Multi modal learning

Focusing On Emotion Recognition

Team members:

Linoy Halifa, Dr. Ezra Ella , Bezalel Itzhaky

Supervisor: Dr. Yehudit Aperstein

Presentation in bullets

- Bentzi
  - Introduction to multimodal in context of deep learning
    - Introducing the modalities
    - Applications
    - Structured and Unstructured data
    - Representations
- Linoy
  - Introduction to multimodal in context of deep learning
    - Fusion strategy
  - Multimodal architectures
    - Image to text
    - Video Description DRL & VQA
- Ezra
  - Article: Multimodality in Emotion Understanding
  - Challenges
  - Our project (include the project challenges)
  - future research

Introduction to multimodal in context of deep learning

Introducing the modalities

Modalities are distinct modes of data transfer

- Each sense is a modality
  - - a unique way of receiving information.
- Together, they help us perceive and understand the world around us

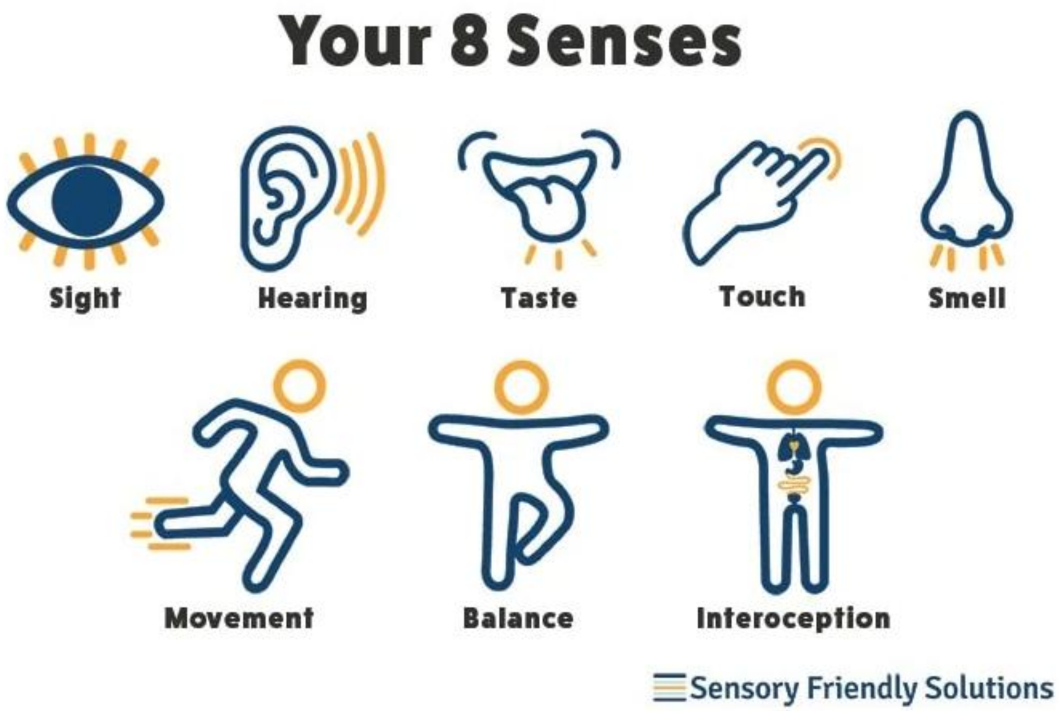


Image source: [Sensory Friendly Solutions]

Multimodal: combining different modalities

- Multimodal means to combine different channels of information simultaneously to understand our surroundings

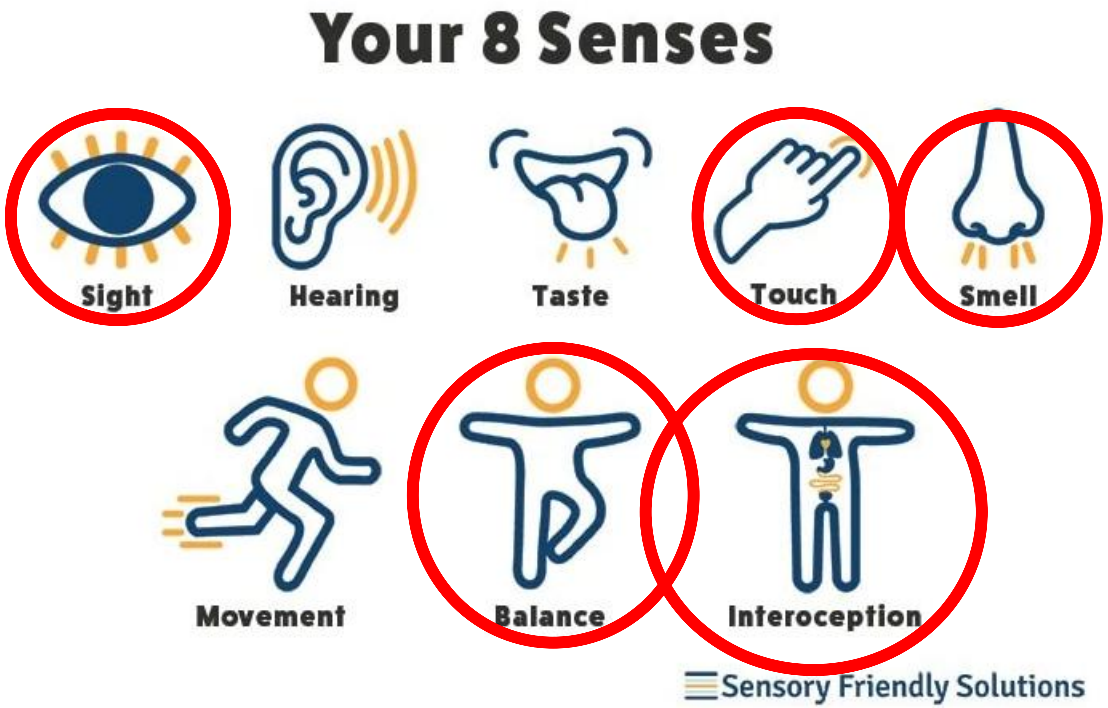


Image source: [Sensory Friendly Solutions]

Computerized modalities cover a wide range of input types

Visual	Auditory	Textual	Haptic	Biometric	Motion	Environ.	GUI	BCI
Images	Sound types	Written text	Tactile input	Heart rate	Movement	Ambient noise levels	click-based interactions	Neural signal decoding
Video	Music	Natural input text	Force feedback	EEG/brainwa ves	Gait analysis	temperatur e	Mouse and keyboard	EEG-based commands
Gestures	Speech		Vibration	GSR	Kinect	Light		
Facial expressions				Eye-tracking		humidity		
				Temperature		GPS & location		

BCI - Brain-Computer Interfaces
GSR - Skin conductance
Kinect- is a motion-sensing device
Haptic - Touch based

Computerized modalities cover a wide range of input types

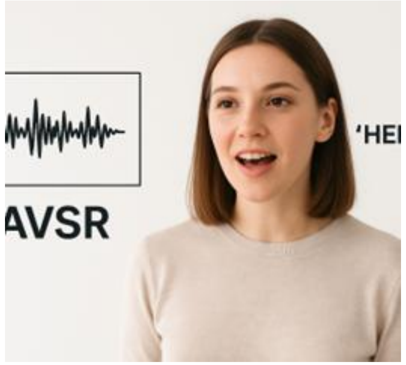
Visual	Auditory	Textual	Haptic	Biometric	Motion	Environ.	GUI	BCI
Images	Sound types	Written text	Tactile input	Heart rate	Movement	Ambient noise levels	click-based interactions	Neural signal decoding
Video	Music	Natural input text	Force feedback	EEG/brainwaves	Gait analysis	temperature	Mouse and keyboard	EEG-based commands
Gestures	Speech		Vibration	GSR	Kinect	Light		
Facial expressions				Eye-tracking		humidity		
				Temperature		GPS & location		

- We are focusing on visual , text and Audio modalities

Introduction to multimodal in context of deep learning

Applications

Applications are wide and almost at any field



Education

Smart platforms adapt to diverse learners by analyzing behavior and responding in real time through text, voice, video, and touch.

