FAST-CAD: Fairness-Aware Self-Supervised Framework for Automated Non-Contact Stroke Diagnosis

Anonymous submission

Abstract

Stroke is an acute cerebrovascular disease, and timely diagnosis significantly improves patient survival. Unfortunately, due to the drawbacks of previous contact and non-contact diagnosis methods, the existing methods have significant limitations in daily applications. Timely diagnosis for stroke patients in daily life is still an open problem, although numerous studies have been performed on automated contact diagnosis of stroke. Therefore, we have turned our attention to automated non-contact stroke diagnosis to explore more feasible methods applicable in daily life. In this work, we constructed a non-contact stroke diagnosis dataset and propose a novel multimodal method inspired by the primary focus areas of non-contact stroke diagnosis. This method selects and integrates specific self-supervised learning models with representation models to pay attention to the key aspects of the diagnosis task. The validation on our collected dataset demonstrates that our method surpasses the currently known baselines in both diagnostic performance and generalization ability. Our work not only provides a benchmark for automated non-contact stroke diagnosis tasks but also offers a highly promising framework for medical automated diagnosis in daily life.

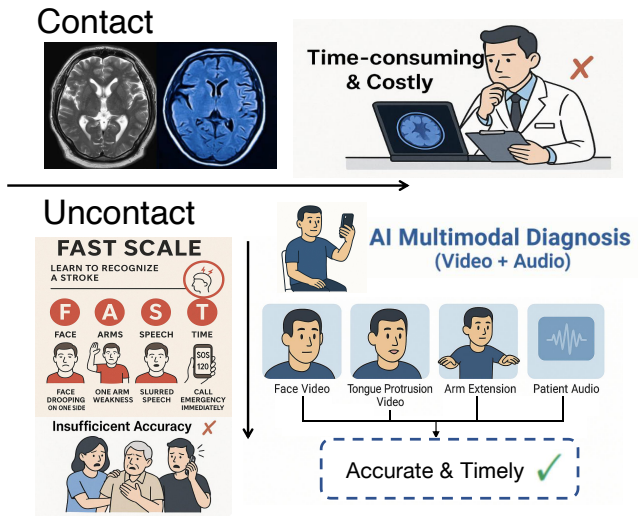

Figure 1: Stroke assessment evolution—from slow, contact-dependent imaging and low-accuracy FAST checks to our novel non-contact method.

Introduction

Stroke is an acute cerebrovascular disease, ranking as the second leading cause of death and the third leading cause of disability worldwide (Johnson et al. 2016). Although patient survival rates have improved in recent years, the incidence of permanent disability remains high, primarily because timely diagnosis within the stroke "golden window" are often not achieved and because emergency care and rehabilitation knowledge are insufficiently disseminated at the community level.

Current stroke diagnosis methods can be categorized into contact and non-contact approaches. Contact methods focus on MRI and CT scan results and are only available in hospitals or specialized medical institutions; they require patients to attend in person and depend on time-consuming, complex procedural workflows. As a result, untimely diagnosis distances patients from the stroke "golden window", compromising the efficacy of subsequent therapeutic interventions. Non-contact methods (such as the FAST scale (Mohd Nor et al. 2004)), while obviating the need for complex workflows and in-person attendance, require emergency care and rehabilitation knowledge that is insufficiently disseminated at the community level and exhibit reduced accuracy when applied by non-medical individuals, thereby limiting their widespread applicability.

In order to overcome the aforementioned limitations, researchers have turned their attention to deep learning technology, which has shown immense potential in certain fields. Current deep learning methods are mainly applied to contact-based stroke diagnosis, aiming to eliminate the manual process of interpreting MRI and CT scan results, thus achieving automated contact-based stroke diagnosis.

However, do these methods address the critical issue of timely diagnosis for stroke patients?

These methods focus on optimizing the extraction of stroke-related information from MRI and CT scans to enable automated contact-based stroke diagnosis. However, they address only a small segment of the contact-based stroke "golden window". By the time patients undergo hospital-based imaging, the utility of automated diagnosis is greatly diminished.

Apparently, these methods fail to address the critical is-

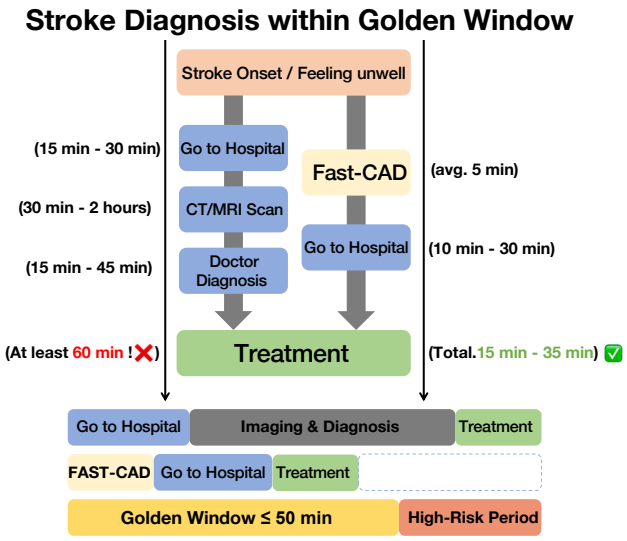

Figure 2: Comparison of the conventional versus FAST-CAD–assisted stroke workflow.

sue of timely diagnosis in stroke patients. Hence, we have turned our attention to automated non-contact stroke diagnosis, aiming to truly optimize diagnosis within the stroke "golden window".

Non-contact medical diagnosis is constrained by stringent privacy, safety, and ethical considerations, as well as concerns about potential harm during data collection. Consequently, there is an acute shortage of high-quality datasets and a lack of innovative data compositions, which has impeded progress in this domain. To overcome these challenges, we have constructed a comprehensive audiovisual dataset tailored for non-contact stroke diagnosis. This dataset not only addresses the critical scarcity of resources but also incorporates multimodal patient data—video and audio recordings captured across diverse postures—thus laying a robust foundation for the development and validation of non-contact diagnostic methods within community and personal contexts.

However, the dataset we have collected differs markedly from previously available public or internal datasets in terms of sample size, modalities, and acquisition methods. Methods developed on large-scale MRI/CT data perform only marginally on our dataset's limited sample size; likewise, conventional video action-classification models fail to guarantee sufficient generalizability or fairness. Therefore, devising approaches that ensure model generalizability, effectiveness, and fairness on our dataset has become a critical and urgent challenge.

We propose a novel framework, FAST-CAD, to sequentially address three principal challenges: (i) the limited size of available datasets, which impedes training from scratch—FAST-CAD's modality module mitigates this issue by leveraging pre-trained self-supervised encoders (SeCo (Yao et al. 2021) for video and a CNN–Transformer–based HuBERT (Hsu et al. 2021) for audio), both pretrained on large-scale general audio–visual corpora and guided by the FAST scale's focus on movement symmetry and speech fluency to extract stroke-relevant features without overfitting to scarce task-specific samples; (ii) the difficulty of aligning interleaved audio and video modalities, which we overcome using an Alternating Dual-Stream Transformer Fusion (ADSTF) architecture, wherein the audio and video streams alternately "query" one another's high-attention regions and dynamically weight their features to preserve unimodal characteristics while reinforcing cross-modal synergy; and (iii) the heterogeneity of data collection methods and substantial environmental interference, which we counteract at the classification stage through adversarial discriminator perturbation—to enforce learning of features invariant to lighting conditions, background variations, and other perturbations—and keypoint-based data augmentation (MMPose Contributors 2020)—to simulate diverse poses and joint configurations, thereby expanding the training distribution and improving generalization.

