arXiv:2601.07014v1 [eess.AS] 11 Jan 2026

DIVINE：用于口面部神经系统疾病评估的多模态解耦表示协同

MohdMujtabaAkhtar[1]*、**Girish**[2]* 和 **Muskaan Singh**[3]†‡Veer Bahadur Singh Purvanchal University，印度[2]UPES，印度[3]阿尔斯特大学，英国

摘要

在本研究中，我们提出了一种多模态框架，通过捕捉声音和面部线索来预测神经面部疾病。我们假设，在多模态基础模型嵌入中明确解耦共享表示与模态特定表示，可以提升临床可解释性和泛化能力。为验证这一假设，我们提出**DIVINE**，这是一种完全解耦的多模态框架，基于从最先进（SOTA）的音频和视频基础模型中提取的表示运行，并结合了层次化变分瓶颈、稀疏门控融合和可学习的症状令牌。**DIVINE**采用多任务学习设置，同时预测诊断类别（健康对照、肌萎缩侧索硬化症、卒中）和严重程度等级（轻度、中度、重度）。该模型使用同步音频和视频输入进行训练，并在多伦多 NeuroFace 数据集上评估，测试条件包括完整的音视频输入以及单一模态输入（仅音频和仅视频）。我们提出的方法**DIVINE** 取得了最先进的结果，其中 DeepSeek-VL2 与 TRILLsson 的组合达到 98.26% 的准确率和 97.51% 的 F1分数。在模态受限的场景下，该框架表现出色，在仅视频或仅音频输入的测试中展现了强大的泛化能力。与单模态模型和基线融合技术相比，该框架始终取得更优的表现。据我们所知，**DIVINE** 是首个结合跨模态解耦、自适应融合和多任务学习，利用同步语音和面部视频全面评估神经系统疾病的框架。

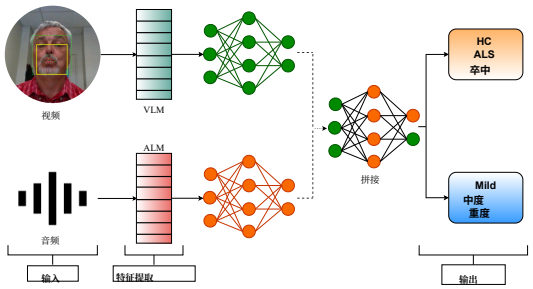

图 1：研究设置概览。

1 引言

神经退行性和神经血管疾病，例如肌萎缩侧索硬化症（ALS）和卒中通常伴随面部运动控制和语音发音障碍——这些症状不仅是诊断指标，也能反映疾病进展（Bandini 等人，2020年；Naeini 等人，2022）。目前，对这些症状的临床评估高度依赖专家的主观判断，这不仅耗费大量人力，在不同评估者之间也存在差异，并且难以扩展到纵向监测。近年来，计算机视觉和语音处理技术的发展，展现出通过分析面部运动学和声音模式进行临床推断的良好潜力。具体而言，利用面部关键点（Gomes 等人，2023年）和声学建模（Migliorelli 等人，2023），能够更加客观地量化神经面部疾病中的运动功能障碍。然而，这些研究通常将每种模态孤立处理，忽视了视听线索的互补性质，以及它们在病理性语音和手势中的时间协同动态。相比之下，多模态架构通过联合利用视觉和声学信息，提供了更稳健、更全面的解决方案。然而，早期的融合策略往往难以将模态特定模式与共享的跨模态表示区分开来。这一局限性削弱了解释性和泛化能力，而这两点对于确保临床可靠性至关重要。如图1所示，我们的任务使用同步语音和面部视频来预测诊断类别及其相关的

*共同作为第一作者
†通讯作者：m.singh@ulster.uk.in

严重程度。

为了解决先前多模态方法的局限性，我们提出 DIVINE（**DI**sentangled**V**ariational **IN**formation**NE**twork），这是一种用于神经面部疾病评估的完全解耦、多任务视听框架。DIVINE 整合了音频和视频模态的预训练基础模型，并采用层次化变分瓶颈，将私有（模态特定）和共享（跨模态）潜在表示解耦。该框架引入了一种稀疏门控融合机制，可动态调节各模态的影响，并采用症状引导的标记化模块，将注意力引向临床显著的口腔运动特征。我们假设，明确解耦共享潜在信息与模态特定潜在信息，既能提升障碍分类和严重程度估计的效果，也能增强模型在多样化临床任务和输入类型上的泛化能力。为验证这一点，我们针对三个临床人群——健康对照、肌萎缩侧索硬化症患者和卒中幸存者——在语音、非语音及混合任务条件下进行了广泛评估。我们的模型采用多任务学习，同时预测障碍类型和五项由临床医生评定的感知严重程度评分。通过系统的消融研究和模态丢弃实验，我们证明**DIVINE**在单模态（仅音频、仅视频）和多模态条件下均能保持顶尖性能，从而为多模态神经面部评估树立了新的基准。

总之，我们的研究主要贡献如下：

- 我们提出了 **DIVINE** （**DI**sentangled**V**ariational**IN**formation **NE**twork），一种完全解耦的音频-视觉变分帧工作框架。该框架采用分层变分瓶颈、跨模态对齐、门控融合块和症状标记模块，提取并整合互补的语音和面部表示，用于神经面部疾病的联合诊断和连续严重程度估计。

