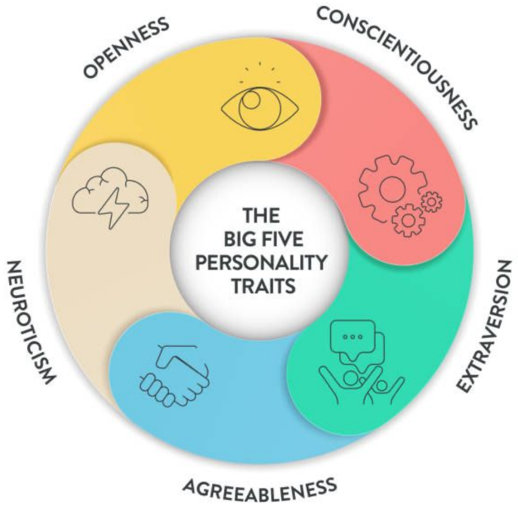
ReMoDiff: Restoring (missing) Modalities with Diffusion

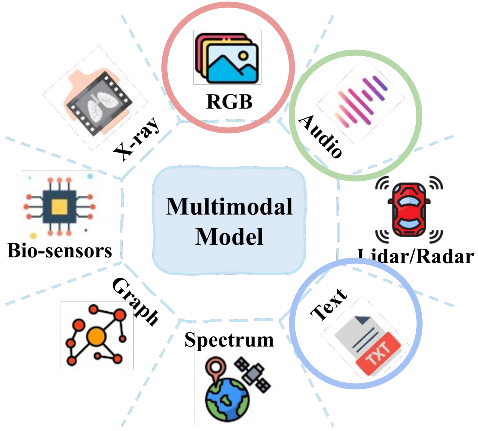


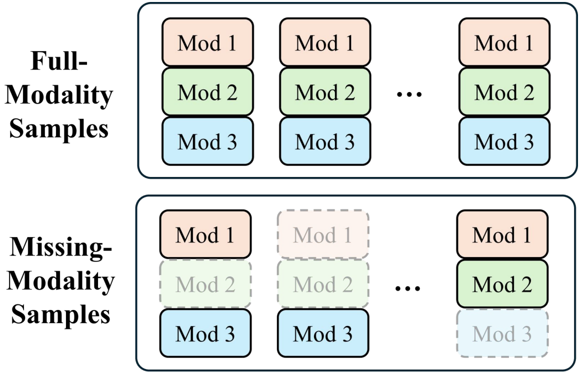
Diffusion-Based Modality Reconstruction for Personality Trait Determination

Arpita Sahu · Rishabh Agrawal · Yash Gawankar

PROBLEM DEFINITION

- **Real world data often contains incomplete or missing modalities:**
  - Sensor malfunctions
  - Environment conditions
  - Privacy concerns
- **Aim**
  - Reconstruct missing modalities + perform a downstream task
- **We are exploring**
  - How well **diffusion models** can reconstruct missing modalities in multimodal datasets
  - The impact of different missing modality scenarios (e.g., missing speech, missing vision) on prediction performance

