POLITECNICO DI TORINO
Repository ISTITUZIONALE

Semantic-aware decoding of covert inner speech: A multimodal EEG–EMG–audio framework

Original
Semantic-aware decoding of covert inner speech: A multimodal EEG–EMG–audio framework / Ahmadi, H., Santamaría-Vázquez, E., Mesin, L., Hornero, R.. - In: BIOMEDICAL SIGNAL PROCESSING AND CONTROL. - ISSN 1746-8094. - 129:(2026). [10.1016/j.bspc.2026.111527]

Availability:
This version is available at: 11583/3015848 since: 2026-09-23T20:08:12Z

Publisher:
Elsevier

Published
DOI:10.1016/j.bspc.2026.111527

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

25 October 2026

Biomedical Signal Processing and Control 129 (2027) 111527

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Semantic-aware decoding of covert inner speech: A multimodal EEG–EMG–audio framework

Hossein Ahmadi [a,*], Eduardo Santamaría-Vázquez [b,c], Luca Mesin [a], Roberto Hornero [b,c]

[a] Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, 10129, Italy
[b] Biomedical Engineering Group, E.T.S. Ingenieros de Telecomunicación, University of Valladolid, Paseo Belén 15, Valladolid, 47011, Spain
[c] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, Madrid, Spain

ARTICLE INFO

Keywords:
Inner speech
Imagined speech
Electroencephalography
Semantic decoding
Multimodal brain–computer interface
Electromyography
Deep learning

ABSTRACT

Non-invasive brain–computer interfaces (BCIs) aim to restore communication by decoding intended messages directly from neural activity, even without audible speech. However, evidence remains limited that non-invasive signals can support subject-held-out decoding of predefined semantic intentions from covert inner speech. We investigate whether a semantic-aware multimodal framework can generalize from overt spoken commands to covert inner speech using scalp electroencephalography (EEG), auxiliary electromyography (EMG), and overt-speech audio supervision. Ten healthy participants produced four everyday commands (*water*, *toilet*, *light*, *pain*) in overt and covert phases. Overt trials included synchronized EEG, EMG, and audio, whereas covert trials were evaluated without audio using EEG with auxiliary EMG. The proposed model learns a shared latent representation from overt multimodal supervision and covert training trials from non-held-out subjects, combining supervised contrastive learning, ArcFace classification, semantic prototype alignment, and overt–covert regularization. Evaluation follows a subject-held-out covert protocol with target-subject overt calibration: covert trials from the held-out participant are never used for training, model selection, or calibration. Across subjects, the model achieves a mean overt-validation accuracy of 0.54 and a mean covert-test accuracy of 0.42 on the four-class task, above the 0.25 chance level. Latent-space and semantic-retrieval analyses indicate that the learned representations retain class-related structure and alignment with text-based semantic prototypes. Ablations reveal that multimodal overt information supports within-subject decoding and representational analysis, whereas EEG-only overt supervision is more robust under held-out-subject covert transfer. These results identify subject variability, modality transfer, and calibration as key challenges for future semantic BCIs.

1. Introduction

Restoring communication for people who cannot speak is a central goal in both neuroscience and assistive technology [1,2]. Loss of speech after stroke, neurodegenerative disease, or severe motor impairment constrains autonomy and social participation, and existing assistive devices rarely match the bandwidth, fluidity, or semantic richness of natural language. Recent intracortical and electrocorticography (ECoG) speech neuroprostheses have demonstrated that rich linguistic content is present in cortical activity and can be decoded in real time, with systems now reconstructing text, phonemes, and continuous speech directly from neural signals [1–3]. These invasive systems demonstrate that high-resolution cortical recordings carry sufficient information to support functional communication rates, but they require neurosurgery, long-term implanted hardware, and specialized clinical infrastructure, severely limiting scalability and access [4].

Scalp electroencephalography (EEG) offers a safer and more portable route toward broader deployment because it is inexpensive, non-surgical, and practical in both clinical and everyday settings. However, its low signal-to-noise ratio (SNR), coarse spatial resolution, and sensitivity to artifacts make language decoding particularly difficult for covert inner speech, that is, internally generated speech without overt articulation [4]. Magnetoencephalography (MEG) provides an informative non-invasive benchmark, achieving high phrase-level performance on closed vocabularies when signal quality and spatial coverage are more favorable [5]. For EEG, recent reviews report steady progress in imagined-speech decoding, but also persistent limitations, including small cue-locked datasets, narrow vocabularies, subject-specific models, and limited online operation [6,7]. Related non-invasive EEG work in the visual domain has shown that preserving scalp-space structure through dense topographic map sequences can

* Corresponding author.
E-mail address: hossein.ahmadi@polito.it (H. Ahmadi).

https://doi.org/10.1016/j.bspc.2026.111527
Received 11 May 2026; Received in revised form 22 July 2026; Accepted 16 September 2026
Available online 23 September 2026
1746-8094/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

improve subject-independent decoding of imagination versus perception, suggesting that explicit spatial representations may help capture subtle cognitive-state differences [8]. Despite advances such as graph-based feature fusion and delay-differential descriptors [9,10], cross-subject generalization and decoding at the level of communicative meaning remain open challenges.

Recent studies have addressed cross-subject imagined-speech EEG through unsupervised domain adaptation, prototypical few-shot learning, zero- and few-shot subject-based meta-learning, and subject-invariant representation learning [11–14]. Large-scale magnetoencephalography and EEG (M/EEG) studies have also extended non-invasive decoding to individual words across participants, devices, tasks, and languages, although during listening and reading rather than covert speech production [15].

Insights from invasive work help explain why this goal is plausible, by clarifying what information is present in speech-related neural activity and how it is distributed across space, time, and frequency. Deep neural models trained on ECoG recordings from the auditory cortex can reconstruct intelligible speech during auditory perception, with performance boosted by combining low-frequency and high-gamma activity and by predicting intermediate vocoder parameters instead of raw spectrograms [16]. Deep convolutional architectures tailored to EEG have enabled more accurate and user-friendly assistive BCIs, for example through inception-style networks designed for event-related potential (ERP) decoding in practical communication settings [17]. High-gamma decoders trained on overt speech often generalize only partially to covert speech, exhibiting a consistent overt–covert gap, particularly in superior temporal regions [18]. Closed-loop stereo-EEG (sEEG) systems have demonstrated real-time audio synthesis during whispered and imagined speech, establishing the feasibility of online feedback and user adaptation [19]. These milestones motivate semantic decoding as a realistic target [20], but the invasiveness of the underlying recordings prevents widespread use. Here, we use semantic decoding to mean recovery of the intended communicative meaning of an inner-speech command rather than its exact acoustic or articulatory form, in line with the broader perspective of semantic communication [21]. In the present study, this notion is operationalized as closed-set classification of four predefined communicative intentions together with alignment to fixed sentence-embedding prototypes; it does not imply open-vocabulary, zero-shot, compositional, or generative language inference.

Crucially, recent analyses illuminate how covert or imagined speech differs from overt production at the neural level. Imagined speech appears to rely more on low-frequency power and cross-frequency coupling than on broadband high-gamma, and it aligns more strongly with perceptual/phonological than purely articulatory representations [22]. Training humans with accurate closed-loop EEG feedback reshapes the neural feature space that is exploited for control and steadily improves online performance, highlighting co-adaptation and the importance of neurofeedback for stabilizing non-invasive BCIs [4]. Invasive studies further suggest that mimed or whispered speech occupies an intermediate neural regime between overt and fully imagined states [23], hinting that intermediate behaviors and additional modalities could help bridge the overt–covert divide.

In parallel, representation learning has shown that contrastive objectives can align heterogeneous signals with powerful inductive biases. Aligning non-invasive MEG/EEG time series to pretrained speech representations via contrastive learning yields significant gains in word and phrase identification, including zero-shot performance, compared to simple regression, suggesting that high-level linguistic targets help non-invasive neural data lock onto semantics [24]. More broadly, contrastive predictive coding and large-scale image–text models demonstrate that InfoNCE-style contrastive alignment across views can induce robust, transferable features without task-specific supervision [25,26]. Non-invasive EEG-based speech reconstruction pipelines are also beginning to emerge, including voice reconstruction from EEG recorded during speech imagination [27]. However, current non-invasive reconstruction approaches still achieve modest intelligibility, often rely on overt audio during training, and show limited generalization beyond fixed vocabularies, especially under purely covert conditions.

Despite this progress, we are not aware of a non-invasive approach that simultaneously (i) leverages overt multimodal signals (EEG, EMG, and audio) to anchor linguistic meaning robustly, (ii) explicitly enforces invariance between overt and covert neural states during training, and (iii) evaluates covert EEG+EMG data without audio in a subject-held-out covert setting with target-subject overt calibration. Existing EEG-based systems typically either classify small command sets, regress acoustic features that are weak or absent during covert production, or align brain signals to audio/text without using articulatory evidence to disambiguate semantics in overt data. As a result, performance tends to drop sharply when models are tested on covert imagination, and there is limited understanding of calibration, semantic alignment, and uncertainty in this regime [6,7,9,10,24,27].

The present work addresses this gap by introducing a multi-phase acquisition protocol and a semantic-aware learning framework for studying overt-to-covert transfer under controlled multimodal supervision. Overt EEG, EMG, and audio are used as available training views, while covert trials from training subjects support overt–covert alignment and fine-tuning; final held-out-subject covert inference is evaluated without audio under the protocol defined in Section 2.6. Importantly, the goal is not to assume that all modalities necessarily improve held-out-subject covert decoding, but to determine how multimodal overt information, semantic anchoring, and overt–covert alignment behave across different generalization regimes.

Accordingly, the objective of this study is not to establish an architecture-level state of the art or to claim that the selected neural network is an optimal decoder. Rather, the emphasis is on establishing a controlled methodological framework for examining multimodal overt supervision, semantic anchoring, overt–covert transfer, calibration, and subject generalization under a common backbone.

This study makes three main contributions:

1. We design and record a controlled imagined-speech protocol with continuous EEG and EMG across perception, overt, covert, and rest phases, together with synchronized audio during overt speech, specifically structured to probe the overt–covert transition.
2. We propose and evaluate a semantic-preserving learning framework that combines supervised classification, multi-view contrastive alignment, semantic prototype alignment, and overt–covert consistency constraints. The framework is designed not only to test multimodal overt supervision, but also to identify when auxiliary EMG/audio information transfers across subjects and when EEG-only representations provide more robust covert decoding.
3. We provide a detailed evaluation of covert imagined-speech decoding across ten subjects, including reliability diagrams, expected calibration error, latent k-nearest-neighbor (k-NN) analysis, and semantic alignment heatmaps, to assess not only accuracy but also the quality and stability of the learned representations.

The remainder of the paper is organized as follows. Section 2 describes the experimental protocol, acquisition setup, preprocessing pipeline, proposed architecture, and training strategy. Section 3 reports quantitative and qualitative results, including subject-wise performance, calibration analyses, and latent-space visualizations. Section 4 discusses the implications of these findings for non-invasive speech neuroprostheses and semantic BCIs, and outlines limitations and future directions. Section 5 concludes the paper.

