NMCNET: A CORTEX-GUIDED EEG-EMG FUSION FRAMEWORK FOR CROSS-SUBJECT DECODING OF MANDARIN INITIALS AND FINALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain-computer interface (BCI) speech decoding offers a promising communication pathway for individuals with severe speech impairments. However, Mandarin initial and final decoding remains underexplored, and existing EEG-EMG fusion methods rarely encode an asymmetric interaction motivated by cortical control and articulatory muscle responses. We propose NMCNet, a cross-subject EEG-EMG fusion framework for audible and silent speech decoding. NMCNet integrates Cortical Drive Gating, which uses EEG features to modulate EMG representations, with Feedback Cross-Modal Fusion, which reintegrates the resulting muscle-response information into the EEG representation. A complementary multi-frequency branch captures frequency-specific dynamics. To improve cross-subject generalization, we further introduce a cross-subject-enhanced supervised contrastive loss that emphasizes same-class samples from different subjects. A Fisher-based channel-selection strategy reduces the EEG input to 15 channels while outperforming the full channel configuration, supporting lightweight acquisition. Across ten participants, NMCNet achieves the best performance under silent speech, with a mean accuracy of 70.25% across the initial and final tasks, each comprising three broad articulatory categories. In a fine-grained setting that treats each individual initial or final within its broad category as a separate class, NMCNet achieves macro-averaged F1 and recall scores of 73.48% and 74.35%, respectively. Ablation studies support the effectiveness of the proposed architectural components and cross-subject learning strategy. These results demonstrate the potential of lightweight cross-subject Mandarin decoding and silent-speech BCIs for individuals with anarthria, severe dysarthria, and other severe motor speech impairments.

1 INTRODUCTION

Brain-computer interface speech decoding offers an important assistive communication pathway for individuals living with anarthria or severe dysarthria following stroke, including brainstem stroke (Willett et al., 2023; Sellers et al., 2014). Mandarin syllables are structured around initials and finals, which convey complementary articulatory information: initials are mainly distinguished by place and manner of articulation, whereas finals are characterized by vowel quality and nasalization. Decoding these components therefore requires sensitivity to both rapid articulatory transitions and more sustained vocal-tract configurations, presenting a finer-grained challenge than tone classification. Nevertheless, neural decoding of Mandarin initials and finals remains largely underexplored.

Electroencephalography (EEG), owing to its high temporal resolution and noninvasive acquisition, is widely used for neural speech decoding because it captures cortical activity related to speech planning and phonological encoding (Sharon et al., 2020; Dash et al., 2021). However, speech production involves coordinated cortical planning and peripheral articulatory execution, and EEG alone provides only a partial representation of this process (Guenther, 2006). Electromyography (EMG) complements EEG by recording orofacial muscle activation associated with articulation (Barrientos Rojas et al., 2022). Our preliminary ERP and ERSP analyses reveal modality-specific and category-dependent temporal and spectral patterns across the initial and final tasks (see Ap-

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

pendix C), suggesting complementary neural and myoelectric information and motivating their joint modeling.

Despite growing interest in multimodal speech decoding, existing work faces four limitations. First, Chinese speech-decoding studies have focused largely on sentences or lexical tones (Qian et al., 2025; Li et al., 2021; Wang et al., 2023), whereas Mandarin initials and finals remain comparatively understudied. Decoding these units requires distinguishing subtle differences in articulatory manner and place, vowel quality, and nasalization, making the task more fine-grained than tone classification. Second, many EEG-EMG fusion methods rely on direct concatenation or symmetric feature interaction, treating the two modalities as equivalent inputs. Such designs do not explicitly encode an asymmetric inductive bias motivated by the directional organization of cortical control and articulatory muscle responses. Third, high-density EEG systems increase hardware complexity and user burden, yet principled channel-reduction strategies that preserve decoding performance remain limited. Finally, substantial inter-subject variability in cortical activity and articulatory patterns continues to hinder generalization under leave-one-subject-out (LOSO) evaluation.

To address these limitations, we introduce NMCNet (Neuromuscular Coupling Network), a cross-subject EEG-EMG fusion framework for Mandarin initial and final decoding. NMCNet combines a temporal branch for rapid articulatory dynamics with a multi-frequency branch for frequency-specific patterns. Within the temporal branch, Cortical Drive Gating (CDG) uses EEG features to modulate EMG representations, and Feedback Cross-Modal Fusion (FCMF) feeds the resulting muscle-response information back into the EEG representation, forming an asymmetric, cortex-guided fusion process. We further apply Fisher-based channel selection to reduce the EEG montage to 15 channels and introduce a cross-subject-enhanced supervised contrastive loss that upweights same-class pairs from different subjects. In LOSO experiments with ten participants, NMCNet achieved a mean accuracy of 70.25% across the two silent-speech tasks. Fine-grained silent-speech decoding yielded macro-averaged F1 and recall scores of 73.48% and 74.35%, respectively.

Our main contributions are as follows: (1) We propose an asymmetric, cortex-guided EEG-EMG fusion framework with an explicitly validated directional ordering, combining EEG-driven EMG modulation with EMG-informed EEG updating. (2) We introduce a cross-subject-enhanced supervised contrastive objective that directly upweights same-class pairs from different subjects without an additional auxiliary loss. (3) We provide cross-subject evaluation across audible and silent speech at broad-category and fine-grained levels, with component, modality, reverse-direction, parameter-matched symmetric controls, and dedicated multimodal baselines.

2 Related Work

2.1 EEG-based Speech Decoding

EEG-based speech decoding has been explored at the phoneme, syllable, and word levels under both overt and imagined speech conditions. Early studies demonstrated the feasibility of classifying imagined vowels using common spatial patterns (DaSalla et al., 2009), distinguishing phonological categories during imagined and articulated speech (Zhao & Rudzicz, 2015), and decoding imagined words from hybrid connectivity features (Qureshi et al., 2018). In Mandarin speech decoding, however, most research has focused on lexical tones. Li et al. (2021) reported tone classification using Riemannian-manifold features, while Wang et al. (2023) developed an end-to-end CNN for cross-subject decoding of Mandarin vowels and tones.

Mandarin initials and finals are fundamental phonological units distinct from lexical tone, yet their neural decoding remains comparatively understudied. Gu et al. (2026) investigated isolated initials and finals using within-subject binary classification. In contrast, our study addresses cross-subject multiclass decoding of naturally coarticulated syllables under both audible and silent speech conditions, covering both broad articulatory categories and fine-grained phonological units.

2.2 Multimodal EEG-EMG Fusion for Speech

EMG records orofacial muscle activity during articulation and therefore provides peripheral information that complements the cortical activity measured by EEG (Janke & Diener, 2017; Wu et al., 2022). Speech production involves a directional organization in which cortical motor commands

drive articulatory muscle activation and peripheral articulatory feedback contributes to ongoing cortical control (Guenther, 2006). This physiological organization motivates fusion strategies that preserve the distinct roles of the two modalities rather than treating them as interchangeable inputs.

Beyond speech decoding, NeuroFusion-Trans uses modality-specific Transformer encoders, learned cross-modality fusion, temporal synchronization, and online adaptation for joint EEG-EMG intent recognition in assistive robotics (Sultan et al., 2025). Previous speech studies have combined EEG and EMG for silent-speech reconstruction through sequence-to-sequence modeling (Li et al., 2023) and for Mandarin tone decoding through cross-attention-based fusion (Zhuang et al., 2026). However, these approaches do not explicitly impose the ordered interaction adopted here. NMCNet introduces a physiologically motivated asymmetric inductive bias by first using EEG features to modulate EMG representations and then feeding the resulting muscle-response information back into the EEG representation.