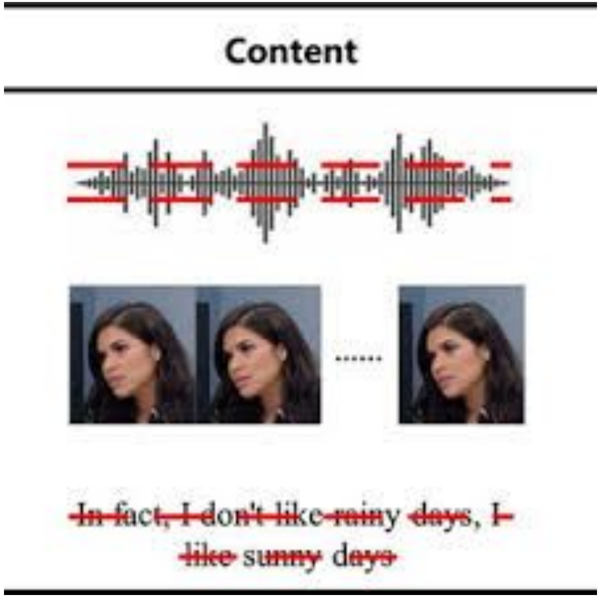




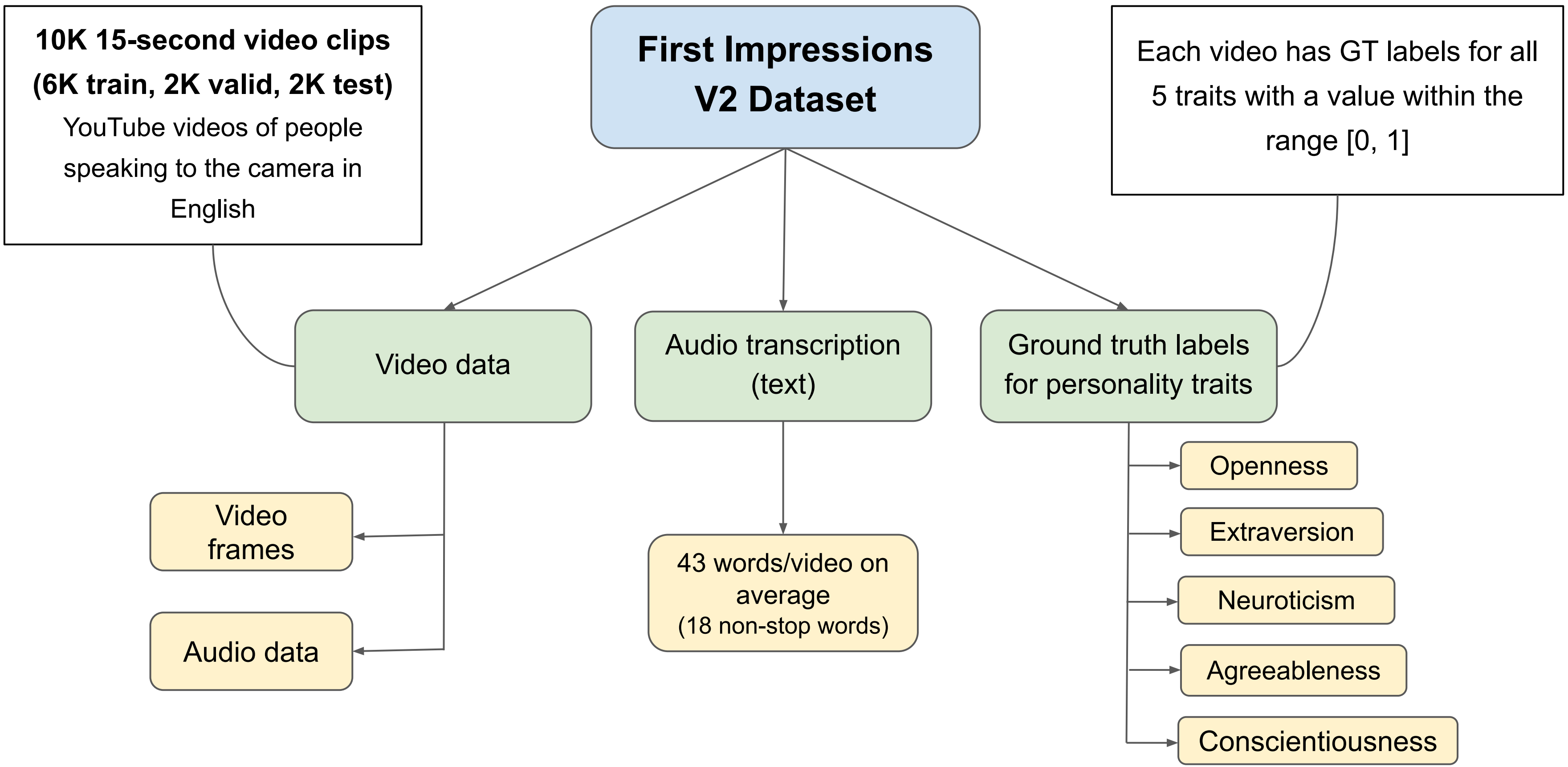
BACKGROUND - Reconstructing Missing Modalities

Deep Multimodal Learning with Missing Modalities (MLMM) methods

1. Data Imputation
   a. Composition - zeroes, random, frame repetition, KNN
2. Data Generation
   i. **VAEs → Our Baseline**
   ii. GANs
   iii. **Diffusion Models → Our Selected Approach**
   iv. Attention + maxpooling
3. Feature Space Engineering - regularization & correlation
4. Architecture Engineering - attention based, distillation & graph based
5. Model Selection - ensembles and dedicated models