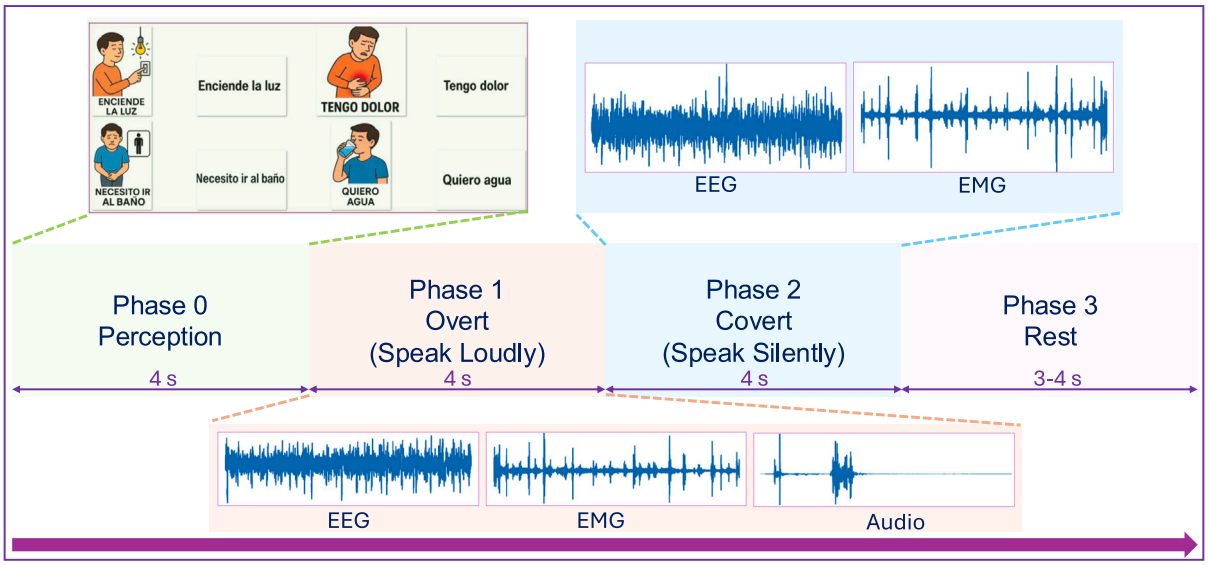

Fig. 1. Examples of the visual stimuli and integrated timeline of a single trial in the experimental protocol. Representative orthographic and pictorial stimuli are shown for the four semantic classes. Each trial consists of four consecutive phases: perception (Phase 0, 4 s), overt speech production (Phase 1, 4 s), covert inner speech (Phase 2, 4 s), and rest (Phase 3, jittered 3–4 s). Representative signal traces illustrate modality availability across phases: EEG and EMG were recorded continuously throughout the trial, whereas audio was available only during the overt phase.

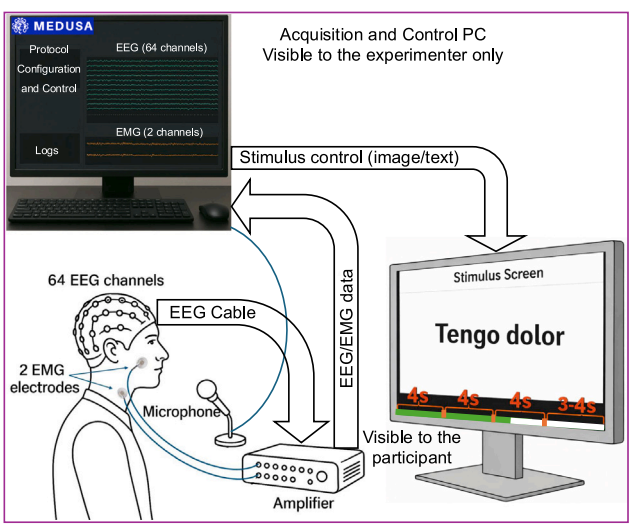

Fig. 2. Overview of the experimental acquisition setup. A 64-channel EEG cap, two EMG electrodes, and a microphone are connected to the amplifier, which streams EEG/EMG data to the acquisition and control PC. The control PC, visible only to the experimenter, runs the protocol, monitors signals, and triggers stimuli. Visual cues are presented on a separate stimulus screen visible to the participant, ensuring synchronized control of perception, overt, and covert phases.

2. Materials and methods

2.1. Dataset and experimental protocol

Ten healthy adults ($N = 10$; 6 men and 4 women; age range: 21–40 years) participated in the study. Nine participants were native Spanish speakers, while the remaining participant had sufficient Spanish proficiency to understand and produce all command sentences used in the protocol. The cohort comprised one bachelor's student, six Ph.D. students, and three postdoctoral researchers. For each participant, the protocol was repeated in 15 recording blocks, each comprising two consecutive runs, for a maximum of 30 runs per participant. In this work, a run denotes one continuous multimodal acquisition file with its associated event markers, whereas a recording block denotes the paired set of two runs within the experimental protocol. All recordings were acquired in a controlled laboratory environment at the Biomedical Engineering Group, University of Valladolid, using the MEDUSA platform for synchronized multimodal acquisition, a Python-based open-source BCI ecosystem that natively supports Lab Streaming Layer (LSL), multimodal sensor integration, and app-based protocol design [28].

The task was built around four semantically meaningful intention classes corresponding to everyday needs or actions, expressed in Spanish: "*Quiero agua*" (I want water), "*Necesito ir al baño*" (I need to go to the toilet), "*Enciende la luz*" (turn on the light), and "*Tengo dolor*" (I am in pain). These define the class set $C = \{\text{water, toilet, light, pain}\}$, which we use throughout the paper for brevity. For each trial, one of these four commands was presented using one of two stimulus modalities: Spanish text on the screen or a representative image depicting the corresponding situation. The stimulus modality for each trial is logged and later used for stratified data splitting and modality-wise analysis. Representative stimuli, the trial timeline, and modality-specific signal availability are shown in Fig. 1.

Each trial comprised a fixed sequence of phases. After a brief visual mask, the perception phase (Phase 0) began, during which the text or image stimulus was displayed for 4 s. In the subsequent overt phase (Phase 1), the participant was cued to pronounce the corresponding sentence aloud for 4 s. This was followed by a covert phase (Phase 2) of 4 s, during which the participant silently repeated the same intention, imagining the sentence without overt speech or deliberate muscle movements. Finally, a rest phase (Phase 3) jittered between 3 and 4 s allowed the signals to return toward baseline while reducing temporal predictability before the next trial. In the present work, we focus on the two intention-bearing phases of interest, overt and covert, and use perception and rest primarily for preprocessing and baseline inspection.

Neurophysiological data were acquired using a 64-channel EEG cap, consisting of 63 scalp EEG electrodes plus one acquisition reference channel, together with two surface EMG channels placed over facial/neck muscles involved in speech production (Fig. 2). The reference channel was used for acquisition and subsequent re-referencing but was not treated as an independent scalp input to the neural encoder. Therefore, the model input contained 63 scalp EEG channels plus 2 EMG

Table 1
Summary of the dataset and acquisition protocol.

Property	Specification
Experimental design	
Participants	N = 10 healthy adults; 6 men and 4 women; age range: 21–40 years; 9 native Spanish speakers; 1 bachelor's student, 6 PhD students, and 3 postdoctoral researchers.
Semantic classes	4 (water, toilet, light, pain)
Stimulus language	Spanish ("Quiero agua", "Necesito ir al baño", "Enciende la luz", "Tengo dolor")
Stimulus modalities	2 (text and image)
Recording blocks per subject	15
Runs per recording block	2 (maximum 30 runs per subject)
Trial structure	
Trial phases	Perception (Phase 0), Overt (Phase 1), Covert (Phase 2), Rest (Phase 3)
Phase durations	Perception: 4 s; Overt: 4 s; Covert: 4 s; Rest: jittered 3–4 s
Trials per class ×modality ×subject	60 trials per intention–modality pair
Trials per subject	480 trials (all phases, all modalities)
Task focus in this work	Overt and covert phases
Recorded signals and acquisition	
EEG acquisition montage	64-channel EEG cap: 63 scalp EEG electrodes + 1 acquisition reference channel
EMG channels	2 surface EMG channels over speech-related muscles
EEG/EMG sampling rate	500 Hz
Audio channel	Close-talking microphone
Audio sampling rate	48 kHz
Synchronization	All modalities time-locked in MEDUSA [28]

channels, yielding 65 EEG–EMG input channels. Two bipolar surface EMG channels were acquired bilaterally from speech-related musculature to capture articulatory recruitment during overt production and any residual articulatory activity during covert production. EMG is treated as an auxiliary articulatory biosignal within the multimodal decoding framework rather than as a stand-alone physiological study of speech-muscle activity.

For each participant, the electrodes were applied by the experimenter using the same bilateral placement procedure over speech-related facial/neck musculature. The two-channel montage was intended to capture broad articulatory activity for exploratory multimodal decoding rather than to isolate individual muscles or provide a standardized clinical EMG assessment.

EEG and EMG were recorded continuously at 500 Hz, whereas 48 kHz microphone audio was stored only during the overt phase. All modalities and trial annotations were synchronized through MEDUSA [28]. Dataset characteristics are summarized in Table 1.

Throughout this work, the target label space is primarily defined at the level of semantic intention rather than stimulus modality. In the main experiments, we model the four intentions as four class labels, and treat textual and pictorial presentations of the same command as different input conditions for the same class. The stimulus modality is therefore stored as metadata and used to construct modality-balanced train/validation splits and to report modality-specific breakdowns in the results. In the ablation analysis (Section 3.5) we also consider an extended 8-class setting that explicitly distinguishes text and image modalities for each intention, in order to probe modality-specific effects.

2.2. Signal preprocessing and epoch construction

2.2.1. EEG and EMG preprocessing

EEG and EMG signals were preprocessed offline using a unified Python pipeline based on NumPy, SciPy, and MNE-Python. All preprocessing was applied to the continuous recordings acquired at 500 Hz before epoch extraction. Filtering was implemented using linear-phase finite-impulse-response (FIR) filters designed with `scipy.signal.firwin` and a Hamming window, and applied in zero-phase mode by forward–backward filtering (`scipy.signal.filtfilt`), yielding a non-causal offline preprocessing chain.

For EEG, raw voltages were first converted to physical units and linearly detrended. The signals were then band-pass filtered between 0.5 and 40 Hz. The lower and upper transition bandwidths were set to 0.5 Hz and 10 Hz, respectively. The band-pass filter was implemented with 1501 taps (order 1500). Discontinuities and numerically invalid samples, including NaN or Inf values, were automatically annotated. Here, "bad segments" refers to intervals automatically flagged because of non-finite samples or robust outlier criteria based on peak-to-peak amplitude and variance; noisy or flat channels were identified using the same signal-quality measures. These annotations represented signal-quality exclusions rather than task-based exclusions, and the marked intervals were omitted when fitting independent component analysis (ICA). The EEG was subsequently re-referenced to the common average and centered by subtracting the temporal mean of each channel.

To attenuate ocular artifacts, ICA was applied to the preprocessed EEG after filtering and bad-segment annotation. For each recording, a FastICA decomposition retaining 99% of the variance was fitted while excluding segments marked as bad. Putative ocular components were identified automatically using MNE's electrooculography (EOG) detection routines based on frontal electrodes, such as Fp1, Fp2, AF7, and AF8. These components were then removed, and the cleaned EEG was reconstructed from the remaining sources.

EMG channels were processed in a separate pipeline. Signals were linearly detrended, band-pass filtered between 20 and 100 Hz, and an additional 50 Hz notch filter was applied to suppress line noise within the retained EMG band. The lower and upper transition bandwidths of the band-pass filter were set to 5 Hz and 20 Hz, respectively. Discontinuities and numerically invalid samples were automatically annotated and repaired by inpainting. No ICA-based artifact correction was applied to EMG.

In the present study, EMG was used as an auxiliary articulatory signal rather than for detailed motor-unit waveform analysis. Surface EMG is commonly used to characterize the magnitude and timing of muscle activation, although its amplitude should not be interpreted as a direct measure of force or specific motor-unit recruitment [29]. Facial EMG has also been used to investigate subtle articulatory activity during inner speech using broad frequency bands beginning around 20 Hz [30]. In addition, the 20–40 Hz portion overlaps the beta range in which EEG–EMG corticomuscular coherence is commonly observed [31]. Accordingly, the 20–100 Hz band was retained as an auxiliary decoding representation and was not interpreted as a direct physiological measure of motor-unit recruitment or cortical drive. Corticomuscular coherence itself was not quantified in the present study.

Subsequently, each channel was robustly scaled independently by subtracting its median and dividing by its interquartile range (IQR). This normalization reduces sensitivity to extreme values and residual artifacts and is more robust to outliers than mean–variance standardization [32,33].