2.3 Cross-Subject Generalization

Cross-subject generalization remains a major challenge in practical BCI systems because neural and articulatory patterns vary substantially across individuals (Chen et al., 2025; Lee & Lee, 2022). Simply pooling data from multiple source participants may also be suboptimal, as it can obscure subject-specific distributions and reduce transferability to unseen users (Zhu et al., 2019). Supervised contrastive learning provides a promising solution by encouraging samples from the same class to form compact representations while separating samples from different classes (Khosla et al., 2020).

Recent work has extended contrastive learning to cross-subject biosignal decoding. Kim et al. (2024), for example, jointly optimized standard supervised contrastive learning with an additional inter-subject similarity objective. In contrast, our approach modifies the supervised contrastive formulation itself by assigning greater weight to same-class pairs from different subjects. This design promotes subject-invariant yet class-discriminative representations without introducing a separate auxiliary loss or requiring target-subject adaptation data.

3 Method

Motivated by the asymmetric interaction between cortical control and articulatory muscle activity, we propose NMCNet, a two-branch EEG-EMG architecture for cross-subject decoding of Mandarin initials and finals (Fig. 1). The temporal-spatial branch encodes EEG and EMG features, after which CDG uses EEG features to modulate EMG features. FCMF transforms the gated EMG feature into a feedback message that refines the EEG representation through a gated residual update. The 64-dimensional refined EEG feature and the 64-dimensional gated EMG feature are concatenated into a 128-dimensional representation. The multi-frequency branch uses Morlet-initialized trainable convolutions and SE gating (Hu et al., 2018) to extract and reweight frequency-specific features from both modalities, producing another 128-dimensional representation. Both outputs are concatenated into a 256-dimensional representation for classification and cross-subject-enhanced supervised contrastive learning. Classifier dimension is adjusted for the three-class broad-category tasks and fine-grained tasks with 4 to 7 classes.

3.1 Signal Inputs

Let $X_i = \{X_i^e, X_i^m\}$ denote the paired EEG and EMG recordings for trial i, where $X_i^e \in \mathbb{R}^{15\times600}$ and $X_i^m \in \mathbb{R}^{5\times600}$. The 600 time points correspond to a 0–1.2 s post-stimulus window sampled at 500 Hz, covering the initial-to-final articulation process. For the broad-category tasks, each trial is assigned either an initial label $y_i^{initial} \in\{1,2,3\}$, representing plosives, fricatives/affricates, or sonorants, or a final label $y_i^{final} \in\{1,2,3\}$, representing monophthongs, diphthongs, or nasal codas. We additionally construct six fine-grained within-category tasks in which each individual initial or final is treated as a separate class, resulting in 4–7 classes depending on the group. Before being fed into the network, each trial is independently standardized for each channel along the temporal dimension using z-score normalization.

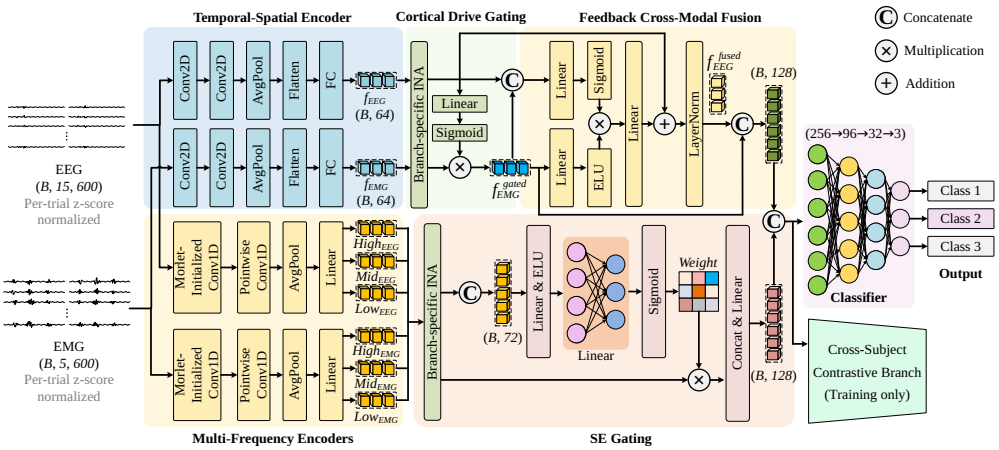


Figure 1: Overall architecture of the proposed NMCNet, illustrated using the three-class setting as an example. The same framework is also used for the fine-grained within-category classification tasks with 4, 5, 6, or 7 classes.

3.2 Temporal-Spatial Encoders

EEG and EMG are processed independently by temporal-spatial encoders with identical architectures but separate parameters. Following the temporal-then-spatial convolution paradigm used in EEG decoding (Lawhern et al., 2018), each encoder applies a temporal convolution with 30 filters of length 25 to capture temporal patterns. A subsequent spatial convolution spans all channels of each modality, integrating information across the 15 EEG channels or 5 EMG channels. Each convolution is followed by batch normalization and ELU activation. Average pooling with a kernel size of 75 and stride of 15 reduces temporal resolution, after which dropout and a fully connected layer project the flattened feature into a 64-dimensional representation. The EEG and EMG features are normalized by branch-specific Instance Normalization Alignment (INA) modules adapted from instance normalization (Ulyanov et al., 2016), yielding $f_{EEG}^{TS} \in \mathbb{R}^{64}$ and $f_{EMG}^{TS} \in \mathbb{R}^{64}$. For a projected feature vector $f_i \in \mathbb{R}^D$, INA computes the mean and variance across its feature dimension and applies a learnable feature-wise affine transform:

$$\mathrm{INA}(f_i)_d = \gamma_d \frac{f_{i,d} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_d, \quad d = 1, \ldots, D. \tag{1}$$

Unlike conventional instance normalization applied to channel-wise spatial or temporal maps, INA is applied after feature projection to each sample's feature vector; EEG and EMG use separate affine parameters.

3.3 Cortical Drive Gating

Given the temporal-spatial features $f_{EEG}^{TS}, f_{EMG}^{TS} \in \mathbb{R}^{64}$, CDG uses the EEG representation to generate a feature-wise gate that modulates the EMG representation:

$$f_{EMG}^{gated} = f_{EMG}^{TS} \odot \sigma\left(W_d f_{EEG}^{TS} + b_d\right) \tag{2}$$

where W_d and b_d are learnable parameters, $\sigma(\cdot)$ denotes the sigmoid function, and $\odot$ denotes element-wise multiplication. The gate selectively enhances or suppresses EMG features according to the EEG context, introducing an asymmetric EEG-to-EMG interaction before cross-modal fusion. This design is motivated by the role of cortical motor representations in speech production (Guenther, 2006), while operating at the feature level rather than explicitly modeling physiological transmission delays.

3.4 Feedback Cross-Modal Fusion

FCMF uses the joint EEG-EMG context to generate feature-wise weights that regulate the contribution of muscle-response information to the EEG representation. It first constructs an EMG feedback message and a joint gate:

$$m = \phi\left(W_m f_{EMG}^{gated} + b_m\right), \tag{3}$$

$$\alpha = \sigma\left(W_g\left[f_{EEG}^{TS}; f_{EMG}^{gated}\right] + b_g\right). \tag{4}$$

The EEG representation is then updated through a gated residual path:

$$f_{EEG}^{fused} = \mathrm{LN}\left(f_{EEG}^{TS} + \mathrm{Dropout}\left(W_r(\alpha \odot m) + b_r\right)\right). \tag{5}$$

Here $\phi(\cdot)$ is ELU, and LN denotes layer normalization. The resulting $f_{EEG}^{fused} \in \mathbb{R}^{64}$ is concatenated with f_{EMG}^{gated} to form the 128-dimensional temporal-spatial representation f_{TS}.