We rigorously evaluated FAST-CAD against reproduced non-contact stroke–diagnosis baselines under identical conditions. Across key metrics (AUC, accuracy, F1), FAST-CAD consistently surpassed all competitors. To test generalization, we assembled an external cohort of 18 patients recorded under entirely different conditions; our framework maintained its performance advantage on this independent set. Comprehensive ablation studies demonstrated both the flexible fusion of modalities and the unique contribution of each branch. Finally, a worst-group risk–based fairness analysis—stratifying by age, gender, and posture—confirmed that diagnostic accuracy remained equitable across all subpopulations.

In summary, our contributions are twofold: (i) we introduce a comprehensive audiovisual dataset for automated non-contact stroke diagnosis, substantially expanding patient coverage and modality diversity beyond existing benchmarks to address the critical shortage of high-fidelity multimodal data; and (ii) we develop a streamlined three-stage pipeline comprising interchangeable video and audio feature-extraction encoders to produce high-fidelity embeddings, an Alternating Dual-Stream Transformer Fusion module that alternately swaps modality roles within self- and cross-attention layers (reinforced by an adversarial discriminator for robust, overfitting-resistant feature learning), and keypoint-based data augmentation (MMPose Contributors 2020) to simulate diverse poses and joint configurations—thereby broadening training diversity and enhancing generalization and fairness.

Related Works

Medical Datasets

Currently, the number of non-contact medical datasets is relatively limited and primarily focus on a few clinical conditions. Most publicly available datasets for visual or audio-based diagnosis primarily target neurological and psychiatric disorders, such as epilepsy, Parkinson's disease (PD), depression, and sleep apnea. For example, no large-scale public dataset for the assessment of Parkinson's dis-

ease (Zhou et al. 2023) has been established; previous video datasets used for PD were small in scale and collected exclusively in clinical settings (typically not released publicly due to patient privacy concerns), thereby exposing deficiencies in data openness and sharing. In the field of depression detection, widely used datasets such as AVEC2013 (Valstar et al. 2013) and DAIC-WOZ (Ringeval et al. 2019) contain only approximately 100 to 300 subject recordings, reflecting the challenges in organizing and sharing sensitive clinical interview data. Similarly, datasets for epilepsy (video-based seizure monitoring) (Guttag 2010) and sleep disorders (audio-based sleep apnea detection) (Korompili et al. 2021)—for instance, an open sleep apnea dataset containing 212 synchronized audio–polysomnography recordings—suffer from limitations in both scale and coverage, and tend to provide only a single modality or lack full synchronization across multiple modalities, thereby restricting the capture of complementary information.

Stroke diagnosis

In stroke diagnosis, hospital-based imaging modalities such as CT (Sanelli et al. 2014) and diffusion-weighted MRI provide rapid detection of hemorrhagic and early ischemic changes but remain constrained by their reliance on clinical settings during the critical "golden window." Non-contact screening methods like the FAST score (Mohd Nor et al. 2004), which assess facial droop, arm weakness, and speech difficulty, enable pre-hospital evaluation yet suffer from limited diagnostic accuracy. To address these limitations, recent studies have turned to deep learning approaches (Inamdar et al. 2021), demonstrating significant potential to accelerate and enhance stroke detection.

Current deep learning methods are primarily applied to contact-based stroke diagnosis, enabling automated analysis of detailed brain images. (Shinohara et al. 2019) proposed a deep convolutional neural network (DCNN) based on CT images, employing 36 feature extraction layers for lesion segmentation. (Lisowska et al. 2017) introduced a model using MRI data and basic clinical information to predict regions of insufficient blood supply. However, while these methods can accelerate expert diagnosis, their substitution degree and practical value are limited, and they do not truly resolve the critical issue of the golden time for stroke diagnosis.

Given the aforementioned limitations, some researchers have explored automatic non-contact stroke diagnosis methods. (Cai et al. 2022) ,M3Stroke(Cai et al. 2024) developed a method to detect stroke occurrences by analyzing video and audio information of potential patients, and reported a detection accuracy as high as 80.85%. Although this level of performance is promising, the approach requires the patient to maintain a strictly controlled posture during image acquisition, which limits its adaptability to the varied poses encountered in real-world applications. Consequently, despite the encouraging diagnostic accuracy, the practical feasibility of this method remains subject to further validation.

Self-supervised learning

In recent years, *self-supervised learning* (SSL) has become a driving force behind representation learning on small-sample or weakly-labelled datasets. In computer vision, contrastive frameworks such as MoCo (He et al. 2020), SimCLR (Chen et al. 2020), and BYOL (Grill et al. 2020) have demonstrated that large, dynamically maintained dictionaries or carefully designed positive–negative pairings enable models to acquire transferable visual features without human annotation. Building on MoCo, SeCo (Yao et al. 2021) extends this paradigm to video by minimising the InfoNCE loss (van den Oord, Li, and Vinyals 2018) over three complementary subtasks, thereby capturing fine-grained spatio-temporal dynamics. Furthermore, VideoMAE (Tong et al. 2022) employs a masked autoencoder architecture tailored for video, randomly masking a high proportion of spatio-temporal cubes and reconstructing them to learn data-efficient video representations.

In natural language processing, BERT (Devlin et al. 2019) generalised the SSL philosophy through masked-language modelling, learning powerful bidirectional token representations that underpin most contemporary NLP systems. Analogously in speech, wav2vec 2.0 (Baevski et al. 2020) pre-trains a Transformer encoder on masked raw audio with a contrastive loss to learn robust speech representations, achieving state-of-the-art results on ASR benchmarks, and HuBERT (Hsu et al. 2021) adopts a mask-and-predict strategy over latent acoustic units, enabling the network to discover rich phonetic structures without transcriptions.

Dataset

Challenge

Constructing an audio-visual dataset for stroke patient diagnosis meeting clinical standards involves several critical challenges: (i) strict compliance with medical ethics and privacy regulations; (ii) ensuring patient safety throughout data collection; (iii) securing patient informed consent and providing a safe and comfortable collection environment; and (iv) thorough anonymization of all collected data, strictly limiting use to research purposes.

Data Collection

To ensure data reliability and validity, standardized data collection protocols and operating procedures were established. High-resolution cameras and sensitive microphones were used to non-invasively record facial expressions, tongue protrusion, arm extension, and speech data.

Specifically, facial movements were recorded from frontal and lateral views to capture subtle muscle activity. Tongue protrusion videos focused on detailed tongue movements essential for accurate diagnostic analysis, while arm extension videos assessed limb coordination and movement stability. Audio data were collected in low-noise environments, with participants performing specific speech tasks to ensure clarity and accuracy.