2.2.2. Audio preprocessing

Speech audio was preprocessed to obtain time-frequency representations aligned with each trial. For each run, trial-wise audio segments corresponding to the semantic task were first extracted based on marker annotations. For every overt trial, the audio segment corresponding to

the 4-s overt phase was initially extracted using the marker timestamps. To match the EEG and EMG analysis window, only the first 3 s were retained for modeling. Each audio segment was therefore trimmed or zero-padded, when necessary, to exactly $T_{\text{audio}} = 3$ s. Audio was kept at the original sampling rate of 48 kHz.

Each trial's audio segment was then converted into a log-mel spectrogram using the `librosa` library [34]. Specifically, a short-time Fourier transform (STFT) with window length $N_{\text{FFT}} = 1024$ samples and hop size of 480 samples (approximately 10 ms) was computed, and 64 mel-frequency bands were retained in the range 50–8000 Hz. The resulting magnitude spectrogram was mapped to the mel scale and converted to log-power. Finally, each spectrogram was standardized by subtracting its mean and dividing by its standard deviation, yielding a normalized log-mel representation of shape $(1, N_{\text{mels}}, T_{\text{frames}})$ per trial. Audio was recorded and processed only for the overt phase; no audio representation is defined for covert trials.

2.2.3. Epoch definition and alignment

Task-related trials were defined using the marker files associated with each run. For every row in the marker table, we read the semantic class, stimulus modality, and phase label. Only trials labeled as overt or covert were retained. For each retained trial, a window of length $T_{\text{trial}} = 4$ s was extracted from the preprocessed EEG and EMG signals starting from the onset timestamp provided in the markers. Although each phase lasted 4 s to provide sufficient time for all participants, participants consistently completed both the overt and covert tasks within the first 3 s. Inspection of multiple recordings from every participant, together with a preliminary comparison using the complete 4-s epochs, confirmed that the final second was predominantly task-irrelevant and did not improve decoding performance. We therefore retained only the first 3 s of all available modalities — EEG, EMG, and overt-phase audio — for modeling. The resulting EEG/EMG epochs contained $T_{\text{EEG/EMG}} = 600$ samples per channel at the resampled rate of 200 Hz, whereas the audio remained at its original sampling rate of 48 kHz and was converted into a 3-s log-mel spectrogram.

Let $X \in \mathbb{R}^{N \times C \times T}$ denote the resulting EEG+EMG data tensor, where N is the number of trials, $C = 65$ is the total number of channels, and $T = 600$. Audio spectrograms are available only for overt trials and are stored in a tensor $A \in \mathbb{R}^{N_{\text{overt}} \times 1 \times N_{\text{mels}} \times T_{\text{frames}}}$, where N_{overt} is the number of overt trials. The labels and metadata are stored in vectors $y \in \{0, 1, 2, 3\}^N$, phase $\in \{\text{overt}, \text{covert}\}^N$, and mode $\in \{\text{text}, \text{image}\}^N$. By construction, EEG and EMG epochs are temporally aligned to the trial markers for all retained trials, while audio is aligned only for overt trials.

2.3. Multimodal neural architecture

The proposed model maps multimodal neurophysiological and audio signals to a shared latent space where discriminative, contrastive, and semantic objectives can be jointly optimized. The core architecture consists of an encoder for EEG and EMG, an audio encoder for overt speech, and a fusion module that projects all modalities into a common embedding.

2.3.1. EEG–EMG encoder

A convolutional encoder inspired by EEGNet [35] processes the concatenated EEG and EMG signals. This compact backbone is used as a common feature extractor and is not introduced as a novel neural architecture. It is retained across the methodological comparisons so that the effects of multimodal supervision, semantic alignment, and overt–covert regularization can be examined without confounding them with changes in backbone design. The encoder receives an input tensor of shape (C, T), where $C = 65$ channels (63 scalp EEG electrodes and 2 EMG channels) and T denotes the number of time samples per trial after preprocessing and epoching. Temporal convolutions capture local temporal patterns in the band-limited signals, while spatial convolutions learn filters across electrodes and EMG sensors.

All 63 scalp EEG channels were retained because covert speech may involve distributed and participant-variable activity beyond a single predefined language region. Retaining the complete montage avoids imposing an a priori regional selection and allows the learned spatial filters to weight informative electrodes while reducing the contribution of less useful channels. Systematic channel-reduction and region-specific analyses remain relevant directions for future work.

Batch normalization, nonlinearities, and pooling layers stabilize training and progressively reduce the temporal resolution. The encoder outputs a fixed-dimensional feature vector $\mathbf{h}_{\text{eeg-emg}} \in \mathbb{R}^D$, obtained by global pooling over time and a final linear projection. Throughout this work, we set the latent dimensionality to $D = 256$.

2.3.2. Audio encoder

Overt speech audio is represented as a log-mel spectrogram (Section 2.2.2), which is processed by a dedicated 2D convolutional encoder. The audio encoder receives an input tensor of shape $(1, N_{\text{mels}}, T_{\text{frames}})$ and applies a sequence of convolution-batch normalization–nonlinearity–pooling blocks over the time–frequency plane. Early layers focus on local spectral patterns, while deeper layers aggregate information over longer temporal contexts. After several blocks, global average pooling and a fully connected layer are used to obtain an audio embedding $\mathbf{h}_{\text{audio}} \in \mathbb{R}^{D_a}$, where D_a denotes the audio-encoder output dimension.

2.3.3. Multimodal fusion

For overt trials, EEG and EMG are first passed through the EEG–EMG encoder to produce $\mathbf{h}_{\text{eeg-emg}}$, and the audio spectrogram is processed by the audio encoder to produce $\mathbf{h}_{\text{audio}}$. The two vectors are concatenated,

$$\mathbf{h}_{\text{concat}} = [\mathbf{h}_{\text{eeg-emg}}; \mathbf{h}_{\text{audio}}],$$

and projected through a fully connected layer with nonlinearity and normalization to obtain the fused latent embedding $\mathbf{z} = f_{\text{fusion}}(\mathbf{h}_{\text{concat}}) \in \mathbb{R}^D$. For covert trials, only EEG and EMG are available. In this case, the MultiModalEncoder reduces to the EEG–EMG encoder, and the latent representation is obtained as $\mathbf{z} = \mathbf{h}_{\text{eeg-emg}}$. This design allows the model to exploit multimodal information when overt speech is present, while maintaining a consistent latent space for covert-only decoding. A concise summary of architectural choices and training settings is provided in Table 2.

2.4. Sentence-transformer semantic space

Semantic supervision is provided by an external sentence-level language model. We use the pre-trained `all-MiniLM-L6-v2` sentence transformer as a fixed encoder $\phi(\cdot)$ that maps each class sentence t_c to a D_{sem}-dimensional embedding, where $D_{\text{sem}} = 384$ in this work [36,37]. The resulting vectors are ℓ_2-normalized to obtain the class-specific prototypes $\{\mathbf{p}_c\}_{c \in \mathcal{C}}$ used by the semantic head. The sentence transformer is not fine-tuned on our data, and no gradients are propagated through $\phi(\cdot)$. As a result, the geometry of the semantic space is determined by the pretrained language model, while the neural encoder and semantic head are trained to align overt and covert neural embeddings with this fixed linguistic structure.

2.5. Heads for contrastive, classification, and semantic objectives

On top of the shared latent embedding $\mathbf{z}$, the model uses three task-specific heads: a contrastive head, a margin-based classification head, and a semantic head.

1. **Contrastive head.** A multilayer perceptron (MLP) maps $\mathbf{z}$ to a lower-dimensional vector $\mathbf{u} \in \mathbb{R}^{d_c}$, followed by ℓ_2-normalization. This head is used for supervised contrastive learning on overt trials and for overt–covert consistency losses.

Table 2
Key hyperparameters and configuration of the proposed multimodal semantic EEG pipeline.

Component	Configuration/setting
Encoders and fusion	
EEG/EMG encoder	1D CNN encoder with four Conv–BN–ReLU blocks followed by global pooling.
EEG/EMG encoder	Layer 1: Conv1d($C \to 64$, kernel 7, stride 2, padding 3), BatchNorm1d, ReLU.
EEG/EMG encoder	Layer 2: Conv1d($64 \to 128$, kernel 5, stride 2, padding 2), BatchNorm1d, ReLU.
EEG/EMG encoder	Layer 3: Conv1d($128 \to 256$, kernel 3, stride 2, padding 1), BatchNorm1d, ReLU.
EEG/EMG encoder	Layer 4: Conv1d($256 \to 256$, kernel 3, stride 2, padding 1), BatchNorm1d, ReLU.
EEG/EMG encoder	Pooling: AdaptiveAvgPool1d(1), followed by flattening; mapped to $D = 256$ in the heads.
Audio encoder	2D CNN encoder applied to log-mel spectrograms.
Audio encoder	Layer 1: Conv2d($1 \to 32$, kernel 3×3, stride 2, padding 1), BatchNorm2d, ReLU.
Audio encoder	Layer 2: Conv2d($32 \to 64$, kernel 3×3, stride 2, padding 1), BatchNorm2d, ReLU.
Audio encoder	Layer 3: Conv2d($64 \to 128$, kernel 3×3, stride 2, padding 1), BatchNorm2d, ReLU.
Audio encoder	Pooling: AdaptiveAvgPool2d(1), followed by flattening.
Multimodal encoder	EEG/EMG and audio embeddings are concatenated and passed through Linear($D_{\text{eeg}} + D_{\text{a}} \to D$), with $D = 256$. For covert trials, the audio branch is disabled and representations are based only on EEG/EMG features.
Heads and semantic space	
Contrastive head	MLP with Linear($D \to D$), BatchNorm1d, ReLU, Linear($D \to d_{\text{c}}$), followed by ℓ_2 normalization; Output dimension: $d_{\text{c}} = 128$.
Semantic head	Linear($D \to D_{\text{sem}}$), producing e_{sem}, followed by ℓ_2 normalization at loss computation; Output dimension: $D_{\text{sem}} = 384$.
Semantic space	Sentence model: `all-MiniLM-L6-v2` from Sentence-Transformers.
Semantic prototypes	Class sentences: "Quiero agua", "Necesito ir al baño", "Enciende la luz", and "Tengo dolor". Prototype vectors are computed as $\mathbf{p}_c = \phi(t_c)/\|\phi(t_c)\|_2$ and kept frozen during training.
Classifier head	ArcFace cosine classifier with additive angular margin.
Training schedule and optimization	
Batching	Batch size of 32 for all stages.
Optimizer	Adam with weight decay 10^{-4}.
Stage 1	Supervised contrastive warm-up on overt trials only; 200 epochs, learning rate 10^{-3}, supervised contrastive (SupCon) temperature $\tau_{\text{sup}} = 0.35$.
Stage 2	Joint multi-task training with overt supervision and overt–covert alignment; 300 epochs, learning rate 10^{-3}. The joint loss is $\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{supcon}} + \mathcal{L}_{\text{CE}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{DA}}\mathcal{L}_{\text{CORAL}}$, with $\lambda_{\text{sem}} = 0.5$ and $\lambda_{\text{DA}} = 0.15$. Model selection uses the best semantic mean reciprocal rank on the overt-validation set.
Stage 3	Covert semantic consistency/overt–covert alignment; 200 epochs, learning rate 10^{-4}. NT-Xent is applied to overt–covert trial pairs when available; otherwise, supervised contrastive loss is applied to covert-training trials. NT-Xent temperature is $\tau_{\text{NTX}} = 0.25$. All encoder parameters remain trainable. Model selection uses the best covert k-NN accuracy on covert-train versus covert-validation embeddings, with $k = 5$.
Stage 4	Covert-only fine-tuning and evaluation; maximum 100 epochs with early stopping, learning rate 10^{-4}. The objective is cross-entropy with label smoothing 0.1. Early stopping uses a patience of 10 epochs on covert-validation accuracy. Best encoder and classifier weights are saved per fold.

