on auditory cortex and visual task data on visual cortex). The memorization task performances MT_a and MT_v and the VARK scores were tested for dependency on the LogReg accuracies (Fig. 5A, *top*) and on the CNN accuracies (Fig. 5A, *bottom*). For the auditory task, the CNN accuracies showed a significant positive association with the kinesthetic $VARK_k$ score ($P < 0.001$, FDR corrected across all 48 comparisons). No significant associations were observed for the visual task in the task-relevant ROI. The respective results for the auditory and visual classification tasks with the data reconstructed in the nontask-relevant ROI (auditory task data on visual cortex and visual task data on auditory cortex) can be seen in Fig. 5B. Here, no significant influence of behavioral data on the model performance can be observed, neither regarding the auditory task nor the visual task.

DISCUSSION

General Model Performances and Regional Specificity

Both the logistic regression and CNN models trained on task-relevant cortical regions achieved above-chance decoding for both the auditory and visual imagery tasks, confirming that modality-specific regions provide informative neural features for imagery decoding. Strikingly, above-chance decoding was also observed for models trained on nontask-relevant regions: for the auditory task, both the CNN ($P < 0.001$) and the LogReg ($P < 0.01$) exceeded the chance level when trained on the visual cortex, while for the visual task, only the CNN significantly exceeded chance when trained on the auditory cortex ($P < 0.001$). For the visual task, although the CNN trained on the nontask-relevant auditory cortex also significantly exceeded the chance level ($P < 0.001$), its performance was significantly lower than that of the task-relevant model ($P < 0.001$, Bonferroni corrected), indicating that the visual cortex provides substantially more informative features for visual imagery decoding than the auditory cortex.

Nonetheless, the above-chance decoding accuracy for the visual task in the auditory cortex indicates that the neural representations underlying imagery are not strictly localized but instead distributed across partially overlapping cortical networks, consistent with the broad literature on cross-modal processing (11, 12, 14). Nishimura et al. (10) reported that visual thinkers activate visual cortical regions even during verbal tasks, indicating that visual processing can accompany language-related activity. Such overlapping and individually variable activation patterns may explain why decoding succeeded even in nontask-relevant regions and why the CNN models, in particular, were sensitive to these distributed neural representations.

Additionally, several experimental factors may contribute to this finding. During the auditory task, participants viewed the written word cue on the screen and were instructed to keep their eyes open and fixate on a central cross throughout the imagery period. Since auditory stimuli were presented as written text rather than acoustically, this cross-modal presentation design likely activated occipital regions, which may have carried stimulus-specific information correlated with the imagery process and thus enabled decoding from the visual cortex. This is a potential explanation for the slightly higher decoding accuracy of the LogReg model as well as the CNN for the auditory task on the nontask-relevant cortex and will be discussed further as a limitation of the present study. In contrast, during the visual task, participants might have silently labeled the presented images (“Herz,” “Stern,” and “Baum”), introducing activity in auditory or language-related areas.

Together, these effects could explain why decoding performance was above chance even in nontask-relevant regions and why cross-modal contributions emerged especially in the CNN models. Unlike the linear classifier, which relies on a single global decision boundary, the CNN can capture nonlinear and spatially distributed feature patterns that may jointly encode information from multiple sensory systems. Subtle and spatially overlapping activations in the auditory cortex, arising from verbal labeling during the visual task, are likely too weak or heterogeneous to be linearly