The entire data collection process was supervised by medical professionals, ensuring compliance with medical ethics

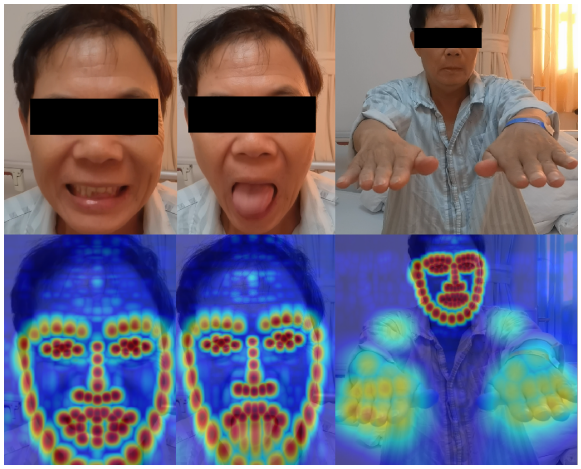

Figure 3: Patient video information in the dataset: Facial video (top left), tongue video (top center), arms video (top right). The lower half shows keypoint heatmap videos corresponding to the respective video information.

and privacy protection standards. Data were anonymized immediately after collection, removing all identifiable information and retaining only features relevant for diagnostic analysis.

Subjects, Labeling, and Data Splitting

The non-contact stroke diagnostic dataset included 243 subjects across diverse gender and age groups, capturing video data of facial expressions, tongue protrusion, arm extension, and speech.

To objectively evaluate the model's ability to diagnose strokes through feature extraction and integration rather than data memorization, the dataset was divided into training and test sets at a 4:1 ratio, with a 5-fold cross-validation method employed within the training set for model training and evaluation.

Table 1: Distribution of Participants by Category, Subcategory, and Count

Category	Subcategory	Count
Age (%)	< 35	65 (26.7%)
	35 – 60	96 (39.5%)
	> 60	82 (33.7%)
Gender (%)	Male	145 (59.7%)
	Female	98 (40.3%)
Posture (%)	Sitting	149 (61.3%)
	Sleeping	94 (38.7%)

Comparison with Existing Datasets

Compared to existing non-contact diagnostic datasets, our dataset demonstrates distinct advantages: (1) richer modalities, integrating facial expressions, tongue movements, arm movements, and speech; (2) larger scale, including 243 subjects, significantly surpassing previous datasets (approximately 151 subjects); (3) greater convenience, economy, and safety compared to traditional contact-based diagnostics (e.g., CT, MRI), making it suitable for rapid, large-scale diagnostic applications.

Overall, our dataset offers significant improvements in scale, modality diversity, and data collection safety, effectively supporting AI-based early stroke diagnosis research.

Methodology

Problem and Challenges

Non-contact stroke diagnosis utilizes multi-modal data—visual signals $x_i \in \mathcal{X}$ and auditory signals $w_i \in \mathcal{W}$—to assess stroke risk via a binary classification framework. For each subject i, our model outputs a probability vector

$$z_i = \begin{bmatrix} z_{i,1} \\ z_{i,2} \end{bmatrix}, \quad z_{i,1} + z_{i,2} = 1, \quad z_{i,1},\, z_{i,2} \in [0,1], \quad (1)$$

where $z_{i,1}$ denotes the probability of a normal condition and $z_{i,2}$ the probability of a stroke condition. The predicted label is then obtained by

$$\hat{y}_i = \arg \max_{j \in \{1,2\}} z_{i,j}, \quad (2)$$

which is compared against the ground truth $y_i \in \{1, 2\}$. While promising, this approach entails several challenges that our proposed method aims to address:

1. **Limited Data Availability:** The relatively small datasets available hinder the training of models, necessitating innovative pre-training or transfer learning strategies.
2. **Multi-modal Fusion Complexity:** Integrating heterogeneous data from video and audio sources requires advanced fusion techniques to effectively capture and exploit the mutual information between these modalities.
3. **Overfitting Risks:** The potential for overfitting, especially due to limited sample sizes, poses a serious threat to the generalizability of the model.

Framework

The overall architecture of our approach for non-contact stroke diagnosis is illustrated in Figure 3. Our framework is designed as a parallel multi-modal processing pipeline that leverages both visual and auditory cues to enhance diagnostic accuracy. Specifically, the process is initiated by two dedicated modules: one for video and one for audio. The visual module processes the raw video input $x_i \in \mathcal{X}$ to extract salient spatiotemporal features capturing dynamic expressions and subtle behavioral cues, while the audio module analyzes the corresponding auditory signals $w_i \in \mathcal{W}$ to derive acoustic features that may reflect neurological anomalies.

These modality-specific features are subsequently integrated within a sophisticated fusion module. By concatenating and further processing the outputs from both modules, the fusion module generates a comprehensive feature representation that encapsulates the complementary information contained in the visual and auditory data. This fused representation is then used to estimate the probability distribution

Notation	Explanation	Notation	Explanation
$\hat{\boldsymbol{x}}_i^{\,v}, \hat{\boldsymbol{x}}_i^{\,r}, \hat{\boldsymbol{x}}_i^{\,t}$	Augmented video-frame sequence of sample i	$\hat{\boldsymbol{w}}_i$	Augmented speech waveform (log-mel spectrogram) of sample i
$f_V(\hat{\boldsymbol{x}}_i^{\,v})$	Video embedding produced by the frozen SeCo encoder	$f_A(\hat{\boldsymbol{w}}_i)$	Audio embedding produced by the frozen HuBERT encoder
$\boldsymbol{p}_{\text{img}}, \boldsymbol{p}_{\text{len}}, \boldsymbol{p}_{\text{wid}}$	Visual positional embeddings attached to video tokens	$\boldsymbol{p}_A$	Positional embedding attached to audio tokens
$\boldsymbol{v}_i^0$	Initial projected & position-encoded video token sequence	$\boldsymbol{a}_i^0$	Initial projected & position-encoded audio token sequence
$\boldsymbol{v}^{(l)}$	Multimodal latent state after the l-th fusion block	$\boldsymbol{a}^{(l)}$	Auxiliary audio state after the l-th fusion block
$\mathcal{D}(\cdot)$	Discriminator that enforces modality-invariant representations	$\sigma(\cdot)$	Data-augmentation operator applied to raw inputs
$\mathcal{L}_{\text{cls}}$	Cross-entropy classification loss for stroke prediction	$\mathcal{L}_{\text{adv}}$	Adversarial loss used to mitigate over-fitting

Table 2: Notations employed by the proposed multimodal stroke-diagnosis framework illustrated in Fig.4.