2. **Classification head.** A margin-based softmax classifier operates on $\mathbf{z}$. The head implements ArcFace [38] with a learnable weight matrix and fixed scale s and margin m. During training, both $\mathbf{z}$ and the class weight vectors are ℓ_2-normalized, and the ground-truth logit is modified by an angular margin. A cross-entropy loss with label smoothing is applied to the ArcFace logits.
3. **Semantic head.** A linear projection maps $\mathbf{z}$ to a semantic embedding $e_{\text{sem}} \in \mathbb{R}^{D_{\text{sem}}}$. This embedding is trained to align with the fixed class prototypes $\{\mathbf{p}_c\}_{c \in \mathcal{C}}$ defined in Section 2.4.

2.6. Data splitting and experimental conditions

Let the set of subjects be $\mathcal{S} = \{1, \dots, N\}$ with $N = 10$. For each fold, one subject $s^* \in \mathcal{S}$ is designated as the held-out test subject, and the remaining subjects $\mathcal{S}_{\text{train}} = \mathcal{S} \setminus \{s^*\}$ form the training pool. Throughout, we refer to this setting as a *subject-held-out covert evaluation with target-subject overt calibration*. In this protocol, covert trials from the held-out subject are never used for training, early stopping, hyperparameter tuning, or temperature calibration, whereas overt trials from that subject may be used in the overt train/validation split. A stricter leave-one-subject-out (LOSO) variant, in which the held-out subject contributes no training data of any kind, is considered separately in the ablation analysis (Section 3.5, X2).

Overt data consist of trials with phase = overt. Overt trials from all subjects, including the held-out subject s^*, are pooled together and split 80/20 into train and validation sets at the trial level. To preserve balance across semantic classes and stimulus modalities, the split is stratified using a key that jointly encodes class index and modality label. This yields $\mathcal{D}^{\text{overt}}_{\text{train}}$ and $\mathcal{D}^{\text{overt}}_{\text{val}}$.

Covert data consist of trials with phase = covert. Covert trials from training subjects are split 80/20 into $\mathcal{D}^{\text{covert}}_{\text{train}}(\mathcal{S}_{\text{train}})$ and $\mathcal{D}^{\text{covert}}_{\text{val}}(\mathcal{S}_{\text{train}})$, stratified by class label. All covert trials from the held-out subject s^* are reserved as the final test set:

$$\mathcal{D}^{\text{covert}}_{\text{test}}(s^*) = \mathcal{D}^{\text{covert}}(s^*).$$

Thus, for each fold, the procedure yields five datasets:

$$\mathcal{D}_{\text{fold}} = \{\mathcal{D}^{\text{overt}}_{\text{train}}, \mathcal{D}^{\text{overt}}_{\text{val}}, \mathcal{D}^{\text{covert}}_{\text{train}}(\mathcal{S}_{\text{train}}), \mathcal{D}^{\text{covert}}_{\text{val}}(\mathcal{S}_{\text{train}}), \mathcal{D}^{\text{covert}}_{\text{test}}(s^*)\}.$$

2.7. Training strategy and loss functions

The training strategy is organized into four stages. For each fold, model parameters are re-initialized, the four stages are executed using the corresponding training and validation sets, and final performance is evaluated on the held-out subject's covert test set.

2.7.1. Supervised contrastive warm-up

In the first stage, only overt trials are used. The encoder and contrastive head are optimized using a supervised contrastive loss [39], which encourages samples from the same class to be close in the contrastive space and samples from different classes to be separated. Batch normalization layers in the encoder are set to evaluation mode to stabilize statistics, and early stopping is applied based on the average warm-up loss on overt validation batches.

2.7.2. Joint multi-task training with domain alignment

In the second stage, overt trials remain the primary supervised source, while covert trials from training subjects are used without labels to regularize the representations. Each training step consumes one batch of overt data and one batch of covert-train data. The joint loss combines supervised contrastive loss, ArcFace cross-entropy, semantic alignment loss, and correlation alignment (CORAL) loss.

For overt sample i with label y_i, the semantic loss is

$$\mathcal{L}_{\text{sem}} = \mathbb{E}_i \left[1 - \cos(e_{\text{sem},i}, \mathbf{p}_{y_i}) \right], \tag{1}$$

where $\mathbf{p}_{y_i}$ is the sentence-transformer prototype of the corresponding Spanish command. To reduce the distributional shift between overt and covert features, we employ a CORAL covariance-matching loss. Let $\mathbf{C}_o$ and $\mathbf{C}_c$ denote the batch covariance matrices of overt and covert embeddings, respectively:

$$\mathcal{L}_{\text{CORAL}} = \|\mathbf{C}_o - \mathbf{C}_c\|_F^2 . \tag{2}$$

The total loss is

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{supcon}} + \mathcal{L}_{\text{CE}} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}} + \lambda_{\text{DA}} \mathcal{L}_{\text{CORAL}} . \tag{3}$$

Optimization is performed with Adam and a cosine-annealing learning rate schedule with warm restarts. Model selection in this stage is based on semantic retrieval mean reciprocal rank (MRR) on the overt validation set.

2.7.3. Covert semantic consistency

The third stage aligns overt and covert embeddings at the level of matched trial instances. Each semantic command trial contains both an overt phase and a covert phase; therefore, within the training pool S_{train}, a paired dataset is built by matching overt and covert epochs that share the same subject, run, trial index, class label, and stimulus modality. Let $D_{\text{pair}} = \{(x_i^{\text{overt}}, x_i^{\text{covert}}, y_i)\}_{i=1}^{N_{\text{pair}}}$ denote this set. A symmetric normalized temperature-scaled cross-entropy (NT-Xent) loss [40] is applied to encourage overt and covert embeddings associated with the same trial-level semantic command to become close, while embeddings from different classes are pushed apart. Model selection is based on a non-parametric k-NN evaluation on covert-train versus covert-validation embeddings.

2.7.4. Covert fine-tuning

In the final stage, the encoder and classification head are fine-tuned using supervised learning on covert trials from training subjects. The best semantic-consistency encoder initializes the network, and the encoder and ArcFace parameters are optimized using cross-entropy with label smoothing. Only real covert-train trials from $D_{\text{train}}^{\text{covert}}(S_{\text{train}})$ are used for training, and $D_{\text{val}}^{\text{covert}}(S_{\text{train}})$ is used for early stopping based on covert classification accuracy. The final model is evaluated on $D_{\text{test}}^{\text{covert}}(s^*)$. Hyperparameters were empirically tuned during preliminary development using only the designated training and validation partitions and were subsequently fixed across all held-out-subject folds and ablation settings; covert test data from the held-out participant were never used for hyperparameter selection. Key optimization hyperparameters are reported in Table 2. All experiments were conducted on a workstation equipped with an AMD Ryzen 7 9700X CPU, 64 GB DDR5 RAM, and an NVIDIA GeForce RTX 5080 GPU with 16 GB of GDDR7 memory. For the main reference configuration, completing the full four-stage training procedure across all ten held-out-subject folds required approximately 2 h 8 min. Because the present study was conducted offline, end-to-end online latency was not formally evaluated.

2.8. Evaluation protocol and metrics

For each fold with held-out subject s^*, the final encoder and classifier are evaluated on the overt validation set and the covert test set. To improve probabilistic calibration, a single temperature parameter T_{cal} is fitted per fold using only overt validation data, by minimizing cross-entropy between temperature-scaled logits and true labels. The resulting T_{cal} is then used to rescale logits for both overt validation and covert test metrics.

Standard classification metrics include accuracy, top-k accuracy, and MRR of the true class in the ranked list of predicted probabilities. Calibration is quantified using the Brier score and expected calibration error (ECE) estimated from reliability diagrams. The structure of the latent space is examined using k-NN analysis with cosine similarity on covert embeddings. Semantic retrieval performance is evaluated through the semantic head by comparing each trial's semantic embedding to the normalized `all-MiniLM-L6-v2` prototypes of the four commands. Unless otherwise stated, metrics are computed by pooling text and image trials in a shared four-class label space.

Because trials were nested within participants and recording blocks, the participant was treated as the independent statistical unit. For the primary covert-decoding result, uncertainty in the mean covert-test accuracy was quantified using a 95% bias-corrected and accelerated (BCa) bootstrap confidence interval obtained by resampling the ten participant-level accuracies with replacement ($B = 100,000$). Statistical evidence against the balanced four-class chance level of 0.25 was assessed using an exact one-sided sign-flipping test on the participant-level accuracy differences, enumerating all 2^{10} possible sign assignments.

For the same-protocol ablation analysis, each variant was compared with the reference configuration B0 using paired participant-level covert-test accuracies. For participant s, the paired effect of ablation a was defined as

$$d_{s,a} = \text{Acc}_{s,a}^{\text{covert}} - \text{Acc}_{s,\text{B0}}^{\text{covert}} . \tag{4}$$

The magnitude of each ablation effect was summarized by the mean paired difference across participants, with a 95% BCa bootstrap confidence interval obtained by resampling the participant-level paired differences using the same procedure described above. Statistical evidence for a difference from B0 was assessed using a two-sided exact paired sign-flipping test on $\{d_{s,a}\}_{s=1}^{10}$, enumerating all 2^{10} possible sign assignments. Family-wise error across the seven same-protocol comparisons (S1-C, S2-C, S3-C, C1, P1, R1, and R2) was controlled using the Holm procedure. The family-wise significance level was set to $\alpha = 0.05$, and Holm-adjusted p-values below 0.05 were considered statistically significant. The participant was treated as the independent unit throughout. Variants that changed the label space or evaluation regime were not included in this family of paired comparisons.

For reproducibility, Table 2 summarizes the main architecture, training objectives, calibration procedure, and evaluation settings used in the proposed framework.

3. Results

3.1. Illustrative neurophysiology of overt and covert inner speech

Before turning to decoding performance, we present representative visualizations of single trials, band-power patterns, and time–frequency dynamics. Figs. 3–5 provide subject-level examples of the signal structure entering the pipeline and are not intended as population-level statistical evidence.

The representative recording in Fig. 4 shows spatially heterogeneous alpha- and beta-band differences between overt and covert phases. Because these maps are based on a single recording block, they are presented as illustrations of signal structure rather than population-level effects.

The spatially structured alpha and beta changes observed here are consistent with previous topomap-based studies, in which imagined content modulated scalp distributions in a class-specific manner during both perception and imagination [8]. Complementing these band-power results, Fig. 5 presents trial-averaged EEG time–frequency log-power for the same participant and representative recording block, shown separately for perception, overt, covert, and rest phases.

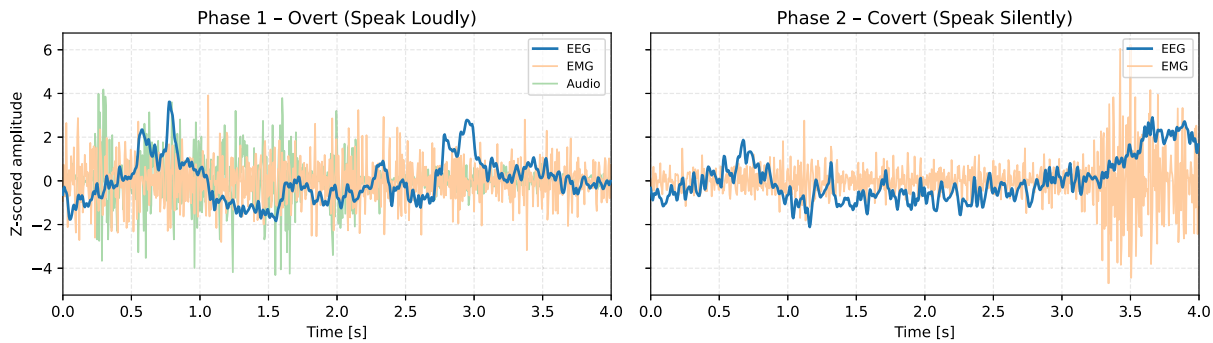

Fig. 3. Representative multimodal trial from participant S02 (run 09). Phase 1 (overt speech) shows z-scored EEG, EMG, and audio waveforms during spoken articulation. Phase 2 (covert inner speech) shows the corresponding EEG and EMG activity during the covert interval, which may still contain low-amplitude residual articulatory activity. Audio is absent in this phase. All signals are z-scored within channel.

