DIVINE : Coordinating Multimodal Disentangled Representations for Oro-Facial Neurological Disorder Assessment

Anonymous ACL submission

Abstract

In this study, we present a multimodal framework for predicting neuro-facial disorders by capturing both vocal and facial cues. We hypothesize that explicitly disentangling shared and modality-specific representations within multimodal foundation model embeddings can enhance clinical interpretability and generalization. To validate this hypothesis, we propose **DIVINE** (DIsentangled Variational INFormation NEtwork), a fully disentangled multimodal framework that operates on representations extracted from state-of-the-art (SOTA) audio and video foundation models, incorporating hierarchical variational bottlenecks, sparse gated fusion, and learnable symptom tokens. DIVINE operates in a multitask learning setup to jointly predict diagnostic categories (Healthy Control, ALS, Stroke) and severity levels (Mild, Moderate, Severe). The model is trained using synchronized audio and video inputs and evaluated on the Toronto NeuroFace dataset under full (audio-video) as well as single-modality (audio-only and video-only) test conditions. Our proposed approach achieves SOTA results, with the DeepSeek-VL2 and TRILLsson combination reaching 98.26% accuracy and 97.51% F1-score. Under modality-constrained scenarios, the framework performs well, showing strong generalization when tested with video-only or audio-only inputs. It consistently yields superior performance compared to unimodal models and baseline fusion techniques. To the best of our knowledge, this is the first framework that combines cross-modal disentanglement, adaptive fusion, and multitask learning to comprehensively assess neurological disorders using synchronized speech and facial video. *Code and model weights will be released upon the completion of the double-blind review process* .

1 Introduction

Neurodegenerative and neurovascular conditions such as Amyotrophic Lateral Sclerosis (ALS) and stroke often arise with impairments in facial motor control and speech articulation—symptoms that are not only diagnostic indicators but also indicative of disease progression (Bandini et al., 2020; Naeini et al., 2022). Current clinical evaluations of these symptoms rely heavily on subjective expert assessments, which are labor-intensive, variable across raters, and difficult to scale for longitudinal monitoring. Recent computer vision and speech processing advances have demonstrated promising capabilities in analyzing facial kinematics and vocal patterns for clinical inference. In particular, leveraging facial landmarks (Gomes et al., 2023) and acoustic modeling (Migliorelli et al., 2023) have enabled more objective quantification of motor dysfunction in neuro-facial disorders. However, these efforts often treat each modality in isolation, neglecting the complementary nature of audiovisual cues and their temporal co-dynamics in pathological speech and gestures. In contrast, multimodal architectures provide a more robust and holistic solution by *jointly leveraging visual and acoustic information*. Nevertheless, earlier fusion strategies frequently struggle to separate modality-specific patterns from shared cross-modal representations. This limitation hampers both interpretability and generalizability, key requirements for ensuring clinical reliability.

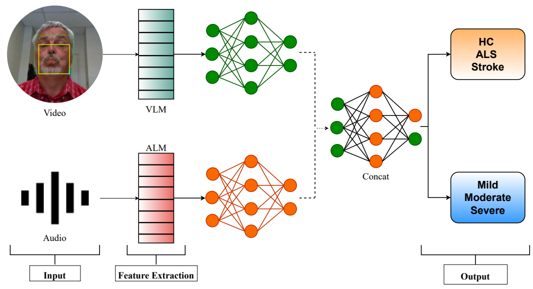


Figure 1: Overview of the **DIVINE** pipeline for clinical diagnosis (HC, ALS, Stroke) and severity prediction(Mild, Moderate, Severe) by encoding synchronized video and audio inputs.

To address the limitations of prior multimodal approaches, we propose DIVINE (**DI**sentangled **V**ariational **IN**formation **NE**twork), a fully disentangled, multitask audio-visual framework for the assessment of neuro-facial disorders. DIVINE integrates pretrained foundation models for both audio and video modalities and employs a hierarchical variational bottleneck to disentangle private (modality-specific) and shared (cross-modal) latent representations. It introduces a sparse gated fusion mechanism that dynamically modulates the influence of each modality and a symptom-guided tokenisation module that directs attention to clinically salient oro-motor features. *We hypothesise that explicitly disentangling shared and modality-specific latent information enhances both disorder classification and severity estimation, while improving generalisation across diverse clinical tasks and input types*. To test this, we conduct extensive evaluations on three clinical populations—HC, ALS, and stroke survivors—across speech, non-speech, and mixed-task conditions. Our model performs multitask learning to jointly predict disorder type and five clinician-rated perceptual severity scores. Through systematic ablations and modality dropout experiments, we demonstrate that DIVINE maintains top performance under unimodal (audio-only, video-only) and multimodal conditions, establishing a new benchmark in multimodal neuro-facial assessment.

To summarize, the main contributions of our study are: (i) We introduced **DIVINE** (**DI**sentangled **V**ariational **IN**formation **NE**twork), a fully disentangled audio–visual variational framework that employs hierarchical variational bottlenecks, cross-modal alignment, gated fusion blocks, and symptom-token modules to extract and integrate complementary speech and facial representations for joint diagnosis and continuous severity estimation of neuro-facial disorders. (ii) We validate our framework on the Toronto NeuroFace dataset under three evaluation settings—full-modality (both audio and video inputs), partial-modality (speech-only or non-speech-only segments), and missing-modality (audio-only or video-only inputs)—and also benchmark over 40 combinations of SOTA audio and vision foundation models. (iii) To the best of our knowledge, **DIVINE** is the first unified framework to combine hierarchical disentangled latent learning, cross-modal alignment losses, and multitask objectives—simultaneously addressing categorical classification (Healthy Control, ALS, Stroke) and regression-style severity prediction—in a single, end-to-end pipeline.

2 Related Work

Early work in oro-facial neurological assessment relied solely on video or images. Researchers used handcrafted spatio-temporal features, such as Improved Dense Trajectories with Fisher Vector encoding, to capture broad facial movements in natural settings (Wang and Schmid, 2013; Afshar and Ali Salah, 2016). (Bandini et al., 2020) introduced the NeuroFace benchmark, showing that standard face-alignment tools can struggle with pathological motion. More recent methods apply deep models: maximisation–differentiation networks for depression screening (de Melo et al., 2021), multiscale CNNs for expression analysis (De Melo et al., 2024), and landmark-aware transformers for estimating expression intensity (Chen et al., 2024). Graph neural networks have also been used to model facial asymmetry and rigidity in ALS patients by treating landmarks as nodes in a facial graph (Gomes et al., 2023). To address video's limitations (occlusion, lighting), simple fusion approaches combine visual and acoustic cues. (Duan et al., 2023) proposes a two-stream system that fuses landmark heat-map volumes with RGB frames via a cross-fusion decoder, improving motion capture. (Neumann et al., 2024) builds a remote dialog system that extracts facial, linguistic, and acoustic biomarkers from ALS patients to track bulbar decline over time. While these methods combine modalities, they treat all features as a single block without separating what each modality contributes. More recent research aims to learn separate, meaningful factors and tackle multiple tasks simultaneously (Duan et al., 2023; Neumann et al., 2024). (Shi et al., 2019) further explores Variational Mixture-of-Experts Autoencoder (MMVAE), which factorises the joint posterior as a mixture of unimodal experts to disentangle shared and private latents and support coherent multi-modal inference. Our work departs from these by introducing a fully disentangled multimodal framework that separates private (audio- or video-specific) and shared representations, and supports joint diagnosis and severity estimation. This approach allows us to quantify each modality's contribution and handle missing or noisy inputs more robustly than previous fusion strategies.