Medical

Combining physiological signals with images or video to detect medical conditions

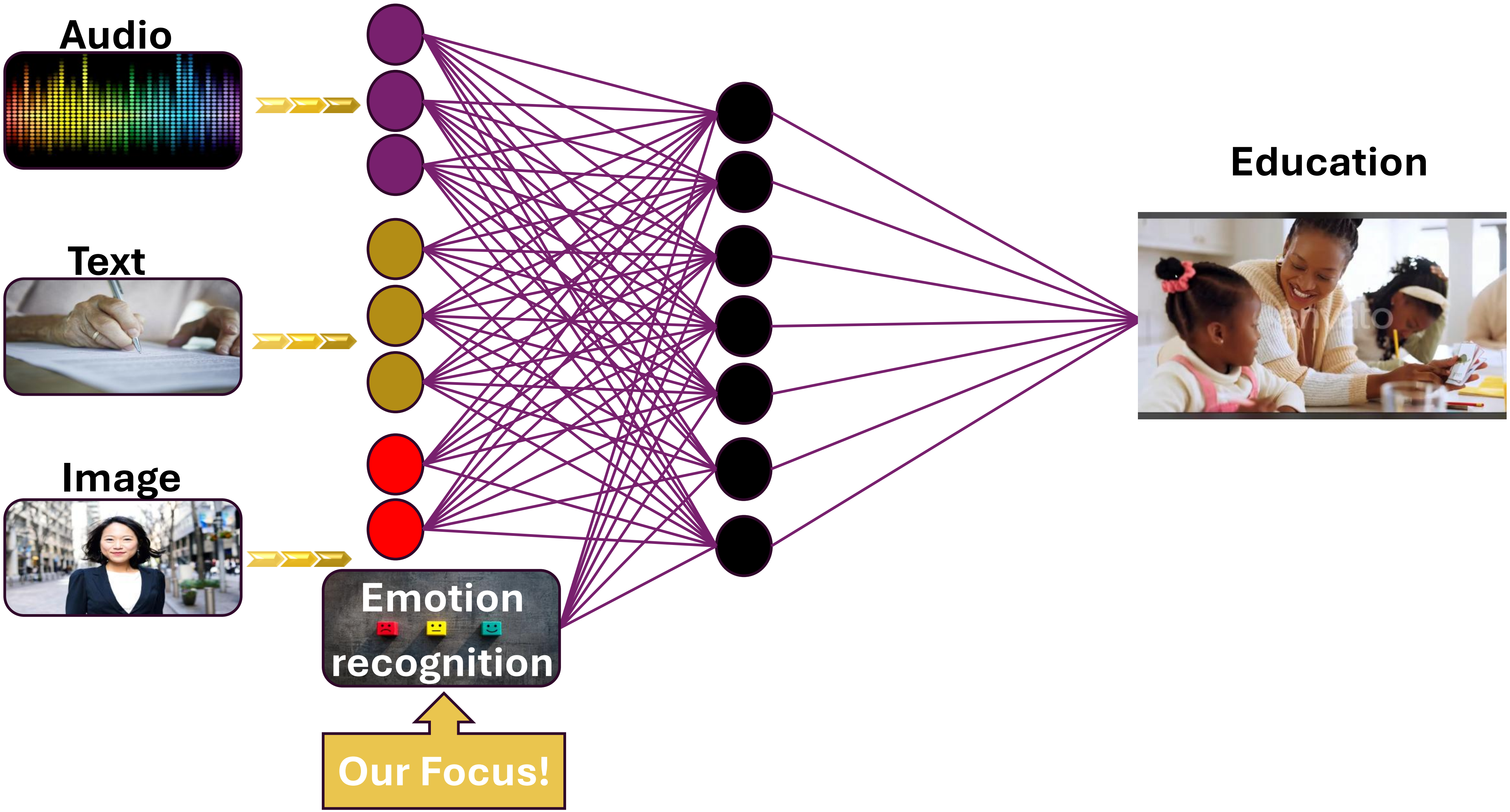
Audio visual speech recognition (AVSR)

combines both audio signals (like voice) and visual cues (like lip movements) to recognize and understand spoken language.

Autonomous driving

Using sensors such as cameras, LIDAR, and radar for accurate and safe environmental analysis.