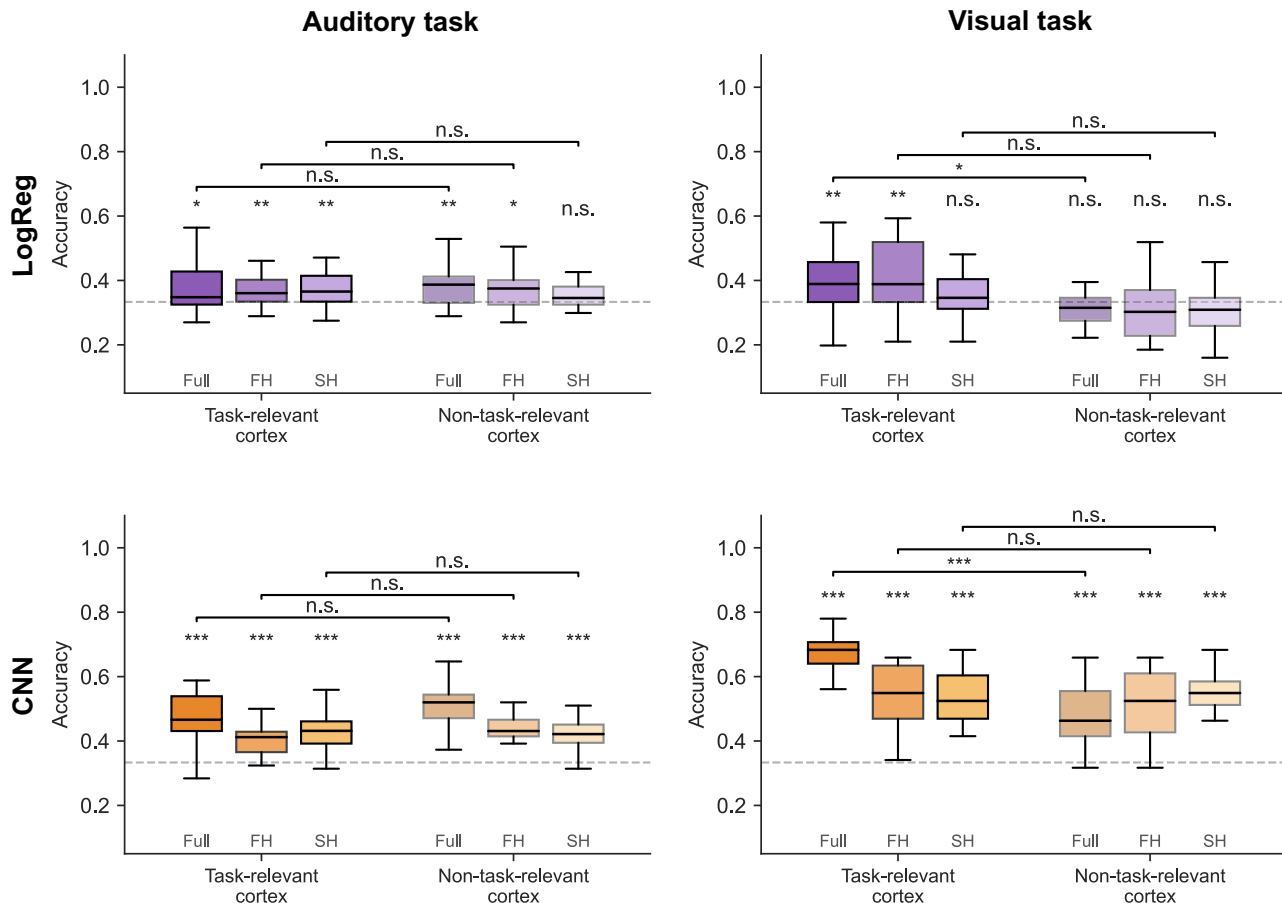


Figure 4. Decoding accuracy across imagery time segments. *Top, bottom, left, and right:* classification accuracies for the logistic regression (*top*, purple) and convolutional neural network (CNN; *bottom*, orange) models, separately for the auditory (*left*) and visual (*right*) tasks. Task-relevant region of interest (ROI) results (*left*, solid) and nontask-relevant ROI results (*right*, transparent) are shown side by side for 3 time segments: full imagery phase (Full, dark), first half (FH; medium), and second half (SH; light). Brackets with significance markers indicate paired comparisons between task-relevant and nontask-relevant ROIs per segment (Bonferroni corrected, $n = 3$). * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$, one-sided t tests against chance level (n.s., not significant). The dashed line marks the three-class chance level (33.3%).

separable. However, the hierarchical convolutional layers of the CNN can integrate these small, distributed signals across sources and time-frequency scales, effectively exploiting multimodal information that the linear model cannot. The results are consistent with previous evidence for functional overlap between visual and auditory processing systems, where visual areas are recruited during inner speech and verbal thinking (7). The ability of deep neural networks to exploit such distributed, cross-modal activity patterns underscores their potential for decoding complex cognitive states that engage multiple sensory systems simultaneously.