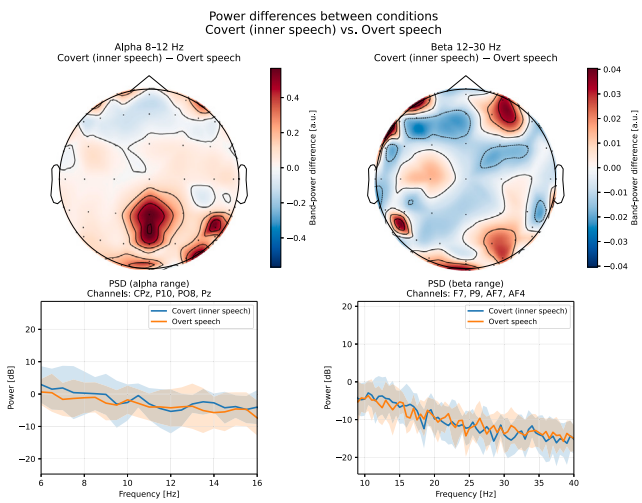

Fig. 4. Representative band-power differences between covert inner speech and overt speech. Top row: scalp topographies of alpha (8–12 Hz, left) and beta (12–30 Hz, right) band-power differences, expressed in arbitrary units (a.u.), for one recording block of participant S02. Warm colors indicate higher power during covert inner speech, and cool colors indicate higher power during overt speech. Bottom row: mean power spectral density (PSD), expressed in decibels (dB), with ±1 standard deviation for the four electrodes showing the largest differences within each band.

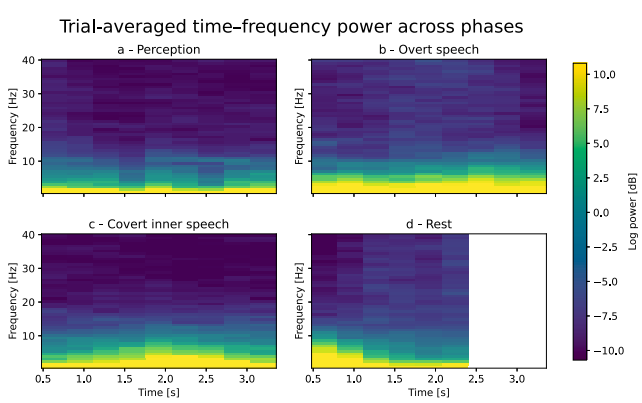

Fig. 5. Trial-averaged time–frequency log-power across phases. For the same participant and representative recording block as in Fig. 4, short-time Fourier transform (STFT)-based power was computed separately for each EEG channel and trial in the $0.5-40$ Hz range, converted to log power in decibels, and then averaged across channels and trials within each phase: (a) perception, (b) overt speech, (c) covert inner speech, and (d) rest. Warmer colors indicate higher log power.

3.2. Decoding accuracy across subjects

Table 3 summarizes per-subject performance after the full training pipeline. Across the ten participants, overt decoding accuracy averaged 0.543, while covert inner-speech decoding averaged 0.416 ± 0.110, where the variability denotes the standard deviation across held-out participants. The 95% BCa bootstrap confidence interval for mean covert-test accuracy was [0.360, 0.492]. An exact one-sided subject-level sign-flipping test confirmed that covert accuracy was above the balanced four-class chance level of 0.25 ($p = 9.77\times10^{-4}$). Several subjects, such as S02, S04, and S06, showed relatively small performance gaps between overt and covert speech, whereas others, such as S01, S05, and S07, exhibited larger drops.

All ten participant-level covert accuracies were numerically above the nominal chance level of 0.25, although performance varied substantially across participants. A dedicated ablation further showed that suppressing covert EMG increased mean covert-test accuracy from 0.42 to 0.46, suggesting that residual covert EMG is not reliably transferable across subjects in the held-out-covert setting. Subject-wise normalized confusion matrices (Fig. 6) show that errors concentrate in subject-specific command pairs rather than being uniformly distributed across classes.

3.3. Calibration of overt and covert decoders

Beyond accuracy, the model's probabilistic outputs should be well calibrated if they are to be used in interactive or safety–critical BCIs. The temperature parameter T_{cal} in Table 3 is learned on overt validation data and then applied to overt and covert predictions. Fig. 7 compares the resulting ECE between overt and covert conditions. On average, overt decoding is moderately calibrated (mean $\text{ECE}_{\text{ov}} = 0.154$), whereas covert decoding shows higher calibration error (mean $\text{ECE}_{\text{ct}} = 0.197$), indicating less reliable probabilities under covert inference. The corresponding reliability diagrams are shown in Figs. 8 and 9. These diagrams provide a subject-wise view of the calibration behavior behind the ECE values in Table 3 and Fig. 7.

3.4. Semantic alignment and latent-space structure

The proposed architecture explicitly encourages latent representations to align with text-based semantic prototypes of the four commands. Fig. 10 visualizes this alignment on the covert-test set via cosine-similarity heatmaps between trial-level semantic embeddings and the corresponding prototype vectors for each subject. Across participants, diagonal entries are generally higher than off-diagonal entries, indicating that the brain-derived semantic embeddings often remain most similar to the intended textual concept even during covert inner speech. Off-diagonal similarities reveal subject-specific representational

Table 3
Per-subject performance metrics for overt and covert phases.

Subject	Acc_{ov}	Acc_{ct}	T_{cal}	ECE_{ov}	ECE_{ct}	Br_{ov}	Br_{ct}	k-NN@5	Sem@ top-1	Sem@ top-3	Sem@ MRR
S01	0.542	0.354	5.733	0.151	0.176	0.140	0.187	0.354	0.500	0.898	0.703
S02	0.688	0.625	2.997	0.193	0.274	0.112	0.117	0.750	0.625	0.979	0.781
S03	0.406	0.259	10.186	0.142	0.237	0.164	0.200	0.375	0.344	0.810	0.622
S04	0.438	0.438	6.565	0.072	0.159	0.164	0.172	0.375	0.375	0.823	0.620
S05	0.604	0.375	7.009	0.161	0.157	0.141	0.184	0.438	0.479	0.869	0.708
S06	0.552	0.562	12.361	0.148	0.141	0.156	0.157	0.625	0.478	0.865	0.675
S07	0.646	0.448	5.800	0.173	0.148	0.127	0.190	0.417	0.375	0.827	0.630
S08	0.458	0.334	5.591	0.164	0.145	0.172	0.188	0.417	0.271	0.784	0.536
S09	0.500	0.344	2.933	0.196	0.313	0.137	0.226	0.375	0.531	0.925	0.719
S10	0.604	0.417	6.140	0.145	0.216	0.134	0.187	0.396	0.604	0.956	0.755
Mean	0.543	0.416	6.531	0.154	0.197	0.145	0.181	0.452	0.458	0.874	0.675

$Acc_{ov/ct}$: overt-validation/covert-test accuracy; T_{cal}: calibration temperature; $ECE_{ov/ct}$: overt-validation/covert-test expected calibration error; Br: Brier score; k-NN@5: 5-NN latent accuracy on covert-test embeddings; Sem@top-1/3/MRR: semantic retrieval metrics computed on overt-validation trials in the sentence space.

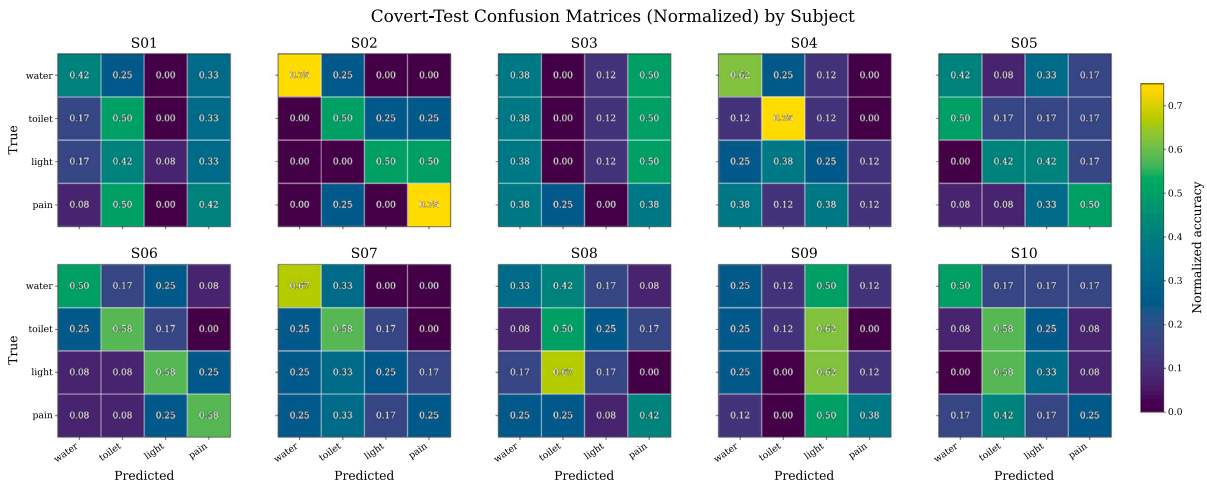

Fig. 6. Normalized confusion matrices for the covert-test set, shown separately for each subject (S01–S10). Rows correspond to the true semantic classes (*water*, *toilet*, *light*, *pain*), and columns to predicted classes. Values indicate row-normalized prediction proportions, highlighting subject-specific error patterns in covert decoding.

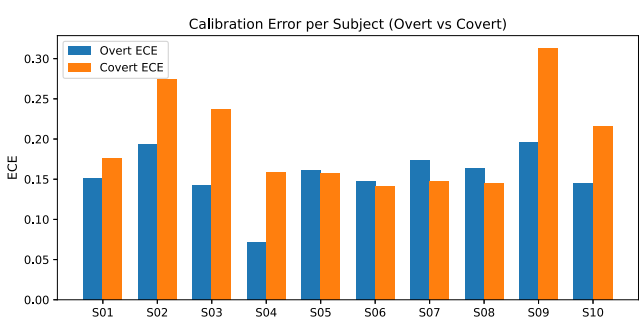

Fig. 7. Comparison of calibration error (ECE) between overt-validation and covert-test conditions for all subjects (S01–S10). Bars report the expected calibration error computed from reliability diagrams; higher values indicate poorer calibration.

overlap among commands, mirroring some of the confusion patterns seen in Fig. 6.

To quantitatively relate semantic structure to latent-space organization, Fig. 11 compares two complementary metrics for each subject: k-NN@5 accuracy computed directly on covert latent embeddings, and semantic retrieval top-1 accuracy on overt trials using the same prototype vectors. The two measures show broadly similar trends, with participants who have well-separated latent clusters also tending to achieve higher semantic retrieval accuracy. Together, Figs. 10–11 and Table 3 indicate that the latent representations retain class-related semantic structure under covert decoding. This alignment is confined to the predefined four-command prototype space.

3.5. Ablation study

To better understand which components of the pipeline are most critical for covert inner-speech decoding, we carried out ablation experiments summarized in Table 4. Each scenario perturbs a specific aspect of the task definition, multimodal supervision signal, semantic machinery, or training regime, while keeping the remaining components fixed.

In this analysis, the full EEG+EMG+audio setting is treated as the reference multimodal configuration rather than as the expected best-performing modality choice for every generalization regime. This distinction is important because the ablations are intended to reveal which sources of information transfer to held-out covert decoding and which sources mainly capture subject-specific or regime-specific variability.

Importantly, B0 and X2 represent two distinct cross-subject evaluation regimes. X2 corresponds to strict calibration-free LOSO generalization, in which the held-out participant contributes no data during training or calibration, whereas B0 evaluates held-out covert decoding supported by overt calibration from the target participant. Mean covert-test accuracy was 0.38 under X2, compared with 0.42 under B0, indicating that strict LOSO transfer remains feasible, although target-participant overt calibration provides a modest performance benefit.