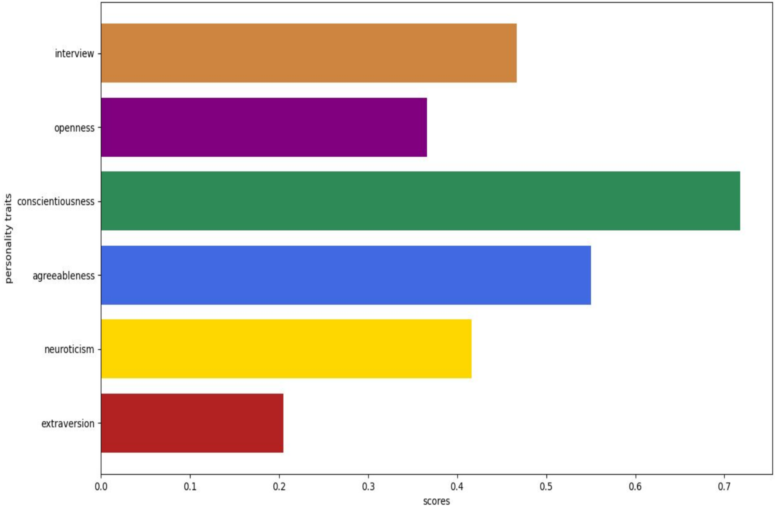


DATASET


10K 15-second video clips
(6K train, 2K valid, 2K test)
YouTube videos of people speaking to the camera in English
First Impressions V2 Dataset
Each video has GT labels for all 5 traits with a value within the range [0, 1]
Video data
Audio transcription (text)
Ground truth labels for personality traits
Video frames
Audio data
43 words/video on average
(18 non-stop words)
Openness
Extraversion
Neuroticism
Agreeableness
Conscientiousness

Ground truth labels

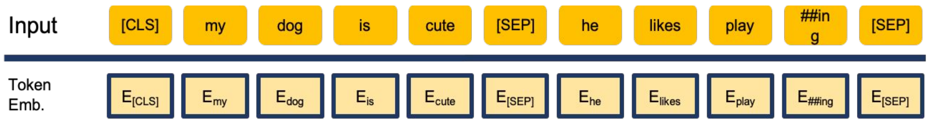
- extraversion 0.20
- neuroticism 0.416
- agreeableness 0.55
- conscientiousness 0.72
- openness 0.367
- interview 0.467



Audio transcription:

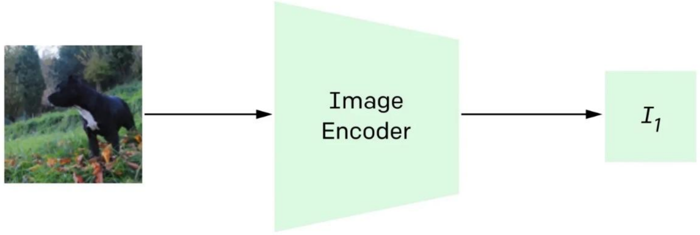
"Procedural safeguards. There's a lot of ... They have to be served very formally. There's certain documents that have to be filed. You have to appear at certain times, and there are very formal documents that have to be filled out. They're very concerned about protecting both sides, and that means there's ..."