LogReg versus CNN: Performance Evaluation

The comparison of a simple linear classifier (LogReg) with a CNN-based approach (AlexNet) revealed a consistent advantage of the CNN across both tasks. In the auditory task, the CNN significantly outperformed LogReg ($P < 0.01$), and in the visual task, the CNN likewise outperformed LogReg ($P < 0.001$), reflecting its ability to capture complex, high-dimensional time-frequency features in both conditions. This result aligns with earlier work demonstrating the suitability of AlexNet for imagined speech decoding (1) and

extends this finding to visual imagery. The CNN's advantage was particularly pronounced in the visual task, where mean decoding accuracy was 66%, with a peak subject-individual performance of 88% likely reflecting the richer and more temporally sustained neural representations associated with visual imagery compared to the shorter imagined speech responses.

One explanation for the CNN's superior performance across both modalities is that the full imagery phase provides ample temporal structure for hierarchical feature learning: the auditory data set comprised 202 source points over 2,500 ms, and the visual data set comprised 107 source points over 4,000 ms. The visual task's 4-s imagery phase, during which participants visualized a static image, likely yielded a higher signal-to-noise ratio and more temporally sustained neural representations compared to the auditory task, where imagined speech is inherently brief and a lower signal-to-noise ratio is likely. The CNN's ability to exploit these richer, more sustained representations may explain the particularly high decoding accuracies observed for visual imagery. Together, these findings indicate that the CNN is the preferred classifier for MEG-based imagery decoding across both

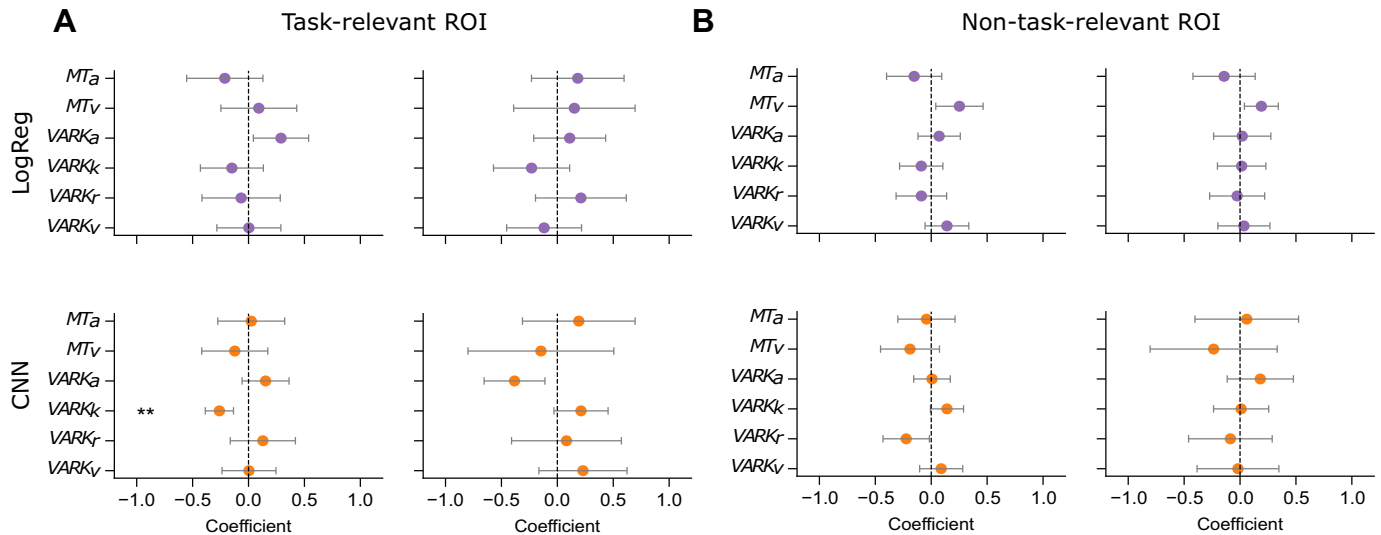


Figure 5. Relationship between behavioral measures and decoding accuracies. **A:** results of linear mixed-effects regression models for the auditory (*left*) and visual (*right*) classification tasks in the respective task-relevant region of interest (ROI): model coefficients ($\pm$ standard error) for the influence of auditory and visual memorization task performance (MT_a and MT_v) and visual, auditory, read-write, and kinesthetic (VARK) subscores ($VARK_a$, $VARK_v$, $VARK_k$, and $VARK_r$) on the logistic regression (LogReg; *top*) and convolutional neural network (CNN; *bottom*) accuracies for the auditory task (*left*) and the visual task (*right*). Significance is indicated: $**P < 0.01$, all corrected for 48 comparisons. **B:** results for the auditory (*left*) and visual (*right*) classification tasks in the respective nontask-relevant ROI.

modalities and that its advantage is especially pronounced when neural representations are temporally sustained as in the case of visual imagery.