3.5 Multi-Frequency Encoders and SE Gating

The multi-frequency branch processes EEG and EMG independently using three parallel encoders whose convolutional kernels are initialized with Morlet waveforms (Torrence & Compo, 1998) centered at 10.5, 21.5, and 40.0 Hz. These center frequencies serve only as initial spectral priors; the convolutional kernels remain trainable and are optimized jointly with the network, allowing their learned frequency responses to adapt to the decoding task. Within each encoder, depthwise temporal convolution captures frequency-specific activity from each channel, while pointwise convolution integrates information across channels. After global average pooling and branch-specific INA, each EEG branch produces a 16-dimensional feature and each EMG branch an 8-dimensional feature. The six modality-frequency features are concatenated into a 72-dimensional vector and passed through an SE gate (Hu et al., 2018), which generates one adaptive weight for each branch. The reweighted features are then projected into the 128-dimensional multi-frequency representation f_{MF}.

3.6 Training Augmentation

During training, Gaussian noise is applied to the EEG and EMG signals to improve resistance to signal perturbations. We then use synchronized Mixup (Zhang et al., 2018), applying the same mixing coefficient and sample permutation to the paired EEG and EMG inputs, class labels, and subject labels. This preserves cross-modal correspondence and produces the soft class and subject labels used by the classification and contrastive objectives. Detailed augmentation settings are provided in the Appendix A.

3.7 Loss Functions

The joint representation f_{joint} is optimized using a class-weighted focal loss and a cross-subject-enhanced supervised contrastive loss. For the soft Mixup target $\widetilde{y}_i$, class probabilities $p_i = \mathrm{softmax}(o_i)$, and class weights a_c, the classification loss is implemented as

$$\mathcal{L}_{focal} = \frac{1}{B}\sum_{i=1}^{B}\left(1 - \sum_c \widetilde{y}_{i,c} p_{i,c}\right)^{\gamma}\left[-\sum_c a_c \widetilde{y}_{i,c}\log p_{i,c}\right], \tag{6}$$

with $\gamma = 2$ (Lin et al., 2017).

Building on supervised contrastive learning (Khosla et al., 2020), we use the soft class labels $\widetilde{y}_i$ and soft subject labels $\widetilde{s}_i$ to assign greater weight to same-class pairs from different subjects. Let $c_{ij} = \widetilde{y}_i^{\top}\widetilde{y}_j$ denote class similarity and $s_{ij} = \widetilde{s}_i^{\top}\widetilde{s}_j$ denote subject similarity. For $i \neq j$, the positive-pair weight is defined as

$$w_{ij} = c_{ij}\left[\gamma_{cs}\left(1 - s_{ij}\right) + s_{ij}\right] \tag{7}$$

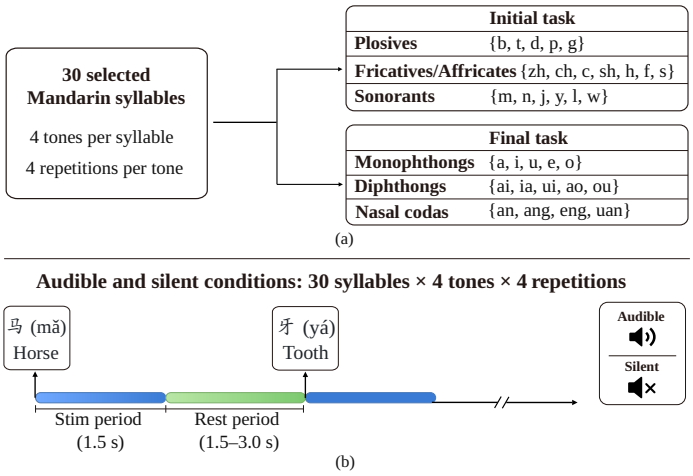


Figure 2: Experimental paradigm. (a) Stimulus design and phonological grouping. (b) Trial timeline for audible and silent speech, with a 1.5 s stimulus followed by a 1.5–3.0 s rest interval.

where γ_{cs} controls the additional emphasis placed on cross-subject same-class pairs. When $\gamma_{cs} = 1$, cross-subject reweighting is removed, and the positive-pair weight reduces to the soft class similarity c_{ij}. Let $z_i = f_{joint,i} / \parallel f_{joint,i} \parallel_2$ denote the normalized representation of sample i. The cross-subject-enhanced contrastive loss is

$$\mathcal{L}_{con} = -\frac{1}{B} \sum_{i=1}^{B} \frac{\sum_{j \neq i} w_{ij} \log \frac{\exp\left(z_i^\top z_j / \tau\right)}{\sum_{k=1}^{B} \exp\left(z_i^\top z_k / \tau\right)}}{\sum_{j \neq i} w_{ij}} \tag{8}$$

where B is the batch size and τ is the temperature parameter. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{focal} + \lambda_{con} \mathcal{L}_{con}, \tag{9}$$

where $\lambda_{con} = 0.2$ for all tasks. Parameters were optimized using Adam (Kingma & Ba, 2015); full settings are provided in Appendix A.

4 Experiments

4.1 Experimental Setup

Dataset Ten healthy native Mandarin-speaking adults aged 24-35 years participated in both audible and silent speech conditions. In silent speech, participants articulated target syllables without audible output. Each condition comprised 480 trials per participant, covering 30 selected Mandarin syllables produced in four tones with four repetitions per tone in pseudorandomized order. The same trials were reorganized to define initial and final decoding tasks (Fig. 2a). Initials were grouped into plosives {b, t, d, p, g}, fricatives/affricates {zh, ch, c, sh, h, f, s}, and sonorants {m, n, j, y, l, w}. Because ai has no overt initial consonant, its trials were excluded from the initial task, leaving 464 trials per participant and condition, with class distributions of 80, 160, and 224. Finals were grouped into monophthongs {a, i, u, e, o}, diphthongs {ai, ia, ui, ao, ou}, and nasal codas {an, ang, eng, uan}. All 480 trials were retained for the final task, including the ai trials, which were assigned to the diphthong class; the corresponding class distributions were 288, 112, and 80. For fine-grained evaluation, six within-category tasks were additionally constructed by treating each individual initial or final as a separate class, resulting in classification problems with 4 to 7 classes.

Data Collection Mandarin syllables containing the target initials and finals were presented under audible and silent conditions (Fig. 2b). Each Chinese character was displayed centrally for 1.5 s, followed by a jittered interstimulus interval of 1.5 to 3.0 s. EEG was recorded using a 64-channel NeuSen W wireless system (Neuracle Technologies, China), together with synchronized EMG from

five orofacial and cervical muscles: the left and right buccinators, left and right cervical trapezius muscles, and the mentalis. Both modalities were sampled at 1000 Hz.

Preprocessing EEG and EMG signals were downsampled to 500 Hz. EEG was band-pass filtered at 0.1–60 Hz, while EMG was high-pass filtered at 10 Hz; both modalities were notch filtered at 48–52 Hz. EEG was rereferenced using the common average. Extended Infomax ICA was applied to the continuous EEG before epoch extraction. Independent components were automatically classified using ICLabel, and components classified as ocular or muscle artifacts with probabilities of at least 0.90 were removed.

Directional Fusion Controls To test whether the proposed asymmetric ordering matters beyond fusion capacity, we evaluated reverse-direction and parameter-matched symmetric controls; the latter uses simultaneous bidirectional EEG-EMG updates with the same number of fusion parameters as CDG and FCMF.