Applications are wide and almost at any field

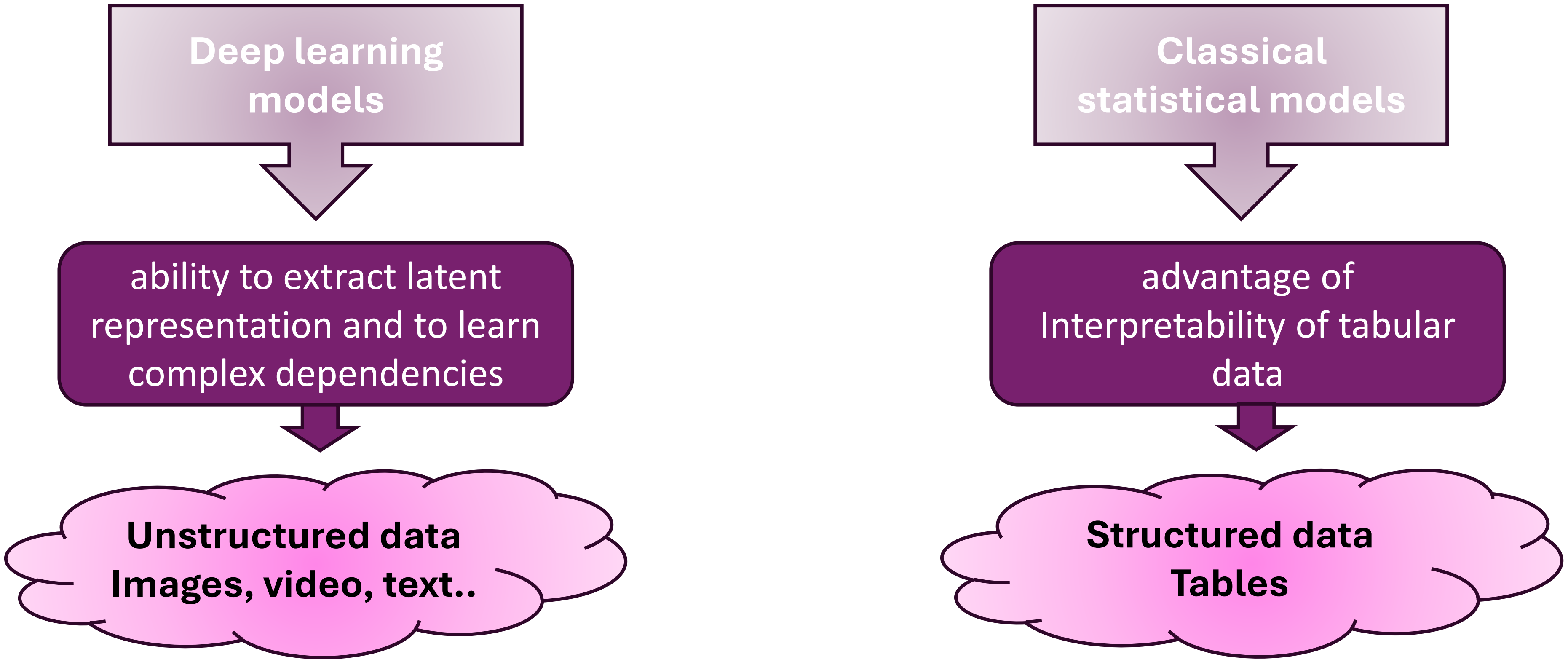


Introduction to multimodal in context of deep learning

Structured and unstructured data

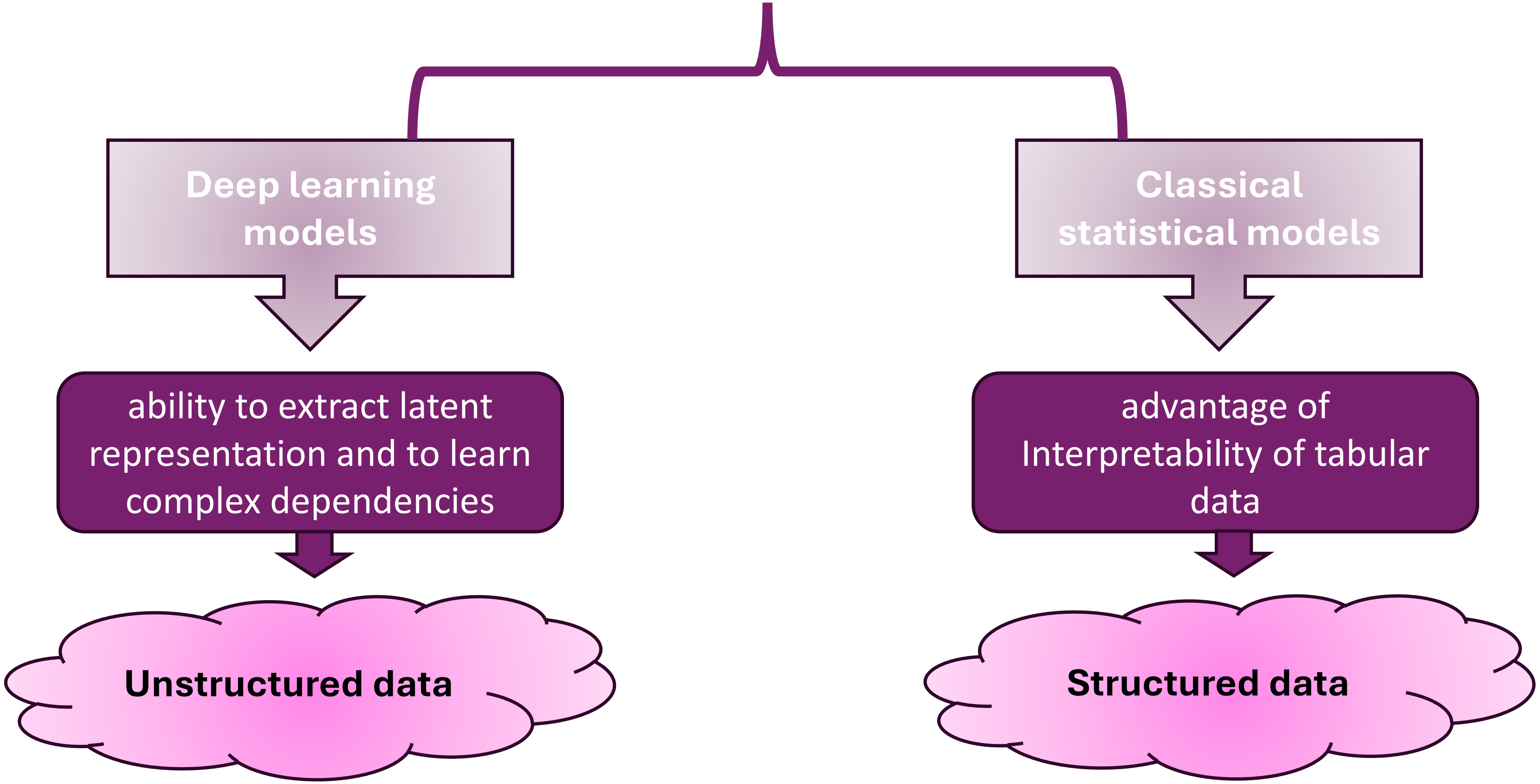
Structured and Unstructured data

- structured and unstructured data substantially differ in certain aspects such as dimensionality and interpretability.
- Require various modeling approaches that are particularly designed for the special characteristics of the data types



Structured and Unstructured data

Discarding one or the other data modality makes it likely to miss out on valuable insights and potential performance improvements



Introduction to multimodal in context of deep learning

Representations

Image Representation

Fig. 11. FACS Action Units for Happiness, Sadness, and Surprise.

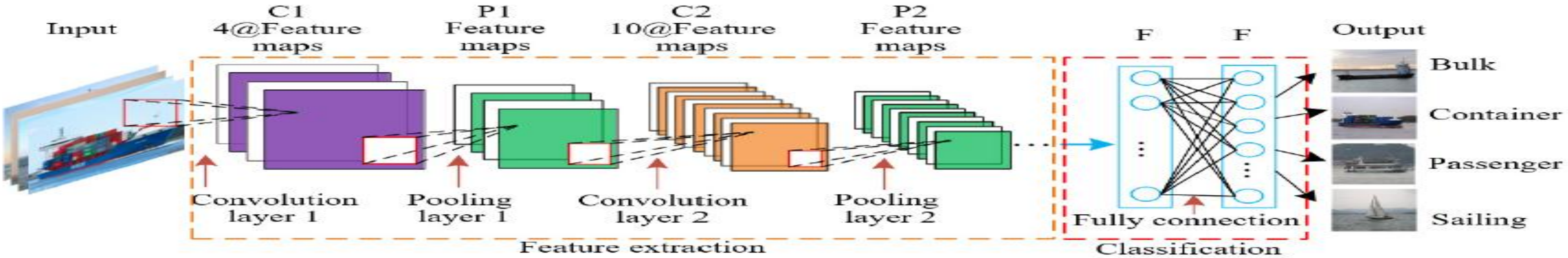


Figure 1. Typical convolutional neural network (CNN) structure.

Bottom Image source: “Multi-Feature Fusion with Convolutional Neural Network for Ship Classification in Optical Images ”/ Yongmei Ren , Jie Yang , Qingnian Zhang and Zhiqiang Guo / 2019

Top Image source: “Survey on multimodal approaches to emotion recognition”/ A. Aruna Gladys *, V. Vetriselvi / 2023

Text Representation
NLP

Fig. 14. Text Representations.

Image source: "Survey on multimodal approaches to emotion recognition"/ A. Aruna Gladys *, V. Vetriselvi / 2023

Representations:

