1 Multimodal Generative Model for Detecting Anxiety & Stress

1.1 Objective

Design a multimodal *generative* model fusing audio, visual, textual, and *optionally* physiological data to detect anxiety and stress.

1.2 Data Representation and Notations

- **Audio:** $X_A \in \mathbb{R}^{T_A \times d_A}$
- **Visual:** $X_V \in \mathbb{R}^{T_V \times H \times W \times C}$
- **Textual:** $X_T \in \mathbb{R}^{T_T \times d_T}$
- **Physio (Optional):** $X_P \in \mathbb{R}^{T_P \times d_P}$

1.3 Model Architecture

1.3.1 Modality-specific Encoders

$$\begin{aligned} h_A &= \text{Encoder}_A(X_A;\theta_A), & h_V &= \text{Encoder}_V(X_V;\theta_V), \\ h_T &= \text{Encoder}_T(X_T;\theta_T), & h_P &= \text{Encoder}_P(X_P;\theta_P) \text{ (optional)} \end{aligned}$$

1.3.2 Multimodal Fusion with Attention

$$H = \begin{cases} [h_A; h_V; h_T; h_P] & \text{if } X_P \text{ available} \\ [h_A; h_V; h_T] & \text{otherwise} \end{cases}, \qquad \alpha = \text{softmax}\big(HW_q(HW_k)^\top / \sqrt{d_h}\big), \quad h^* = \frac{1}{M}\sum_{i=1}^{M}(\alpha H W_v)_i.$$

1.3.3 Generative Component (VAE)

$$\mu, \sigma^2 = f_{\text{enc}}(h^*;\phi), \quad z \sim \mathcal{N}(\mu, \sigma^2 I), \quad \hat{X}_m = f_{\text{dec}_m}(z;\psi_m).$$

1.3.4 Prediction Head

$$\text{Classification: } y = \text{softmax}(W_c h^* + b_c), \qquad \text{Regression: } y = W_r h^* + b_r.$$

1.4 Proposed Novel Contributions

1. **Cross-Modal Diffusion Imputation (CDI).** A conditional diffusion process $q_\theta(X_{\text{miss}} | X_{\text{avail}})$ learns to generate *entire missing modalities* given the available ones, replacing the VAE decoder when data are incomplete. Diffusion-based imputation has not yet been explored for stress/anxiety analytics, offering state-of-the-art perceptual quality and calibrated uncertainty.

2. **Uncertainty-Aware Attention Fusion (UAF).** Each modality embedding is accompanied by a variance estimate σ_m^2 (obtained via Monte-Carlo dropout). Fusion weights are set to $\tilde{\alpha}_m \propto \alpha_m/\sigma_m^2$, down-weighting noisy or low-confidence modalities *on-the-fly*.

3. **Self-Supervised Cross-Modal Pre-training (SCP).** Encoders are first trained with a contrastive objective $\mathcal{L}_{\text{SCP}} = -\log \frac{\exp(\langle h_i^{(m)}, h_i^{(n)} \rangle/\tau)}{\sum_j \exp(\langle h_i^{(m)}, h_j^{(n)} \rangle/\tau)}$ across all unordered modality pairs (m, n). This leverages *unlabelled* video-interview corpora before fine-tuning for stress labels, a step missing from prior work.

4. **Subject-Level Bayesian Adaptation (SBA).** For longitudinal use, a lightweight Bayesian linear head updates $p(y \mid h^*, \kappa)$ via conjugate priors when a user contributes a few calibration samples, providing personalised baselines without re-training the entire network.

Why Novel? To the best of our knowledge, *no existing anxiety/stress system combines diffusion-based cross-modal imputation, uncertainty-aware fusion, self-supervised contrastive pre-training,* and *personalised Bayesian adaptation in a single framework.* Each component individually moves beyond StressNet, MuSe Transformer, and other baselines; together they form a distinctive, practically valuable pipeline.

1.5 Algorithm Including Novel Components

Algorithm 1 Multimodal Generative Anxiety/Stress Detection with Novel Additions

Input: Available modalities $\{X_m\}_{m \in \mathcal{M}_{\text{avail}}}$
Output: Prediction y, reconstructed/ imputed $\hat{X}_m$
for $m \in \mathcal{M}_{\text{avail}}$ **do** ▷ Encode & uncertainty
    $h_m, \sigma_m^2 \leftarrow \text{Encoder}_m(X_m)$
end for
if $\exists$ missing modality **then**
    $\hat{X}_{\text{miss}} \leftarrow \textbf{CDI}\big(X_{\text{avail}}\big)$ ▷ Diffusion imputation
    $h_{\text{miss}}, \sigma_{\text{miss}}^2 \leftarrow \text{Encoder}_{\text{miss}}(\hat{X}_{\text{miss}})$
end if
$H \leftarrow \text{stack}(\{h_m\}), \quad \alpha \leftarrow \textbf{UAF}(H, \{\sigma_m^2\})$
$h^* \leftarrow \alpha H W_v$ ▷ Uncertainty-aware attention
$(\mu, \sigma^2) \leftarrow f_{\text{enc}}(h^*)$, $z \sim \mathcal{N}(\mu, \sigma^2 I)$
for $m \in \mathcal{M}$ **do** $\hat{X}_m \leftarrow f_{\text{dec}_m}(z)$
end for
$y \leftarrow \text{BayesHead}_{\textbf{SBA}}(h^*)$
Compute $\mathcal{L} = \mathcal{L}_{\text{pred}} + \beta\text{KL} + \gamma\mathcal{L}_{\text{recon}} + \lambda\mathcal{L}_{\text{SCP}}$
Update parameters $\{\theta, \phi, \psi, W\}$ via back-prop

1.6 Real Research-Paper Baselines

Baseline	**Modalities**	**Reference**
StressNet	Physio, Audio	He et al. 2022
MuSe Transformer	Audio, Visual, Text	Stappen et al. 2021
AVEC CNN-LSTM	Audio, Visual	Valstar et al. 2016
DAIC-WOZ Transformer	Audio, Visual, Text	Mallol-Ragolta et al. 2019
WESAD CNN-LSTM	Physiological	Schmidt et al. 2018

1.7 Datasets & Metrics

Datasets: DAIC-WOZ, WESAD, MuSE, ForDigitStress.
Metrics: Accuracy, F1, AUC-ROC, RMSE, FID, etc.

1.8 Interpretability & Ethics

Attention maps and diffusion uncertainty visualisations aid interpretability; privacy-preserving storage is mandatory for physiological/video data.