3 Pre-trained Models

Speech Models Our speech encoders include monolingual models—*Wav2Vec2.0* (Baevski et al., 2020) and *WavLM* (Chen et al., 2022)—trained on large-scale English corpora using contrastive and masked prediction objectives. We also leverage *HuBERT* (Hsu et al., 2021), which predicts latent acoustic units via masked prediction, capturing long-range dependencies in speech. We also include multilingual models such as *Whisper* (Radford et al., 2023), trained on 680k hours of cross-lingual data, trained on 128 languages. For prosodic variation and speaker-dependent cues, we use *TRILLsson* (Shor and Venugopalan, 2022) and *x-vector* (Snyder et al., 2018), both known for their robustness in paralinguistic speech tasks.

Vision Models For facial video modeling, we utilize transformer-based models including *VideoMAE* (Tong et al., 2022), *VideoMAE-V2* (Wang et al., 2023), and *ViViT* (Arnab et al., 2021), all employing spatiotemporal encoding strategies. We further use *DeepSeek-VL2* (Wu et al., 2024), a vision-language model with a dynamic tiling and token aggregation mechanism. As structured baselines, we include handcrafted kinematic features from OpenFace (Baltrusaitis et al., 2018) and temporal attention features extracted using a ResNet18+TANN pipeline. Additional details regarding the above PTMs are provided in Appendix A.1.

4 Modeling

We consider two downstream networks, i.e., a fully connected network (FCN) and a CNN with individual PTM representations applied independently to each audio and video foundation model representation. The FCN model consists of three dense layers with 256, 128, and 64 neurons, followed by the output layer. The CNN model comprises two convolution blocks, each containing a 1D convolutional layer followed by batch normalization and a max-pooling operation, then a flattening step and a dense FCN block with the same configuration as above. Detailed hyperparameter settings and model configurations are described in Appendix A.4.

DIVINE: We propose **DIVINE**, a novel multimodal learning framework tailored for neuro-facial disorder assessment. It is built upon a fully disentangled pipeline that incorporates *hierarchical latent modeling, gated cross-modal fusion, and clinical token-aware dense reasoning over synchronized audio and video inputs*. The overall architecture of the proposed framework is illustrated in Figure 2. We extract foundational audio and video representations from raw inputs using frozen pretrained models. Let the raw video and audio inputs be denoted as

$$v \in \mathbb{R}^{T_v \times H \times W \times C}, \quad a \in \mathbb{R}^{T_a}.$$

We extract frozen representations using pretrained foundation models:

$$\begin{aligned} X_v &= \mathrm{VFM}(v) \in \mathbb{R}^{T_v \times d_v}, \\ X_a &= \mathrm{SFM}(a) \in \mathbb{R}^{T_a \times d_a}. \end{aligned}$$

Local Temporal Refinement We first refine the local temporal structure for each modality using CNN-based feature transformation. For each modality $m \in \{v, a\}$, we apply a temporal refinement stage:

$$X'_m = \mathrm{CNN}_m(X_m) \in \mathbb{R}^{T'_m \times d'_m}, \qquad (1)$$

where CNN_m consists of a 1D Convolution, Batch Normalization, ReLU activation, and Max Pooling.

Local VAE (VAE_window) We apply a local VAE over temporally refined segments. For each temporal index $t = 1, \ldots, T''$ and modality $m \in \{v, a\}$, the local variational encoding and decoding steps are:

$$\begin{aligned} (\mu_w^m(t), \log \sigma_w^m(t)) &= f_{\mathrm{enc}}^w(X'_m[t]), \\ z_{\mathrm{sig}}^m(t) &= \mu_w^m(t) + \exp\left(\tfrac{1}{2} \log \sigma_w^m(t)\right) \odot \epsilon, \\ \epsilon &\sim \mathcal{N}(0, I), \\ \hat{X}'_m[t] &= f_{\mathrm{dec}}^w(z_{\mathrm{sig}}^m(t)). \end{aligned} \qquad (2)$$

The local VAE loss is defined as:

$$\begin{aligned} \mathcal{L}_w^m = \frac{1}{T''} \sum_{t=1}^{T''} \left\| X'_m[t] - \hat{X}'_m[t] \right\|^2 \\ + \mathrm{KL}\left(\mathcal{N}(\mu_w^m(t), \sigma_w^m(t)^2) \,\|\, \mathcal{N}(0, I)\right) \end{aligned} \qquad (3)$$

Global Average Pooling We summarize local latent variables across time via global average pooling to obtain fixed-length utterance-level embeddings.

$$\bar{z}^m = \frac{1}{T''} \sum_{t=1}^{T''} z_{\mathrm{sig}}^m(t) \in \mathbb{R}^{d_w}. \qquad (4)$$

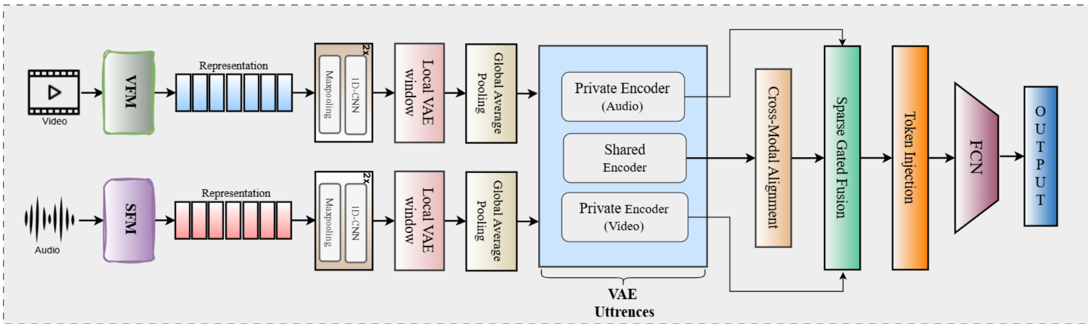


Figure 2: Proposed modeling architecture : **DIVINE**

Utterance-Level VAE (VAE_utterance) We disentangle modality-invariant (shared) and modality-specific (private) representations at the utterance level using two parallel variational autoencoders (VAEs). For each modality $m \in \{v, a\}$, the shared encoder is weight-tied across modalities and maps the global latent representation $\bar{z}^m$ to the parameters of a Gaussian distribution, producing a mean μ_s^m and log-variance $\log \sigma_s^m$. A shared latent variable is sampled using the reparameterization trick as

$$z_{\text{shared}}^m = \mu_s^m + \exp\left(\tfrac{1}{2}\log\sigma_s^m\right) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

In parallel, a modality-specific private encoder $f_{\text{enc}}^{p,m}$, which is unique to each modality, generates the private latent representation by producing μ_p^m and $\log \sigma_p^m$, from which the private vector is sampled as

$$z_{\text{priv}}^m = \mu_p^m + \exp\left(\tfrac{1}{2}\log\sigma_p^m\right) \odot \epsilon.$$

To regularize shared and private encodings, we define the utterance-level VAE loss as the sum of a reconstruction term and KL divergence penalties. The total loss is represented as:

$$\begin{aligned} \mathcal{L}_u^m = \mathcal{L}_{\text{rec}}^m &+ \beta_s \,\mathrm{KL}\left(\mathcal{N}(\mu_s^m, \sigma_s^{m2}) \,\|\, \mathcal{N}(0, I)\right) \\ &+ \beta_p \,\mathrm{KL}\left(\mathcal{N}(\mu_p^m, \sigma_p^{m2}) \,\|\, \mathcal{N}(0, I)\right), \end{aligned} \tag{5}$$

where β_s and β_p are hyperparameters controlling the relative importance of the shared and private KL divergence terms.