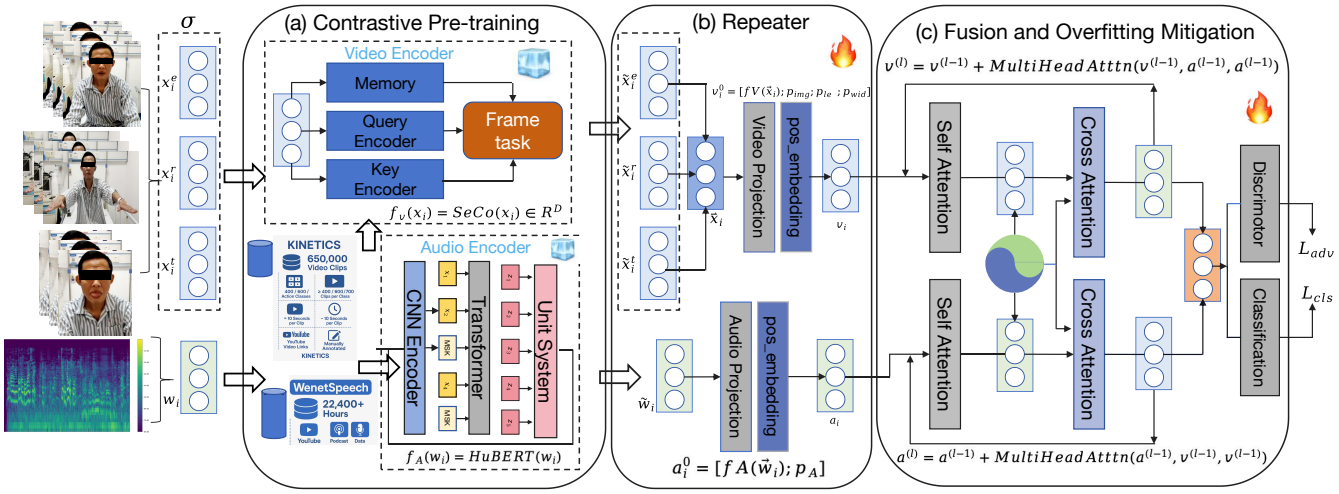

Figure 4: The snowflake symbol indicates that the model is frozen, while the spark symbol represents that the model is trainable. The left portion of the figure illustrates our method's process, which is divided into three parts from left to right: data processing, modality module, and fusion module. The right portion of the figure illustrates the pre-training process used in our approach.

z_i (see Section 4.1 for its definition). Finally, the predicted label is computed as

$$\hat{y}_i = \arg\max_{j \in \{1,2\}} z_{i,j}.$$

By centralizing the mathematical definitions in Section 4.1, we focus here on the fusion process itself and avoid redundant formulae.

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{2} y_{i,j} \log\left(z_{i,j}\right), \tag{3}$$

where $y_{i,j}$ is a binary indicator of whether class j is the correct label for sample i, and N is the total number of samples.

Data Processing Module

Limited data availability poses a significant challenge by hindering the training of models. To overcome this, we propose a unified data processing module that leverages self-supervised transfer learning to import robust representations from large-scale pretraining into our small-sample regime. Below, we describe our video and audio processing submodules in detail.

In the video domain, our objective is to extract and refine motion symmetry cues from patient videos x_i. To this end, we employ the SeCo model (Yao et al. 2021), pretrained on the extensive Kinetics dataset (Kay et al. 2017) using proxy tasks including intra-frame instance discrimination, inter-frame instance discrimination, and temporal order verification. For each video x_i, the SeCo model projects it into a feature space as follows:

$$\boldsymbol{f}_V(x_i) = \text{SeCo}(x_i) \in \mathbb{R}^D. \tag{4}$$

Here, D denotes the dimensionality of the video features. To further enhance these features, we incorporate three distinct positional encodings that capture spatial configurations: image, length, and width positional encodings. Specifically, let

$$\boldsymbol{p}_{\text{img}} \in \mathbb{R}^{d_1}, \quad \boldsymbol{p}_{\text{len}} \in \mathbb{R}^{d_2}, \quad \boldsymbol{p}_{\text{wid}} \in \mathbb{R}^{d_3}, \tag{5}$$

be the encodings corresponding to the maximum indices M_1, M_2, M_3, respectively. The combined video feature is then formulated as:

$$\boldsymbol{v}_i = \left[\boldsymbol{f}_V(x_i);\ \boldsymbol{p}_{\text{img}};\ \boldsymbol{p}_{\text{len}};\ \boldsymbol{p}_{\text{wid}}\right], \tag{6}$$

where $[\cdot;\cdot]$ denotes vector concatenation. This enriched feature representation is then passed through a dedicated representation model $\Phi_V(\cdot)$ to capture the intrinsic spatio-temporal relationships:

$$\boldsymbol{F}_V(x_i) = \Phi_V\left(\boldsymbol{v}_i\right). \tag{7}$$

For the audio modality, our focus is on capturing speech fluency—a key indicator in non-contact stroke diagnosis. We utilize the HuBERT model (Hsu et al. 2021) (pretrained as detailed in (Zhang et al. 2022)) to extract high-fidelity audio representations from patient recordings w_i. The audio feature extraction is defined as:

$$\boldsymbol{f}_A(w_i) = \mathrm{HuBERT}(w_i) \in \mathbb{R}^{D_a}. \tag{8}$$

Here, D_a represents the dimensionality of the audio features. To integrate temporal information, we apply sinusoidal positional encodings:

$$\begin{aligned} \mathrm{PE}(pos, 2i) &= \sin\Big(\tfrac{pos}{10000^{2i/d}}\Big), \\ \mathrm{PE}(pos, 2i+1) &= \cos\Big(\tfrac{pos}{10000^{2i/d}}\Big). \end{aligned} \tag{9}$$

where d is the total dimensionality of the positional encoding. Denote the resulting audio positional encoding vector by

$$\boldsymbol{p}_A \in \mathbb{R}^{d_a}.$$

We then form the complete audio feature by concatenating the HuBERT feature with the positional encoding:

$$\boldsymbol{a}_i = \big[\boldsymbol{f}_A(w_i);\ \boldsymbol{p}_A\big]. \tag{10}$$

This composite vector is input to an audio representation model $\Phi_A(\cdot)$ to extract fluency-related characteristics:

$$\boldsymbol{F}_A(w_i) = \Phi_A\big(\boldsymbol{a}_i\big). \tag{11}$$

Fusion Module

Integrating video and audio modalities is challenging because capturing their complex, complementary dynamics is nontrivial. Traditional methods concatenate modality features and apply a transformer encoder, but this often fails to fully exploit intermodal correlations. To address this, we propose an ADSTF(Alternating Dual-Stream Transformer Fusion) Module that successively swaps modality roles within the transformer, thereby more effectively integrating their complementary information.

Video Fusion Stream Let the positional-encoded video features be denoted by

$$\boldsymbol{\beta} = \mathrm{PE}\big(\boldsymbol{f}_V\big) \tag{12}$$

and the positional-encoded audio features by

$$\boldsymbol{\alpha} = \mathrm{PE}\big(\boldsymbol{f}_A\big) \tag{13}$$

Here, $\boldsymbol{\beta}$ serves as the *query*, while $\boldsymbol{\alpha}$ provides the *keys* and *values*. The cross-modal attention update is then

$$z_V = \mathrm{Softmax}\Big(\frac{W_Q\,\beta\,(W_K\,\alpha)^\top}{\sqrt{d_k}}\Big)\,W_V\,\alpha, \tag{14}$$

where W_Q, W_K, and W_V are the shared learnable projection matrices, and d_k is the key dimension used for scaling.