- 我们在多伦多 NeuroFace 数据集上，通过三种评估设置验证了该框架——全模态（同时输入音频和视频）、部分模态（仅语音或非语音片段）以及模态缺失（仅音频或纯视频输入）；此外，我们还对 40 多种最先进的音频与视觉基础模型组合进行了基准测试。

- 据我们所知，**DIVINE** 是首个统一框架，将分层解耦潜在表示学习、跨模态对齐损失和多任务目标结合起来，在单一端到端流程中同时实现分类（健康对照、肌萎缩侧索硬化症、卒中）和回归式严重程度预测。

项目资源（代码、模型检查点和评估脚本）可在此处获取：[1]

2 相关工作

早期的口面部神经学评估完全依赖视频或图像。研究人员使用手工设计的时空特征，例如结合费舍尔向量编码的改进密集轨迹，以捕捉自然环境中的广泛面部运动（Wang and Schmid, 2013；Afshar and Ali Salah, 2016）。（Bandini 等人，2020年）引入了 NeuroFace 基准，表明标准人脸对齐工具在处理病理性运动时可能表现不佳。较新的方法应用深度模型：用于抑郁症筛查的最大化–微分网络（de Melo et al., 2021）、用于表情分析的多尺度卷积神经网络（De Melo et al., 2024），以及用于估计表情强度的关键点感知 Transformer（Chen et al., 2024）。图神经网络也被用于建模肌萎缩侧索硬化症患者的面部不对称和僵硬，将关键点视为面部图中的节点（Gomes et al., 2023）。为应对视频的局限性（遮挡、光照），简单的融合方法结合了视觉和声学线索。相关的视听语音识别研究还表明，通过基于注意力的融合，明确利用视觉特征对强大的音频模型进行条件控制，可以提高其在噪声声学环境下的鲁棒性（Rouditchenko et al., 2024；Choi et al., 2025）。（Duan 等，2023）提出了一种双流系统，通过交叉融合解码器将关键点热图体与 RGB 帧融合，从而改善运动捕捉效果。（Neumann 等，2024）构建了一个远程对话系统，从肌萎缩侧索硬化症患者中提取面部、语言和声学生物标志物，以追踪延髓功能随时间的衰退。尽管这些方法融合了多种模态，但它们将所有特征作为单一整体处理，没有分离每种模态所作出的贡献。较新的研究旨在学习相互独立且具有意义的因素，并同时处理多个

[1] https://github.com/Helixometry/SIGNAL.git

任务（Duan 等，2023；2023年；Neumann 等，2024；2024年）。Shi 等人，2019）进一步探索了变分混合专家自编码器（MM-VAE），该模型将联合后验分解为单模态专家的混合，以解耦共享潜变量和私有潜变量（模态特定），并支持连贯的多模态推理。我们的工作通过引入一个完全解耦的多模态框架区别于这些方法，该框架分离私有表示（音频或视频特定）与共享表示，并支持联合诊断和严重程度估计。这种方法使我们能够量化每种模态的贡献，并且比以往的融合策略更稳健地处理缺失或含噪输入。

3 表示

语音模型我们的语音编码器包括单语模型——*Wav2Vec2.0* (Baevski 等人，2020年）和 *WavLM* (Chen 等人，2022）——使用对比和掩码预测目标在大规模英语语料库上训练。我们还利用*HuBERT* (Hsu 等人，2021年），它通过掩码预测来预测潜在声学单元，从而捕捉语音中的长程依赖关系。我们还纳入了多语种模型，例如 *Whis-per* (Radford 等人， 2023年），其使用 68 万小时的跨语言数据进行训练，涵盖 128 种语言。为捕捉韵律变化和说话人相关线索，我们使用*TRILLsson* (Joel Shor 和 Subhashini Venu- gopalan， 2022）和*x-vector* (Snyder 等人， 2018），两者均以在副语言语音任务中的稳健性而闻名。

视觉模型 在面部视频建模方面，我们采用了基于 Transformer 的模型，包括*Video-MAE* (Tong 等人，2022）、*VideoMAE-V2* (Wang 等人， 2023年）和*ViViT* (Arnab 等人，2021年），它们均采用时空编码策略。我们还使用 *DeepSeek-VL2* (Wu 等人， 2024年），这是一种具有动态切块和令牌聚合机制的视觉语言模型。作为结构化基线，我们纳入了从 OpenFace 提取的手工运动学特征（Baltrusaitis 等人， 2018），以及使用 ResNet 提取的时间注意力特征18+ TANN 流程。有关上述预训练模型（PTM）的更多详情见附录A.1。

4 建模

我们考虑两个下游网络，即全连接网络（FCN）和卷积神经网络（CNN），分别独立应用于各个 PTM 表示音频和视频基础模型表示。FCN 模型由三个全连接层组成，分别包含 256、128 和 64 个神经元，后接输出层。CNN 模型包含两个卷积块，每个卷积块均包含一个一维卷积层，后接批归一化和最大池化操作，然后执行展平步骤，并连接一个与上述配置相同的全连接网络块。详细的超参数设置和模型配置见附录A.4。