Model Performance across Tasks

Overall, decoding accuracies were higher in the visual task compared to the auditory condition, particularly and significantly for the CNN model ($P < 0.001$). Several factors likely contribute to this difference. Visual imagination occurred within the same modality as stimulus presentation (image to image), whereas auditory imagination required a cross-modal transformation from visual cues to internal auditory representations, likely imposing additional cognitive demands. In terms of working memory, visual imagery is naturally supported by the visuospatial sketchpad, whereas auditory imagery relies on the phonological loop, a distinction consistent with Baddeley's model of working memory (29), with the phonological loop potentially being more effortful to engage given the cross-modal nature of the auditory task.

We additionally note that, although all epochs were time-locked to imagery onset, the choice of analysis window length may influence decoding performance; analyses across the full imagery phase, its first half, and its second half confirmed that above-chance decoding was consistent across all time windows (Fig. 3). Kilmarx et al. (18) found that visual imagery following direct visual presentation produced stronger EEG signatures than spontaneous imagery from memory, suggesting that short-term, cue-based visual imagery elicits robust and classifiable neural responses. The sustained visual engagement during our trials may have similarly facilitated stable neural representations and enhanced model performance.

Interestingly, while the CNN outperformed LogReg in both tasks, it also showed a stronger task-related difference,

achieving substantially higher accuracy in the visual than in the auditory condition. This suggests that the CNN is particularly effective at exploiting the richer and more stable neural representations associated with visual imagery, while still capturing sufficient information from auditory imagery to significantly exceed chance level. The observation that CNNs could decode imagery from nonrelevant regions further supports the idea that imagined perception involves widespread cortical engagement, possibly mediated by top-down feedback between sensory areas. Such cross-modal representational overlap may also imply that combining auditory and visual imagery in future paradigms could enhance decoding performance by exploiting these shared neural substrates.

Decoding across Imagery Time Segments

The analyses across full, first-half, and second-half imagery windows (Fig. 3) reveal several important patterns. First, above-chance CNN decoding was obtained across all time windows and both ROIs, confirming that imagery-related information is distributed throughout the entire imagery period rather than being confined to its onset. Second, the finding that the second half of the imagery phase also yielded above-chance decoding is particularly noteworthy: it provides strong evidence that the classification reflects genuine mental imagery representations rather than residual stimulus processing. Third, for the LogReg model, decoding tended to be highest for the full imagery window, consistent with the idea that the linear model benefits from averaging over a longer signal epoch. In contrast, CNN performance was more stable across time segments, suggesting that the CNN extracts temporally local features that remain discriminative throughout the imagery period.

Correlation between Memorization Task Performance, VARK Scores, and Model Accuracies

Beyond general decoding performance, exploratory analyses revealed one significant association between behavioral measures and model accuracies after FDR correction across all 48 comparisons: CNN accuracies for the auditory task were negatively correlated with kinesthetic learning preference (VARK_k, $P < 0.001$). This suggests that participants who engage more embodied or motor-related cognitive strategies may generate poorer auditory neural patterns that deep models can exploit. However, given the exploratory nature of these analyses and the small sample size ($n = 18$), these findings should be interpreted with caution.

The fact that behavioral correlations emerged for only the CNN, not the LogReg model, suggests that each classifier type may capture partly different aspects of the underlying neural signal. The linear model appears to reflect more direct, modality-specific associations in the auditory task, such as auditory preference improving imagined speech decoding, whereas the CNN, by modeling more complex signal patterns, may emphasize other dimensions of individual variability. Importantly, the absence of overlapping effects between models indicates that these relationships are not driven by subject-individual learning preferences or task performances but by how each architecture extracts information from the MEG data. This model-specific sensitivity underlines that behavioral-neural correlations should be interpreted carefully in light of the decoding approach used, as different algorithms may highlight distinct facets of the same cognitive process. Notably, no significant behavioral effects were found in models trained on nontask-relevant regions.