4.2 Channel Selection

To reduce electrode count while preserving discriminative information, we performed channel-wise one-way ANOVA separately for the audible-initial, audible-final, silent-initial, and silent-final tasks using the mean amplitude within the predefined analysis window. For each channel, the four condition-specific p values were combined using Fisher's method (Fisher, 1925), and channels were ranked according to the combined statistic. We then evaluated the top 5, 10, 15, and 20 channels against the full-channel configuration and compared the fixed 15-channel ranking with fold-wise LOSO selection.

As shown in Table 1, the fixed 15-channel configuration achieved 67.26%, 71.46%, 67.61%, and 73.06% accuracy for audible-initial, audible-final, silent-initial, and silent-final decoding, respectively. Its overall average accuracy was 69.85%, exceeding the full-channel configuration by 1.34%. Increasing the number of electrodes did not further improve overall performance. Therefore, the 15 highest-ranked channels were used in all subsequent experiments to balance decoding performance and electrode efficiency. The ranked channels are listed in Appendix D.

Table 1: Performance across EEG channel counts (%). A.I., A.F., S.I., and S.F. denote audible-initial, audible-final, silent-initial, and silent-final. Ranked channels are listed in Appendix D.

Ch	A.I.	A.F.	S.I.	S.F.	Avg.
5	65.67	69.44	66.73	69.96	67.95
10	65.84	70.48	66.10	71.85	68.57
15	**67.26**	**71.46**	**67.61**	**73.06**	**69.85**
20	65.60	70.83	66.68	71.08	68.55
Full	67.00	70.36	66.77	69.90	68.51

4.3 Baseline Comparisons

We compared NMCNet with eight representative methods: ETE-CNN (Wang et al., 2023), VLAAI (Accou et al., 2023), EEGNet (Lawhern et al., 2018), DeepConvNet (Schirrmeister et al., 2017), GAT (Song et al., 2023), DRDA (Zhao et al., 2021), NeuroFusion-Trans (Sultan et al., 2025), and CAT-Net (Zhuang et al., 2026). CAT-Net and NeuroFusion-Trans were included as dedicated EEG-EMG fusion baselines. NMCNet and all baseline methods were evaluated in a leave-one-subject-out setting, with each baseline implemented according to its standard cross-subject protocol. Accuracy and Cohen's Kappa were used as evaluation metrics.

Overall Cross-Subject Performance Under cross-subject evaluation, Table 2 compares NMCNet with the baseline methods across the four speech conditions. NMCNet achieved the highest accuracy in all conditions, reaching 67.80%, 72.19%, 67.18%, and 73.31% for A.I., A.F., S.I., and S.F., respectively. Compared with the second-best methods, NMCNet improved accuracy by 2.54% and 0.45% in the two initial tasks and by 7.50% and 10.52% in the two final tasks. CAT-Net and

NeuroFusion-Trans, the dedicated EEG-EMG fusion baselines, achieved overall average accuracies of 58.69% and 50.81%, respectively. The larger margins on the final tasks indicate that the proposed representation is particularly effective for cross-subject final decoding while remaining consistently competitive across all four conditions.

Table 2: Cross-subject decoding accuracy across the four speech conditions (%). The best results are shown in bold, and the second-best results are underlined.

Method	A.I.	A.F.	S.I.	S.F.	Avg.	Avg. Kappa
EEGNet	48.73	47.48	48.90	49.08	48.55	0.1970
DeepConvNet	<u>65.26</u>	56.31	<u>66.73</u>	62.67	<u>62.74</u>	<u>0.3836</u>
VLAAI	49.36	55.67	47.03	54.56	51.66	0.1145
DRDA	58.82	<u>64.69</u>	59.03	<u>62.79</u>	61.33	0.3125
ETE-CNN	50.28	55.63	46.38	55.29	51.89	0.1417
GAT	45.97	53.94	44.31	53.13	49.34	0.1199
NeuroFusion-Trans	47.80	54.19	45.95	55.31	50.81	0.1245
CAT-Net	54.50	62.35	55.82	62.06	58.69	0.2425
NMCNet	**67.80**	**72.19**	**67.18**	**73.31**	**70.12**	**0.4812**

Silent-Speech Performance Given the potential of silent-speech decoding for assistive communication, we averaged the silent-initial and silent-final results for each held-out participant. As shown in Table 3, NMCNet achieved the highest overall accuracy of 70.25%, outperforming DeepConvNet by 5.55%. NMCNet ranked first for nine of the ten participants, while CAT-Net achieved the highest accuracy for S1. Across all participants, NMCNet achieved a mean improvement of 4.50% relative to the participant-wise strongest baselines. The largest gains were observed for S10, S9, and S5 at 11.22%, 9.03%, and 7.10%, respectively. These results indicate that the complementary EEG-EMG representation remains effective during silent articulation across most held-out participants rather than being driven by only a few subjects.

Table 3: Subject-wise accuracy averaged across the two silent-speech conditions (%).

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
EEGNet	35.93	47.53	50.00	52.37	48.96	50.48	53.38	49.81	48.50	52.98	48.99
DeepConvNet	49.26	56.13	70.09	61.39	64.75	71.44	74.83	70.07	66.26	62.77	64.70
VLAAI	53.82	47.67	51.38	50.64	51.78	51.64	51.09	51.23	49.77	48.95	50.79
DRDA	50.16	59.29	60.13	63.34	61.86	62.92	64.23	60.91	64.89	61.40	60.91
ETE-CNN	43.54	56.08	31.69	56.03	54.97	51.71	54.04	52.91	55.63	51.77	50.84
GAT	45.74	44.60	48.13	49.36	49.46	50.90	54.71	47.40	49.42	47.48	48.72
NeuroFusion-Trans	45.71	49.99	48.74	49.58	51.40	55.00	54.14	48.46	49.56	53.74	50.63
CAT-Net	**54.68**	54.40	56.46	58.10	57.66	63.72	63.90	60.80	61.16	58.54	58.94
NMCNet	49.65	**63.31**	**72.87**	**69.88**	**71.85**	**74.97**	**75.48**	**75.19**	**75.29**	**73.99**	**70.25**

Fine-Grained Silent-Speech Decoding To examine whether the advantages of NMCNet extend beyond the three broad phonological categories, we compared NMCNet with DeepConvNet and DRDA on six within-category tasks containing 4–7 classes. As shown in Table 4, NMCNet achieved the highest F1 score in all six groups and the highest recall in three groups. It obtained macro-averaged F1 and recall scores of 73.48% and 74.35%, outperforming DeepConvNet by 3.00% and 1.34%, respectively. The largest F1 gains over the strongest baseline were observed for sonorants, nasal codas, and plosives, reaching 5.17%, 4.42%, and 3.72%, respectively. Although DeepConvNet achieved slightly higher recall for fricatives/affricates, monophthongs, and diphthongs, NMCNet obtained the best overall macro F1 and recall. These results show that the learned representation captures discriminative information at both broad-category and fine-grained phonological levels.

4.4 Ablation Study

We assessed statistical significance using two-sided Wilcoxon signed-rank tests over the ten held-out subjects. Table 5 evaluates cross-subject reweighting. Relative to the otherwise identical variant

Table 4: Fine-grained silent-speech decoding performance within six phonological groups (%). F and R denote macro-averaged F1 score and recall, respectively. Macro Avg. is the unweighted mean across the six tasks. Fric./Affr. denotes fricatives/affricates.