The paired participant-level analysis in Table 5 shows that the observed same-protocol differences were not explained solely by a

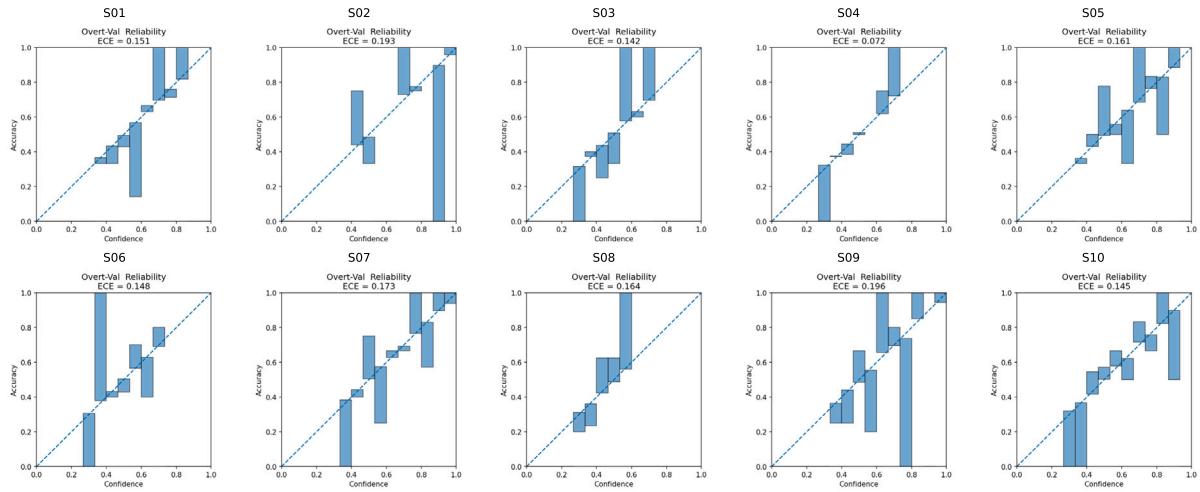

Fig. 8. Reliability diagrams for overt-validation predictions. Each panel corresponds to one held-out-subject fold. The diagonal line indicates ideal calibration, and the bars compare empirical accuracy and confidence within probability bins after temperature scaling.

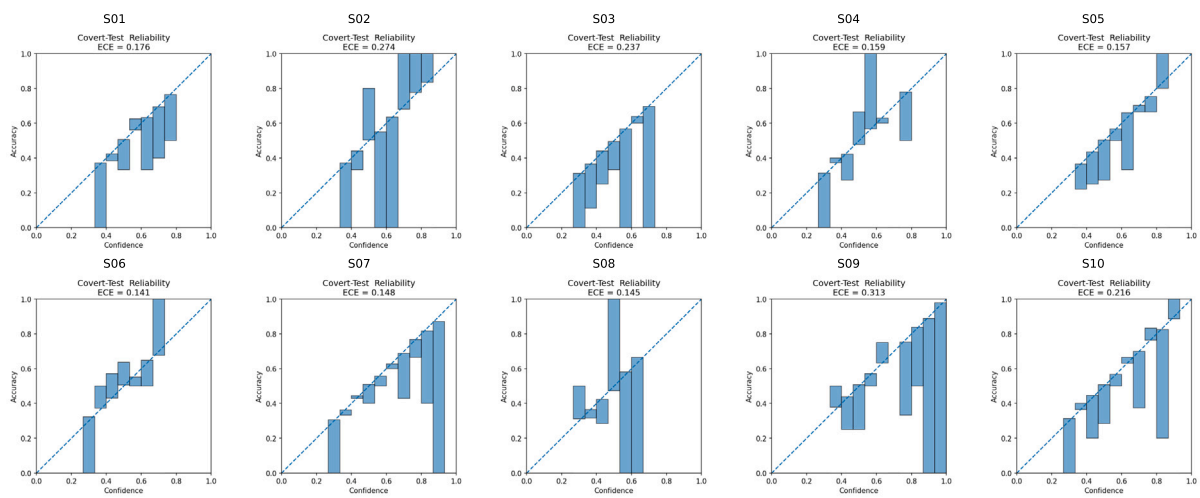

Fig. 9. Reliability diagrams for covert-test predictions. Each panel corresponds to one held-out-subject fold. Compared with overt validation, the covert-test diagrams show larger deviations from the ideal calibration line, consistent with the higher covert ECE values reported in Table 3 and Fig. 7.

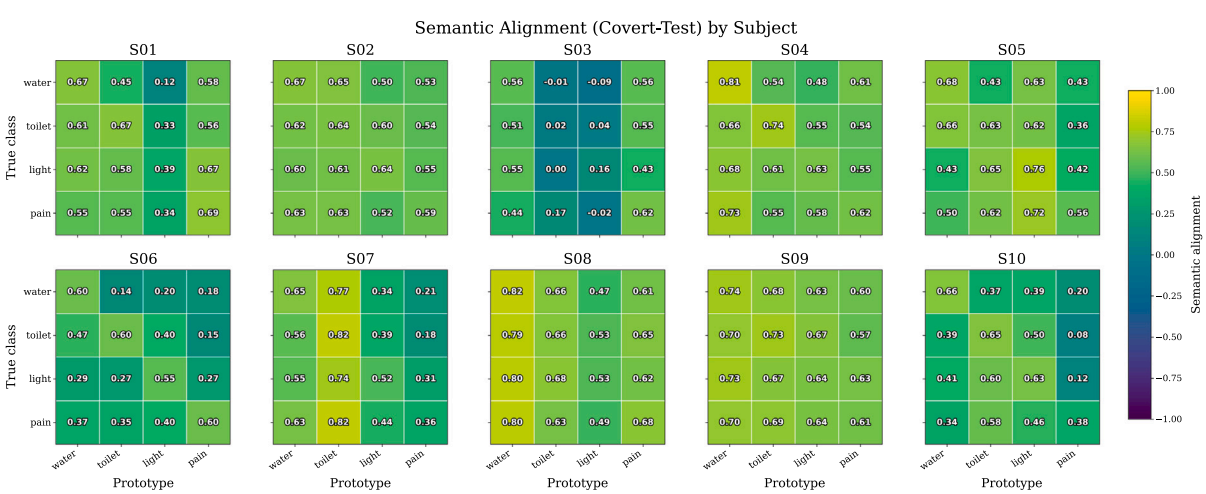

Fig. 10. Semantic alignment between trial-level semantic embeddings and text-based class prototypes on the covert-test set. Each panel (S01–S10) shows a cosine-similarity heatmap where rows denote true classes and columns denote semantic prototypes. Higher values on the diagonal indicate stronger alignment between brain-derived semantic representations and the intended textual concepts, while off-diagonal similarities reveal class-level representational overlap.

Table 4
Ablation analysis on covert-test accuracy under different design variants. Reported values are observed mean accuracies relative to the reference multimodal configuration ($\mathrm{Acc}_{ct} = 0.42$ in Table 3). Formal participant-level paired comparisons were restricted to the same-protocol ablations S1-C, S2-C, S3-C, C1, P1, R1, and R2 and are reported in Table 5. Variants M1, X1, X2, S1-L, S2-L, and S3-L alter the label space or evaluation regime and are therefore reported descriptively rather than compared inferentially with B0.

ID	Scenario	Observed $\overline{\mathrm{Acc}}_{ct}$	Δ vs. baseline	Main observation
B0	Reference multimodal configuration: 4 semantic classes, overt EEG+EMG+audio	0.42	0	Reference configuration with full semantic and alignment machinery enabled.
M1	8-class labels: text/image variants treated as separate classes	0.23	−0.19	Lower accuracy under a larger modality-specific label space; the effects of increased class number and modality-specific labeling cannot be separated.
S1-C	No EMG in overt training (EEG+audio only)	0.45	+0.03	Removing overt-phase EMG produced a modest improvement relative to B0, suggesting that overt EMG did not transfer reliably to held-out covert decoding when audio was retained.
S2-C	No audio in overt training (EEG+EMG only)	0.43	+0.01	Removing overt-phase audio produced a small improvement relative to B0, suggesting limited transfer of acoustic supervision to held-out covert decoding when EMG was retained.
S3-C	EEG-only overt training	0.52	+0.10	Best performance in the held-out-covert setting; EEG-only overt supervision yielded the most robust cross-subject decoder among the evaluated modality configurations.
C1	No covert EMG; overt branch unchanged	0.46	+0.04	Suppressing covert-phase EMG improved held-out-subject covert decoding, indicating that covert EMG did not transfer reliably across participants.
P1	No sentence-transformer prototypes or semantic loss	0.39	−0.03	Removing semantic prototype supervision produced a modest reduction in covert-test accuracy; semantic retrieval is not directly available in this variant.
R1	Stage 3 semantic-consistency alignment removed	0.37	−0.05	Removing Stage 3 overt–covert consistency alignment reduced covert-test accuracy, supporting its contribution to cross-phase transfer.
R2	No CORAL-based domain alignment	0.40	−0.02	Removing CORAL produced a small reduction in covert-test accuracy, consistent with a modest contribution from phase-domain alignment.
X1	Within-subject leave-one-run-out (LORO)	0.63	+0.21	Removing inter-subject variability substantially simplifies covert decoding.
X2	Strict LOSO, no target-subject overt calibration	0.38	−0.04	Lack of subject-specific overt anchors makes cross-subject transfer harder but still feasible.
S1-L	No EMG in overt training, LORO	0.53	+0.11	Under within-subject LORO evaluation, EEG+audio achieved substantially higher accuracy than the cross-subject reference configuration.
S2-L	No audio in overt training, LORO	0.55	+0.13	EEG+EMG achieved the highest accuracy among the LORO modality variants, suggesting that EMG was more informative within subject than across subjects.
S3-L	EEG-only overt training, LORO	0.47	+0.05	EEG-only decoding also improved under LORO relative to the cross-subject reference, although it remained below the LORO variants containing an auxiliary modality.

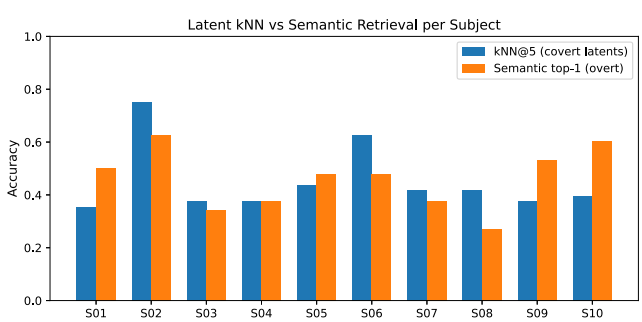

Fig. 11. Latent-space evaluation across subjects using two complementary metrics. Blue bars show k-NN@5 accuracy computed on covert latent embeddings, while orange bars report semantic retrieval top-1 accuracy on overt trials using class prototypes in the semantic space. The alignment between the two metrics highlights how well the learned latent space preserves class structure and semantic consistency across subjects.

Table 5
Paired participant-level comparisons of the same-protocol ablations with the reference configuration B0. $\Delta\mathrm{Acc}_{ct} = \mathrm{Acc}_a - \mathrm{Acc}_{B0}$. Confidence intervals are 95% BCa bootstrap intervals, and p_{Holm} denotes Holm-adjusted exact two-sided sign-flipping p-values.

Variant	$\Delta\mathrm{Acc}_{ct}$	95% CI	p_{Holm}
S1-C	+0.034	[0.018, 0.048]	0.0156
S2-C	+0.014	[0.004, 0.024]	0.0391
S3-C	+0.104	[0.084, 0.125]	0.0137
C1	+0.044	[0.034, 0.054]	0.0137
P1	−0.026	[−0.032, −0.019]	0.0137
R1	−0.046	[−0.054, −0.039]	0.0137
R2	−0.016	[−0.018, −0.013]	0.0137

small number of high-performing participants. The direction of the paired effects was also highly consistent: S3-C and C1 improved covert-test accuracy for all ten participants, whereas P1, R1, and R2 reduced

Table 6
Representative speech/inner-speech BCIs compared with the present work. "Perf." reports one representative performance metric from each study; because the underlying tasks differ substantially, these values are informative but not directly comparable across rows. "Gen." summarizes the main evaluation setting.