FEATURE EXTRACTION

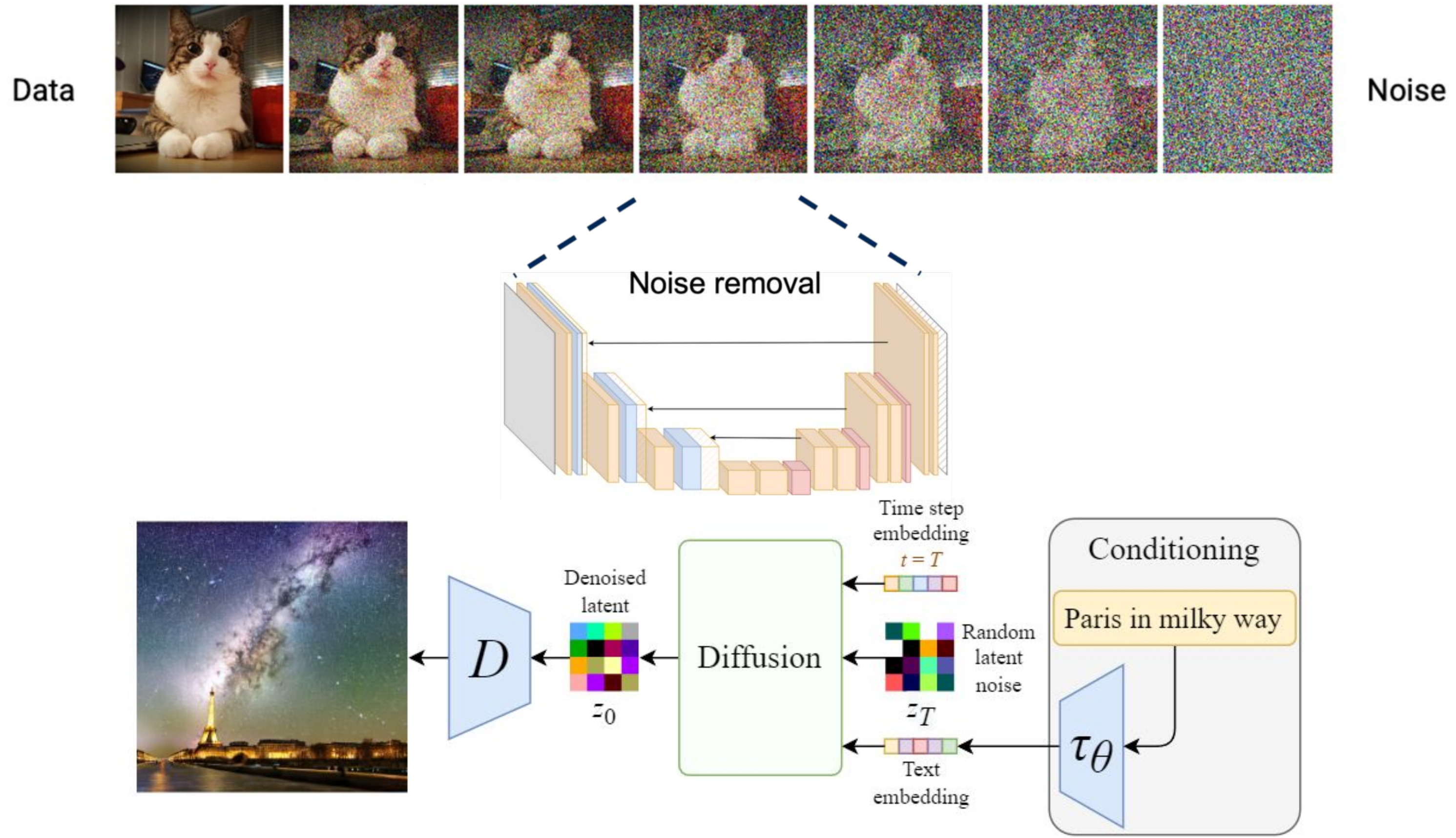


Each input video is divided into 5 segments.

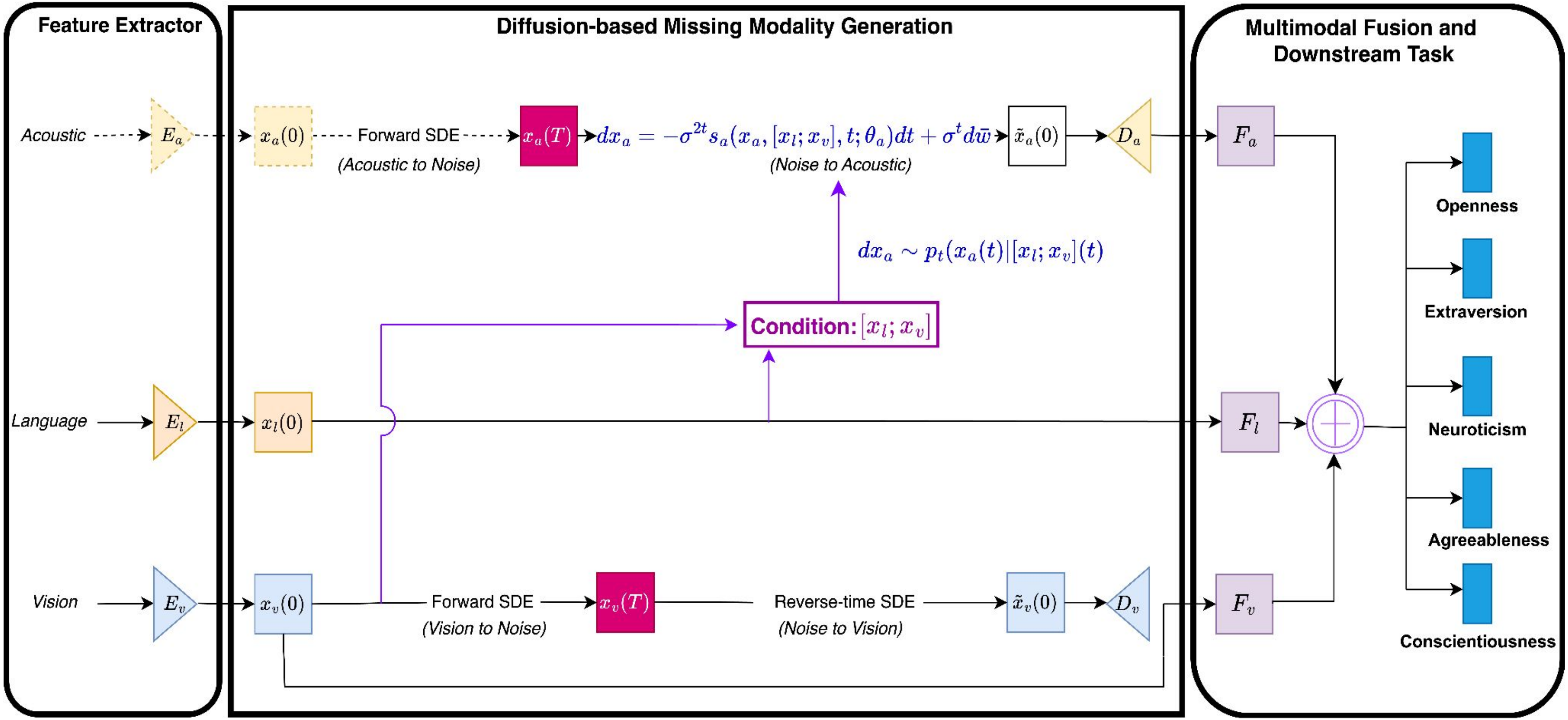
1. **Text** → **RoBERTa** fine-tuned for emotion recognition
   - Feature embedding for text = **768 dim**
2. **Audio** → **wav2vec2** fine-tuned on Speech Emotion Recognition (SER) task
   - Size audio segment's feature embedding = **149 x 1024 dim**
   - Number of timesteps/clip = 149
   - Feature vector size/timestep = 1024
3. **Video frames** → **CLIP ViT**
   - Each frame's feature embedding = **512 dim**