DIVINE：我们提出了**DIVINE**，这是一种专为神经面部障碍评估设计的新型多模态学习框架。它建立在完全解耦的流程之上，融合了分层潜在建模、门控跨模态融合，以及对同步音频和视频输入进行的临床令牌感知密集推理。直观而言，DIVINE 将语音和面部运动视为同一潜在运动过程的两种含噪观测。局部变分自编码器压缩短时发音动态，而话语级变分自编码器则从共享的神经运动特征中分离出模态特定噪声。随后，稀疏融合门平衡视觉与声学证据，临床令牌则促进跨模态的可解释聚合。所提出框架的整体架构如图2所示。DIVINE 结合局部和话语级变分自编码器，以实现（i）短时模式压缩，以及（ii）模态特定线索与模态共享线索的分解。轻量级跨模态解码器利用惩罚项对齐表示，以限制子空间间的信息泄漏；稀疏门控融合则引导来自更可靠流的信息。最终，症状令牌提供了一个紧凑的网格，便于进行具有临床意义的聚合；为确保在某一模态受损或完全缺失时仍具鲁棒性，必须避免过度依赖单个强大的编码器。

我们使用冻结的预训练模型，从原始输入中提取基础音频和视频表示。将原始视频和音频输入记为

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

我们使用预训练基础模型提取冻结的表示：

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

局部时间细化 首先，我们使用基于卷积神经网络的特征变换，对每种模态的局部时间结构进行细化。对于每种

	FCN		CNN		FCN		CNN		CNN			
	A ↑	F1 ↑	A ↑	F1 ↑	M ↓	R ↓	M ↓	R ↓	A ↑	F1 ↑	M ↓	R ↓
	C				R				M			
	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**

表 1：在分类、回归和多任务设置下，使用 FCN 和卷积神经网络头对语音-视频样本进行评估时，各个视频基础模型（VFMs）和语音基础模型（SFMs）的性能。**最佳**和次佳结果以高亮显示（表 2中采用相同规则）。**缩写：** 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）。注：表 1 中使用的缩写在表 2、3、7、8 和 9中均保持一致。

模态 $m \in \{v, a\}$，我们应用一个时间细化阶段：

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

其中，CNN_m 由一维卷积、批归一化、ReLU 激活和最大池化组成。

局部变分自编码器（变分自编码器_窗口）我们在经时间细化的片段上应用局部变分自编码器。对于每个时间索引 $t = 1, \ldots, T''$和模态 $m \in \{v, a\}$，局部变分编码和解码步骤如下：

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

局部变分自编码器损失定义如下：

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

全局平均池化 我们通过全局平均池化汇总随时间变化的局部潜变量，从而获得固定长度的话语级嵌入。

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

话语级变分自编码器 (变分自编码器_话语) 我们使用两个并行的变分自编码器（VAE），在话语级别解耦模态不变（共享）表示和模态特定（私有）表示。对于每个模态 $m \in \{v, a\}$，共享编码器在各模态间共享权重，并将全局潜在表示 $\bar{z}^m$ 映射为高斯分布的参数，生成均值 μ_s^m和对数方差 $\log \sigma_s^m$。随后，使用重参数化技巧对共享潜变量进行采样，形式如下：

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

与此同时，模态专属的私有编码器$f_{\mathrm{enc}}^{p,m}$为每种模态所独有，通过生成 μ_p^m和 $\log \sigma_p^m$来生成私有潜在表示，并据此对私有向量进行采样：

$$z_{\mathrm{priv}}^m = \mu_p^m + \exp\big(\tfrac{1}{2}\log \sigma_p^m\big) \odot \epsilon$$

为规范共享编码和私有编码，我们将话语级变分自编码器损失定义为重构项与 KL散度惩罚项之和。总损失表示为：

$$\begin{aligned}\mathcal{L}_u^m = \mathcal{L}_{\mathrm{rec}}^m + \beta_s\, \mathrm{KL}\big(\mathcal{N}\big(\mu_s^m, (\sigma_s^m)^2\big) \,\|\, \mathcal{N}(0, I)\big)\\ + \beta_p\, \mathrm{KL}\big(\mathcal{N}\big(\mu_p^m, (\sigma_p^m)^2\big) \,\|\, \mathcal{N}(0, I)\big)\end{aligned}$$

其中，β_s 和 β_p 是控制共享与私有KL散度项相对重要性的超参数。

组合	语音视频				仅测试视频				仅测试音频			
	A ↑	F1 ↑	R ↓	M ↓	A ↑	F1 ↑	R ↓	M ↓	A ↑	F1 ↑	R ↓	M ↓
拼接												
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

表 2：在三种设置（语音视频、仅视频和仅音频）下，VFM 与语音基础模型通过简单拼接进行组合的性能。所有分数均以百分比（%）表示，并取 5 折交叉验证的平均值。

跨模态对齐 我们将视频共享表示解码到音频共享潜在空间中：

$$\hat{z}_a = D_a(z^v_{\text{shared}}),$$
$$\mathcal{L}_{\text{cycle}} = \|\hat{z}_a - z^a_{\text{shared}}\|_2^2$$

稀疏门控融合 我们对特定于模态和共享嵌入进行稀疏、可学习的融合，从而动态地对音频和视频线索加权。

$$g_v = \sigma\big(W_v z^v_{\text{priv}} + b_v\big),$$
$$g_a = \sigma\big(W_a z^a_{\text{priv}} + b_a\big)$$

融合潜在表示的计算方式如下：

$$h_{\text{fused}} = g_v \odot z^v_{\text{shared}} + g_a \odot z^a_{\text{shared}} \in \mathbb{R}^{d_s}.$$

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

令牌注入与全连接层设$T_1, \ldots, T_K \in \mathbb{R}^{d_s}$为可学习的临床症状令牌。拼接后输入全连接层：

$$S = [T_1, \ldots, T_K, h_{\text{fused}}] \in \mathbb{R}^{(K+1)\times d_s},$$

$$H_{\text{out}} = \text{Dense}(S), \quad H_{\text{out}} \in \mathbb{R}^{(K+1)\times d_s}$$