Cross-Modal Alignment We decode the video-shared representation into the audio-shared latent space:

$$\begin{aligned} \hat{z}_a &= D_a(z_{\text{shared}}^v), \\ \mathcal{L}_{\text{cycle}} &= \|\hat{z}_a - z_{\text{shared}}^a\|_2^2 . \end{aligned} \tag{6}$$

(Optionally, add the reverse term $\|\hat{z}_v - z_{\text{shared}}^v\|_2^2$.)

Sparse Gated Fusion We compute a sparse, learnable fusion of modality-specific and shared embeddings to dynamically weigh audio and video cues.

$$\begin{aligned} g_v &= \sigma(W_v z_{\text{priv}}^v + b_v), \\ g_a &= \sigma(W_a z_{\text{priv}}^a + b_a). \end{aligned} \tag{7}$$

The fused latent representation is computed as:

$$h_{\text{fused}} = g_v \odot z_{\text{shared}}^v + g_a \odot z_{\text{shared}}^a \in \mathbb{R}^{d_s}. \tag{8}$$

Sparsity penalty:

$$\mathcal{L}_{\text{sparse}} = \|g_v\|_1 + \|g_a\|_1. \tag{9}$$

Token Injection and Dense Layer Let $T_1, \ldots, T_K \in \mathbb{R}^{d_s}$ be learnable clinical symptom tokens. Concatenate and input to dense layer:

$$\begin{aligned} S &= [T_1, \ldots, T_K, h_{\text{fused}}] \in \mathbb{R}^{(K+1)\times d_s}, \\ H_{\text{out}} &= \mathrm{Dense}(S), \quad H_{\text{out}} \in \mathbb{R}^{(K+1)\times d_s} \end{aligned} \tag{10}$$

Add token specialization regularization term $\mathcal{L}_{\text{token}}$.

Output Heads Finally, we derive diagnosis and severity predictions from the fused representation using softmax or linear heads. Let $\mathbf{h} = H_{\text{out}}[K+1]$ denote the fused output. The classification and severity predictions are:

$$\begin{aligned} \hat{y}_{\text{cls}} &= \mathrm{softmax}(W_{\text{cls}}\mathbf{h} + b_{\text{cls}}), \\ \hat{y}_{\text{sev}} &= \mathrm{softmax}(W_{\text{sev}}\mathbf{h} + b_{\text{sev}}). \end{aligned} \tag{11}$$

All non-frozen parameters are optimized end-to-end using the Adam optimizer with early stopping.

Joint Loss Function The function combines classification, severity, reconstruction, and regularization terms:

$$\begin{aligned} \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha\,\mathcal{L}_{\text{sev}} + \epsilon\,(\mathcal{L}_{\text{cycle}} + \mathcal{L}_{\text{sparse}} + \lambda\,\mathcal{L}_{\text{token}}) \\ + \sum_{m\in\{v,a\}} (\mathcal{L}_w^m + \mathcal{L}_u^m). \end{aligned} \tag{12}$$

	C				R				M			
	FCN		CNN		FCN		CNN		CNN			
	A ↑	F1 ↑	A ↑	F1 ↑	M ↓	R ↓	M ↓	R ↓	A ↑	F1 ↑	M ↓	R ↓
VFM												
Vi	80.54	78.57	83.78	81.35	10.82	8.99	9.76	7.28	82.89	82.62	10.25	8.64
V2	82.16	81.65	85.69	83.58	9.29	7.38	8.73	6.84	85.38	84.81	9.45	7.65
VV	79.16	77.28	82.59	81.27	11.26	9.52	10.22	8.63	81.53	80.16	11.86	9.32
DS	85.23	84.61	88.94	86.57	9.22	7.26	8.58	6.81	88.33	86.15	9.38	7.58
KI	72.29	71.61	76.16	74.51	11.82	9.58	10.50	8.89	75.45	73.98	11.55	9.88
TA	78.31	77.18	79.56	77.06	10.58	8.97	9.96	8.05	78.11	76.65	11.09	9.11
SFM												
WV	78.29	77.19	80.83	79.37	8.38	9.70	7.61	8.51	82.09	80.35	6.88	7.60
W2	74.02	73.77	76.37	74.32	8.54	9.89	7.66	8.38	82.13	81.98	6.38	7.55
WR	80.61	79.49	82.34	81.06	8.47	9.36	7.18	8.16	85.94	83.56	6.22	7.29
XV	85.85	83.96	86.29	85.81	8.16	8.59	6.94	7.61	89.27	87.64	6.15	7.14
HT	77.39	76.28	79.62	78.09	9.72	10.12	8.68	9.87	80.11	79.51	6.85	7.74
TR	86.06	84.64	87.58	86.64	7.50	7.88	6.83	7.25	90.51	88.69	6.12	7.01

Table 1: Performance of individual Video Foundation Models (VFMs) and Speech Foundation Models (SFMs) across classification, regression, and multitask tasks on speech-video samples using FCN and CNN backbones.; **Abbreviations:** VFMs – Vi (VideoMAE), V2 (VideoMAE V2), VV (ViViT), DS (DeepSeek-VL2), KI (Kinematic), TA (Temporal); SFMs – WV (WavLM), W2 (Wav2Vec2), WR (Whisper), XV (X-vector), HT (HuBERT), TR (TRILLsson). Note: The abbreviations used in Table 1 are consistent across Tables 2, 3,7,8 and 9.

5 Experiments

Benchmark Dataset: We conduct our experiments on the Toronto NeuroFace (TNF) dataset (Bandini et al., 2020), which contains synchronized audio and video recordings from cognitively intact adults across three clinical groups: ALS, stroke, and healthy controls. We follow a 5-fold cross-validation protocol across all experimental settings. Detailed information on the dataset, task design, and annotation procedures is provided in Appendix A.2,A.3.

Training Details: We use softmax activation in the output layers for both classification and severity prediction heads to produce probability distributions. All models are trained using the Adam optimizer with a learning rate of 10^{-3}, a batch size 32, and categorical cross-entropy loss. Training is performed for 50 epochs with early stopping and dropout regularization to mitigate overfitting. For all **DIVINE** experiments, we fix the hyperparameters: $\alpha = 2$, $\epsilon = 0.1$, and $\lambda = 0.4$, selected based on preliminary validation performance. These values are kept consistent across all fusion and ablation experiments.