METHODOLOGY - Diffusion Models



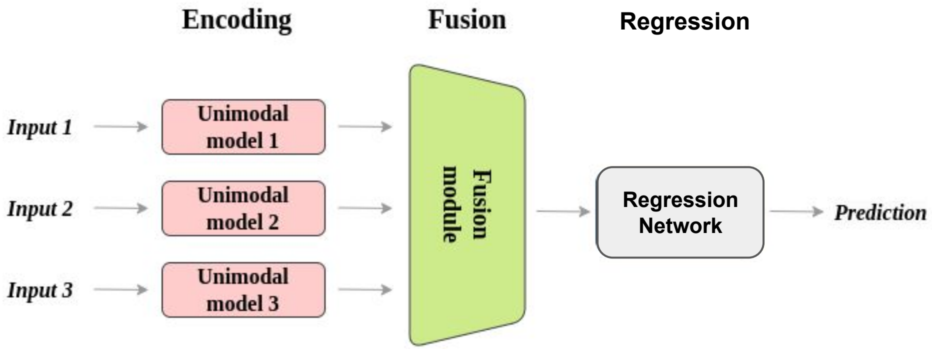
METHODOLOGY - Our Approach



Methodology: Data Fusion & Downstream Inference

Reconstructions

a. Conditional VAE - Baseline
   i. C(A + T) → V’
   ii. C(V + T) → A’
b. Diffusion
   i. D(A + T) → V’’
   ii. D(V + T) → A’’



Downstream Inference

1. Expected Upper Bound
   a. Original (A + V + T)
2. Reconstructions
   a. O(A + T) + V’/V’’
   b. O(V + T) + A’/A’’
3. Expected Lower Bound
   a. Original (A + T)
   b. Original (V + T)