添加令牌专门化正则化项$\mathcal{L}_{\text{token}}$。

输出头 最后，我们使用 Softmax 或线性头，从融合表示中得出诊断和严重程度预测。设 $\mathbf{h} = H_{\text{out}}[K+1]$ 表示融合输出。分类和严重程度预测为：

$$\hat{y}_{\text{cls}} = \text{softmax}(W_{\text{cls}}\mathbf{h} + b_{\text{cls}}),$$
$$\hat{y}_{\text{sev}} = \text{softmax}(W_{\text{sev}}\mathbf{h} + b_{\text{sev}})$$

所有未冻结参数均使用 Adam 优化器进行端到端优化，并采用早停机制。

JointLoss 函数 该函数结合了分类、严重程度、重建和正则化项：

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

5 实验

基准数据集： 我们在多伦多神经面部（TNF）数据集（Ban- dini 等人，2020年）上开展实验。该数据集包含来自三个临床组别（肌萎缩侧索硬化症、卒中和健康对照组）的认知功能完好成人的同步音频和视频录音录像。我们采用按受试者划分的 5 折交叉验证：各折按参与者划分，每次运行中，一名参与者的所有录音录像只会出现在一个数据划分（训练集/验证集/测试集）中，从而避免不同划分之间发生说话人泄漏。有关数据集、任务设计和标注流程的详细信息见附录 A.2、A.3。

训练详情： 分类头和严重程度预测头的输出层均采用 softmax 激活，以生成概率分布。所有模型均使用 Adam 优化器进行训练，学习率为 10^{-3}，批量大小为 32，并采用分类交叉熵损失。训练共进行 50 个训练轮次，并结合早停和 dropout 正则化来缓解过拟合。对于所有**DIVINE** 实验，我们固定超参数：$\alpha = 2$、$\epsilon = 0.1$ 和 $\lambda = 0.4$，这些参数是根据初步验证性能选定的。所有融合实验和消融实验均采用这些一致的参数值。

实验结果： 表1 展示了各视频基础模型（VFM）和语音基础模型（SFM）在 TNF 语音–视频样本上的表现，采用了全连接网络（FCN）和卷积神经网络（CNN）骨干。在视觉基础模型（VFM）中，DeepSeek-VL2（DS）表现最佳，采用 CNN 时准确率为 88.94%，F1 分数为 86.57%，并取得最低的回归误差（平均绝对误差 = 8.58，均方根误差 = 6.81）以及最低的多任务误差（平均绝对误差 = 7.58，均方根误差 =9.38）。VideoMAE V2 紧随其后（准确率 85.69%，F1 分数 83.58%；平均绝对误差 = 8.73，均方根误差 = 6.84）。手工设计的运动学和时间特征表现较弱（采用 CNN 时准确率为 76–79%），凸显了预训练视觉编码器的价值。在音频领域，TRILLsson（TR）表现最佳：其在多任务设置下准确率达到 90.51%，F1 分数达到 88.69%，平均绝对误差 = 6.12，均方根误差 = 7.01。Wav2Vec 2.0 和 Whisper 也表现出色（例如，Wav2Vec 2.0 的准确率达到 89.27%，F1 分数达到 87.64%），而 WavLM 和 X-vector 的回归一致性较弱。总体而言，卷积神经网络骨干优于全连接网络，证实了其在捕捉局部时间模式方面的优势。

接下来，我们使用简单的嵌入拼接来融合视觉基础模型和语音基础模型（表2）。在这里，DS +TR 在完整语音-视频输入下达到 94.65% 的准确率和 93.87% 的 F1 分数；即使仅提供视频，准确率仍为 86.33%，仅提供音频时则为 82.01%。VideoMAE V2 + X-vector 的表现也很出色（准确率为 93.22%，F1 分数为 92.55%）。这些结果表明，即使采用简单直接的嵌入融合，也能利用互补的模态信息，并在某一模态不可用时平稳地降低性能。最后，表 3 展示了我们的 DIVINE解耦融合结果。表现最佳的组合 DS +TR 在同时提供音频和视频嵌入时，达到**98.26%** 的准确率和 **97.51%** 的 F1 分数。仅使用视频嵌入进行评估时，DS +TR 的准确率仍达 **89.27%**（F1 = 88.23）；仅使用音频嵌入进行评估时，准确率则为**84.34%**（F1 = 83.20）。其他表现出色的组合包括 VideoMAE V2 + X-vector（准确率为 96.41%，F1 分数为 95.68%）和 ViViT + TR（准确率超过 90%）。所有结果均为五个随机种子和五折交叉验证的平均值。我们报告均值 ± ± 标准差，并采用配对威尔科克森检验与最强基线进行比较（$p < 0.05$）。结果在不同随机种子下保持稳定，证实了其可靠性。

为评估 DIVINE 处理纯视觉输入的能力，我们在非语音视频上进行测试(这些实验的详细结果见（附录 *A.5*，表7–9）。表7显示，DS 单独使用时准确率达到 89.26%，F1 分数为 88.29%（平均绝对误差= 6.02，均方根误差 = 8.06）。当我们仅将 VFM 和 SFM 的嵌入进行拼接时（表8），DS +X-vector 的准确率仍达到 87.24%，F1 分数为 86.01%，表明预先计算的音频特征有助于仅视频推理。采用我们的 **DIVINE**框架进行融合后（表 9），DS +TR 的准确率升至 92.58%，F1 分数达到 91.63%（平均绝对误差 = 3.84，均方根误差 = 5.55），这证实模型仅使用视觉信息也能保持出色性能。更多详情请参阅（附录 A.5）。此外，我们还在图3 中展示了关键配置的混淆矩阵（附录 A.6.1）。