Limitations and Conclusion

Although the classification accuracies were above chance level, it must be acknowledged that the overall decoding performance remained relatively low in the auditory task, which could be due to several different limitations in our work. One potential aspect impeding classification performance during the auditory task could be the challenge of accurately identifying the timing of the imagined event. Temporal misalignment in the recorded neural signals can introduce noise, potentially reducing the effectiveness of the classification models. We accounted for this by using the full imagery phase (auditory: 0–2,500 ms; visual: 0–4,000 ms) as the main analysis window, and additionally validated results on the first and second half of the imagery phase. However, since the length and, thus, articulation duration of the words (Absicht, Idee, and Ehre) differ from subject to subject and even between trials, especially the data where the shorter words had to be imagined, are likely to still contain significant noise.

A further methodological consideration concerns the stimulus presentation modality in the auditory task. Rather than being presented acoustically, the auditory stimuli (German words) were displayed as written text on the screen, and participants were instructed to mentally rehearse their pronunciation. This cross-modal presentation, a visual cue prompting auditory imagery, may have activated visual cortical regions during the auditory task, potentially contributing to the above-chance decoding

observed in the nontask-relevant visual ROI. Similarly, the sustained visual engagement required to read the written cues may have introduced systematic visual cortical activity that is partially stimulus specific, thereby facilitating cross-modal decoding. This confound could be addressed in future studies by using genuinely auditory stimulus presentation, which would allow a cleaner dissociation between auditory and visual cortical contributions during imagery.

Moreover, methodological considerations concern the choice of source reconstruction algorithm. The LCMV beamformer was selected because it operates in the time domain and preserves single-trial temporal structure, which was required for the subsequent CWT decomposition. However, as a broadband spatial filter, it does not optimally suppress interference from out-of-band neural activity compared to a frequency-specific alternative such as DICS. Future studies employing frequency-resolved classification pipelines may benefit from applying separate, frequency-band-specific beamformers before feature extraction.

We wish to clarify that all analysis epochs were time locked to imagery onset rather than stimulus onset, ensuring that stimulus-evoked activity does not directly contribute to the decoding results. The main analyses reported here use the full imagery phase. To further validate robustness, we additionally trained models on the first half and second half of the imagery phase; above-chance decoding was obtained across all time windows and both classifier types. Notably, the second half of the imagery phase, which contains no stimulus-evoked activity, also yielded above-chance CNN decoding, confirming that the results reflect genuine imagery-related neural representations. A systematic time-resolved decoding analysis such as that done by Dash et al. (1) for speech imagery, which would capitalize on MEG's high temporal resolution to track the dynamics of imagery-related neural representations over time, remains an important direction for future work, ideally with a larger number of trials per condition enabled by a modality-specific experimental design.

An important limitation of the present study is the absence of a compliance check for the behavioral task during the MEG session. Participants were instructed to engage in unimodal imagery (e.g., to avoid verbally labeling visual images), but adherence to these instructions cannot be verified from MEG data alone. It therefore cannot be excluded that decoding performance partly reflects cross-modal cognitive strategies rather than purely modality-specific imagery. Future studies should incorporate within-trial behavioral probes (e.g., a forced-choice recognition task after each imagery period) to provide an objective measure of task engagement and compliance.

A further limitation concerns participant fatigue and attentional fluctuations. The 1-h task duration was demanding, and no direct behavioral measures of concentration were collected during the breaks. Incorporating brief random control checks or task-related recall prompts, as suggested by Nieto et al. (19), could help track engagement and improve data quality in future studies.

One should keep in mind that the number of participants in the study ($n = 18$) may further limit the power to detect statistical significance, making it more challenging to

identify meaningful correlations or effects, especially when it comes to individual learning preferences. A larger sample size in a future study may provide more statistical power, potentially revealing more robust correlations between behavioral measures and decoding performance. A larger number of subjects would also increase the available amount of training data for the decoding algorithm, presumably further boosting mental imagery detection.