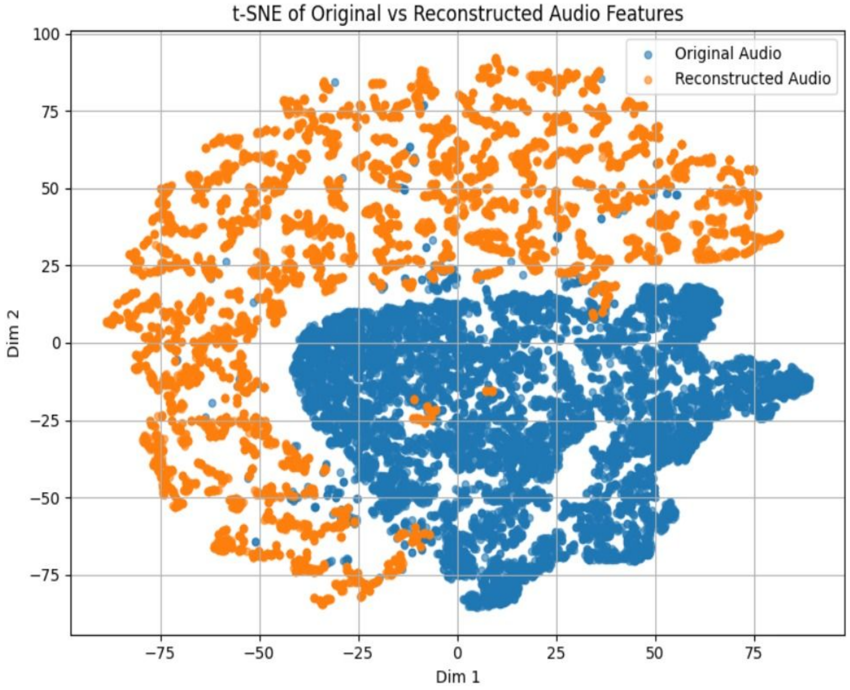
→

1. Temporal Modeling - LSTM
   a. For early & Late fusions
2. Data Fusion
   a. Early
   b. Late
   c. Transformer

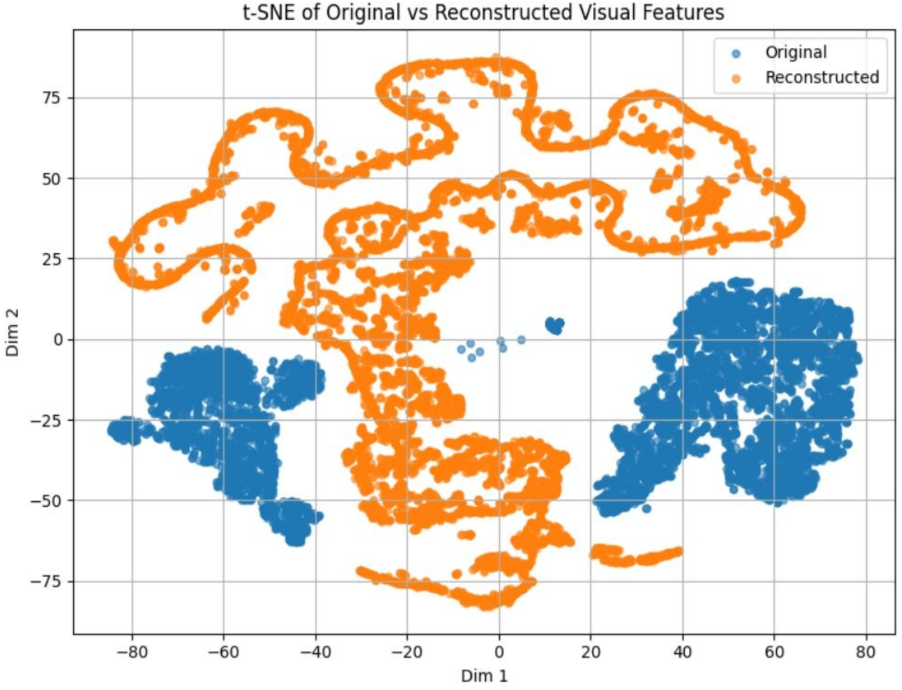
   +

3. MLP Regressor

RESULTS: CVAE Reconstruction

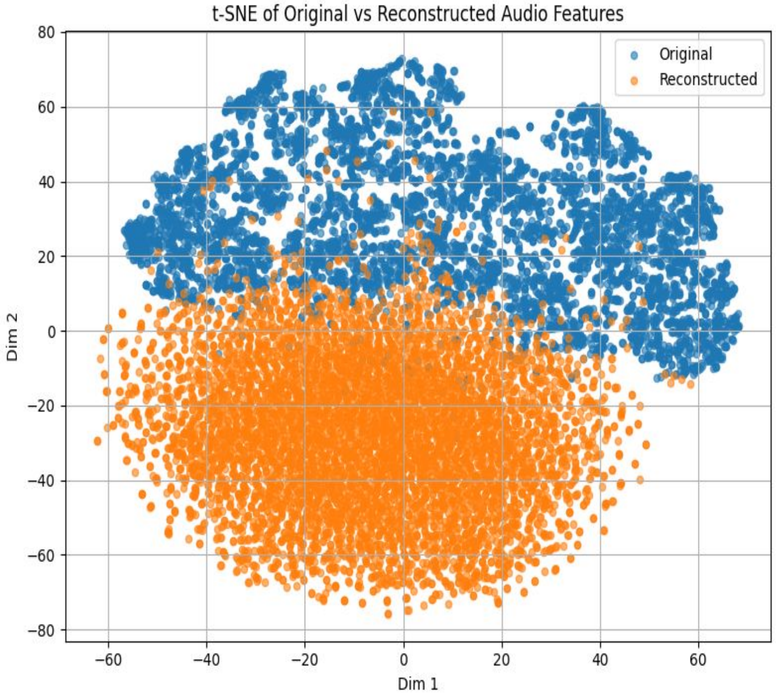


Audio

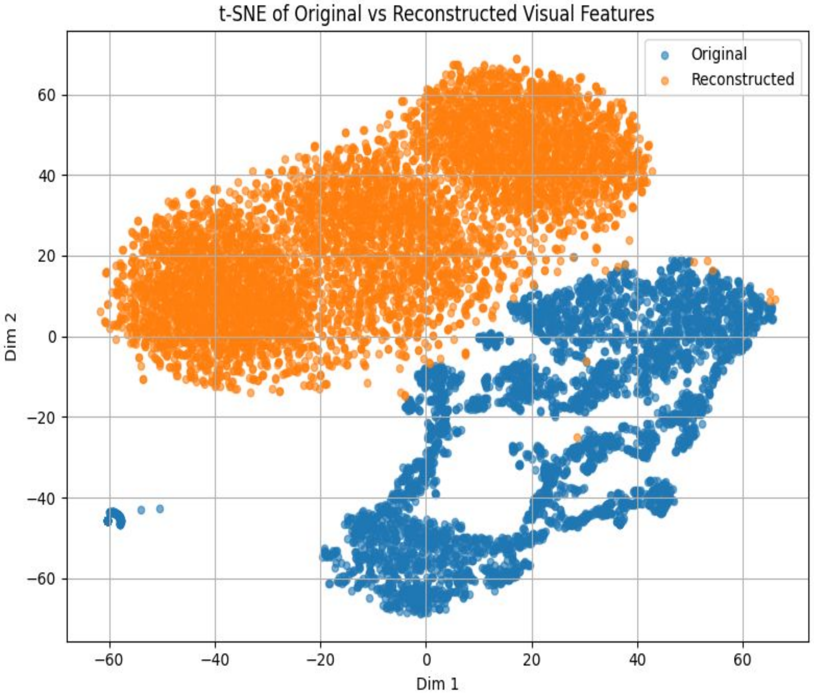


Visual

RESULTS: Diffusion Model Reconstruction



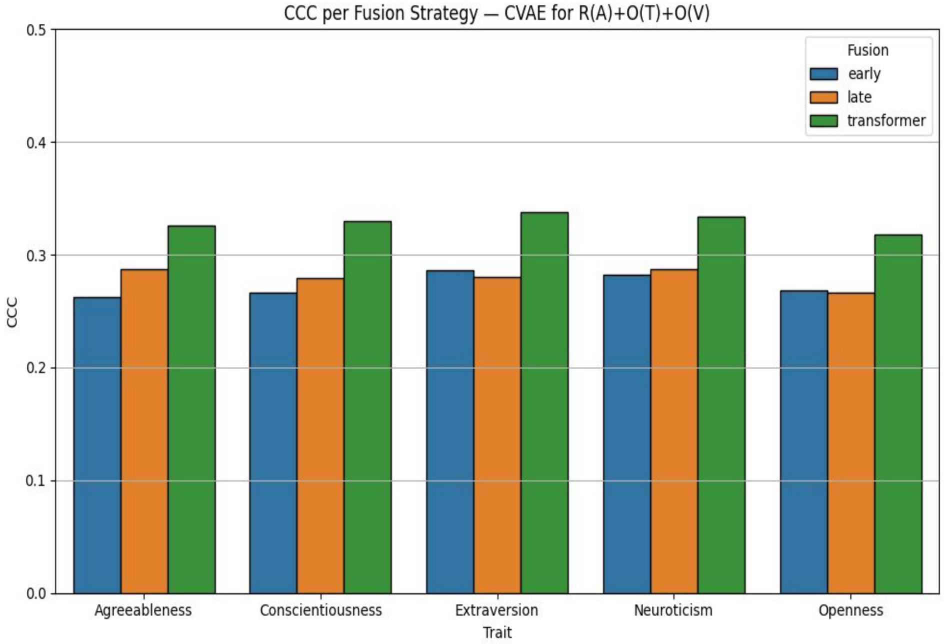
Audio