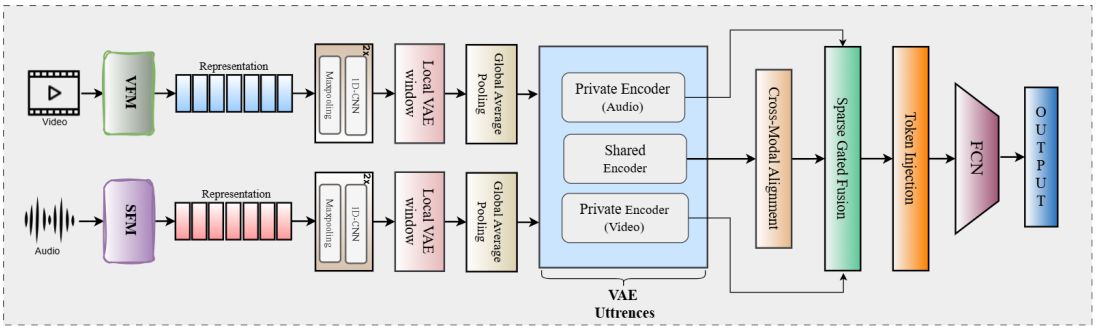

图2：提出的建模架构：**DIVINE**

组合	语音视频				仅测试视频				仅测试音频			
	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

表3：在三种设置下，DIVINE 框架中 VFM 与语音基础模型的组合性能：语音视频、仅视频和仅音频。所有数值均以百分比（%）表示，并取 5折交叉验证的平均值。

5.1 消融研究

为评估所提出框架中各关键组件的作用，我们沿以下三个方面进行了详细的消融研究：

5.1.1 模态的作用

我们此前已在第 5 节讨论过单模态性能。此处我们重新审视这些结果，以分离各模态的独立贡献。我们保留完整模型，但移除音频或推理时的视频输入。

设置	A ↑	F1 ↑	M ↓	R ↓
DIVINE（音频 + 视频）	**98.26**	**97.51**	**1.12**	**1.93**
仅音频	89.27	88.23	5.02	3.44
仅视频	84.34	83.20	4.80	3.31

表 4：体现模态作用的性能。

5.1.2 正则化的作用

我们比较 DIVINE 中的三个正则化组件：循环一致性（CC）损失、稀疏门控（SG）和词元重建（TR）损失。我们分别移除每个组件，以评估其对性能的影响。

配置	A ↑	F1 ↑	M ↓	R ↓
DIVINE	**98.26**	**97.51**	**1.12**	**1.93**
不含循环一致性	96.14	94.95	1.68	2.37
不含稀疏门控	95.83	94.21	1.84	2.65
不含词元重建	95.62	93.89	1.90	2.71

表5：正则化作用的性能表现。

5.1.3 潜在空间解耦的作用

DIVINE 基于解耦表示构建，采用彼此独立的模态不变潜在空间和模态特定潜在空间。我们将这一设计与更简单的变体进行比较：**平坦融合**和 **单层潜在**。

架构变体	A ↑	F1 ↑	M ↓	R ↓
DIVINE（2 层变分自编码器）	**98.26**	**97.51**	**1.12**	**1.93**
平坦融合（无瓶颈）	93.87	92.10	2.11	2.88
单层潜在融合	95.22	93.80	1.85	2.62

表 6：子空间解耦作用的性能表现。

6 结论

总之，我们提出了 **DIVINE**，这是一个用于联合分类和神经面部疾病严重程度估计的解耦多模态框架。该方法基于分层潜在建模、稀疏门控融合和可学习的症状令牌，能够有效解耦并整合来自口面部视频和语音模态的临床线索。我们在多伦多 NeuroFace 数据集上针对言语和非言语任务、单模态和多模态条件，以及模态缺失场景开展了广泛实验。结果表明，我们的框架始终优于单独的音频/视频模型和基线融合技术。值得注意的是，通过 **DIVINE** 拼接 DeepSeek-VL2 和 TRILLsson，达到了最先进的性能。

局限性与未来工作

尽管我们在 TNF 上进行了广泛的域内评估，结果显示 DIVINE 性能出色，但要进行全面的跨数据集验证，仍取决于能否获取合适的外部语料库。在最终定稿版本中，我们计划在数据可用的前提下，分别在外部单模态基准上评估音频和视频编码器，因为目前没有合适的语料库同时提供同步音视频录制。

伦理声明

本研究使用了经适当机构审批并取得参与者同意后访问的非公开临床数据。所有录音录像均已匿名化，以确保隐私。所提出的框架仅用于研究，尚未经过临床验证，不可用于诊断。

社会影响

DIVINE 旨在作为研究和临床筛查工作流程中的决策支持工具，而非独立的诊断系统。录制条件、人口统计因素或不同机构间的分布偏移都可能导致预测产生偏差；若在缺乏保障措施的情况下使用，错误输出可能引发不当的临床行动。因此，任何部署都必须有临床医生监督，清楚传达不确定性，并进行人在环验证。我们还提醒，不应在缺乏透明审计和持续监测的情况下，将其滥用于自动化分诊或资格判定。

参考文献

Sadaf Afshar 和 Albert Ali Salah。2016。使用改进密集轨迹和费舍尔向量编码进行野外面部表情识别。载于*IEEE*计算机视觉与模式识别会议研讨会论文集，第 66–74 页。