Experimental Results: Table 1 shows how each Video Foundation Model (VFM) and Speech Foundation Model (SFM) performs on the TNF speech–video samples, using both FCN and CNN backbones. Among the VFMs, DeepSeek-VL2 (DS) leads with a CNN accuracy of 88.94% and F1 of 86.57%, and achieves the lowest regression errors (MAE = 8.58, RMSE = 6.81) as well as the lowest multitask errors (MAE = 7.58, RMSE = 9.38). VideoMAE V2 follows closely (85.69 % accuracy, 83.58 % F1; MAE = 8.73, RMSE = 6.84). Handcrafted kinematic and temporal features lag behind (76–79 % accuracy with CNN), highlighting the value of pretrained vision encoders. In the audio domain, TRILLsson (TR) is top: it records 90.51 % accuracy and 88.69 % F1 in the multitask setting, with MAE = 6.12 and RMSE = 7.01. Wav2Vec 2.0 and Whisper also perform well (e.g. Wav2Vec 2.0 reaches 89.27 % accuracy, 87.64 % F1), while WavLM and X-vector show weaker regression consistency. Overall, CNN backbones outperform FCNs, confirming their strength at capturing local temporal patterns. Next, we fuse VFMs and SFMs using a simple embedding concatenation (Table 2). Here, DS + TR achieves 94.65 % accuracy and 93.87 % F1 on full speech–video inputs, while still holding 86.33 % accuracy when only video is available and 82.01 % when only audio is available. VideoMAE V2 + X-vector also performs strongly (93.22 % accuracy, 92.55 % F1). These results show that even a straightforward fusion of embeddings leverages complementary modality information and degrades gracefully when one modality is unavailable. Finally, Table 3 reports our `DIVINE` disentangled fusion. The best pair, DS + TR, reaches **98.26 %** accuracy and **97.51 %** F1 when both audio and video embeddings are provided. When evaluated with only video embeddings, DS + TR still scores **89.27 %** accuracy (F1 = 88.23), and when evaluated with only audio embed-

Combinations	Speech Videos				Testing Only Video				Testing Only Audio			
	A ↑	F1 ↑	R ↓	M ↓	A ↑	F1 ↑	R ↓	M ↓	A ↑	F1 ↑	R ↓	M ↓
Concatenation												
Vi + WV	84.55	83.64	4.82	3.96	79.25	79.11	11.72	10.27	74.46	73.61	11.66	10.29
Vi + W2	83.41	82.61	4.86	3.74	78.73	77.79	12.25	9.69	72.31	71.55	12.13	9.65
Vi + WR	87.25	86.23	4.75	3.87	79.22	78.75	11.55	9.91	78.20	77.42	11.68	10.04
Vi + XV	91.64	90.85	4.68	3.91	80.68	79.68	12.13	9.52	81.16	80.29	12.31	10.55
Vi + HT	85.32	84.64	4.80	3.98	78.41	77.79	12.20	9.60	74.66	73.90	12.10	10.55
Vi + TR	92.65	91.11	4.29	3.52	80.65	79.23	11.61	9.88	82.27	81.44	10.52	8.78
V2 + WV	86.36	85.29	4.86	3.49	83.08	82.17	11.96	10.59	72.61	71.81	11.89	9.71
V2 + W2	85.27	84.56	4.81	3.45	82.18	81.16	11.72	9.74	73.28	72.27	11.86	9.62
V2 + WR	87.21	86.21	4.67	3.38	83.52	82.39	11.70	10.46	74.63	73.80	11.84	9.39
V2 + XV	93.22	92.55	3.72	2.75	83.34	82.51	11.09	10.03	80.04	79.21	10.09	8.15
V2 + HT	87.65	86.08	4.78	3.42	83.21	82.65	11.65	8.39	74.95	74.10	11.82	9.35
V2 + TR	90.99	89.24	3.76	2.69	83.54	82.08	11.65	9.87	81.88	81.01	10.31	9.61
VV + WV	82.19	81.65	6.29	5.16	77.69	79.11	13.44	10.81	72.09	71.38	13.52	11.43
VV + W2	81.54	79.69	6.23	4.77	76.82	75.17	13.19	10.58	71.22	70.11	13.66	11.59
VV + WR	85.47	84.43	6.39	5.23	78.23	76.86	13.39	10.73	76.21	75.09	13.99	11.71
VV + XV	89.36	88.14	6.38	4.29	79.28	78.35	13.25	10.59	79.14	78.16	13.52	11.29
VV + HT	83.17	82.64	6.85	4.12	77.15	76.38	13.11	10.34	72.68	71.24	13.25	11.05
VV + TR	90.35	89.15	6.16	4.85	78.61	77.29	12.05	10.27	81.53	80.17	11.23	9.28
DS + WV	91.58	90.09	4.59	3.36	86.08	85.23	11.03	9.15	74.28	73.46	10.93	8.20
DS + W2	89.25	88.34	4.52	3.23	85.56	84.27	11.49	9.04	72.44	71.66	11.42	10.23
DS + WR	92.66	91.01	4.36	3.08	86.09	85.37	11.31	10.70	78.34	77.51	11.40	10.64
DS + XV	92.69	91.14	3.89	2.77	84.53	83.20	10.01	9.70	80.10	79.29	10.07	9.32
DS + HT	90.27	89.64	4.47	3.15	85.88	85.17	11.27	9.99	74.83	74.03	11.37	10.11
DS + TR	94.65	93.87	3.73	2.61	86.33	85.27	12.06	10.10	82.01	81.15	10.12	9.19
KI + WV	81.63	80.52	5.98	4.78	72.07	70.61	14.70	12.37	72.58	71.76	14.84	13.44
KI + W2	79.64	78.11	5.91	4.70	71.96	70.55	14.35	12.01	74.89	74.09	14.25	12.92
KI + WR	84.25	83.64	5.92	4.69	72.25	70.55	14.64	12.10	74.71	73.88	14.52	13.04
KI + XV	85.66	84.29	5.24	4.16	74.20	72.92	12.88	10.14	82.05	81.22	13.01	12.09
KI + HT	80.56	79.65	5.85	4.62	71.48	70.76	14.76	11.98	80.13	79.30	14.71	13.12
KI + TR	86.19	85.35	5.17	4.04	85.34	84.28	12.85	10.14	75.39	74.25	12.94	11.14
TA + WV	82.26	81.93	5.66	4.38	74.35	73.62	14.25	10.55	72.36	71.54	14.44	13.55
TA + W2	80.52	79.67	5.43	4.27	74.55	73.64	13.59	10.42	74.72	73.89	13.78	12.53
TA + WR	83.15	82.65	5.76	4.51	74.56	73.62	13.88	11.12	74.54	73.73	14.07	13.08
TA + XV	86.74	85.51	5.11	4.01	76.77	75.91	13.07	10.36	81.96	81.12	13.17	12.51
TA + HT	82.12	81.68	5.49	4.23	75.39	75.25	13.40	10.23	80.07	79.24	13.56	12.33
TA + TR	90.52	89.58	5.05	3.86	77.15	75.84	12.58	10.73	78.15	77.33	12.57	9.75

Table 2: Performance on combinations of VFM and SFM on simple concatenation combinations across three settings: speech videos, video-only, and audio-only. All scores are reported in percentage (%) and averaged over 5-fold cross-validation.

dings, it achieves **84.34 %** accuracy (F1 = 83.20). Other strong pairs include VideoMAE V2 + X-vector (96.41 % accuracy, 95.68 % F1) and ViViT + TR (over 90 % accuracy).