Method	Plosives		Fric./Affr.		Sonorants		Monophthongs		Diphthongs		Nasal Codas		Macro Avg.	
	F	R	F	R	F	R	F	R	F	R	F	R	F	R
DeepConvNet	74.56	75.88	65.96	**69.49**	63.52	67.46	56.70	**60.33**	80.65	**82.71**	81.51	82.19	70.48	73.01
DRDA	64.73	65.25	57.32	56.88	61.48	60.95	47.36	46.41	74.86	74.38	81.54	81.95	64.55	64.30
NMCNet	**78.28**	**78.88**	**67.20**	68.24	**68.69**	**69.64**	**58.98**	60.22	**81.77**	82.54	**85.96**	**86.56**	**73.48**	**74.35**

with $\gamma_{cs} = 1$, emphasizing same-class pairs from different subjects significantly improved Cohen's Kappa ($p < 0.05$), supporting the contribution of cross-subject pair reweighting.

Table 5: Ablation of cross-subject reweighting under LOSO evaluation. RW denotes cross-subject reweighting. $^{*}p < 0.05$.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg. Kappa
NMCNet w/o RW	0.1404	0.3535	0.4893	0.4341	0.5229	0.5761	0.5579	0.5699	0.5847	0.5155	0.4744*
NMCNet	**0.1575**	**0.3560**	0.4860	**0.4487**	**0.5268**	0.5741	**0.5584**	**0.5768**	**0.5956**	**0.5325**	**0.4812**

Table 6 examines the fusion components, directional ordering, and input modalities. Removing FCMF caused the largest architectural degradation, reducing average Kappa from 0.4812 to 0.3909, compared with 0.4276 and 0.4329 without MF and CDG, respectively. Reversing EEG-to-EMG-to-EEG to EMG-to-EEG-to-EMG reduced Kappa to 0.4397; the parameter-matched symmetric bidirectional control reduced it to 0.4301. Both control variants significantly reduced Kappa, supporting the proposed asymmetric ordering over reversed or simultaneous bidirectional interaction. The parameter-matched symmetric control further indicates that the improvement is not explained by greater fusion capacity. Removing EEG and EMG reduced Kappa to 0.3316 and 0.2380, respectively, confirming complementary information from both modalities. All component, modality, and fusion-control variants significantly reduced Kappa ($p < 0.01$).

Table 6: Component, directional-control, and modality ablation results under LOSO evaluation. MF denotes the multi-frequency branch. Reverse Fusion reverses the interaction order; Symmetric Fusion is the parameter-matched bidirectional control. Avg. Kappa is averaged across the four conditions. $^{**}p < 0.01$.

Method	MF	CDG	FCMF	EEG	EMG	Avg. Kappa
NMCNet w/o CDG	✓		✓	✓	✓	0.4329**
NMCNet w/o FCMF	✓	✓		✓	✓	0.3909**
NMCNet w/o MF		✓	✓	✓	✓	0.4276**
NMCNet w/o EEG	✓				✓	0.3316**
NMCNet w/o EMG	✓			✓		0.2380**
NMCNet w/ Reverse Fusion	✓	✓	✓	✓	✓	0.4397**
NMCNet w/ Symmetric Fusion	✓			✓	✓	0.4301**
NMCNet	✓	✓	✓	✓	✓	**0.4812**

5 Conclusion

We proposed NMCNet, a cortex-guided EEG-EMG framework with asymmetric fusion for cross-subject decoding of Mandarin initials and finals. Using 15 EEG channels, NMCNet achieved 70.25% mean accuracy across the two silent-speech tasks under LOSO; fine-grained decoding reached macro-averaged F1 and recall scores of 73.48% and 74.35%. These results support broad-category and fine-grained phonological decoding and motivate evaluation in larger, more diverse clinical populations across sessions and real-world settings.

AI USE STATEMENT

Generative AI tools were used to aid or polish the writing of this manuscript and assist with statistical analysis. All AI-assisted text and statistical analyses were reviewed and verified by the authors for accuracy, clarity, and consistency. The authors take full responsibility for the final content.

ETHICS STATEMENT

The study involved human participants and was approved by the institutional ethics committee. All experiments were conducted in accordance with the Declaration of Helsinki (World Medical Association, 2013).

REPRODUCIBILITY STATEMENT

The model architecture, preprocessing pipeline, training objectives, evaluation protocol, and principal experimental settings are described in the main paper. Complete hyperparameter settings, EEG and EMG channel information, channel rankings, additional subject-wise results, and modality attribution analyses are provided in the appendix.

REFERENCES

Bernd Accou, Jonas Vanthornhout, Hugo Van Hamme, and Tom Francart. Decoding of the speech envelope from EEG using the VLAAI deep neural network. *Scientific Reports*, 13:812, 2023.

S. J. Barrientos Rojas, R. Ramirez-Valencia, D. Alonso-Vazquez, R. Caraza, H. R. Martinez, O. Mendoza-Montoya, and J. M. Antelis. Recognition of grammatical classes of overt speech using electrophysiological signals and machine learning. In *Proceedings of the 2022 IEEE 4th International Conference on BioInspired Processing (BIP)*, pp. 1–6, 2022.

T. Y.-H. Chen, Y. Chen, P. Soederhaell, S. Agrawal, and K. Shapovalenko. Decoding EEG speech perception with transformers and VAE-based data augmentation. *arXiv preprint arXiv:2501.04359*, 2025.

C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike. Single-trial classification of vowel speech imagery using common spatial patterns. *Neural Networks*, 22(9):1334–1339, 2009.

D. Dash, P. Ferrari, and J. Wang. Role of brainwaves in neural speech decoding. In *Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 1357–1361, 2021.

R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

J. Gu, J. Jiang, Q. Cai, and H. Wang. Imagined chinese speech decoding based on initials and finals from EEG activity. *IET Signal Processing*, 2026(1):5451362, 2026.

F. H. Guenther. Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39(5):350–365, 2006.

J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

M. Janke and L. Diener. EMG-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2375–2385, 2017.

P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

H. Kim, J. Kim, and S. B. Kim. Cross-subject generalizable representation learning with class-subject dual labels for biosignals. *Knowledge-Based Systems*, 295:111855, 2024.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

Y.-E. Lee and S.-H. Lee. EEG-transformer: Self-attention from transformer architecture for decoding EEG of imagined speech. In *Proceedings of the 2022 10th International Winter Conference on Brain-Computer Interface (BCI)*, pp. 1–4, 2022.

H. Li, M. Wang, H. Gao, S. Zhao, G. Li, and Y. Wang. Hybrid silent speech interface through fusion of electroencephalography and electromyography. In *Proceedings of Interspeech 2023*, pp. 1184–1188, 2023.

M. Li, S. H. Pun, and F. Chen. Mandarin tone classification in spoken speech with EEG signals. In *Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 1163–1166, 2021.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774, 2017.

Y. Qian, C. Liu, P. Yu, X. Ran, S. Li, Q. Yang, Y. Liu, L. Xia, Y. Wang, J. Qi, E. Zhou, J. Lu, Y. Li, T. H. Tao, Z. Zhou, and J. Wu. Real-time decoding of full-spectrum chinese using brain-computer interface. *Science Advances*, 11(45):eadz9968, 2025.

M. N. I. Qureshi, B. Min, H. J. Park, D. Cho, W. Choi, and B. Lee. Multiclass classification of word imagination speech with hybrid connectivity features. *IEEE Transactions on Biomedical Engineering*, 65(10):2168–2177, 2018.

R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

E. W. Sellers, D. B. Ryan, and C. K. Hauser. Noninvasive brain-computer interface enables communication after brainstem stroke. *Science Translational Medicine*, 6(257):257re7, 2014.

R. A. Sharon, S. S. Narayanan, M. Sur, and H. A. Murthy. Neural speech decoding during audition, imagination and production. *IEEE Access*, 8:149714–149729, 2020.