Anurag Arnab、Mostafa Dehghani、Georg Heigold、Chen Sun、Mario Lu˘ci´c 和 Cordelia Schmid。2021年。ViViT：一种视频视觉变换器。载于*IEEE/CVF*国际计算机视觉会议论文集，第6836–6846页。

Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。2020年。Wav2Vec 2.0：一种用于语音表征自监督学习的框架。神经信息处理系统进展，33：12449–12460。

Tadas Baltrusaitis、Amir Zadeh、Yao Chong Lim 和 Louis-Philippe Morency。2018年。OpenFace 2.0：面部行为分析工具包。载于*2018*年 *IEEE* 第*13*届国际自动人脸与手势识别会议（*FG 2018*），第 59–66 页。

Andrea Bandini、Sia Rezaei、Diego L Guarín、Madhura Kulkarni、Derrick Lim、Mark I Boulos、Lorne Zinman、Yana Yunusova 和 Babak Taati。2020年。一项针对神经系统疾病患者的面部运动分析新数据集。*IEEE* 生物医学与健康信息学杂志，25(4)：1111–1119。

Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao及另外1位作者。2022。Wavlm：面向全栈语音处理的大规模自监督预训练。*IEEE* 信号处理精选主题期刊，16(6):1505–1518。

Yin Chen、Jia Li、Shiguang Shan、Meng Wang和Richang Hong。2024年。从静态到动态：将地标感知图像模型适配于视频中的面部表情识别。*IEEE* 情感计算汇刊。

Kwanghee Choi、Eunjung Yeo、Kalvin Chang、Shinji Watanabe和David R Mortensen。2025年。利用自监督语音模型中的音位变体进行非典型发音评估。载于*2025*年美洲计算语言学协会分会会议论文集：人类语言技术（第*1*卷：长篇论文），第2613–2628页，美国新墨西哥州阿尔伯克基。计算语言学协会。

Wheidima Carneiro de Melo、Eric Granger 和 Miguel Bordallo Lopez。2021 年。Mdn：一种用于时空抑郁症检测的深度最大化-微分网络。*IEEE* 情感计算汇刊，14(1):578–590。

Wheidima Carneiro De Melo、Eric Granger 和 Miguel Bordallo Lopez。2024年。利用分解式多尺度时空网络进行面部表情分析。《专家系统与应用》，236:121276。

Shuchao Duan、Amirhossein Dadashzadeh、Alan Whone 和 Majid Mirmehdi。2023年。Qafe-net：利用地标热图评估面部表情质量。*arXiv* 预印本 *arXiv:2312.00856*。

Nícolas BarbosaGomes、Arissa Yoshida、MateusRoder、Guilherme Camargo de Oliveira 和 João Paulo Papa。2023年。用于识别肌萎缩侧索硬化症的面部关键点图。*arXiv* 预印本 *arXiv:2307.12159*。

Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdel-rahman Mohamed。2021年。HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习。*IEEE/ACM* 音频、语音与语言处理汇刊，29：3451–3460。

Joel Shor 和 Subhashini Venugopalan。2022 年。TRILLs- son：蒸馏通用副语言语音表征。载于 *I nterspeech 2022*，第 356–360 页。

Brian McFee。2025年。librosa/librosa：0.11.0。

Lucia Migliorelli、Daniele Berardini、Kevin Cela、Michela Coccia、Laura Villani、Emanuele Frontoni 和 Sara Moccia。2023年。一种基于云端存储转发的远程监控系统，通过视频录制分析自动评估神经系统疾病患者构音障碍的演变。《生物学与医学中的计算机》，163:107194。

Saeid Alavi Naeini、Leif Simmatis、Deniz Jafari、Diego L Guarin、Yana Yunusova 和 Babak Taati。2022。口面部评估视频的自动时间分割。发表于*2022 IEEE-EMBS* 国际生物医学与健康信息学会议（*BHI*），第01–06页。IEEE。

Michael Neumann、Hardik Kothare 和 Vikram Ramanarayanan。2024年。用于远程监测肌萎缩侧索硬化症疾病进展的多模态语音生物标志物。生物学与医学中的计算机，180：108949。

Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。2023年。通过大规模弱监督实现稳健语音识别。发表于国际机器学习会议，第 28492–28518 页。PMLR。

Andrew Rouditchenko、Yuan Gong、Samuel Thomas、Leonid Karlinsky、Hilde Kuehne、Rogerio Feris 和 James Glass。2024年。Whisper-Flamingo：将视觉特征集成到 Whisper 中，用于视听语音识别与翻译。发表于*Interspeech2024*，第 2420–2424 页。

Yuge Shi、Brooks Paige、Philip Torr 和其他 1 位作者。2019年。用于多模态深度生成模型的变分混合专家自编码器。神经信息处理系统进展，第 32 卷。

David Snyder、Daniel Garcia-Romero、Gregory Sell、Daniel Povey 和 Sanjeev Khudanpur。2018年。X 向量：用于说话人识别的稳健深度神经网络嵌入。*2018 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 5329–5333 页。

Zhan Tong、Yibing Song、Jue Wang 和 Limin Wang。2022年。VideoMAE：掩码自编码器是用于自监督视频预训练的数据高效学习器。神经信息处理系统进展，35：10078–10093。