To assess DIVINE's ability to handle purely visual input, we test on non-speech videos *(Detailed results for these experiments are presented in (Appendix A.5, Tables 7–9*). Table 7, DS individually achieves 89.26 % accuracy and 88.29 % F1 (MAE = 6.02, RMSE = 8.06). When we simply concatenate VFM and SFM embeddings (Table 8), DS + X-vector still reaches 87.24 % accuracy and 86.01 % F1, showing that pre-computed audio features can aid video-only inference. With our **DIVINE** framework fusion (Table 9), DS + TR climbs to 92.58 % accuracy and 91.63 % F1 (MAE = 3.84, RMSE = 5.55), confirming that the model maintains strong performance using only visual information. Refer to (Appendix A.5) for more detail. Additionally, we also present confusion matrices of key configurations in Figure 3 (Appendix A.6.1).

5.1 Ablation Study

To assess the contribution of key components in the proposed framework, we conduct a detailed ablation study along three axes:

5.1.1 Role of Modalities

While unimodal performance was previously discussed in Section 5. We revisit these results here to isolate the individual contribution of each modality. We retain the full model but remove the audio or video input at inference time.

5.1.2 Role of Regularization

We compare the three regularization components in DIVINE: Cycle-consistency (CC) loss, sparse gating (SG), and token reconstruction (TR) loss. Each component is removed independently to evaluate its influence on performance.

Combinations	Speech Videos				Testing Only Video				Testing Only Audio			
	A ↑	F1 ↑	R ↓	M ↓	A ↑	F1 ↑	R ↓	M ↓	A ↑	F1 ↑	R ↓	M ↓
DIVINE												
Vi + WV	86.99	86.11	2.60	2.10	81.69	82.26	6.32	5.31	76.54	75.43	6.13	5.09
Vi + W2	85.23	84.39	2.89	1.92	81.20	80.75	6.64	4.92	74.79	73.90	6.08	4.73
Vi + WR	89.45	88.55	2.84	1.95	81.75	81.42	6.25	5.19	80.55	79.54	6.06	5.02
Vi + XV	93.06	92.17	2.47	2.11	83.77	82.35	6.08	5.20	83.67	82.65	5.90	4.89
Vi + HT	87.25	86.42	2.81	2.33	80.98	80.55	6.37	5.23	76.63	75.76	6.04	5.13
Vi + TR	94.51	93.63	2.41	1.78	83.38	81.82	5.69	4.26	83.61	82.06	5.51	4.43
V2 + WV	88.04	87.23	2.88	1.76	85.89	84.87	6.47	4.67	74.75	73.97	6.31	4.47
V2 + W2	88.54	84.68	2.59	2.01	84.95	83.99	6.27	4.53	75.30	74.36	6.22	4.21
V2 + WR	89.48	88.59	2.35	1.75	86.51	85.64	6.19	4.54	76.98	75.91	5.86	4.31
V2 + XV	**96.41**	**95.68**	**2.16**	**1.51**	86.59	85.48	**4.95**	3.71	**83.71**	**82.27**	**4.68**	**3.45**
V2 + HT	89.85	88.97	2.58	2.04	86.26	85.91	6.36	4.51	77.07	76.22	6.06	4.29
V2 + TR	95.16	94.68	**2.08**	**1.39**	86.93	84.83	**5.06**	**3.50**	**84.03**	**83.08**	**4.84**	**3.41**
VV + WV	85.48	84.60	2.71	2.21	79.94	80.61	8.43	6.90	74.22	73.46	8.11	6.60
VV + W2	83.72	82.81	3.09	2.03	79.62	78.57	8.27	6.32	73.21	72.08	7.85	6.14
VV + WR	87.89	87.04	3.03	2.06	79.62	78.88	8.46	6.95	78.29	77.14	8.30	6.58
VV + XV	91.35	90.32	2.64	2.24	81.33	79.88	8.33	5.64	81.44	80.32	8.03	5.38
VV + HT	86.11	85.27	2.99	2.46	78.73	78.13	8.97	5.53	74.66	73.06	8.61	5.25
VV + TR	93.47	92.49	2.50	1.86	81.75	79.99	8.02	6.53	82.24	81.31	7.90	6.23
DS + WV	93.83	92.91	2.47	1.97	88.64	87.49	6.14	4.37	76.45	75.47	5.75	4.22
DS + W2	91.11	90.28	2.52	1.84	88.55	87.45	6.05	4.21	74.72	73.78	5.70	4.05
DS + WR	95.48	94.55	2.49	1.74	**89.01**	**88.66**	5.74	4.02	80.71	79.78	5.57	3.89
DS + XV	**95.56**	**94.63**	2.26	1.61	**88.85**	**88.02**	5.25	**3.61**	82.24	81.29	5.06	3.49
DS + HT	92.99	92.11	2.55	1.72	87.89	86.53	5.88	4.49	76.97	76.13	5.81	4.32
DS + TR	**98.26**	**97.51**	**1.93**	**1.12**	**89.27**	**88.23**	**5.02**	**3.44**	**84.34**	**83.20**	**4.80**	**3.31**
KI + WV	83.26	82.41	3.43	2.55	75.11	73.26	8.02	6.39	74.77	73.91	7.62	6.08
KI + W2	81.22	80.45	3.53	2.39	74.94	72.97	7.94	6.25	77.04	76.21	7.62	5.90
KI + WR	86.07	85.23	3.37	2.57	75.26	73.29	7.80	6.24	76.86	75.90	7.55	5.88
KI + XV	87.19	86.28	2.79	2.31	76.99	75.85	6.30	5.42	82.23	81.36	6.71	5.37
KI + HT	82.14	81.37	3.44	2.54	73.94	73.16	7.77	6.23	82.26	81.33	7.54	5.97
KI + TR	88.39	87.63	2.94	2.38	88.25	87.30	6.82	5.40	77.95	76.56	6.47	5.21
TA + WV	84.97	84.13	3.35	2.54	76.80	76.44	7.40	5.85	74.57	73.63	7.36	5.58
TA + W2	82.88	82.07	2.93	2.17	76.96	76.42	7.07	5.61	76.94	75.98	6.94	5.47
TA + WR	85.99	85.15	3.43	2.56	77.01	76.39	7.76	5.92	76.77	75.85	7.38	5.65
TA + XV	88.63	87.78	2.57	2.10	79.24	78.75	6.72	5.29	82.10	81.22	6.50	5.14
TA + HT	84.35	83.48	2.78	2.16	77.91	78.08	7.41	5.63	81.19	80.30	7.13	5.48
TA + TR	93.20	92.32	2.75	2.24	80.06	78.71	6.75	5.19	80.43	79.48	6.52	4.84

Table 3: Performance on combinations of VFM and SFM on `DIVINE` framework across three settings: speech videos, video-only, and audio-only. All values are reported in percentage (%) and averaged over 5-fold cross-validation.

Setting	A ↑	F1 ↑	M ↓	R ↓
DIVINE (Audio + Video)	**98.26**	**97.51**	**1.12**	**1.93**
Audio only	89.27	88.23	5.02	3.44
Video only	84.34	83.20	4.80	3.31

Table 4: Performance representing the role of modality.