Audio Representation

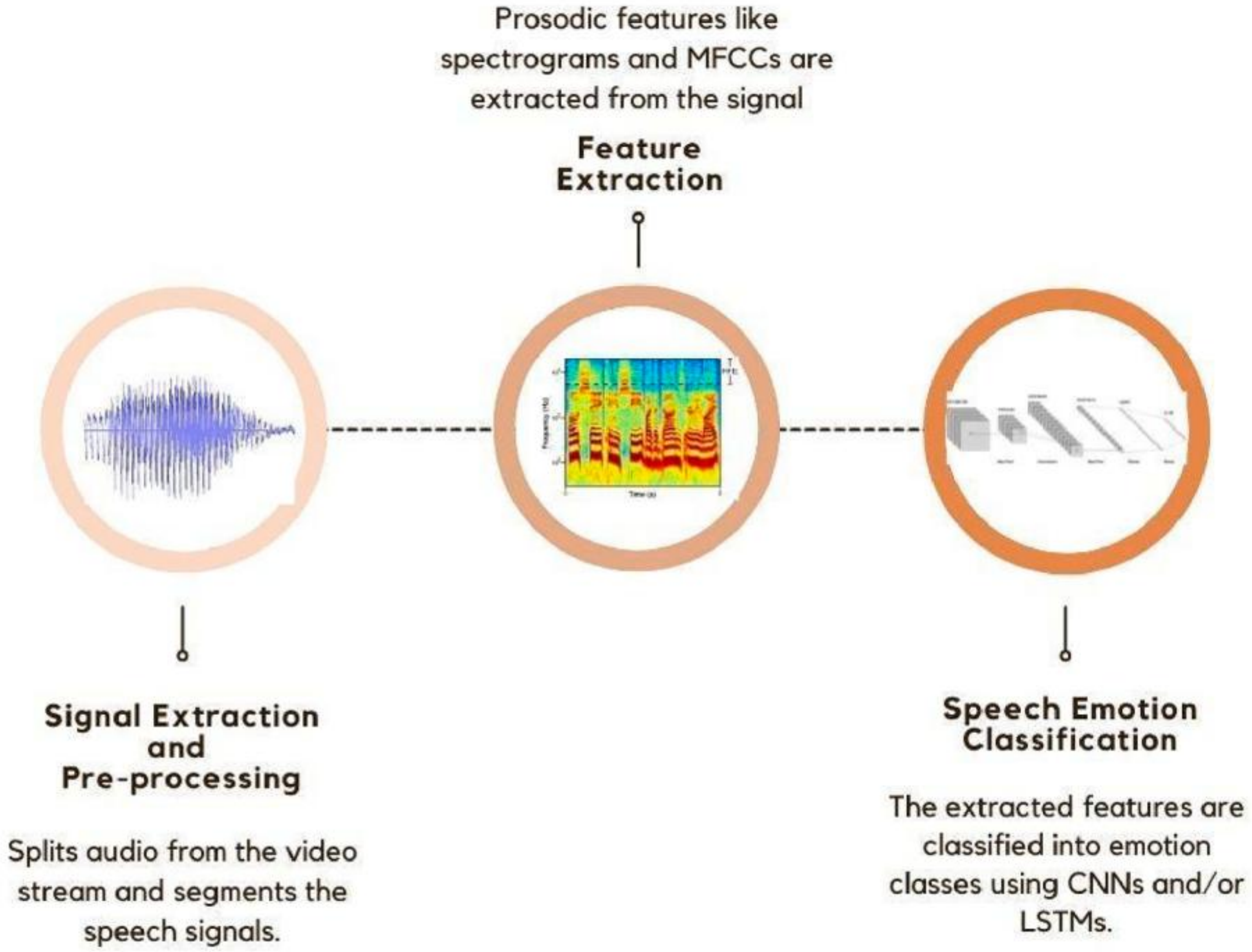
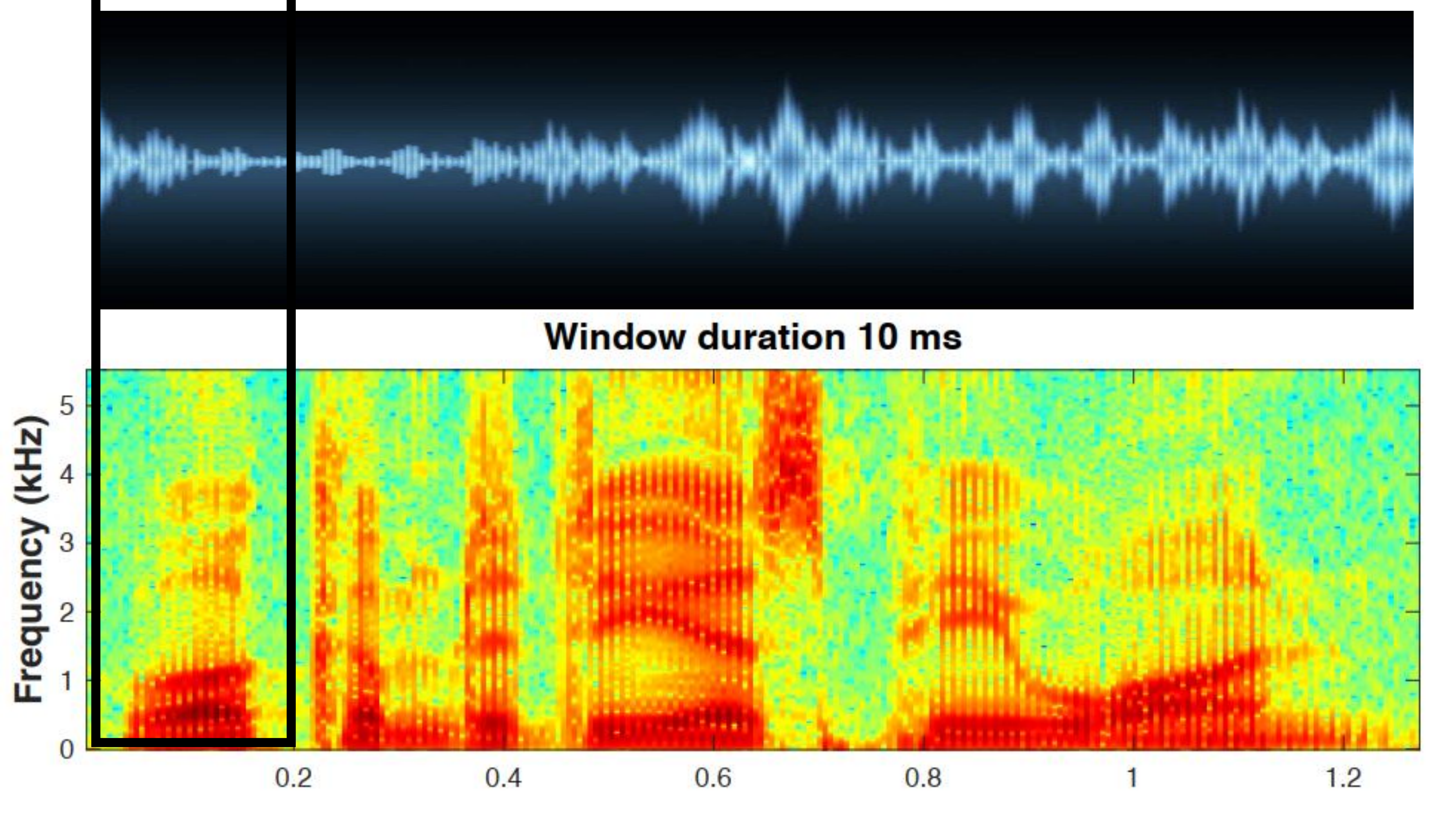
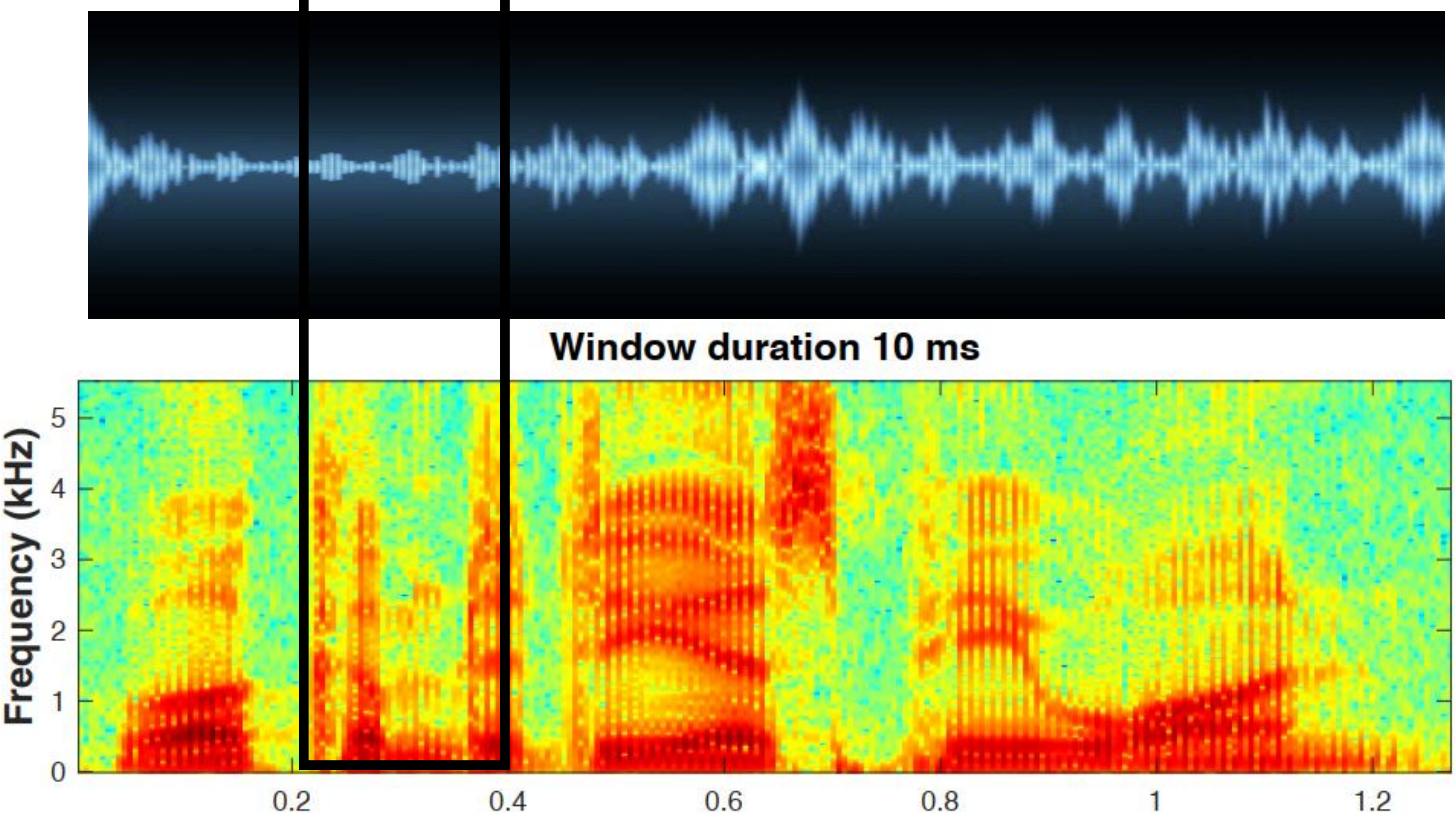