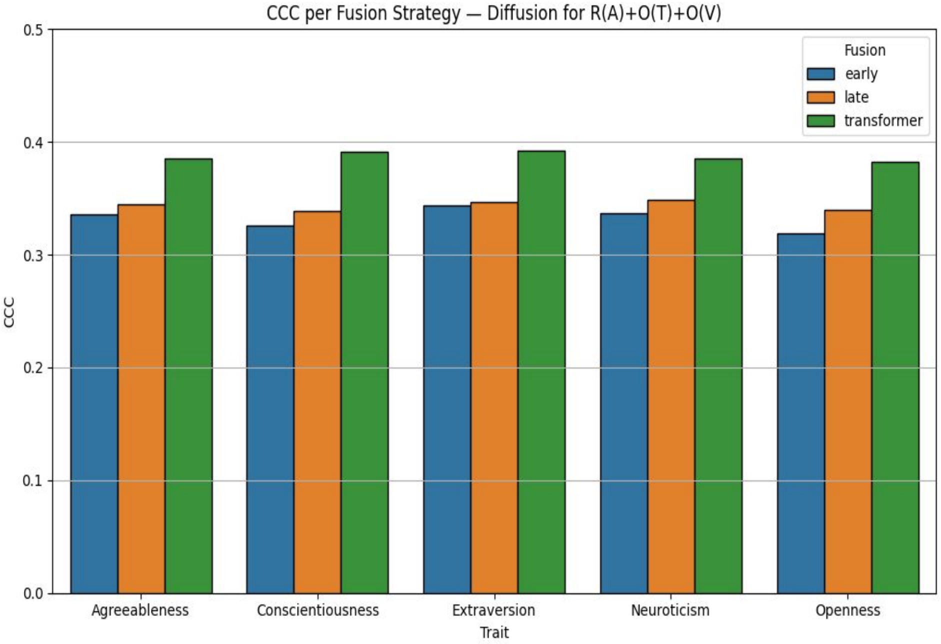
Visual

RESULTS: Comparison of Different Fusion Methods

Task: Original Text and Original Visual Features + Generated Audio Features

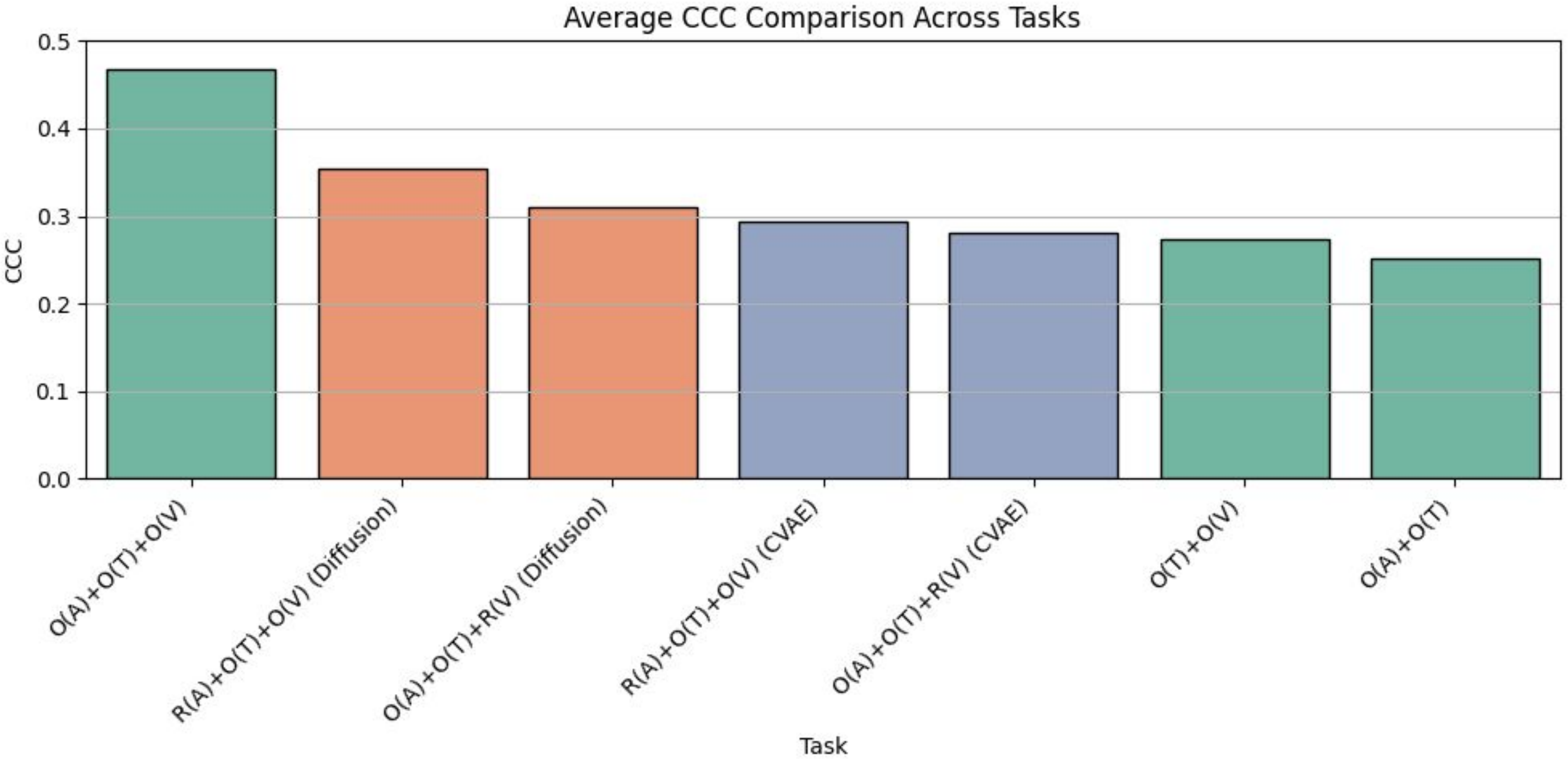


CVAE



Diffusion

RESULTS: Downstream comparison across Tasks



SUMMARY & CONTRIBUTIONS

- **Yash Gawankar**
  - Conditional Variational AutoEncoder (CVAE) - Baseline
    - Development & Training
  - Data Fusion & Downstream Inference
    - Original, CVAE, Diffusion reconstructions with Early and Late Fusion
- **Arpita Sahu**
  - Dataset
    - Exploration, Feature Extraction
  - Data Fusion & Downstream Inference
    - CVAE reconstructions with Model Level (Transformer) Fusion
- **Rishabh Agrawal**
  - Diffusion
    - Development & Training
  - Data Fusion & Downstream Inference
    - Diffusion Reconstructions with Transformer Fusion

Thank you! Any questions?