Finally, extending the analysis to higher order multisensory association cortices (e.g., superior temporal sulcus, inferior frontal gyrus) or to whole-brain source estimates could yield more robust and informative decoding signals and represents an important direction for future work, although facing a high computational effort.

Our findings advance the understanding of how cognitive processes such as memory and imagination are represented in the brain and demonstrate the potential of modern deep learning for their decoding. The CNN (AlexNet) achieved robust above-chance performance across modalities and regions, suggesting that deep models can exploit distributed neural information and are less dependent on precise anatomical targeting. This robustness may prove valuable for future BCI development, where recording coverage and user-specific variability often limit decoding performance. Finally, the observed relationship between individual learning preferences and decoding performance points toward the possibility of personalized BCIs that adapt to a user's cognitive strengths, emphasizing visual or auditory channels according to individual processing preferences. Such approaches could enhance the efficiency, robustness, and accessibility of next-generation BCI systems.

DATA AVAILABILITY

Data will be made available upon reasonable request.

GRANTS

This project was supported by the German Federal Ministry of Education and Research (Cluster4Future SEMECO, Project No. 03ZU1210FB) and the German Science Foundation (DFG, Project No. 523344822).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

A.S. and T.R. conceived and designed research; A.S. and J.R. performed experiments; A.S., C.J., and M.S. analyzed data; A.S., C.J., M.S., and T.R. interpreted results of experiments; A.S. prepared figures; A.S. drafted manuscript; C.J., M.S., J.R., and T.R. edited and revised manuscript; A.S., C.J., M.S., J.R., and T.R. approved final version of manuscript.

REFERENCES

1. Dash D, Ferrari P, Wang J. Decoding imagined and spoken phrases from non-invasive neural (MEG) signals. *Front Neurosci* 14: 290, 2020. doi:10.3389/fnins.2020.00290.
2. Berg BV, Donkelaar SV, Alimardani M. Inner speech classification using EEG signals: a deep learning approach. In: *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*. IEEE, 2021, p. 1–4. doi:10.1109/ICHMS53169.2021.9582457.
3. Abiri R, Borhani S, Sellers EW, Jiang Y, Zhao X. A comprehensive review of EEG-based brain-computer interface paradigms. *J Neural Eng* 16: 011001, 2019. doi:10.1088/1741-2552/aaf12e.
4. Rashid M, Sulaiman N, Abdul Majeed AP, Musa RM, Nasir AF, Bari BS, Khatun S. Current status, challenges, and possible solutions of EEG-based brain-computer interface: a comprehensive review. *Front Neurobot* 14: 25, 2020. doi:10.3389/fnbot.2020.00025.
5. Panachakel JT, Ramakrishnan AG. Decoding covert speech from EEG—a comprehensive review. *Front Neurosci* 15: 642251, 2021. doi:10.3389/fnins.2021.642251.
6. Bocquelet F, Hueber T, Girin L, Chabardès S, Yvert B. Key considerations in designing a speech brain-computer interface. *J Physiol Paris* 110: 392–401, 2016. doi:10.1016/j.jphysparis.2017.07.002.
7. Amit E, Hoeflin C, Hamzah N, Fedorenko E. An asymmetrical relationship between verbal and visual thinking: converging evidence from behavior and fMRI. *Neuroimage* 152: 619–627, 2017. doi:10.1016/j.neuroimage.2017.03.029.
8. Kosmyna N, Lindgren JT, Lécuyer A. Attending to visual stimuli versus performing visual imagery as a control strategy for EEG-based brain-computer interfaces. *Sci Rep* 8: 13222, 2018. doi:10.1038/s41598-018-31472-9.
9. Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. *Nature* 381: 520–522, 1996. doi:10.1038/381520a0.
10. Nishimura K, Aoki T, Inagawa M, Tobinaga Y, Iwaki S. Brain activities of visual thinkers and verbal thinkers: a MEG study. *Neurosci Lett* 594: 155–160, 2015. doi:10.1016/j.neulet.2015.03.043.
11. Ghazanfar AA, Schroeder CE. Is neocortex essentially multisensory? *Trends Cogn Sci* 10: 278–285, 2006. doi:10.1016/j.tics.2006.04.008.
12. Vetter P, Smith FW, Muckli L. Decoding sound and imagery content in early visual cortex. *Curr Biol* 24: 1256–1262, 2014. doi:10.1016/j.cub.2014.04.020.
13. Pollicina G, Müller SA, Dalton P, Vetter P. Decoding semantic sound categories in early visual cortex. *Cereb Cortex*, 35: bhaf208, 2025. doi:10.1093/cercor/bhaf208.
14. Gu J, Liu B, Li X, Wang P, Wang B. Cross-modal representations in early visual and auditory cortices revealed by multi-voxel pattern analysis. *Brain Imaging Behav* 14: 1908–1920, 2020. doi:10.1007/s11682-019-00135-2.
15. Iamshchinina P, Karapetian A, Kaiser D, Cichy RM. Resolving the time course of visual and auditory object categorization. *J Neurophysiol* 127: 1622–1628, 2022. doi:10.1152/jn.00515.2021.
16. Lee SH, Lee M, Jeong JH, Lee SW. Towards an EEG-based intuitive BCI communication system using imagined speech and visual imagery. In: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, p. 4409–4414. doi:10.1109/SMC.2019.8914645.
17. Marion G, Di Liberto GM, Shamma SA. The music of silence: part I: responses to musical imagery encode melodic expectations and acoustics. *J Neurosci* 41: 7435–7448, 2021. doi:10.1523/JNEUROSCI.0183-21.2021.
18. Kilmarx J, Tashev I, del R Millán J, Sulzer J, Lewis-Peacock J. Evaluating the feasibility of visual imagery for an EEG-based brain-computer interface. *IEEE Trans Neural Syst Rehabil Eng* 32: 2209–2219, 2024. doi:10.1109/TNSRE.2024.3410870.
19. Nieto N, Peterson V, Rufiner HL, Kamienkowski JE, Spies R. Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition. *Sci Data* 9: 52, 2022. doi:10.1038/s41597-022-01147-2.
20. Ahlfors SP, Mody M. Overview of MEG. *Organ Res Methods* 22: 95–115, 2019. doi:10.1177/1094428116676344.
21. Fleming ND, Mills C. Not another inventory, rather a catalyst for reflection. *To Improve Acad* 11: 137–155, 1992. doi:10.1002/j.2334-4822.1992.tb00213.x.
22. Schüller A, Schilling A, Krauss P, Reichenbach T. The early subcortical response at the fundamental frequency of speech is temporally separated from later cortical contributions. *J Cogn Neurosci* 36: 475–491, 2024. doi:10.1162/jocn_a_02103.
23. Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK. PsychoPy2: experiments in behavior