Audio Fusion Stream Analogously, taking

$$\boldsymbol{\alpha} = \mathrm{PE}\big(\boldsymbol{f}_A\big), \quad \boldsymbol{\beta} = \mathrm{PE}\big(\boldsymbol{f}_V\big), \tag{15}$$

we let $\boldsymbol{\alpha}$ act as the *query*, and $\boldsymbol{\beta}$ as the *keys* and *values*. The update becomes

$$z_A = \mathrm{Softmax}\Big(\frac{W_Q\,\alpha\,(W_K\,\beta)^\top}{\sqrt{d_k}}\Big)\,W_V\,\beta, \tag{16}$$

where W_Q, W_K, and W_V are again shared across both streams, all using the same key dimension d_k.

Baseline Variants for Ablation. To quantify the specific contributions of our proposed Alternating Dual-Stream Transformer Fusion (ADSTF), we construct two simplified baseline variants:

- General Transformer: features from both modalities are concatenated and passed through a 6-layer Transformer encoder, without any cross-modal attention.
- Cross Attention: an additional single unidirectional cross-attention block (visual as query, audio as key/value) is applied prior to the same 6-layer encoder.

In contrast, Our method performs alternating bidirectional cross-attention at each encoder layer and incorporates discriminator-guided adversarial regularization on the fused representation to mitigate overfitting.

Overfitting Mitigation Module

Limited sample sizes may cause the model to memorize complete feature vectors instead of extracting diagnostic-relevant features, thereby undermining its generalization. To mitigate this, we adopt a two-fold strategy. First, the concatenated multimodal features $\{m_i\}$ are position-encoded and passed through a representation model followed by an MLP block to produce diagnostic embeddings $z_{i,1/2}$, which encourages the extraction of task-relevant information.

Second, to further regularize the model, the feature vectors $\{m_i\}$ are evenly segmented along the temporal dimension and then concatenated with segments from the same or different batches to form $h_{i,j}$ and $h_{i,k}$. These are then fed into a discriminator to discourage mere memorization. The discriminator is optimized via:

$$L_{\mathrm{adv},1} = -\sum_i \Big(\|\mathrm{D}\big(\boldsymbol{h}_{i,j}\big)\|_2 + \|1 - \mathrm{D}\big(\boldsymbol{h}_{i,k}\big)\|_2\Big), \tag{17}$$

$$L_{\mathrm{adv},2} = -\sum_i \Big(\|\tfrac{1}{2} - \mathrm{D}\big(\boldsymbol{h}_{i,j}\big)\|_2 + \|\tfrac{1}{2} - \mathrm{D}\big(\boldsymbol{h}_{i,k}\big)\|_2\Big). \tag{18}$$

$$\mathrm{Loss} \;=\; -L_{\mathrm{cls}} \;+\; \lambda\Big(L_{\mathrm{adv}}(\mathrm{D}) + L_{\mathrm{adv}}(\mathrm{E})\Big). \tag{19}$$

In addition, we incorporate keypoint heatmap videos x_i' generated via a MMPOSE (MMPose Contributors 2020) configuration to augment our data. This extra modality further enriches the feature space and substantially enhances the model's generalization.

Experiment

Evaluation metrics

We assess the model performance using three metrics: Accuracy, F1 Score, and Area Under the Receiver Operating Characteristic Curve (AUC).

Accuracy (Acc)

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{20}$$

where TP, TN, FP, and FN are the counts of true positives, true negatives, false positives, and false negatives, respectively.

F1 Score

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{21}$$

which balances Precision (the ratio of correctly predicted positive samples to all predicted positives) and Recall (the ratio of correctly predicted positive samples to all actual positives).

Area Under the ROC Curve (AUC): AUC measures the ability of the model to distinguish between classes by computing the area under the ROC curve, which plots the True Positive Rate against the False Positive Rate over various thresholds.

Metrics	AUC (%)	Acc (%)	F1 (%)
I3D (Carreira and Zisserman 2017)	68.1±9.7	70.9±10.6	75.8±9.7
MViT (Fan et al. 2021)	78.0±9.0	**81.0±7.0**	**86.0±6.0**
SlowFast (Feichtenhofer et al. 2019)	72.9±9.3	75.9±8.6	81.0±9.0
TimeSformer (Bertasius, Wang, and Torresani 2021)	74.4±7.2	79.9±6.2	85.4±6.3
Swin Transformer(Liu et al. 2021)	74.6±7.2	72.1±4.8	80.0±3.4
VideoMAE (Tong et al. 2022)	81.0±3.2	78.2±5.6	82.7±4.8
DeepStroke (Cai et al. 2022)	84.5±5.6	76.2±5.9	82.1±4.5
M3Stroke (Cai et al. 2024)	**86.3±4.3**	79.2±3.9	84.2±4.2
wav2vec 2.0 (Baevski et al. 2020)	63.1±3.7	71.6±4.7	73.4±6.8
Maximum ↑	**86.3±4.3**	**81.0±7.0**	**86.0±6.0**
General Transformer*	84.6±3.1	80.5±4.3	84.9±3.1
Cross Attention*	88.6±2.2	83.1±4.8	87.2±3.4
Ours*	**91.2±1.5**	**87.2±3.1**	**90.8±2.3**

Table 3: Comparison of performance evaluation results for our method under different settings and baselines. * indicates that the input data were processed through feature extraction. The performance metrics are reported as mean±standard deviation across multiple runs.

Implementation details

We employ MMPOSE (MMPose Contributors 2020), configured via the official repository[1], to generate heatmap videos of patients' facial keypoints. Our audio features are extracted using HuBERT (Hsu et al. 2021), pretrained on the WenetSpeech Mandarin corpus[2] (Zhang et al. 2022). The video encoder is initialized with SeCo weights inherited from MoCoV2 (He et al. 2020), pretrained on the Kinetics dataset[3] (Kay et al. 2017).

For representation learning, we adopt a Transformer (Vaswani et al. 2017) with six layers ($L = 6$) and eight attention heads ($M = 8$). To encode spatial and temporal positions, we employ three learned embeddings: image position indices up to $M_1 = 128$ map to $d_1 = 32$-dimensional vectors, height indices up to $M_2 = 305$ map to $d_2 = 20$-dimensional vectors, and width indices up to $M_3 = 200$ map to $d_3 = 20$-dimensional vectors.

All components are trained for 20 epochs with a batch size of 32, using the AdamW optimizer (Loshchilov and Hutter 2017). The main network uses a learning rate of 1×10^{-5}, while the adversarial discriminator is optimized at 1×10^{-6}. We set the adversarial weighting factor λ to 0.1 to balance the generator and discriminator objectives.

Upon acceptance of this manuscript, we will publicly release the complete source code and pretrained model weights to facilitate transparency and reproducibility.

Comparison

As shown in Table 3, our method achieves 91.2% AUC, 87.2% Acc, and 90.8% F1—corresponding to relative improvements of 13.2%, 6.2%, and 4.8% over MViT's 78.0%, 81.0%, and 86.0%, respectively. Compared with the multimodal baseline DeepStroke (84.5% AUC, 76.2% Acc, 82.1% F1), our approach yields gains of 6.7%, 11.0%, and 8.7% in AUC, Acc, and F1.