Configuration	A ↑	F1 ↑	M ↓	R ↓
DIVINE	**98.26**	**97.51**	**1.12**	**1.93**
w/o Cycle Consistency	96.14	94.95	1.68	2.37
w/o Sparse Gating	95.83	94.21	1.84	2.65
w/o Token Reconstruction	95.62	93.89	1.90	2.71

Table 5: Performance representing the role of regularization.

5.1.3 Role of Latent Space Disentanglement

DIVINE is built on disentangled representations using separate modality-invariant and modality-specific latent spaces. We compare this design against simpler variants: **Flat Fusion** and **Single-Level Latent**.

Architecture Variant	A ↑	F1 ↑	M ↓	R ↓
DIVINE (2-Level VAE)	**98.26**	**97.51**	**1.12**	**1.93**
Flat Fusion (No Bottleneck)	93.87	92.10	2.11	2.88
Single-Level Latent Fusion	95.22	93.80	1.85	2.62

Table 6: Performance representing the role of subspace disentanglement.

6 Conclusion

In conclusion, we introduced **DIVINE**, a disentangled multimodal framework for joint classification and severity estimation of neuro-facial disorders. The approach is built upon hierarchical latent modeling, sparse gated fusion, and learnable symptom tokens, enabling effective disentanglement and integration of clinical cues from orofacial video and speech modalities. We conduct extensive experiments on the Toronto NeuroFace dataset across speech and non-speech tasks, unimodal and multimodal conditions, and scenarios with missing modalities. Performance demonstrates that our framework consistently outperforms individual audio/video models and baseline fusion techniques. Notably, the concatenation of DeepSeek-VL2 and TRILLsson through **DIVINE** achieves SOTA performance.

Limitations and Future Work While our extensive in-domain evaluation on TNF demonstrates DIVINE's strong performance, full cross-dataset validation is contingent on access to suitable external corpora. In the camera-ready version, we plan—subject to data availability—to evaluate our audio and video encoders separately on external unimodal benchmarks, since no suitable corpus provides both synchronized audio–video recordings.

Ethical Statement This study uses non-public clinical data accessed with appropriate institutional approvals and participant consent. All recordings were anonymized to ensure privacy. The proposed framework is intended for research purposes and is not clinically validated for diagnostic use.

References

Sadaf Afshar and Albert Ali Salah. 2016. Facial expression recognition in the wild using improved dense trajectories and fisher vector encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 66–74.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66.

Andrea Bandini, Sia Rezaei, Diego L Guarín, Madhura Kulkarni, Derrick Lim, Mark I Boulos, Lorne Zinman, Yana Yunusova, and Babak Taati. 2020. A new dataset for facial motion analysis in individuals with neurological disorders. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1111–1119.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. 2024. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *IEEE Transactions on Affective Computing*.

Wheidima Carneiro de Melo, Eric Granger, and Miguel Bordallo Lopez. 2021. Mdn: A deep maximization-differentiation network for spatio-temporal depression detection. *IEEE transactions on affective computing*, 14(1):578–590.

Wheidima Carneiro De Melo, Eric Granger, and Miguel Bordallo Lopez. 2024. Facial expression analysis using decomposed multiscale spatiotemporal networks. *Expert Systems with Applications*, 236:121276.

Shuchao Duan, Amirhossein Dadashzadeh, Alan Whone, and Majid Mirmehdi. 2023. Qafe-net: Quality assessment of facial expressions with landmark heatmaps. *arXiv preprint arXiv:2312.00856*.

Nícolas Barbosa Gomes, Arissa Yoshida, Mateus Roder, Guilherme Camargo de Oliveira, and João Paulo Papa. 2023. Facial point graphs for amyotrophic lateral sclerosis identification. *arXiv preprint arXiv:2307.12159*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 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.

Brian McFee. 2025. librosa/librosa: 0.11.0.

Lucia Migliorelli, Daniele Berardini, Kevin Cela, Michela Coccia, Laura Villani, Emanuele Frontoni, and Sara Moccia. 2023. A store-and-forward cloud-based telemonitoring system for automatic assessing dysarthria evolution in neurological diseases from video-recording analysis. *Computers in Biology and Medicine*, 163:107194.

Saeid Alavi Naeini, Leif Simmatis, Deniz Jafari, Diego L Guarin, Yana Yunusova, and Babak Taati. 2022. Automated temporal segmentation of orofacial assessment videos. In *2022 IEEE-EMBS international conference on biomedical and health informatics (BHI)*, pages 01–06. IEEE.

Michael Neumann, Hardik Kothare, and Vikram Ramanarayanan. 2024. Multimodal speech biomarkers for remote monitoring of als disease progression. *Computers in Biology and Medicine*, 180:108949.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Yuge Shi, Brooks Paige, Philip Torr, and 1 others. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems*, 32.

Joel Shor and Subhashini Venugopalan. 2022. Trillsson: Distilled universal paralinguistic speech representations. *arXiv preprint arXiv:2203.00236*.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

A Appendix

In the Appendix, we provide:
Section A.1: Detailed Information of Pre-trained Models.
Section A.2: Benchmark Dataset.
Section A.3: Data Preprocessing.
Section A.4: Hyperparameters and System Configurations.
Section A.5: Result on Non-Speech Video Samples.
Section A.6: Visualization Analysis.

A.1 Detailed Information of Pre-trained Models

In this section, we detail the pretrained encoders used in our study. We employ pretrained speech models covering self-supervised, supervised, and multilingual training paradigms. All models are used in a frozen setting to extract utterance-level acoustic representations.

Speech Foundation Models

WavLM (Chen et al., 2022)[1]: is a self-supervised speech representation model designed to support full-stack speech processing. It is pretrained using a masked prediction and denoising objective over a diverse 94k-hour dataset composed of public English corpora.

Wav2Vec2.0 (Baevski et al., 2020)[2]: learns contextualized speech representations via contrastive prediction in the latent space. It combines a convolutional encoder with a Transformer network, masking parts of the input and optimizing discrimination against negative samples.

Whisper (Radford et al., 2023)[3]: is a multilingual encoder-decoder model pretrained on 680k hours of weakly labeled internet audio for transcription, translation, and speech activity detection. We use the encoder features from the base model.

x-vector (Snyder et al., 2018)[4]: is a time-delay neural network (TDNN) trained for speaker classification using the VoxCeleb dataset. The extracted vectors are speaker-discriminative and widely adopted in speaker verification and spoof detection tasks.

HuBERT (Snyder et al., 2018)[5]: is a self-supervised speech representation model that combines masked prediction with offline k-means clustering. Pretrained on large-scale datasets (e.g., LibriSpeech 960h, Libri-Light 60k), it performs state-of-the-art speech recognition and paralinguistic tasks. It is available in multiple configurations (BASE, LARGE, X-LARGE), and we use the BASE variant in frozen mode for extracting utterance-level embeddings.

TRILLsson (Shor and Venugopalan, 2022)[6]: is a

[1] https://huggingface.co/microsoft/wavlm-base
[2] https://huggingface.co/facebook/wav2vec2-base
[3] https://huggingface.co/openai/whisper-base
[4] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb
[5] https://huggingface.co/facebook/hubert-base-ls960
[6] https://www.kaggle.com/models/google/

lightweight self-supervised speech model designed specifically for paralinguistic speech tasks, such as emotion recognition, speaker identification, and synthetic speech detection. It is created using knowledge distillation from the CAP12 Conformer model, which was trained on 900K hours of YouTube speech data. It was trained on 58K hours of public speech data from Libri-Light and AudioSet.