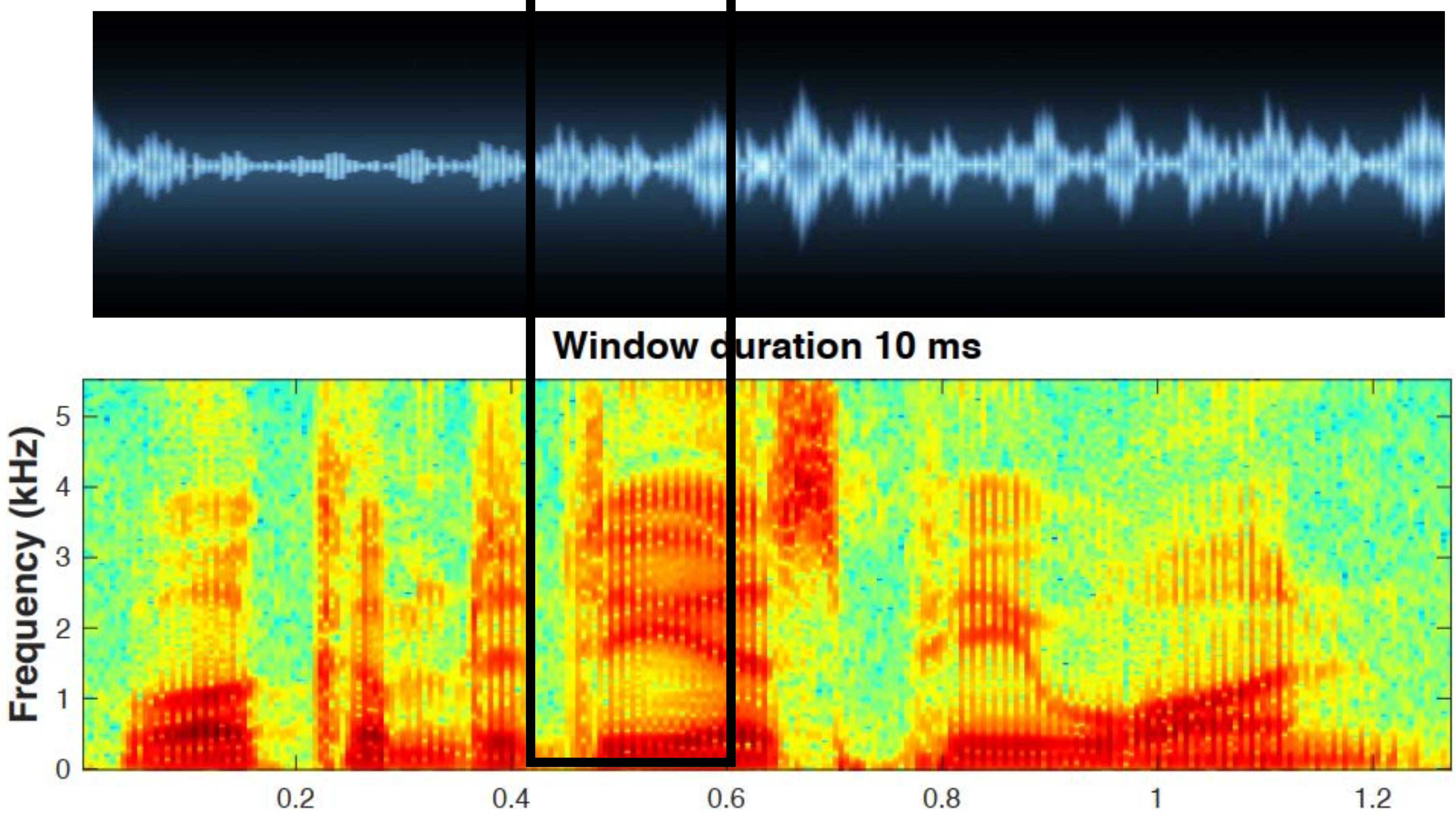
Fig. 13. Steps in Speech Emotion Recognition.