Furthermore, when evaluated on feature representations extracted by SeCo and HuBERT, the concatenation-only baseline General Transformer attains 84.6% AUC, 80.5% Acc, and 84.9% F1. Relative to this, our method improves AUC, Acc, and F1 by 6.6%, 6.7%, and 5.9%, respectively. Likewise, the single cross-attention variant Cross Attention, a%lied on the same SeCo/HuBERT feature space, achieves 88.6% AUC, 83.1% Acc, and 87.2% F1; our approach yields further improvements of 2.6% in AUC, 4.1% in Acc, and 3.6% in F1.

In summary, by integrating pre-trained self-supervised backbones such as SeCo and HuBERT with our dual-stream recurrent cross-attention architecture, our framework not only captures motion symmetry and speech fluency cues critical for non-contact stroke diagnosis but also demonstrates strong scalability, robust multimodal integration, and efficient real-time inference—making it well-suited for practical clinical deployment.

Model Fairness Evaluation

To ensure our approach boosts overall performance without sacrificing fairness for any demographic subgroup, we evaluate it using the worst-group-risk criterion (Sagawa et al. 2020). Three models—Maximum, General Transformer*, and Our Method—are then benchmarked across nine sensitive subpopulations defined by AGE, GENDER, and POSTURE. As shown in Fig.5, which reports the mean±standard deviation of AUC, ACC, and F1, Our method achieves the best macro-level scores (AUC 91.0%, Acc 87.0%, F1 90.8%), surpassing the strong baseline General Transformer* by +6.6, +6.9, and +6.1 percentage points, respectively. Even in the most challenging cohort (Age>60, Female, Sleeping posture), Our method retains AUC 89.5%, Acc 85.0%, and F1 89.5%, whereas both baselines fall below 82% on at least one metric.

We further quantify fairness using the range of subgroup means, $\Delta_{\max - \min}$. Relative to Maximum, Our method reduces the AUC gap from 8.0 to 3.0 points (–62%), the Acc

[1]https://github.com/open-mmlab/mmpose/tree

[2]https://wenet.org.cn/WenetSpeech/

[3]https://github.com/cvdfoundation/kinetics-dataset

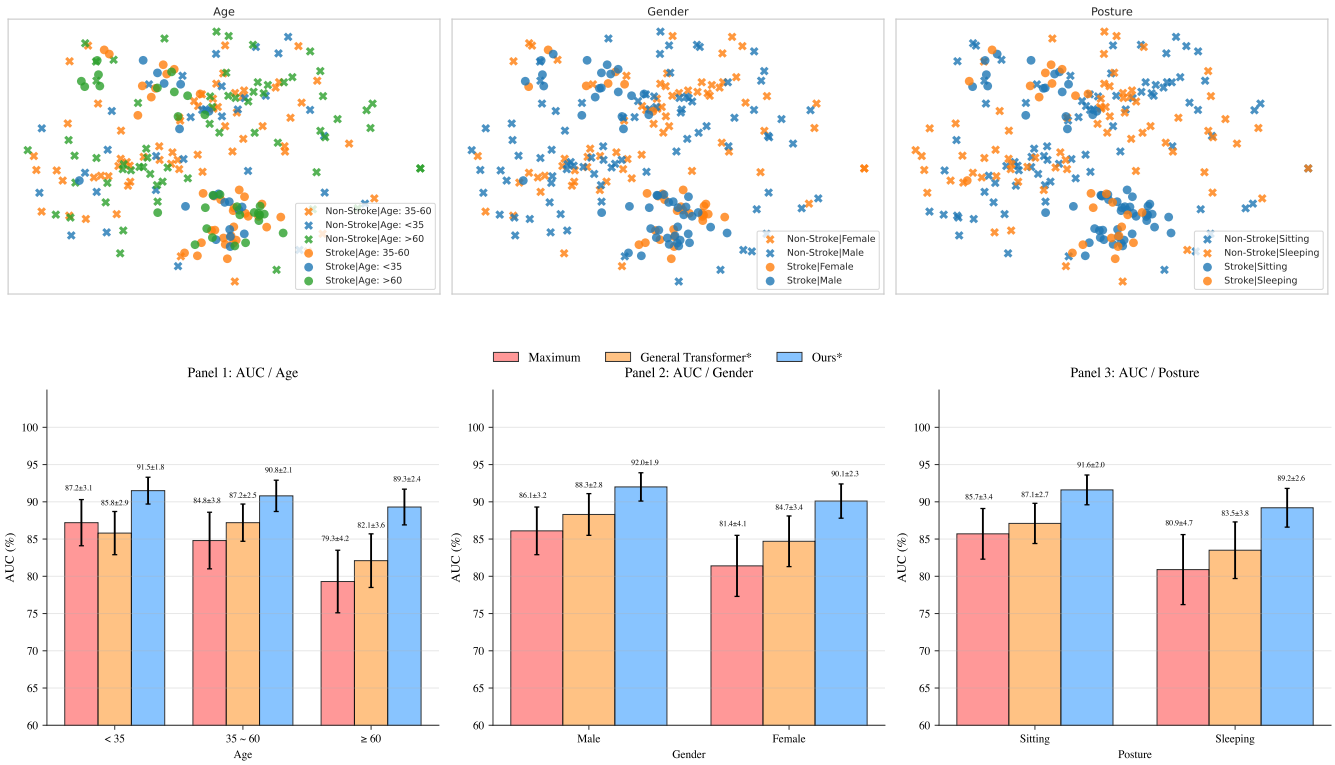

Figure 5: Top: t-SNE embeddings coloured by stroke label and faceted by Age, Gender, and Posture. Bottom: mean±SD AUC (Panels 1–3) for the three models (full results—including Acc and F1—are reported in Appendix A); Our method consistently leads every subgroup, demonstrating improved performance without sacrificing fairness.

gap from 8.0 to 4.0 points (–50%), and the F1 gap from 7.0 to 2.5 points (–64%). These reductions demonstrate that our cross-modal regularization not only elevates the overall decision boundary but also balances predictive performance across heterogeneous clinical conditions, thereby enhancing both accuracy and fairness.

Modality Combinations

To validate the effectiveness of our framework in utilizing multimodal information, we designed a modality fusion ablation experiment to study the impact and scalability of the framework under different modality combinations. As shown in Table 4, when only the Face modality (1a) was used, our method achieved AUC, Acc, and F1 scores of 92.1%, 86.6%, and 90.2%, respectively. After introducing the Audio modality, some metrics slightly decreased, indicating that the Audio modality introduced some interference in stroke diagnosis within this combination. However, when the Tongue and Body modalities (3a) were added, compared to using only the Face modality (1a), the AUC, Acc, and F1 scores increased by 3.3%, 3.2%, and 2.0%, respectively. These experiments suggest that our framework is highly sensitive to modality combinations. To further validate this, we removed the Audio modality and observed that in settings 2b and 3b, all metrics showed a certain degree of decline. Despite the sensitivity of our method to modality combinations, the integration of multimodal information significantly enhances the model's generalization capability and diagnostic performance. This further demonstrates that our framework has high scalability and flexibility in effectively handling and integrating multimodal information from diverse sources.

Generalization ability

To further validate the generalization ability of our framework, we used audio and video data from 18 stroke patients collected under conditions different from those of the original dataset as a generalization test set. Under the same experimental settings, the generalization performance of each method is shown in Table 5. Our model consistently outperformed other baselines in terms of Acc across all settings. This indicates that the model specialized by our framework for this task has strong generalization capabilities, effectively adapting to different collection conditions and being deployable in new scenarios. This again demonstrates the strong versatility of our framework.