Heng Wang 和 Cordelia Schmid。2013。利用改进轨迹进行动作识别。载于*IEEE*计算机视觉国际会议论文集，第3551–3558页。

Limin Wang、Bingkun Huang、Zhiyu Zhao、Zhan Tong、Yinan He、Yi Wang、Yali Wang 和 Yu Qiao。2023年。VideoMAE V2：通过双重掩码扩展视频掩码自编码器。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第14549–14560页。

ZhiyuWu、Xiaokang Chen、Zizheng Pan、Xingchao Liu、Wen Liu、Damai Dai、Huazuo Gao、Yiyang

Ma、Chengyue Wu、Bingxuan Wang 和另外 1 位作者。2024年。DeepSeek-VL2：用于高级多模态理解的混合专家视觉语言模型。*arXiv* 预印本 *arXiv:2412.10302*。

A 附录

在附录中，我们提供：
章节A.1：预训练模型的详细信息。章节A.2：基准数据集。章节A.3：数据预处理。章节A.4：超参数和系统配置。章节 A.5：非语音视频样本的结果。

章节A.6：可视化分析。

A.1 预训练模型的详细信息

本节详细介绍了我们研究中使用的预训练编码器。我们采用了涵盖自监督、监督和多语言训练范式的预训练语音模型。所有模型均在冻结设置下使用，以提取话语级声学表示。

语音基础模型

WavLM(Chen 等人，2022)[2]：这是一种自监督语音表示模型，旨在支持全栈语音处理。它在一个包含多种公开英语语料库、总计 9.4 万小时的数据集上进行预训练，采用掩码预测和去噪目标。

Wav2Vec2.0 (Baevski 等人，2020年)[3]：通过在潜在空间中进行对比预测，学习上下文化语音表示。它将卷积编码器与 Transformer 网络相结合，对输入的部分内容进行掩码，并优化模型，使其能够区分正样本和负样本。

Whisper (Radford 等人，2023年)[4]：这是一种多语言编码器-解码器模型，在 68 万小时的弱标注互联网音频上进行预训练，用于转录、翻译和语音活动检测。我们使用基础模型的编码器特征。

X-vector （Snyder 等人，2018）[5]：一种时延用于说话人分类的神经网络（TDNN），采用 VoxCeleb 数据集进行训练。提取的向量具有说话人区分能力，广泛应用于说话人验证和欺骗检测任务。

HuBERT (Hsu 等人，2021年)[6]：是一种自监督语音表征模型，将掩码预测与离线 k 均值聚类相结合。该模型在大规模数据集（例如 LibriSpeech 960 小时、Libri-Light 6万小时）上进行预训练，在语音识别和副语言任务方面达到了当前最先进的水平。它提供多种配置（基础版、大型版、超大型版），我们使用基础版的冻结模式来提取话语级嵌入。

TRILLsson(Joel Shor 和 Subhashini Venu-gopalan，2022)[7]：是一种轻量级自监督语音模型，专为副语言语音任务设计，例如情绪识别、说话人识别和合成语音检测。该模型通过对 CAP12 Conformer 模型进行知识蒸馏而创建，后者曾使用 YouTube 上 90 万小时的语音数据进行训练。它使用来自 Libri-Light 和 AudioSet 的 5.8万小时公开语音数据进行训练。

视觉基础模型

Video-MAE (Tong 等人，2022)[8]：采用掩码自编码策略，以较高的掩码比例（90–95%）处理时空立方体。该模型使用原始 ViT 主干网络作为编码器，并通过重建这一自监督前置任务进行训练。

VideoMAEV2 (Wang 等人，2023年)[9]：是一种可扩展的自监督视频预训练框架，通过采用双重掩码策略，同时对编码器和解码器的令牌进行掩码，以降低内存和计算负载。该框架采用渐进式训练，先在百万级无标签视频语料库上进行无监督学习，再在带标签混合数据集上进行后训练。我们在冻结设置下使用 ViT-B 变体提取片段级面部特征。

[2]https://huggingface.co/microsoft/wavlm-base[3]https://huggingface.co/facebook/wav2vec2-base[4]https://huggingface.co/openai/whisper-base[5]https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

[6]https://huggingface.co/facebook/hubert-base-ls960[7]https://www.kaggle.com/models/google/trillsson[8]https://huggingface.co/docs/transformers/en/model_doc/videomae[9]https://huggingface.co/OpenGVLab/VideoMAEv2-基础版

ViViT (Arnab 等人，2021年)[10]：是一种纯 Transformer 架构，使用 tubelet 嵌入对空间和时间进行因式分解自注意力。我们采用从 ViT 图像权重初始化的 ViViT-B/16×2 变体。

Deepseek-VL2 (Wu 等人，2024年)[11]：这是一种混合专家（MoE）视觉语言模型，旨在实现先进的多模态理解。该模型针对多种任务进行训练，涵盖视觉语言对齐、多模态预训练和监督微调，包括视觉定位、光学字符识别和文档理解。本研究采用冻结模式下的视觉编码器，从面部视频片段中提取时间对齐的视觉嵌入。

Kinematic[12]：使用 OpenFace2.0 工具包提取（Baltrusaitis 等人，2018），该工具包可提供三维标志点位置、头部姿态（偏航角、俯仰角、翻滚角）、注视方向和面部动作单元（AU）。

Temporal：使用在 ImageNet 上预训练的 ResNet18 模型提取帧级外观嵌入。随后在这些特征之上采用时序注意力网络（TANN），以建模帧间依赖关系。