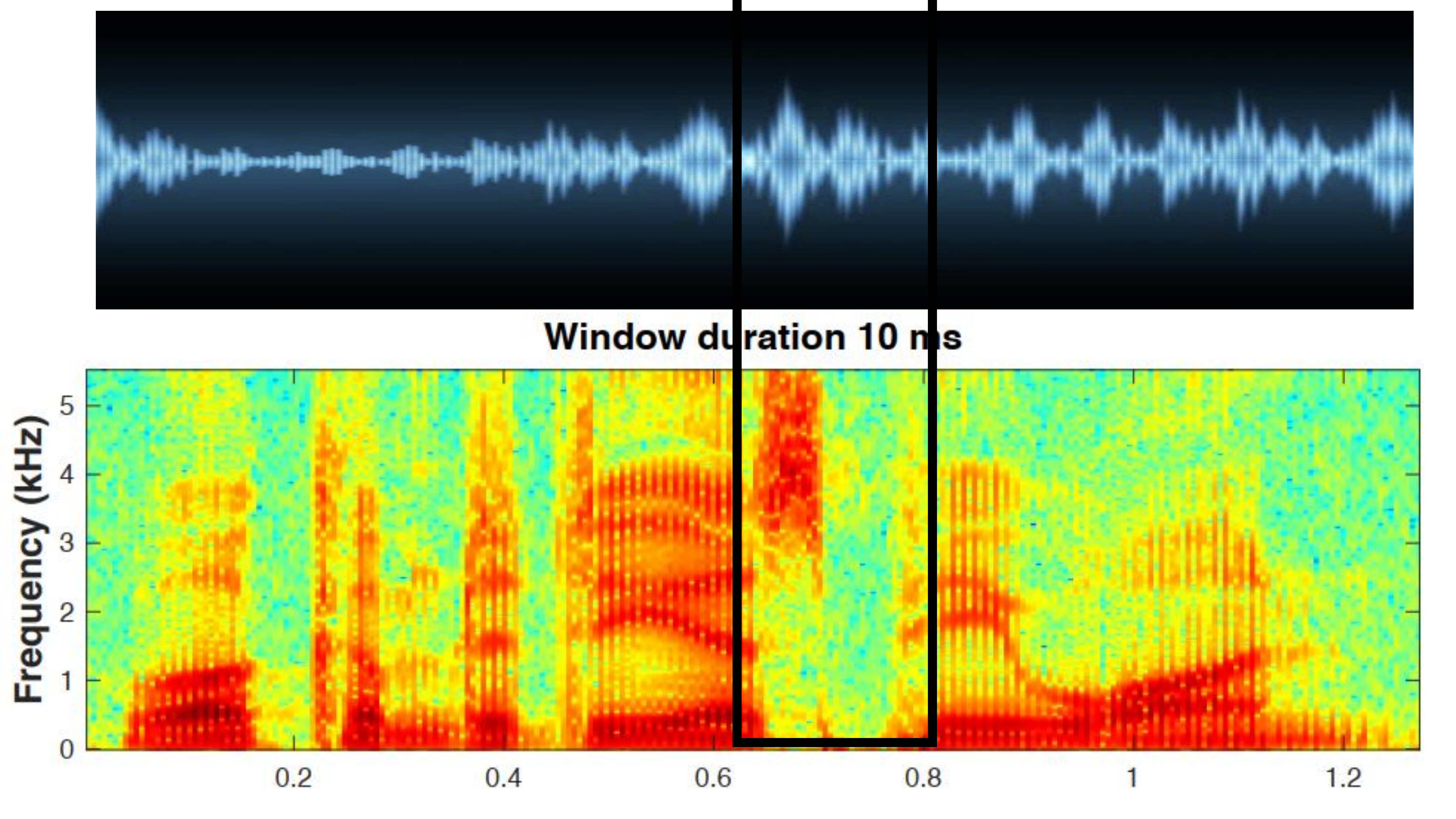
Image source: “Survey on multimodal approaches to emotion recognition”/ A. Aruna Gladys *, V. Vetriselvi / 2023