Study	Modality/invasiveness	Task/output	Vocab/units	Perf.	Gen.
Willett et al. [1]	Intracortical spikes; highly invasive	Attempted continuous speech to text	50-word and 125k-word vocabularies	9.1% WER; 23.8% WER; 62 WPM	Single participant, within-subject, real-time
Metzger et al. [41]	ECoG; highly invasive	Silent spelling via code words	26 NATO code words; 1152-word sentence vocabulary	6.13% CER; 10.53% WER; 29.4 CPM	Single participant, within-subject, real-time
Chaudhary et al. [3]	Intracortical microelectrodes; highly invasive	Auditory speller/binary selection	2 responses for letter-group navigation	1.08 characters/min; 5.2 bits/min ITR	Single participant, within-subject, real-time
Dash et al. [5]	MEG; non-invasive, non-portable	Spoken and imagined phrase classification	5 phrases	Up to 93% imagined, 96% spoken accuracy	Subject-dependent
Défossez et al. [24]	MEG + EEG; non-invasive	Brain–speech segment matching	> 1000 candidate segments	MEG top-1 accuracy up to 41%	Large cohort; cross-subject
Typical EEG imagined-speech BCIs [4,6,7,9,10]	Scalp EEG; non-invasive	Imagined vowels/words classification	2–9 classes	Typically ~60%–80% in binary tasks; ~20%–50% in more complex tasks	Mostly subject-dependent; limited LOSO
Lee et al. [27]	Scalp EEG + audio; non-invasive	Spoken/imagined speech to voice reconstruction	12 words/phrases + silent phase	Imagined EEG: 68.3% CER; MOS 2.78	Subject-dependent; small-*N* LOSO explored
This work	**Scalp EEG + EMG + overt audio; non-invasive**	**Overt and covert semantic command classification**	**4 everyday inner-speech commands**	**Overt-val acc. 0.54; covert-test acc. 0.42**	**10 participants; subject-held-out covert evaluation with target-subject overt calibration; strict LOSO ablation**

Note: WER = word error rate; CER = character error rate; WPM = words per minute; CPM = characters per minute; ITR = information transfer rate; MOS = mean opinion score; NATO = North Atlantic Treaty Organization.

accuracy for all ten participants. S1-C improved accuracy for eight participants, produced one tie, and reduced accuracy for one participant, while S2-C improved accuracy for seven participants, produced one tie, and reduced accuracy for two participants. Relative to B0, EEG-only overt training (S3-C) produced the largest improvement, increasing mean covert-test accuracy by 0.104, or 10.4 percentage points (95% BCa CI [0.084, 0.125]). Suppressing covert-phase EMG (C1) increased accuracy by 4.4 percentage points (95% BCa CI [0.034, 0.054]). S1-C and S2-C produced smaller improvements of 3.4 and 1.4 percentage points, respectively. Conversely, removing the semantic prototypes and semantic loss (P1), Stage 3 overt–covert consistency alignment (R1), or CORAL domain alignment (R2) reduced mean accuracy by 2.6, 4.6, and 1.6 percentage points, respectively. All seven comparisons remained significant after Holm correction, although the absolute effects for S2-C, P1, and R2 were small.

Overall, the ablations reveal three patterns. First, the eight-class formulation was harder than the four-class task, although the effects of doubling the class count and introducing modality-specific labels cannot be separated. Second, auxiliary modalities were more beneficial within subject than under held-out-subject transfer. Third, removing semantic prototypes, overt–covert consistency, or CORAL alignment produced small but consistent accuracy reductions and weakened representational organization.

4. Discussion

The framework was designed to isolate the contribution of multimodal supervision, semantic anchoring, overt–covert alignment, and calibration within a single reproducible pipeline. For this reason, the backbone and hyperparameters were kept fixed across ablations rather than retuned separately for each variant. Consequently, the contribution of the individual components should not be evaluated solely through their effect on aggregate classification accuracy. Some components, particularly semantic prototype alignment, primarily influence semantic retrieval and latent-space organization, whereas others target overt–covert consistency, domain transfer, or calibration. The ablations are therefore intended to characterize these complementary roles within a common pipeline, rather than to demonstrate that the complete configuration is necessarily the simplest or highest-performing possible decoder.

Overall, the results show that the proposed pipeline supports above-chance covert command decoding and preserves structured latent representations across overt and covert phases under the evaluation protocol used here. The above-chance covert performance, together with the pronounced subject-wise variability and structured confusion patterns, suggests that the model captures intention-related information but does not yet provide uniformly reliable separation of the four commands across participants.

Calibration error is higher in covert than in overt decoding, indicating that probability estimates become less reliable when the model is transferred from overt-calibrated training to audio-free covert inference. This supports the use of explicit calibration and uncertainty-aware decision rules in deployment. Semantic retrieval and prototype-similarity heatmaps indicate structured alignment between latent representations and command prototypes, while the P1/R1/R2 ablations show that removing semantic supervision mainly disrupts this representational structure and only modestly changes four-class accuracy.

The participant-level paired analysis provides additional support for the ablation interpretation. The improvement obtained with EEG-only overt supervision was observed for all ten participants and was the largest same-protocol effect, while suppressing covert EMG also improved performance consistently across participants. The smaller changes associated with S2-C, P1, and R2 remained significant after multiplicity correction but amounted to only approximately 1.4–2.6 percentage points. Thus, statistical significance in these comparisons should not be conflated with a large practical effect, particularly given the small cohort and the stochastic nature of deep-network optimization.

The ablations reveal a regime-dependent role for auxiliary modalities. EMG and audio were informative in within-subject analyses but

transferred less reliably to held-out participants, likely because they contained subject-specific articulatory and acoustic variability. Consistently, EEG-only overt supervision and suppression of covert EMG improved held-out-subject covert accuracy. Because the EMG montage was designed for exploratory multimodal decoding, however, these results should not be interpreted as a physiological characterization of residual muscle activity or as evidence that EEG was free of myogenic contributions.

4.1. Comparison with previous work

Speech and inner-speech BCIs span a broad design space, ranging from highly invasive intracortical and ECoG neuroprostheses for continuous text generation or spelling, to non-invasive MEG and EEG systems for phrase classification, segment matching, or small-vocabulary imagined-speech decoding [1–3,16,18,19,22]. Because these studies target substantially different outputs and use different evaluation protocols, a meaningful comparison should not rely on a single raw performance number. Instead, it is more informative to compare them jointly in terms of sensing modality and invasiveness, output granularity, representative reported performance, and generalization regime.

These studies also show that subject-independent evaluation encompasses different regimes, including unlabeled target-domain adaptation, few-shot support, zero-calibration meta-learning, and subject-invariant training [11–14]. Their reported accuracies are therefore not directly comparable with either our target-subject overt-calibrated protocol or the strict X2 LOSO condition; likewise, large-scale M/EEG word-decoding results obtained during listening or reading address a different task from covert command production [15].

Taken together, Table 6 places the present framework between two extremes. Compared with invasive speech neuroprostheses, it offers substantially lower communication bandwidth but avoids neurosurgery and is evaluated on multiple participants rather than a single implanted case. Compared with many non-invasive imagined-speech EEG studies, the present work places greater emphasis on cross-subject covert evaluation, semantic structure, and calibration analysis. The combination of multimodal overt supervision, explicit overt–covert alignment, and sentence-embedding prototypes distinguishes the present approach from conventional EEG-based imagined-speech classification pipelines and from purely acoustic reconstruction pipelines.

From a practical BCI perspective, the present system should be considered an offline proof of concept rather than a communication-ready interface. Using the standard four-class information-transfer-rate (ITR) formulation, the mean covert accuracy of 0.416 and the 3-s decoding window correspond to a nominal upper-bound rate of approximately 1.9 bits/min. This estimate excludes cue presentation, rest intervals, overt calibration, preprocessing, and rejected predictions and therefore should not be interpreted as end-to-end online communication throughput. Practical deployment will require online validation, shorter interaction cycles, simplified acquisition, and evaluation against established stimulus-driven EEG communication interfaces.

4.2. Limitations and future directions

Several limitations concern the scope and generalizability of the present study. The vocabulary was restricted to four discrete semantic commands expressed in a single language, and the experiments were conducted with healthy volunteers in a controlled laboratory environment. Although this setting is appropriate for investigating overt–covert transfer and semantic alignment, it does not capture the linguistic richness, user variability, or environmental complexity of real-world assistive communication. The current framework also targets closed-set command classification rather than continuous or generative message decoding.

The evaluation was entirely offline and involved a relatively small cohort, a single acquisition setup, and fixed hyperparameter choices. In addition, each experimental configuration was represented by only one training realization per held-out-participant fold. The paired statistical analysis therefore quantifies the consistency of the observed accuracy differences across participants, but it does not separately estimate variability arising from random initialization, minibatch ordering, or stochastic optimization. Repeated experiments with multiple random seeds, larger cohorts, and independent datasets will be required to determine whether the smaller ablation effects remain stable. The current evaluation also does not account for fatigue, attention fluctuations, or long-term user-decoder co-adaptation.

The calibration results motivate selective decoding strategies in which low-confidence predictions are rejected and the user is asked to repeat the command, thereby trading decision coverage and information-transfer rate for improved reliability. However, confidence thresholds and rejection policies were not evaluated prospectively, and their effects on usability, latency, and communication rate remain unknown. These aspects should be assessed in future closed-loop experiments.

The physiological interpretation of the EMG measurements is also limited. The setup used only two surface EMG channels and a preprocessing scheme designed for multimodal decoding rather than detailed characterization of articulatory activity. Consequently, the present data cannot comprehensively quantify residual muscle activation or EMG contamination during covert speech. Future studies should report electrode placement in greater anatomical detail, preserve the original 500 Hz EMG signal when addressing EMG physiology, and examine dedicated EEG–EMG coupling measures such as corticomuscular coherence. Moreover, because microphone data were not recorded during covert phases, inadvertent whispering or vocalization could not be retrospectively screened acoustically. Continuous microphone monitoring should therefore be included in future protocols. Decoding-oriented analyses should also be complemented by population-level investigations of the spatial, spectral, temporal, and cross-frequency neural processes that distinguish overt from covert speech.

Taken together, these limitations motivate several directions for further research: expanding the protocol to larger vocabularies and graded intention hierarchies; validating the framework in clinical, cross-lingual, and cross-device settings; integrating the decoder into closed-loop BCIs with uncertainty-aware feedback; combining inner-speech decoding with task-independent semantic feature extraction strategies [20]; and linking semantic anchoring with generative language models or secure semantic communication frameworks [42, 43].

5. Conclusion

We introduced a semantic-aware framework for four-class covert inner-speech decoding using EEG, auxiliary EMG, and overt-speech audio supervision. Under subject-held-out covert evaluation with target-subject overt calibration, the framework achieved a mean covert accuracy of 0.42, significantly above the 0.25 chance level, while preserving class-related semantic structure. Ablations showed that semantic and overt–covert alignment contributed to the learned representations, but the transfer value of auxiliary modalities depended on the evaluation regime: EMG and audio were useful within subject, whereas EEG-only supervision generalized more robustly across participants. These results establish a controlled framework for studying semantic overt-to-covert transfer while highlighting subject variability, calibration, and cross-modal transfer as major obstacles to practical non-invasive speech BCIs.

CRediT authorship contribution statement

Hossein Ahmadi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Eduardo**

Santamaría-Vázquez: Writing – review & editing, Visualization, Supervision, Software, Project administration, Methodology, Data curation, Conceptualization. **Luca Mesin:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition. **Roberto Hornero:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Ethical considerations and informed consent

All procedures involving human participants were conducted in accordance with the Declaration of Helsinki and with the applicable institutional procedures of the University of Valladolid. The study involved healthy adult volunteers participating in a low-risk, non-invasive EEG–EMG–audio experimental protocol. No clinical intervention was performed, no biological samples were collected, and no vulnerable populations were involved. All participants were informed about the nature, purpose, and procedures of the study and provided written informed consent before participation. Participation was voluntary, and participants were free to withdraw from the study at any time. A signed institutional clarification letter from the senior investigator at the host institution, explaining the institutional handling of this low-risk non-invasive study, is provided as supporting documentation.