A.2 基准数据集

在本研究中，我们使用了由（Ban- dini 等人，2020年）收集的多伦多神经面部（TNF）数据集中的数据。该数据集汇集了经过精心采集的高质量视频录制资料，记录了健康成年人和患有神经功能障碍者的口面部动作。36名认知功能完整的志愿者（11名肌萎缩侧索硬化症患者、14名卒中后患者、11名年龄匹配对照组）分别完成了一组包含9项言语和非言语任务的测试——从快速音节重复（“/pa/”“/pa-ta-ka/”）和绕口令“Buy Bobby a Puppy”，到最大张口、噘唇动作和表情性微笑——均在标准化光照和摄像机距离（30–60cm，640 × 480像素，~50帧/秒）条件下进行。两名言语语言病理学专家使用5分制，对每次试验的对称性、运动范围、速度、变异性和疲劳程度进行评分，从而形成了一套可靠的临床评分（总范围5–25；评定者间$\kappa = 0.33$–0.61）。对于3300多帧经过精心选取的帧，人工标注了68个面部关键点（评定者间归一化均方根误差 = $1.36 \pm 0.46\%$），并精确地提取了面部边界框。数据集还随录音录像提供了丰富的元数据，包括受试者人口统计学信息、任务标签、视频时间信息和临床医生评分。通过将受控采集流程与全面的真实值标注和临床评估相结合。尽管该数据集并未公开发布，我们获准访问。据我们所知，这是目前唯一已知的、包含与神经面部疾病相关的、由专家进行临床标注的同步高质量面部视频和音频录音资源。

A.3 数据预处理

我们执行预处理步骤，以确保音频和视频流的数据质量、一致性和对齐。对于面部视频，我们使用二维人脸对齐网络（2D-FAN）[13] 检测每帧中的 68 个关键点。这有助于清晰地识别人脸，并确保其处于可见且居中的位置。对于音频，我们在话语级别应用振幅归一化和强制对齐，并使用片段级时间戳；预处理通过 `librosa`（McFee，2025年）实现。

A.4 超参数和系统配置

用于单模态建模的 CNN 架构以两个一维卷积块为起始。第一个卷积块使用 256 个滤波器，核大小为 3，随后进行批归一化和最大池化（池化大小 = 2）。第二个卷积块使用 128 个滤波器，核大小同样为 3，随后进行批归一化和最大池化（池化大小= 2）。展平后的输出被输入一个由三个全连接层组成的全连接网络，这些层分别包含 256、128 和 64 个神经元，最后接一个任务专用输出层（Softmax 或回归头）。使用单独的预训练表征的 CNN 模型，其可训练参数量为 80 万至 120 万，具体取决于提取的嵌入维度。由于增加了 Transformer 和融合层，融合实验的参数量增至 350 万至 650 万。我们使用 TensorFlow 框架实现所有模型，并在 NVIDIA RTX 4050 GPU 上进行训练和评估。

A.5 非语音视频样本上的结果

我们展示了所有 VFM+SFM 组合在非语音视频

[10]https://huggingface.co/docs/transformers/en/ 模型_doc/vivit[11]https://github.com/deepseek-ai/DeepSeek-VL2[12] https://github.com/TadasBaltrusaitis/OpenFace

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

TNF 数据集样本上的完整评估结果。表7报告了采用全连接网络和卷积神经网络骨干网络的各视频基础模型的分类和回归指标，其中 DeepSeek-VL2 的准确率最高（89.26 %），F1 分数也最高（88.29 %）。表8展示了视频基础模型与预先计算的语音基础模型在纯视频输入上的简单嵌入拼接结果，表明 DS +X-vector 即使没有音轨，也能达到 87.24 % 的准确率和 86.01 % 的 F1 分数。最后，表9 给出了我们的DIVINE 融合框架在所有模型配对上的完整结果，其中 DS +TR 表现最佳，准确率达 92.58 %，F1 分数达 91.63 %（平均绝对误差 = 3.84，均方根误差 = 5.55）。这些表格详细展示了模型在纯视觉条件下的表现，并对正文中的简要总结作了补充。

A.6 可视化分析

A.6.1 混淆矩阵

图 3展示了在我们的 TNF 测试场景中评估的八种代表性 DIVINE 配置的混淆矩阵：(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；以及 (h) DeepSeek-VL2 + WavLM。这些矩阵展示了关键模型配对之间的分类一致性和错误模式。

	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

表7：视频基础模型（VFMs）在非语音视频样本上的性能。所有数值均以百分比（%）表示，并取5折交叉验证的平均值。

组合	非语音测试视频			
	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

表8：在非
语音测试视频上的所有 VFM+语音基础模型组合的简单拼接性能。所有数值均以百分比（%）表示，并取平均值采用5折交叉验证。

组合	非语音测试视频			
	A ↑	F1 ↑	R ↓	M ↓
	VFM +语音基础模型			
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

表9：所提出的**DIVINE**框架在非语音测试视频上针对 VFM+SFM 的组合表现。所有数值均以百分比（%）报告，并取 5折交叉验证的平均值。



图3：DIVINE 配置表示。每个矩阵展示了综合诊断和严重程度类别中的真实类别分布与预测类别分布：(a) DeepSeek-VL2+TRILLsson；(b) DeepSeek-VL2+X-vector；(c) DeepSeek-VL2+X-vector（仅测试视频）；(d) DeepSeek-VL2+TRILLsson（仅测试音频）；(e) ViViT（多任务）；(f) WavLM；(g) Kinematic（多任务）；以及 (h) Kinematic（分类）。这些矩阵突出了每种融合配对的分类一致性和错误模式。