- made easy. *Behav Res Methods* 51: 195–203, 2019. doi:[10.3758/s13428-018-01193-y](https://doi.org/10.3758/s13428-018-01193-y).
24. **Fischl B.** FreeSurfer. *Neuroimage* 62: 774–781, 2012. doi:[10.1016/j.neuroimage.2012.01.021](https://doi.org/10.1016/j.neuroimage.2012.01.021).
25. **Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, Hämäläinen MS.** MEG and EEG data analysis with MNE-Python. *Front Neurosci* 7: 267, 2013. doi:[10.3389/fnins.2013.00267](https://doi.org/10.3389/fnins.2013.00267).
26. **Larson E, Gramfort A, Engemann DA, Leppakangas J, Brodbeck C, Jas M, Brooks TL, Sassenhagen J, McCloy D, Luessi M, King JR, Höchenberger R, Goj R, Brunner C, Favelier G, van Vliet M, Wronkiewicz M, Rockhill A, Holdgraf C, et al.** Mne-python (v1.9.0). 2024. doi:[10.5281/zenodo.592483](https://doi.org/10.5281/zenodo.592483). [2026 Aug 8].
27. **Krizhevsky A, Sutskever I, Hinton GE.** Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 25, 2012.
28. **Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al.** SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*, 17: 261–272, 2020 [Erratum in *Nat Methods* 17: 352, 2020]. doi:[10.1038/s41592-019-0686-2](https://doi.org/10.1038/s41592-019-0686-2).
29. **Baddeley A.** Working memory: looking back and looking forward. *Nat Rev Neurosci* 4: 829–839, 2003. doi:[10.1038/nrn1201](https://doi.org/10.1038/nrn1201).