Window duration 10 ms
Frequency (kHz)
0
1
2
3
4
5
0.2
0.4
0.6
0.8
1
1.2

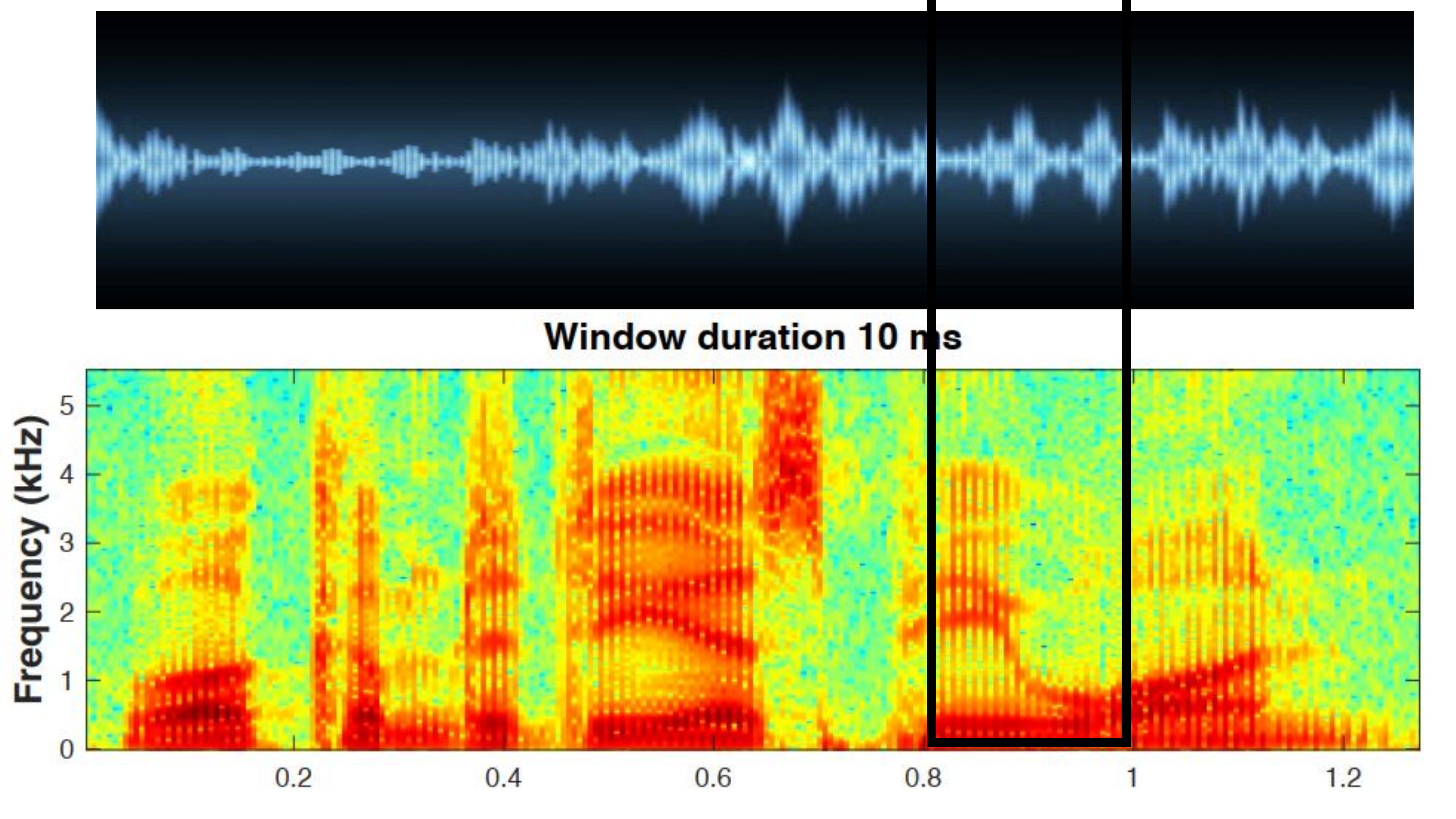


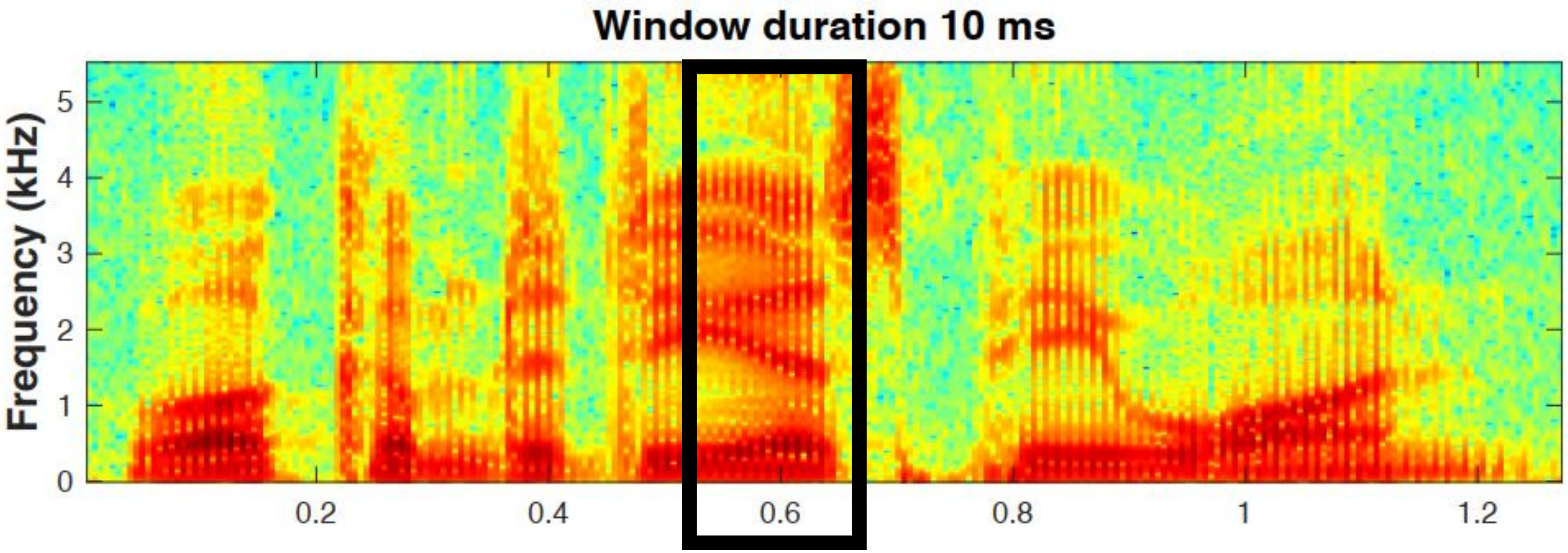
Window duration 10 ms
Frequency (kHz)
0
1
2
3
4
5
0.2
0.4
0.6
0.8
1
1.2


Window duration 10 ms
Frequency (kHz)
0
1
2
3
4
5
0.2
0.4
0.6
0.8
1
1.2



Window duration 10 ms
Frequency (kHz)
0
1
2
3
4
5
0.2
0.4
0.6
0.8
1
1.2


Window duration 10 ms
Frequency (kHz)
0
1
2
3
4
5
0.2
0.4
0.6
0.8
1
1.2



$s(f,\tau)$ – intensity of f.

$$S(f,\tau) = \sum_{t=0}^{N-1} x(t)\omega(t-\tau)e^{-j2\pi ft}$$

$s(f,\tau)$ – intensity of f.

$\sum_{t=0}^{N-1}$ A summation over a window of length N samples analyzing only a small segment of the signal at each step

$x(t)$ – *the origin signal*

$\omega(t-\tau)$ – *the window centered around* τ

$e^{-j2\pi ft}$ – a Fourier basis function. It tests how much of frequency f is present in the current time window. This component extracts the frequency content

Equation source: “Condition based maintenance of machine tools—A review ”/ Deepam Goyal *, B.S. Pabla 1 / 2015

Multimodal Learning Challenges

1. multimodal representation (MMR).
   vector form several media
2. multimodal translation (MMT).
   mapping information from one modality to another
3. multimodal alignment (MMA).
   Align the same event from two data sources / different medias.
4. multimodal fusion (MMF).
   perform regression or classification from two data sources / different medias.
5. multimodal co-learning (MMC).
   transmitting information/knowledge among modalities

Introduction to multimodal in context of deep learning

Fusion

Uni to Multi

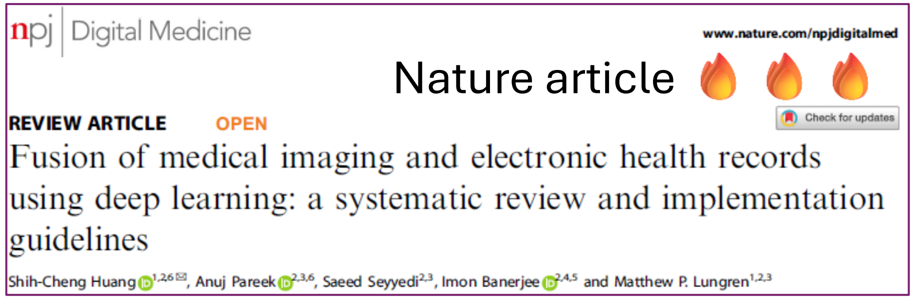
Structured and Unstructured data: Fusion Strategies

- Fusion strategies are used to merge data modalities into a single model
- There are many ways to fuse data, which can be categorized into three distinct strategies

Early Fusion

Late Fusion

Hybrid Fusion


npj | Digital Medicine

www.nature.com/npjdigitalmed

Nature article

REVIEW ARTICLE OPEN

Check for updates

Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines

Shih-Cheng Huang[1,2,6], Anuj Pareek[2,3,6], Saeed Seyyedi[2,3], Imon Banerjee[2,4,5] and Matthew P. Lungren[1,2,3]

Fusion Strategies

Early Fusion: Input-Level Integration of Modalities

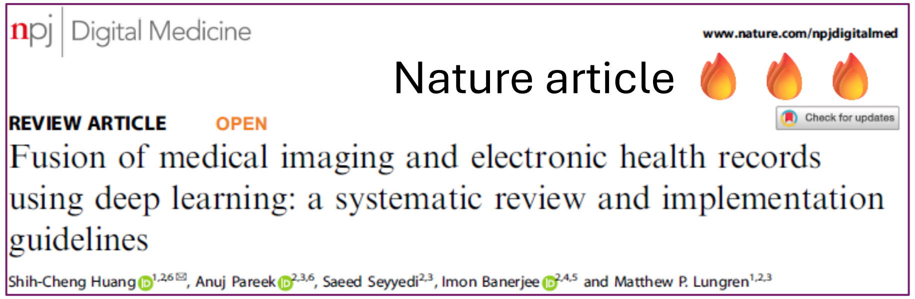
- **Early fusion** refers to the integration of modalities at a stage close to the input, resulting in a single unified representation.

merging data modalities into a common
feature vector already at the input layer

Late fusion

Hybrid fusion

Nature article 🔥 🔥 🔥


npj | Digital Medicine

www.nature.com/npjdigitalmed

REVIEW ARTICLE OPEN

Check for updates

Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines

Shih-Cheng Huang[1,2,6], Anuj Pareek[2,3,6], Saeed Seyyedi[2,3], Imon Banerjee[2,4,5] and Matthew P. Lungren[1,2,3]