Component Ablation

To dissect the contributions of individual architectural components, we perform an ablation study over the following

Setting	Face	Tongue	Body	Audio	AUC (%)	Change (%)	Acc (%)	Change (%)	F1 (%)	Change (%)
1.a	✓	✗	✗	✗	82.1±5.1	–	75.2±5.3	–	81.4±4.3	–
1.b	✓	✗	✗	✓	83.4±5.2	+1.3	75.9±5.1	+0.7	81.7±4.7	+0.3
2.a	✓	✓	✗	✓	87.1±3.1	+5.0	82.3±3.3	+6.4	86.8±2.7	+5.4
3.a	✓	✓	✓	✓	91.2±1.5	+9.1	87.2±3.0	+12.0	90.8±2.3	+9.4
2.b	✓	✓	✗	✗	85.7±4.7	+3.6	78.7±4.7	+3.5	83.6±3.9	+2.2
3.b	✓	✓	✓	✗	88.3±2.2	+5.5	86.1±4.1	+10.9	87.6±2.5	+6.2

Table 4: Performance of our method with different modality combinations (%). The table shows the AUC, accuracy (Acc), and F1-score for various combinations of input modalities: Face, Tongue, Body, and Audio. The performance metrics are reported as mean ± standard deviation across multiple runs.

Ours	Acc (%)	Baselines	Acc (%)
1.a	86.9±2.1	M3stroke	67.9±6.7
1.b	87.9±3.2	Deepstroke	52.9±7.2
2.b	94.5±3.8	TimeSformers	86.2±6.1
3.b	94.0±5.8	Mvit	84.8±9.8
2.a	94.7±4.9	Slowfast	82.7±8.3
3.a	98.4±2.3	I3D	64.8±9.6

Table 5: The generalization performance of our method and other methods (%). The table compares the accuracy (Acc) of our method in different settings, as in Table 4, against the baselines. The performance metrics are reported as mean ± standard deviation across multiple runs.

Variant	**Configuration**
Full Model (Ours)	–
VideoMAE	SeCo → VideoMAE
VGGish	HuBERT → VGGish
No Keypoint	without keypoint branch
No Discriminator	without adversarial regularization
General Transformer*	no cross-modal attention or discriminator
Cross Attention*	single cross-attention, no discriminator

Table 6: Component ablation variants used in our experiments.

variants Table 6, with the results presented in Table 7.

Replacing SeCo with VideoMAE incurs a 6.9% decrease in AUC and a 3.7% drop in F1, underscoring the efficacy of our contrastive video pretraining. Substituting HuBERT with VGGish produces declines of 4.6% in AUC, 4.3% in accuracy, and 7.7% in F1, affirming the superior diagnostic granularity of our chosen audio encoder. Ablating the keypoint branch yields a 2.0% AUC reduction, while omitting the adversarial discriminator imposes a 3.2% AUC penalty, thereby validating their roles in feature enrichment and regularization. Finally, the General Transformer and Cross Attention variants trail our full model by up to 6.6% in AUC and 5.9% in F1, confirming the indispensable value of our alternating dual-stream fusion with adversarial regularization.

Setting	**AUC (%)**	**Acc (%)**	**F1 (%)**
Full Model	91.2±1.5	87.2±3.1	90.8±2.3
VideoMAE	84.3±3.0	83.6±4.0	87.1±3.2
VGGish	86.6±3.5	82.9±4.2	83.1±3.7
No Keypoint	89.2±2.7	83.5±3.4	87.3±2.8
No Discriminator	88.0±3.2	84.1±3.9	86.7±3.5
General Transformer*	84.6±3.1	80.5±4.3	84.9±3.1
Cross Attention*	88.6±2.2	83.1±4.8	87.2±3.4

Table 7: Ablation over key components: video encoder, audio encoder, keypoint branch, and adversarial discriminator. * indicates that the input data have been processed through feature extraction. All models are evaluated with alternating dual-stream cross-attention unless otherwise noted.

tion.

Conclusion

In this study, we first constructed a non-contact, multimodal stroke assessment dataset—encompassing facial expressions, tongue protrusion, limb movements, and speech—which fills a critical gap in non-contact diagnostic resources. Building on this dataset, we propose FAST-CAD, a novel method for automated non-contact stroke diagnosis whose self-supervised learning component addresses the challenge of limited medical data. Extensive experiments validate the effectiveness of our approach; ablation studies and generalization tests further confirm the soundness of each module's design and the strong generalizability of our method; and fairness evaluations demonstrate that our method maintains superior diagnostic performance across different age groups, genders, and patient postures. In future work, we will explore extending our approach to additional modalities to facilitate its application in community and clinical settings.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space–Time Attention All You Need for Video Understanding? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, 10040–10050.

Cai, T.; Ni, H.; Yu, M.; Huang, X.; Wong, K.; Volpi, J.; Wang, J. Z.; and Wong, S. T. 2022. DeepStroke: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning. *Medical Image Analysis*, 80: 102522.

Cai, T.; Wong, K.; Wang, J. Z.; Huang, S.; Yu, X.; Volpi, J. J.; and Wong, S. T. 2024. M3 Stroke: Multi-Modal Mobile AI for Emergency Triage of Mild to Moderate Acute Strokes. –.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6299–6308.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv preprint arXiv:2002.05709.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, volume 1, 4171–4186.

Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; and Feichtenhofer, C. 2021. Multiscale Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 337–347.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6201–6210.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; Piot, B.; Kavukcuoglu, K.; and Munos, R. 2020. Bootstrap your own latent: A new approach to self-supervised Learning. arXiv preprint arXiv:2006.07733.

Guttag, J. 2010. CHB-MIT Scalp EEG Database. Online; available at PhysioNet.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3451–3460.

Inamdar, M. A.; Raghavendra, U.; Gudigar, A.; Chakole, Y.; Hegde, A.; Menon, G. R.; Barua, P.; Palmer, E. E.; Cheong, K. H.; Chan, W. Y.; CiAccio, E. J.; and Acharya, U. R. 2021. A Review on Computer Aided Diagnosis of Acute Brain Stroke. *Sensors (Basel, Switzerland)*, 21(24): 8507.

Johnson, W.; et al. 2016. Stroke: a global response is needed. *Bulletin of the World Health Organization*, 94(9): 634–634A.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, A.; Suleyman, M.; and Zisserman, A. 2017. The Kinetics Human Action Video Dataset. arXiv preprint arXiv:1705.06950.

Korompili, G.; Amfilochiou, A.; Kokkalas, L.; Mitilineos, S. A.; Tatlas, N.-A.; Kouvaras, M.; Kastanakis, E.; Maniou, C.; and Potirakis, S. M. 2021. PSG-Audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies. –.

Lisowska, A.; O'Neil, A. Q.; Dilys, V.; Daykin, M.; Beveridge, E.; Muir, K. W.; McLaughlin, S.; and Poole, I. 2017. Context-Aware Convolutional Neural Networks for Stroke Sign Detection in Non-contrast CT Scans. In *Proceedings of the Annual Conference on Medical Image Understanding and Analysis*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

MMPose Contributors. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. Online; https://github.com/open-mmlab/mmpose.