Y. Song, Q. Zheng, Q. Wang, X. Gao, and P.-A. Heng. Global adaptive transformer for cross-subject enhanced EEG classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:2767–2777, 2023.

T. Sultan, G. Liu, P. Sikorski, S. Alshathri, W. El-Shafai, and M. Babaiasl. NeuroFusion-Trans: A novel transformer-based EEG-EMG fusion model for assistive robotics. *IEEE Access*, 13: 103245–103263, 2025.

C. Torrence and G. P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, 1998.

D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

X. Wang, M. Li, H. Li, S. H. Pun, and F. Chen. Cross-subject classification of spoken mandarin vowels and tones with EEG signals: A study of end-to-end CNN with fine-tuning. In *Proceedings of the 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 535–539, 2023.

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, and J. M. Henderson. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

World Medical Association. World medical association declaration of helsinki: Ethical principles for medical research involving human subjects. *JAMA*, 310(20):2191–2194, 2013.

J. Wu, Y. Zhang, L. Xie, Y. Yan, X. Zhang, S. Liu, X. An, E. Yin, and D. Ming. A novel silent speech recognition approach based on parallel inception convolutional neural network and mel frequency spectral coefficient. *Frontiers in Neurorobotics*, 16:971446, 2022.

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations*, 2018.

H. Zhao, Q. Zheng, K. Ma, H. Li, and Y. Zheng. Deep representation-based domain adaptation for nonstationary EEG classification. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):535–545, 2021.

S. Zhao and F. Rudzicz. Classifying phonological categories in imagined and articulated speech. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 992–996, 2015.

Y. Zhu, F. Zhuang, and D. Wang. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5989–5996, 2019.

Y. Zhuang, C. Huang, Z. Yu, Y. Zou, and J. Ju. CAT-Net: A cross-attention tone network for cross-subject EEG-EMG fusion tone decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18090–18098, 2026.

A HYPERPARAMETER SETTINGS

Table A.1: Hyperparameter settings used for all NMCNet experiments. A.I., A.F., S.I., and S.F. denote audible-initial, audible-final, silent-initial, and silent-final, respectively.

Hyperparameter	Value
Optimizer	Adam
Learning Rate	1×10^{-3}
Loss Function	Weighted Focal Loss + Cross-Subject-Enhanced SupCon
Focal Loss Parameter	$\gamma = 2.0$
Contrastive Loss Weight	$\lambda_{\text{con}} = 0.2$
Contrastive Temperature	$\tau = 0.1$
Cross-Subject Pair Weight	1.35 / 1.75 / 1.75 / 1.40 for A.I. / A.F. / S.I. / S.F.
Batch Size	64
Dropout Rates	Encoders: 0.5; FCMF: 0.1; MF projection: 0.5; classifier: 0.6
Epochs	500
Early Stopping Patience	50
LR Scheduler	ReduceLROnPlateau (factor=0.5, patience=25)
Mixup Parameter	$\alpha = 0.4$
Gaussian Noise	std=0.05, probability=0.5
EEG/EMG Channels Used	15 and 5
Random Seed	42

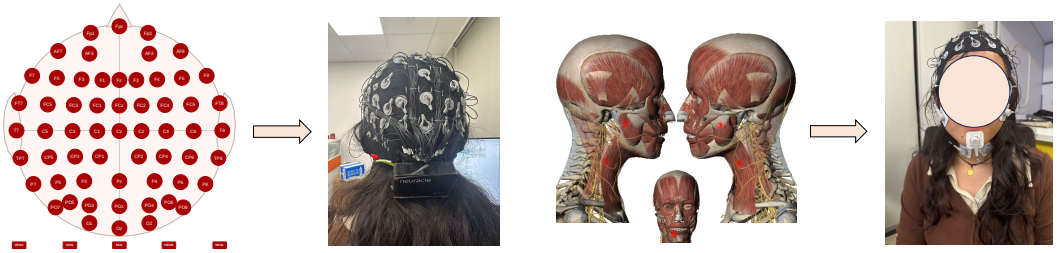
Algorithm 1 Single training iteration of NMCNet

Require: Augmented EEG and EMG mini-batch $\widetilde{X}_{EEG}, \widetilde{X}_{EMG}$, soft class labels $\widetilde{Y}$, soft subject labels $\widetilde{S}$

$f_{EEG}^{TS} \leftarrow E_{EEG}^{TS}(\widetilde{X}_{EEG})$
$f_{EMG}^{TS} \leftarrow E_{EMG}^{TS}(\widetilde{X}_{EMG})$
$f_{EMG}^{gated} \leftarrow G_{CDG}(f_{EEG}^{TS}, f_{EMG}^{TS})$
$f_{EEG}^{fused} \leftarrow F_{FCMF}(f_{EEG}^{TS}, f_{EMG}^{gated})$
$f_{TS} \leftarrow \text{Concat}(f_{EEG}^{fused}, f_{EMG}^{gated})$
$f_{MF} \leftarrow G_{SE}(\text{Concat}(E_{EEG}^{MF}(\widetilde{X}_{EEG}), E_{EMG}^{MF}(\widetilde{X}_{EMG})))$
$f_{joint} \leftarrow \text{Concat}(f_{TS}, f_{MF})$
$\widehat{Y} \leftarrow C(f_{joint})$
$\mathcal{L}_{focal} \leftarrow \text{WeightedFocalLoss}(\widehat{Y}, \widetilde{Y})$
$\mathcal{L}_{con} \leftarrow \text{CrossSubjectSupCon}(f_{joint}, \widetilde{Y}, \widetilde{S}, \tau, \gamma_{cs})$
$\mathcal{L} \leftarrow \mathcal{L}_{focal} + \lambda_{con}\mathcal{L}_{con}$
Back-propagate $\mathcal{L}$ and update parameters with Adam
return $\mathcal{L}$

B EEG AND EMG CHANNEL PLACEMENTS

We collected synchronized EEG and EMG signals from all participants. EEG was recorded using a high-density cap, while EMG was acquired from five muscles involved in orofacial and cervical movements. The corresponding sensor locations are summarized in Fig. B.1.



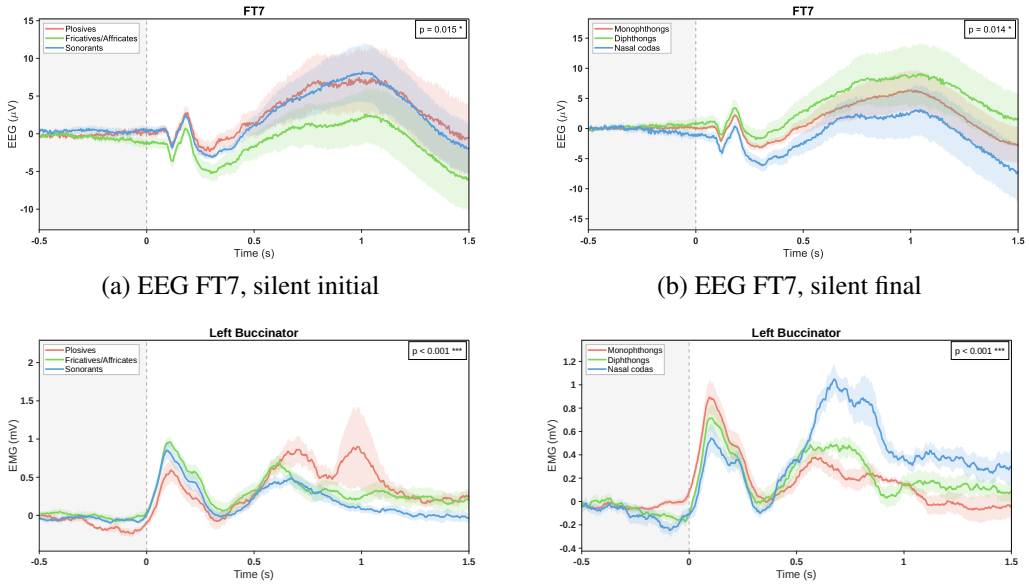
(a) EEG electrode placement (b) EMG sensor placement

Figure B.1: EEG and EMG sensor placements. EMG sensors were positioned over (1) right buccinator, (2) right cervical trapezius, (3) left buccinator, (4) left cervical trapezius, and (5) mentalis.