Vision Foundation Models

Video-MAE (Tong et al., 2022)[7]: follows a masked autoencoding strategy with high masking ratios (90–95%) applied to spatiotemporal cubes. A vanilla ViT backbone is used as the encoder, and the model is trained using reconstruction as a self-supervised pretext task.

VideoMAE V2 (Wang et al., 2023)[8]: is a scalable self-supervised video pretraining framework that extends VideoMAE with a dual masking strategy, masking both encoder and decoder tokens to reduce memory and computational load. It adopts progressive training, starting with unsupervised learning on a million-level unlabeled video corpus, followed by post-training on a labeled hybrid dataset. We employ the ViT-B variant in a frozen setting to extract clip-level facial features.

ViViT (Arnab et al., 2021)[9]: is a pure-transformer architecture that performs factorized self-attention over space and time using tubelet embeddings. We employ the ViViT-B/16×2 variant initialized from ViT image weights.

Deepseek-VL2 (Wu et al., 2024)[10]: is a Mixture-of-Experts (MoE) vision-language model designed for advanced multimodal understanding. The model is trained across vision-language alignment, multimodal pretraining, and supervised fine-tuning stages on diverse tasks including visual grounding, OCR, and document understanding. Our study uses its vision encoder in a frozen mode to extract temporally aligned visual embeddings from facial video clips.

Kinematic[11]: are extracted using the OpenFace 2.0 toolkit (Baltrusaitis et al., 2018), which provides 3D landmark positions, head pose (yaw, pitch, roll), gaze direction, and facial Action Units (AUs).

Temporal: use a ResNet18 model pretrained on ImageNet to extract frame-level appearance embeddings. A Temporal Attention Network (TANN) is employed on top of these features to model inter-frame dependencies.

A.2 Benchmark Dataset

This study used data from the Toronto NeuroFace (TNF) dataset collected by (Bandini et al., 2020), which brings together meticulously collected, high-quality video recordings of oro-facial gestures in healthy adults and individuals living with neurological impairment. Thirty-six cognitively intact volunteers (11 with ALS, 14 post-stroke, 11 age-matched controls) each performed a battery of nine speech and non-speech tasks—ranging from rapid syllable repetitions ("/pa/," "/pa-ta-ka/") and the tongue-twister "Buy Bobby a Puppy," to maximum jaw openings, lip puckers, and expressive smiles—under standardized lighting and camera distance (30–60cm, 640 $\times$ 480px, $\sim$50fps). Two expert speech-language pathologists rated every trial on symmetry, range of motion, speed, variability, and fatigue using a 5-point scale, yielding a robust set of clinical scores (total range 5–25; inter-rater κ = 0.33–0.61). For over 3300 carefully chosen frames, 68 facial landmarks were hand-annotated (inter-rater nRMSE = 1.36 $\pm$ 0.46%), and precise face-bounding boxes were derived. Rich metadata—including subject demographics, task labels, video timing, and clinician ratings—is provided alongside the recordings. By combining controlled acquisition protocols with thorough ground-truth annotations and clinical assessments. Although the dataset is not publicly available, we were granted access. To our knowledge, it is the only known resource containing synchronized, high-quality facial video and audio recordings with expert clinical annotations specific to neuro-facial disorders.

trillsson

[7] https://huggingface.co/docs/transformers/en/model_doc/videomae

[8] https://huggingface.co/OpenGVLab/VideoMAEv2-Base

[9] https://huggingface.co/docs/transformers/en/model_doc/vivit

[10] https://github.com/deepseek-ai/DeepSeek-VL2

[11] https://github.com/TadasBaltrusaitis/OpenFace

A.3 Data Preprocessing

We perform preprocessing steps to ensure data quality, consistency, and alignment across audio and video streams. For facial videos, we use the 2D Face Alignment Network (2D-FAN) [12] to detect 68 landmarks on each frame. This helps identify the face clearly, which is visible and centrally positioned. For audio, we apply amplitude normalization and forced alignment at the utterance level using segment-level timestamps, implemented via `librosa`(McFee, 2025) for preprocessing.

A.4 Hyperparameters and System Configurations

The CNN architecture used for unimodal modeling begins with two 1D convolutional blocks. The first convolutional block uses 256 filters with a kernel size of 3, followed by batch normalization and max pooling (pool size = 2). The second block applies 128 filters, again with a kernel size of 3, followed by batch normalization and max pooling (pool size = 2). The flattened outputs are passed to an FCN comprising three dense layers with 256, 128, and 64 neurons, respectively, and a final task-specific output layer (either softmax or regression head). The trainable parameters for CNN models using individual pretrained representations range from 0.8 to 1.2 million, depending on the dimensionality of the extracted embeddings. This increases to 3.5–6.5 million parameters for fusion experiments due to additional transformers and fusion layers. We implement all models using the TensorFlow framework and conduct training and evaluation on an NVIDIA RTX 4050 GPU. Code and model weights will be made publicly available upon acceptance.

A.5 Result on Non-Speech Video Samples

We present the complete evaluation of all VFM+SFM combinations on non-speech video samples from the TNF dataset. Table 7 reports the classification and regression metrics for each Video Foundation Model using both FCN and CNN backbones, where DeepSeek-VL2 achieves the highest accuracy (89.26 %) and F1 (88.29 %). Table 8 shows the results of simple embedding concatenation between VFMs and pre-computed SFMs on video-only inputs, demonstrating that DS + X-vector attains 87.24 % accuracy and 86.01 % F1 even without an audio track. Finally, Table 9 provides the full results of our `DIVINE` fusion framework across all model pairings, with DS + TR leading at 92.58 % accuracy and 91.63 % F1 (MAE = 3.84, RMSE = 5.55). These tables offer a detailed view of model performance under purely visual conditions, complementing the concise summary in the main text.

A.6 Visualization Analysis

A.6.1 Confusion Matrices

Figure 3 presents confusion matrices for eight representative DIVINE configurations evaluated across our TNF test scenarios: (a) DeepSeek-VL2 + TRILLsson; (b) DeepSeek-VL2 + X-vector; (c) VideoMAE-V2 + TRILLsson; (d) VideoMAE-V2 + X-vector; (e) ViViT + TRILLsson; (f) ViViT + X-vector; (g) DeepSeek-VL2 + Wav2Vec 2.0; and (h) DeepSeek-VL2 + WavLM. These matrices illustrate classification consistency and error patterns across our key model pairings.

[12] https://github.com/1adrianb/face-alignment

	FCN		CNN		FCN		CNN		CNN			
	A ↑	**F1 ↑**	**A ↑**	**F1 ↑**	**M ↓**	**R ↓**	**M ↓**	**R ↓**	**A ↑**	**F1 ↑**	**M ↓**	**R ↓**
	C				**R**				**M**			
	VFM											
Vi	81.69	80.25	83.71	82.05	7.86	9.84	7.05	9.12	84.26	83.31	8.10	9.38
V2	82.08	81.67	86.04	85.22	6.98	8.82	6.26	7.59	87.47	86.09	7.13	8.62
VV	79.65	78.16	81.54	80.46	9.43	11.29	8.29	10.65	83.37	82.09	9.17	10.99
DS	86.85	85.98	89.26	88.29	6.86	8.35	6.02	8.06	90.05	89.84	6.97	8.76
KI	73.84	72.26	75.65	74.57	8.81	10.57	7.99	9.73	76.95	75.21	8.39	10.51
TA	78.26	77.23	80.69	78.08	7.24	9.83	7.11	9.29	80.33	79.62	7.69	10.03

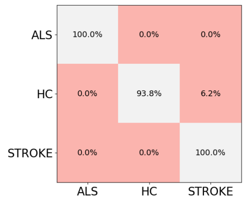
Table 7: Performance on video-foundation models (VFMs) on non-speech video samples. All values are reported in percentage (%) and averaged over 5-fold cross-validation.