Mohd Nor, A.; McAllister, C.; Louw, S.; Dyker, A.; Davis, M.; Jenkinson, D.; and Ford, G. 2004. Agreement Between Ambulance Paramedic- and Physician-Recorded Neurological Signs With Face Arm Speech Test (FAST) in Acute Stroke Patients. *Stroke*, 35(6): 1355–1359.

Ringeval, F.; Schuller, B.; Valstar, M.; Cummins, N.; Cowie, R.; Tavabi, L.; Schmitt, M.; Alisamir, S.; Amiriparian, S.; Messner, E.-M.; Song, S.; Liu, S.; Zhao, Z.; Mallol-Ragolta, A.; Zhao, R.; Soleymani, M.; and Pantic, M. 2019. AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. In *Proceedings of the AVEC 2019 Workshop*.

Sagawa, S.; Koh, P. W.; Hashimoto, T.; and Liang, P. 2020. Distributionally Robust Neural Networks for Group Shifts. In *International Conference on Machine Learning (ICML)*, 9117–9126. PMLR.

Sanelli, P.; Sykes, J.; Ford, A.; Lee, J.-M.; Vo, K.; and Hallam, D. 2014. Imaging and Treatment of Patients with Acute Stroke: An Evidence-Based Review. *American Journal of Neuroradiology*, 35(6): 1045–1051.

Shinohara, Y.; Takahashi, N.; Lee, Y.; Ohmura, T.; and Kinoshita, T. 2019. Development of a deep learning model to identify hyperdense MCA sign in patients with acute ischemic stroke. *Japanese Journal of Radiology*, 38: 112–117.

Tong, X.; Li, W.; Li, L.; Loy, C. C.; and Lin, D. 2022. VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Advances in Neural Information Processing Systems*.

Valstar, M.; Schuller, B.; Smith, K.; Eyben, F.; Jiang, B.; Bilakhia, S.; Schnieder, S.; Cowie, R.; and Pantic, M. 2013.

AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the AVEC 2013 Challenge*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. –.

Yao, T.; Zhang, Y.; Qiu, Z.; Pan, Y.; and Mei, T. 2021. SeCo: Exploring Sequence Supervision for Unsupervised Representation Learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

Zhang, B.; Lv, H.; Guo, P.; Shao, Q.; Yang, C.; Xie, L.; Xu, X.; Bu, H.; Chen, X.; Zeng, C.; Wu, D.; and Peng, Z. 2022. WenetSpeech: A 10000+ Hours Multi-domain Mandarin Corpus for Speech Recognition. In *ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing*, 6182–6186.

Zhou, A.; Li, S.; Sriram, P.; Li, X.; Dong, J.; Sharma, A.; Zhong, Y.; Luo, S.; Jaromin, M.; Kindratenko, V.; Heintz, G.; Zallek, C.; and Wang, Y.-X. 2023. YouTubePD: A Multimodal Benchmark for Parkinson's Disease Analysis. –.

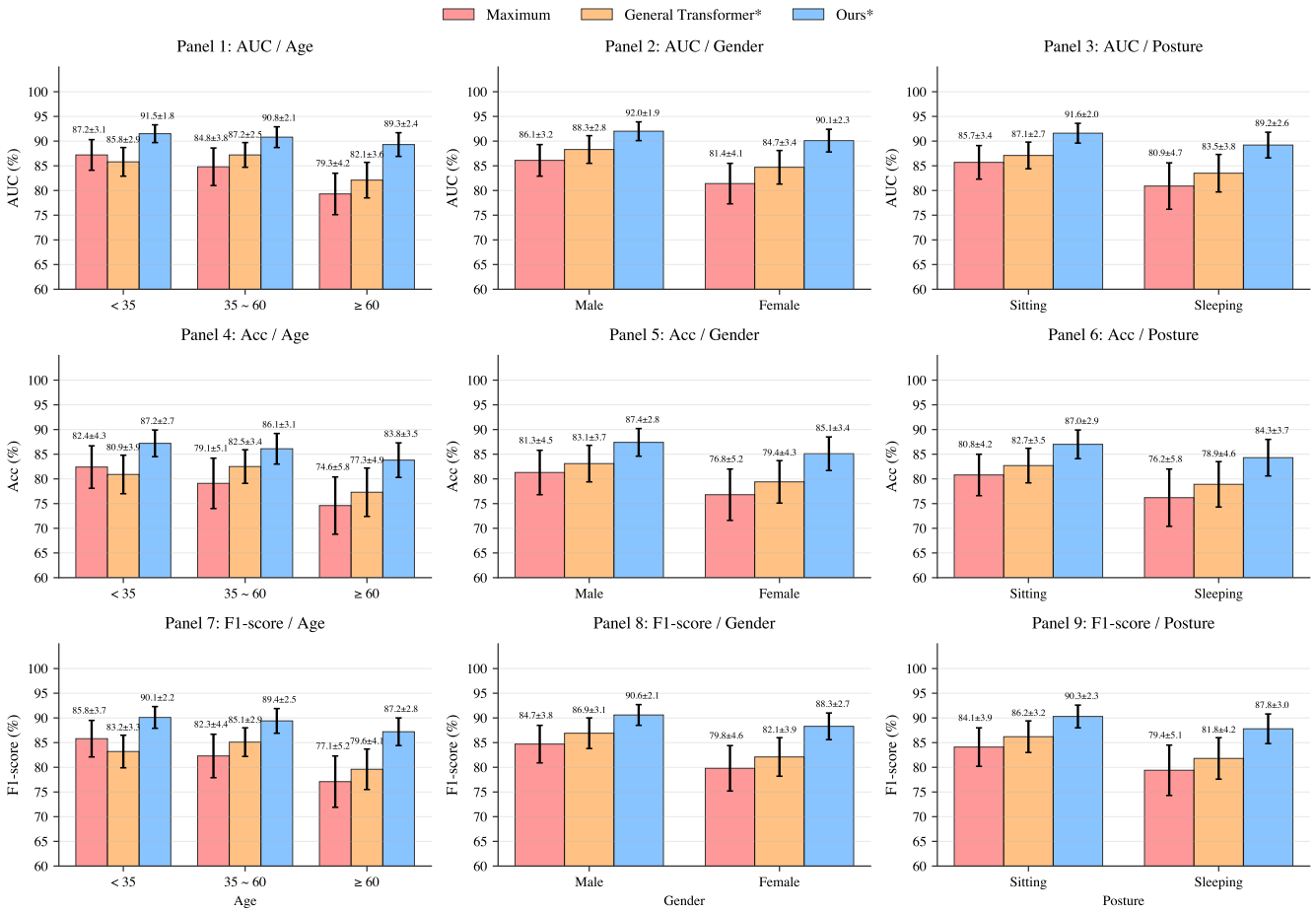

Figure 6: Top: t-SNE embeddings coloured by stroke label and split by Age, Gender, and Posture. Bottom: mean±SD AUC (Panels 1–3) and Acc (Panels 4–6), F1(Panels 7-9) for three models; *Ours** consistently tops all subgroups, indicating performance gains without fairness loss.