C Feature Visualization

C.1 Visualization of Temporal Features

To examine temporal characteristics during silent Mandarin speech production, we present representative EEG and EMG responses for both initial and final decoding. The two modalities show category-dependent yet different temporal profiles. These observations indicate that cortical and muscular signals capture complementary aspects of articulation, providing empirical support for their joint modeling and fusion within the NMCNet framework.



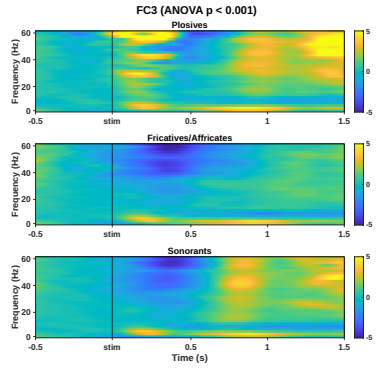
(a) EEG FT7, silent initial (b) EEG FT7, silent final

(c) EMG Left Buccinator, silent initial (d) EMG Left Buccinator, silent final

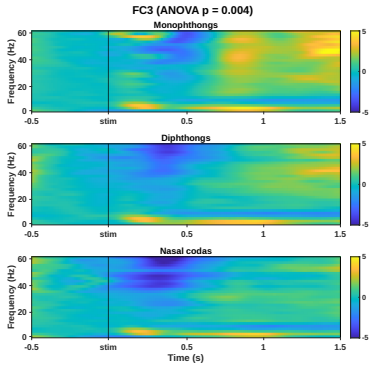
Figure C.1: Representative temporal features used to examine modality- and category-dependent response patterns during silent speech. All four panels are shown at a consistent display scale.

C.2 Visualization of Time-Frequency Features

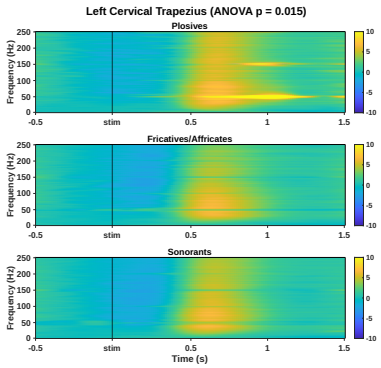
To examine spectral dynamics during silent articulation, we visualize baseline-corrected time-frequency representations from representative EEG and EMG channels in initial and final tasks. The modalities display distinct temporal-spectral patterns across phonological categories, indicating complementary frequency-specific information and motivating multi-frequency representation learning in NMCNet.



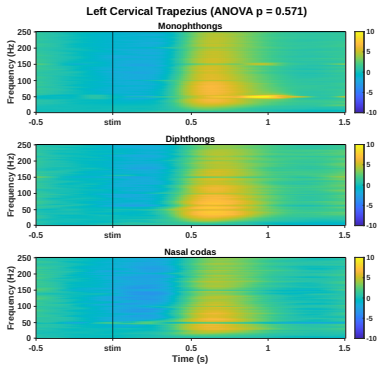
(a) EEG FC3, silent initial



(b) EEG FC3, silent final



(c) EMG Left Cervical Trapezius, silent initial



(d) EMG Left Cervical Trapezius, silent final

Figure C.2: Representative time-frequency features during silent speech. All four panels use a consistent display box to avoid size differences caused by source aspect ratios.

D Fisher-Based EEG Channel Selection

Table D.1: EEG channels ranked by $-\log_{10}(p)$, where p is the Fisher-combined value across the four decoding conditions (Fisher, 1925).

Ch	$-\log_{10}(p)$	Ch	$-\log_{10}(p)$	Ch	$-\log_{10}(p)$
Fp1	4.10	FC5	1.66	Fz	0.96
Pz	3.79	F4	1.63	FC3	0.92
C6	3.26	Oz	1.62	F2	0.89
PO7	3.24	C4	1.54	FC1	0.87
AF7	3.15	Cz	1.52	F3	0.83
PO8	2.78	AF8	1.39	FCz	0.80
PO6	2.68	C3	1.37	CP6	0.78
F5	2.56	P3	1.36	FC6	0.73
PO3	2.53	P4	1.35	FC2	0.67
TP8	2.52	POz	1.28	CP2	0.54
FT7	2.46	O1	1.29	P6	0.51
F7	2.37	AF4	1.21	P5	0.51
CP4	2.27	C1	1.18	CP5	0.50
CP3	2.26	FC4	1.08	F6	0.46
T7	2.17	F8	1.07	PO5	0.37
Fp2	2.14	P7	1.03	T8	0.13
C2	2.13	P8	1.01	TP7	0.13
CP1	2.11	O2	1.00	FT8	0.12
PO4	1.95	Fpz	0.99	C5	0.01
AF3	1.68	F1	0.99		

E BASELINE METHODS

ETE-CNN (Wang et al., 2023). An end-to-end convolutional architecture that extracts temporal representations directly from physiological signals for cross-subject speech decoding.

VLAAI (Accou et al., 2023). A modular neural architecture composed of repeated convolutional blocks with output-context modeling and skip connections for decoding speech envelopes from EEG.

EEGNet (Lawhern et al., 2018). A compact network combining temporal convolution, depthwise spatial filtering, and separable convolution for efficient EEG representation learning.

DeepConvNet (Schirrmeister et al., 2017). A deep convolutional architecture that progressively extracts hierarchical spatial-temporal representations through stacked convolution and pooling blocks.

GAT (Song et al., 2023). A transformer-based architecture that models spatial and temporal dependencies through adaptive attention mechanisms.

DRDA (Zhao et al., 2021). A domain-adaptation framework that jointly optimizes a feature extractor, classifier, and domain discriminator through adversarial learning, with center loss used to promote invariant representations.

NeuroFusion-Trans (Sultan et al., 2025). A Transformer-based EEG-EMG fusion framework with modality-specific encoders, learned cross-modality fusion, temporal synchronization, and online user adaptation.

CAT-Net (Zhuang et al., 2026). A cross-attention EEG-EMG fusion architecture that independently encodes the two modalities and models their bidirectional interactions for cross-subject Mandarin speech decoding.

F SUBJECT-WISE CROSS-SUBJECT RESULTS

Table F.1: Subject-wise accuracy averaged across the four speech conditions (%).

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
EEGNet	36.40	47.82	46.92	50.08	49.88	50.69	53.97	49.86	48.50	51.36	48.55
DeepConvNet	49.59	54.93	63.47	61.52	63.55	67.80	71.07	67.84	65.92	61.72	62.74
VLAAI	51.92	50.85	51.64	50.71	51.50	51.75	53.07	53.82	50.29	50.97	51.65
DRDA	53.66	58.12	60.23	63.85	62.19	63.65	64.08	62.04	64.14	61.35	61.33
ETE-CNN	45.84	53.66	45.05	55.97	54.68	49.35	55.01	54.85	51.61	52.93	51.89
GAT	49.28	44.31	46.39	51.18	49.09	51.38	54.41	47.41	50.44	49.49	49.34
NeuroFusion-Trans	46.03	49.28	47.82	49.74	53.10	52.40	54.38	51.19	50.93	53.24	50.81
CAT-Net	54.44	55.07	56.21	58.26	59.83	62.28	61.90	60.35	60.89	57.62	58.69
NMCNet	52.80	62.58	70.86	68.36	72.19	75.08	74.64	75.45	76.34	72.89	70.12

Table F.2: Subject-wise accuracy under the silent-initial condition (%).