Combinations	**Non-speech Testing Videos**			
	A ↑	**F1 ↑**	**R ↓**	**M ↓**
VFM + SFM				
Vi + WV	81.49	80.08	13.23	11.56
Vi + W2	80.84	79.31	14.02	11.77
Vi + WR	80.65	79.44	13.63	10.89
Vi + XV	81.22	79.97	14.51	11.58
Vi + HT	82.39	81.14	14.77	11.98
Vi + TR	81.87	80.48	13.94	10.89
V2 + WV	84.52	83.17	13.55	11.73
V2 + W2	82.65	81.24	13.01	10.97
V2 + WR	83.12	81.96	13.34	11.67
V2 + XV	84.88	83.64	12.67	11.31
V2 + HT	83.99	82.71	12.58	9.56
V2 + TR	84.16	82.99	13.35	10.96
VV + WV	80.13	78.58	15.32	12.36
VV + W2	78.24	77.05	15.15	12.27
VV + WR	79.05	77.29	16.34	12.21
VV + XV	80.26	79.35	15.08	12.60
VV + HT	81.18	79.31	16.86	12.85
VV + TR	79.36	78.62	13.89	11.44
DS + WV	85.84	84.71	12.27	10.16
DS + W2	87.89	86.53	13.25	11.13
DS + WR	86.99	85.62	13.65	11.91
DS + XV	87.24	86.01	11.09	10.71
DS + HT	86.19	84.92	12.49	11.46
DS + TR	86.64	85.29	13.44	11.47
KI + WV	73.79	72.63	16.58	14.03
KI + W2	74.56	73.32	15.40	13.68
KI + WR	74.38	73.11	15.88	13.61
KI + XV	74.12	72.96	14.52	11.67
KI + HT	73.76	72.54	16.65	13.26
KI + TR	83.05	82.58	14.55	11.68
TA + WV	76.94	75.58	15.78	12.29
TA + W2	76.31	74.92	16.66	12.93
TA + WR	77.26	75.88	15.26	12.54
TA + XV	76.64	75.28	14.79	11.43
TA + HT	77.04	75.84	15.13	11.61
TA + TR	74.06	72.89	14.13	12.92

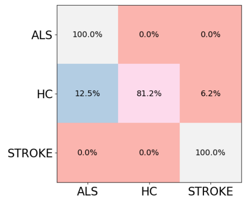
Table 8: Simple concatenation performance on non-speech testing videos for all VFM+SFM combinations. All values are reported in percentage (%) and averaged over 5-fold cross-validation.

Combinations	**Non-speech Testing Videos**			
	A ↑	**F1 ↑**	**R ↓**	**M ↓**
VFM + SFM				
Vi + WV	86.46	85.59	7.01	5.88
Vi + W2	85.94	85.11	7.33	5.50
Vi + WR	87.12	86.25	6.91	5.79
Vi + XV	86.29	85.44	6.70	5.64
Vi + HT	87.05	86.18	7.11	5.58
Vi + TR	85.78	84.95	6.35	4.72
V2 + WV	88.89	88.04	7.26	5.25
V2 + W2	88.21	87.93	7.26	5.32
V2 + WR	89.26	88.38	6.85	5.09
V2 + XV	88.74	87.89	**5.51**	4.12
V2 + HT	89.55	88.64	7.10	5.07
V2 + TR	89.12	88.24	5.63	**3.91**
VV + WV	86.46	85.59	9.37	7.55
VV + W2	85.94	85.11	9.19	7.09
VV + WR	87.12	86.25	9.41	7.61
VV + XV	86.29	85.44	9.26	6.32
VV + HT	87.05	86.18	9.96	7.27
VV + TR	85.78	84.95	8.97	7.19
DS + WV	91.89	91.02	6.80	4.91
DS + W2	91.63	90.77	6.70	4.72
DS + WR	**92.18**	**91.23**	6.34	4.48
DS + XV	**92.41**	**91.47**	**5.80**	**4.06**
DS + HT	91.76	90.91	6.53	5.13
DS + TR	**92.58**	**91.63**	**5.55**	**3.84**
KI + WV	77.81	76.98	8.92	7.16
KI + W2	78.63	77.79	8.84	7.01
KI + WR	78.27	77.41	8.69	6.99
KI + XV	78.45	77.66	7.03	6.04
KI + HT	77.92	77.09	8.67	6.98
KI + TR	88.15	87.08	7.53	6.01
TA + WV	80.88	80.03	8.27	6.56
TA + W2	81.66	80.81	7.92	6.29
TA + WR	81.44	80.59	8.64	6.63
TA + XV	81.39	80.52	7.52	5.90
TA + HT	80.97	80.12	8.35	6.32
TA + TR	78.52	77.65	7.54	5.83

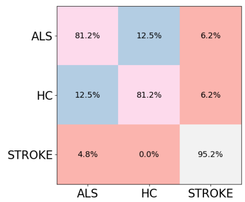
Table 9: Performance on combinations of the proposed **DIVINE** framework across non-speech testing videos for VFM+SFM. All values are reported in percentage (%) and averaged over 5-fold cross-validation.



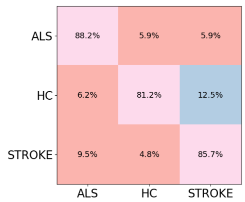
(a)



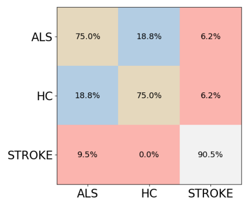
(b)



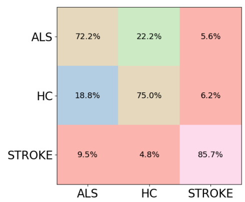
(c)



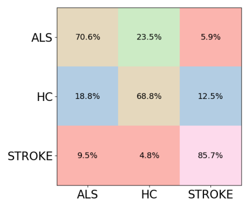
(d)



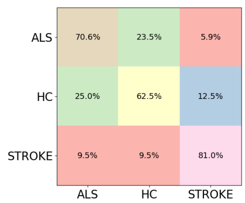
(e)



(f)



(g)



(h)

Figure 3: Representing **DIVINE** configurations. Each displays true versus predicted class distributions across the combined diagnosis and severity categories: (a) DeepSeek-VL2+TRILLsson; (b) DeepSeek-VL2+X-vector; (c) DeepSeek-VL2+X-vector (testing only video); (d) DeepSeek-VL2+TRILLsson (testing only audio); (e) ViViT (Multitask); (f) WavLM; (g) Kinematic (Multitask); and (h) Kinematic (Classification). These matrices highlight classification consistency and error patterns for each fusion pairing.