Funding

The authors declare that no specific funding was received for this work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the Grupo de Ingeniería Biomédica (GIB) at the University of Valladolid for hosting the experimental sessions and providing access to the MEDUSA facilities used to acquire the dataset.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.bspc.2026.111527.

Data availability

The authors do not have permission to share data.

References

[1] F.R. Willett, E.M. Kunz, C. Fan, D.T. Avansino, G.H. Wilson, E.Y. Choi, F. Kamdar, M.F. Glasser, L.R. Hochberg, S. Druckmann, K.V. Shenoy, J.M. Henderson, A high-performance speech neuroprosthesis, Nature 620 (2023) 1031–1036, http://dx.doi.org/10.1038/s41586-023-06377-x.

[2] S.L. Metzger, K.T. Littlejohn, A.B. Silva, D.A. Moses, M.P. Seaton, R. Wang, M.E. Dougherty, J.R. Liu, P. Wu, M.A. Berger, I. Zhuravleva, A. Tu-Chan, K. Ganguly, G.K. Anumanchipalli, J. Chartier, E.F. Chang, A high-performance neuroprosthesis for speech decoding and avatar control, Nature 620 (2023) 1037–1046, http://dx.doi.org/10.1038/s41586-023-06443-4.

[3] U. Chaudhary, I. Vlachos, J.B. Zimmermann, A. Espinosa, A. Tonin, A. Jaramillo-Gonzalez, M. Khalili-Ardali, H. Topka, J. Lehmberg, G.M. Friehs, A. Woodtli, J.P. Donoghue, N. Birbaumer, A. Gharabaghi, Spelling interface using intracortical signals in a completely locked-in patient enabled via auditory neurofeedback training, Nat. Commun. 13 (1) (2022) 1236, http://dx.doi.org/10.1038/s41467-022-28859-8.

[4] K. Bhadra, A.-L. Giraud, S. Marchesotti, Learning to operate an imagined speech brain-computer interface involves the spatial and frequency tuning of neural activity, Commun. Biol. 8 (2025) 271, http://dx.doi.org/10.1038/s42003-025-07464-7.

[5] D. Dash, P. Ferrari, J. Wang, Decoding imagined and spoken phrases from non-invasive neural (MEG) signals, Front. Neurosci. 14 (2020) 290, http://dx.doi.org/10.3389/fnins.2020.00290.

[6] J.T. Panachakel, A.G. Ramakrishnan, Decoding covert speech from EEG—a comprehensive review, Front. Neurosci. 15 (2021) 642251, http://dx.doi.org/10.3389/fnins.2021.642251.

[7] D. Lopez-Bernal, D. Balderas, P. Ponce, A. Molina, A state-of-the-art review of EEG-based imagined speech decoding, Front. Hum. Neurosci. 16 (2022) 867281, http://dx.doi.org/10.3389/fnhum.2022.867281.

[8] H. Ahmadi, L. Mesin, Decoding visual imagination and perception from EEG via topomap sequences, in: 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, 2025, pp. 1–7, http://dx.doi.org/10.1109/EMBC58623.2025.11251641.

[9] A. Einizade, M. Mozafari, S. Jalilpour, S. Bagheri, S.H. Sardouie, Neural decoding of imagined speech from EEG signals using the fusion of graph signal processing and graph learning techniques, Neurosci. Informatics 2 (3) (2022) 100091, http://dx.doi.org/10.1016/j.neuri.2022.100091.

[10] V.R. Carvalho, E.M.A.M. Mendes, A. Fallah, T.J. Sejnowski, L. Comstock, C. Lainscsek, Decoding imagined speech with delay differential analysis, Front. Hum. Neurosci. 18 (2024) 1398065, http://dx.doi.org/10.3389/fnhum.2024.1398065.

[11] M. Jiménez-Guarneros, P. Gómez-Gil, Standardization-refinement domain adaptation method for cross-subject EEG-based classification in imagined speech recognition, Pattern Recognit. Lett. 141 (2021) 54–60, http://dx.doi.org/10.1016/j.patrec.2020.11.013.

[12] A. Hernandez-Galvan, G. Ramirez-Alonso, J. Ramirez-Quintana, A prototypical network for few-shot recognition of speech imagery data, Biomed. Signal Process. Control. 86 (2023) 105154, http://dx.doi.org/10.1016/j.bspc.2023.105154.

[13] H.W. Ng, C. Guan, Subject-independent meta-learning framework towards optimal training of EEG-based classifiers, Neural Netw. 172 (2024) 106108, http://dx.doi.org/10.1016/j.neunet.2024.106108.

[14] S. Dörterler, E. Şahin, D. Özdemir, A dynamic subject-invariant fragment mixing strategy to suppress subject variability in EEG imagined speech classification, Neuroscience 608 (2026) 18–45, http://dx.doi.org/10.1016/j.neuroscience.2026.05.004.

[15] S. d'Ascoli, C. Bel, J. Rapin, H. Banville, Y. Benchetrit, C. Pallier, J.-R. King, Towards decoding individual words from non-invasive brain recordings, Nat. Commun. 16 (1) (2025) 10521, http://dx.doi.org/10.1038/s41467-025-65499-0.

[16] H. Akbari, B. Khalighinejad, J.L. Herrero, A.D. Mehta, N. Mesgarani, Towards reconstructing intelligible speech from the human auditory cortex, Sci. Rep. 9 (1) (2019) 874, http://dx.doi.org/10.1038/s41598-018-37359-z.

[17] E. Santamaria-Vazquez, V. Martinez-Cagigal, F. Vaquerizo-Villar, R. Hornero, EEG-Inception: A novel deep convolutional neural network for assistive ERP-based brain-computer interfaces, IEEE Trans. Neural Syst. Rehabil. Eng. 28 (12) (2020) 2773–2782, http://dx.doi.org/10.1109/TNSRE.2020.3048106.

[18] S. Martin, P. Brunner, C. Holdgraf, H.-J. Heinze, N.E. Crone, J. Rieger, G. Schalk, R.T. Knight, B.N. Pasley, Decoding spectrotemporal features of overt and covert speech from the human cortex, Front. Neuroeng. 7 (2014) 14, http://dx.doi.org/10.3389/fneng.2014.00014.

[19] M. Angrick, M.C. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, J. Saal, A.J. Colon, L. Wagner, D.J. Krusienski, P.L. Kubben, T. Schultz, C. Herff, Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity, Commun. Biol. 4 (1) (2021) 1055, http://dx.doi.org/10.1038/s42003-021-02578-0.

[20] H. Ahmadi, L. Mesin, Universal semantic feature extraction from EEG signals: A task-independent framework, J. Neural Eng. 22 (3) (2025) 036003, http://dx.doi.org/10.1088/1741-2552/add08f.

[21] D. Gündüz, Z. Qin, I.E. Aguerri, H.S. Dhillon, Z. Yang, A. Yener, K.-K. Wong, C.-B. Chae, Beyond transmitting bits: Context, semantics, and task-oriented communications, IEEE J. Sel. Areas Commun. 41 (1) (2023) 5–41, http://dx.doi.org/10.1109/JSAC.2022.3223408.

[22] T. Proix, J. Delgado Saa, A. Christen, S. Martin, B.N. Pasley, R.T. Knight, X. Tian, D. Poeppel, W.K. Doyle, O. Devinsky, et al., Imagined speech can be decoded from low- and cross-frequency intracranial EEG features, Nat. Commun. 13 (1) (2022) 48, http://dx.doi.org/10.1038/s41467-021-27725-3.

[23] J. Kwon, G. Lee, Y.J. Park, E.J. Lee, C.K. Chung, Mimed speech as an intermediary state between overt and imagined speech production in an electrocorticography study, Sci. Rep. 15 (1) (2025) 33393, http://dx.doi.org/10.1038/s41598-025-18537-2.

[24] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, J.-R. King, Decoding speech perception from non-invasive brain recordings, Nat. Mach. Intell. 5 (2023) 1097–1107, http://dx.doi.org/10.1038/s42256-023-00714-5.

[25] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv preprint arXiv:1807.03748, URL https://arxiv.org/abs/1807.03748.

[26] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 8748–8763, URL https://proceedings.mlr.press/v139/radford21a.html.

[27] Y.-E. Lee, S.-H. Lee, S.-H. Kim, S.-W. Lee, Towards voice reconstruction from EEG during imagined speech, Proc. AAAI Conf. Artif. Intell. 37 (5) (2023) 6030–6038, http://dx.doi.org/10.1609/aaai.v37i5.25745.

[28] E. Santamaría-Vázquez, V. Martínez-Cagigal, D. Marcos-Martínez, V. Rodríguez-González, S. Pérez-Velasco, S. Moreno-Calderón, R. Hornero, MEDUSA©: A novel Python-based software ecosystem to accelerate brain-computer interface and cognitive neuroscience research, Comput. Methods Programs Biomed. 230 (2023) 107357, http://dx.doi.org/10.1016/j.cmpb.2023.107357.

[29] A.D. Vigotsky, I. Halperin, G.J. Lehman, G.S. Trajano, T.M. Vieira, Interpreting signal amplitudes in surface electromyography studies in sport and rehabilitation sciences, Front. Physiol. 8 (2018) 985, http://dx.doi.org/10.3389/fphys.2017.00985.

[30] L. Nalborczyk, R. Grandchamp, E.H.W. Koster, M. Perrone-Bertolotti, H. Lœvenbruck, Can we decode phonetic features in inner speech using surface electromyography? PLoS One 15 (5) (2020) e0233282, http://dx.doi.org/10.1371/journal.pone.0233282.

[31] J. Liu, Y. Sheng, H. Liu, Corticomuscular coherence and its applications: A review, Front. Hum. Neurosci. 13 (2019) 100, http://dx.doi.org/10.3389/fnhum.2019.00100.

[32] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, 1987, http://dx.doi.org/10.1002/0471725382.

[33] S. Hartmann, M. Baumert, Subject-level normalization to improve a-phase detection of cyclic alternating pattern in sleep EEG, in: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, 2023, pp. 1–4, http://dx.doi.org/10.1109/EMBC40787.2023.10340124.

[34] B. McFee, C. Raffel, D. Liang, D.P.W. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in Python, in: Proceedings of the 14th Python in Science Conference, third ed., 2015, pp. 18–24, http://dx.doi.org/10.25080/Majora-7b98e-003.

[35] V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance, Eegnet: A compact convolutional neural network for EEG-based brain-computer interfaces, J. Neural Eng. 15 (5) (2018) 056013, http://dx.doi.org/10.1088/1741-2552/aace8c.

[36] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992, http://dx.doi.org/10.18653/v1/D19-1410, URL https://aclanthology.org/D19-1410/.

[37] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, Adv. Neural Inf. Process. Syst. 33 (2020) 5776–5788, URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[38] J. Deng, J. Guo, N. Xue, S. Zafeiriou, ArcFace: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699, http://dx.doi.org/10.1109/CVPR.2019.00482.

[39] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 18661–18673, URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

[40] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 1597–1607, URL https://proceedings.mlr.press/v119/chen20j.html.

[41] S.L. Metzger, J.R. Liu, D.A. Moses, M.E. Dougherty, M.P. Seaton, K.T. Littlejohn, J. Chartier, G.K. Anumanchipalli, A. Tu-Chan, K. Ganguly, E.F. Chang, Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis, Nat. Commun. 13 (1) (2022) 6510, http://dx.doi.org/10.1038/s41467-022-33611-3.

[42] H. Ahmadi, A. Kuhestani, L. Mesin, Adversarial neural network training for secure and robust brain-to-brain communication, IEEE Access 12 (2024) 39450–39469, http://dx.doi.org/10.1109/ACCESS.2024.3376657.

[43] H. Ahmadi, A. Kuhestani, M. Keshavarzi, L. Mesin, Securing brain-to-brain communication channels using adversarial training on SSVEP EEG, IEEE Access 13 (2025) 14358–14378, http://dx.doi.org/10.1109/ACCESS.2025.3528770.