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
EEGNet	36.85	51.51	49.57	48.06	50.00	47.41	53.23	51.29	47.41	53.66	48.90
DeepConvNet	43.10	67.46	67.46	64.44	65.95	73.71	76.94	73.06	69.18	65.95	66.72
VLAAI	47.63	47.84	45.47	44.61	43.97	48.06	46.34	48.71	48.49	49.14	47.03
DRDA	46.98	57.11	57.76	62.72	61.42	62.50	60.13	60.99	61.64	59.05	59.03
ETE-CNN	37.50	52.37	38.79	49.35	48.06	45.91	48.28	49.78	50.00	43.75	46.38
GAT	37.72	44.61	43.75	42.89	42.24	47.63	51.29	43.97	46.34	42.67	44.31
NeuroFusion-Trans	42.67	43.32	42.89	43.32	46.34	50.00	48.49	44.83	48.49	49.14	45.95
CAT-Net	49.14	51.51	56.25	55.17	53.02	60.78	58.84	60.13	57.11	56.25	55.82
NMCNet	41.38	61.21	71.77	67.46	66.81	73.06	72.63	73.71	73.28	70.47	67.18

Table F.3: Subject-wise accuracy under the silent-final condition (%).

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
EEGNet	35.00	43.54	50.42	56.67	47.92	52.71	53.54	48.33	49.58	52.29	49.08
DeepConvNet	55.42	44.79	72.71	58.33	63.54	69.17	72.71	67.08	63.33	59.58	62.67
VLAAI	60.00	47.50	57.29	56.67	59.58	55.21	55.83	53.75	51.04	48.75	54.56
DRDA	53.33	61.46	62.50	63.96	62.29	63.33	68.33	60.83	68.13	63.75	62.79
ETE-CNN	49.58	59.79	24.58	62.71	61.88	57.50	59.79	56.04	61.25	59.79	55.29
GAT	53.75	44.58	52.50	55.83	56.67	54.17	58.13	50.83	52.50	52.29	53.13
NeuroFusion-Trans	48.75	56.67	54.58	55.83	56.46	60.00	59.79	52.08	50.62	58.33	55.31
CAT-Net	60.21	57.29	56.67	61.04	62.29	66.67	68.96	61.46	65.21	60.83	62.06
NMCNet	57.92	65.42	73.96	72.29	76.88	76.88	78.33	76.67	77.29	77.50	73.31

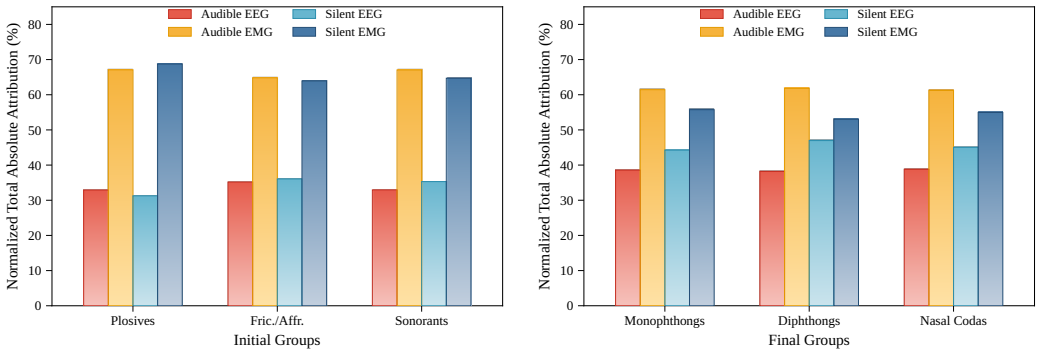
Overall Results Table F.1 reports the subject-wise accuracy averaged across the four decoding conditions. NMCNet achieved the highest average accuracy of 70.12%, outperforming the strongest baseline, DeepConvNet, by 7.38%. Except for S1, where NMCNet was 1.64% lower than CAT-Net, it outperformed the strongest subject-specific baseline on the remaining nine subjects by an average of 7.23%. Particularly large improvements were observed for S10 and S9, reaching 11.17% and 10.42%, respectively. These results demonstrate consistent cross-subject generalization across participants.

Silent-Initial Results As shown in Table F.2, NMCNet achieved an average accuracy of 67.18%, exceeding DeepConvNet by 0.46%. It obtained the highest accuracy on six subjects, namely S3, S4, S5, S8, S9, and S10, but performed below the strongest baseline on S1, S2, S6, and S7. The largest improvements were observed for S10, S3, and S9, reaching 4.52%, 4.31%, and 4.10%, respectively. Overall, the advantage of NMCNet in silent-initial decoding was relatively modest and varied across participants.

Silent-Final Results A substantially larger improvement was observed under the silent-final condition. As shown in Table F.3, NMCNet achieved an average accuracy of 73.31%, exceeding the strongest baseline, DRDA, by 10.52%. It obtained the highest accuracy on nine of the ten subjects. The only exception was S1, where NMCNet was 2.29% lower than CAT-Net. Across S2 to S10, NMCNet surpassed the strongest subject-specific baseline by an average of 8.08%. Particularly large improvements were observed for S10 and S5, reaching 13.75% and 13.34%, respectively, while gains above 9% were also obtained for S8 and S9. Moreover, NMCNet achieved accuracies above 70% for all subjects from S3 to S10.

Compared with the relatively small average gain in silent-initial decoding, the broader and more consistent improvements in silent-final decoding indicate that the benefit of NMCNet is more pronounced for the final task. Mandarin finals involve sustained cues related to vowel quality and nasalization, which may allow the temporal-spatial and multi-frequency branches to capture more stable temporal, spectral, and neuromuscular patterns across subjects. The gains observed across nine participants suggest that its complementary EEG-EMG representation is particularly effective for cross-subject silent-final decoding, rather than the overall improvement being driven by only a small number of subjects.

G SHAP-Based Modality Attribution



(a) Initial groups (b) Final groups

Figure G.1: SHAP attribution of EEG and EMG across Mandarin (a) initial and (b) final groups.

We used SHAP (Lundberg & Lee, 2017) to quantify the relative attribution of EEG and EMG across the three initial groups and three final groups under audible and silent speech. For each sample, absolute SHAP values were summed over all channels and time points within each modality and normalized by the combined EEG–EMG attribution. The resulting ratios were averaged across the ten LOSO folds and represent model attribution rather than causal contributions to decoding performance. As shown in Fig. G.1(a), EMG accounted for 63.94%–68.74% of the total attribution across the initial groups, while EEG accounted for 31.26%–36.06%. Fig. G.1(b) shows a more balanced pattern for final decoding. EEG attribution ranged from 38.18% to 38.78% under audible speech and increased to 44.25%–46.97% under silent speech.

This pattern is consistent with the asymmetric roles of the two modalities in NMCNet. CDG uses EEG as the conditioning signal to gate EMG features, while retaining the gated EMG representation as execution-related evidence. FCMF then converts this representation into a feedback message that updates EEG before the two features are combined. Thus, the larger EMG attribution reflects the strong class information carried by muscle responses, whereas the substantial EEG attribution reflects its role in guiding and refining cross-modal fusion. The more balanced attribution in silent-final decoding suggests that NMCNet relies more evenly on cortical and muscular cues in this task.