Early Fusion Input-Level Integration

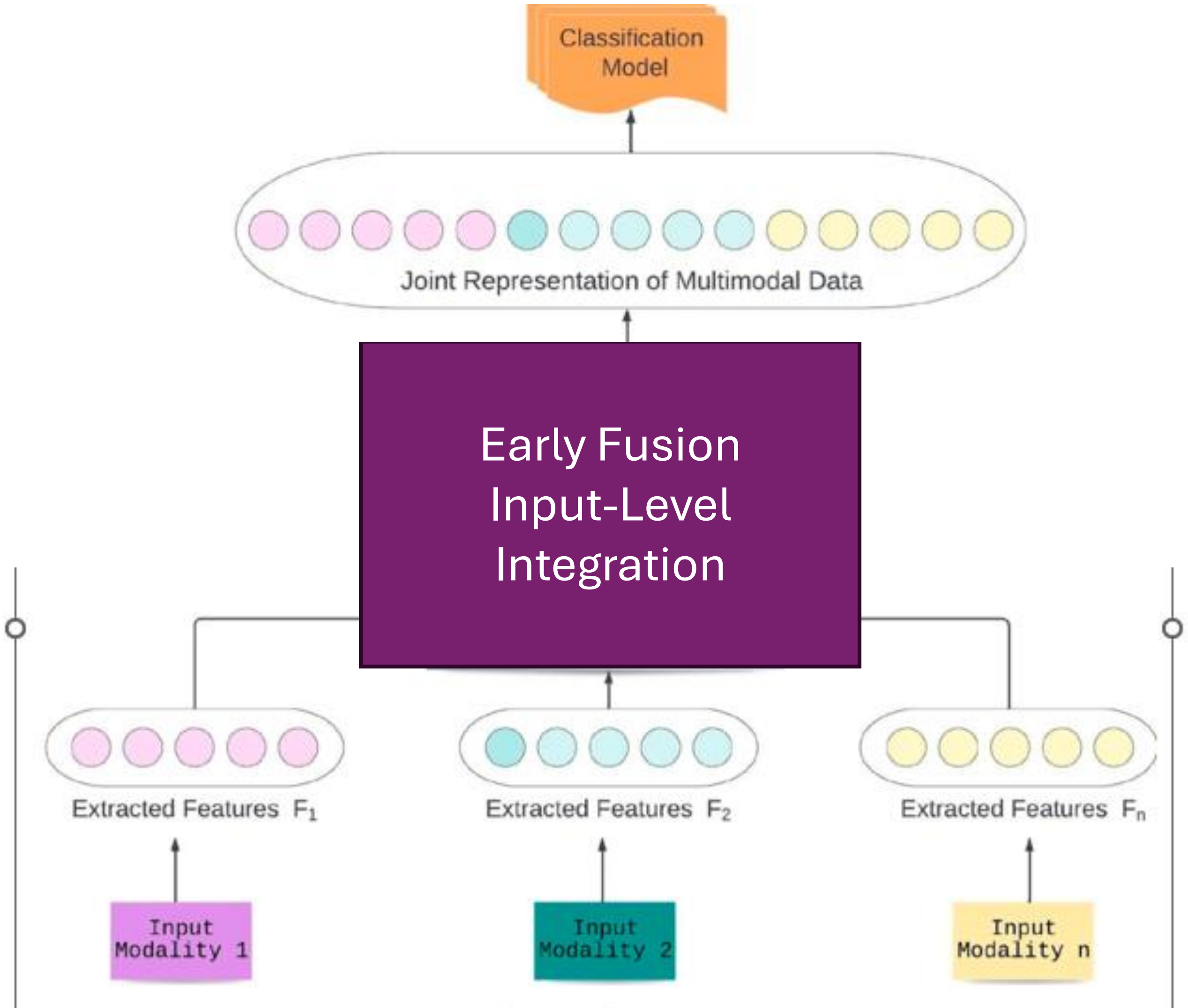


Fig. 17. Early Fusion.

Image source: “Survey on multimodal approaches to emotion recognition”/ A. Aruna Gladys *, V. Vetriselvi / 2023

Fusion Strategies

Late Fusion: Decision-Level Integration of Modalities

- **Late fusion** integrates outputs from separate unimodal models after processing, combining decisions or features at a higher level while allowing independent feature learning per modality.

Early fusion

fusing the **predictions** of multiple models that have been trained on each data modality separately

Hybrid fusion

npj | Digital Medicine

Nature article

www.nature.com/npjdigitalmed

REVIEW ARTICLE OPEN

Check for updates

Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines

Shih-Cheng Huang[1,2,6], Anuj Pareek[2,3,6], Saeed Seyyedi[2,3], Imon Banerjee[2,4,5] and Matthew P. Lungren[1,2,3]

Late Fusion
Decision-Level

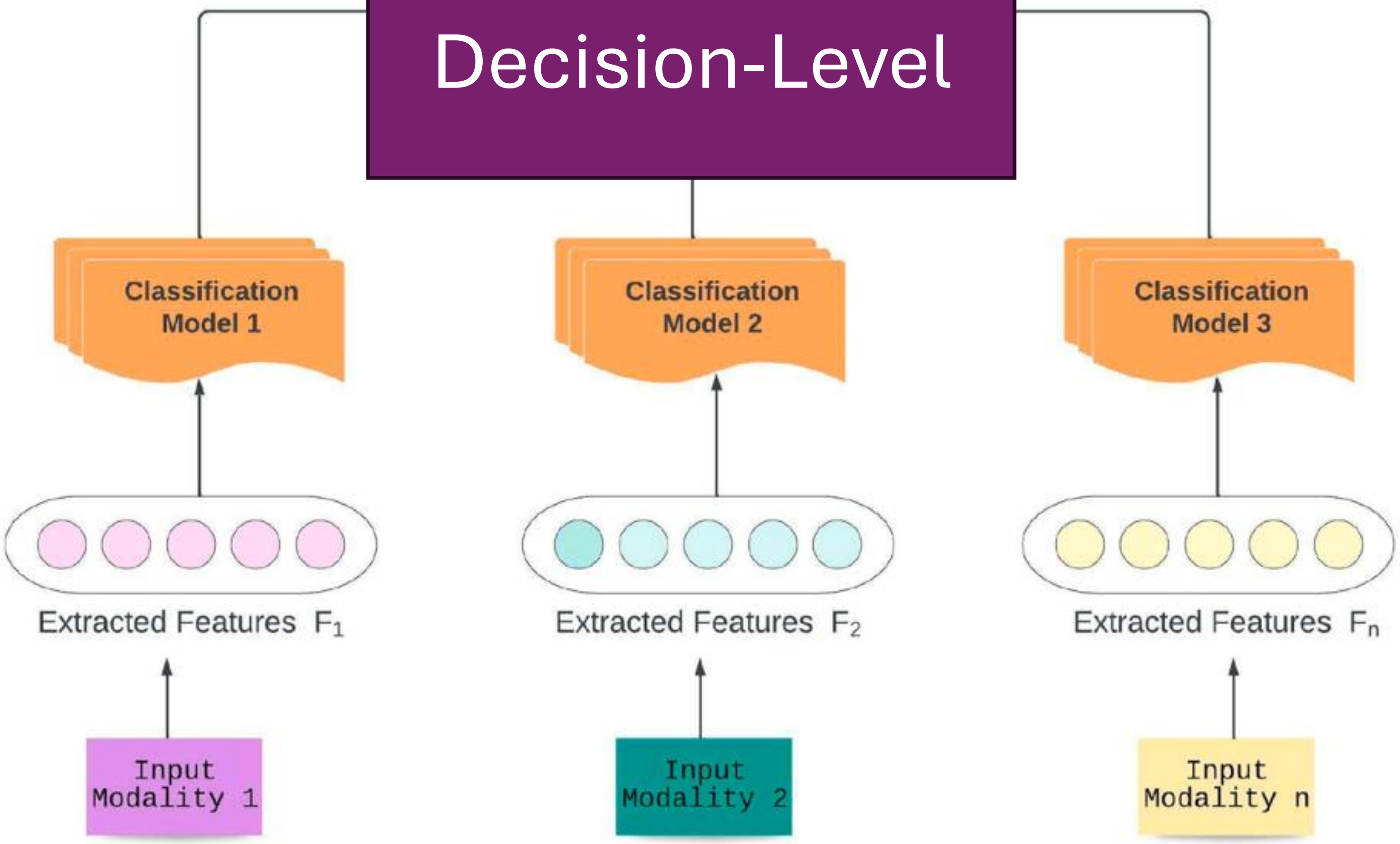


Fig. 18. Late Fusion.

Image source: “Survey on multimodal approaches to emotion recognition”/ A. Aruna Gladys *, V. Vetriselvi / 2023

Fusion Strategies

Hybrid Fusion: Combining Early and Late Fusion Strategies

- Hybrid fusion combines early and late fusion by integrating both feature-level and decision-level information, leveraging the strengths of each to enhance multimodal learning.

Early fusion

Late Fusion

flexibility to merge the modalities at different depths of the model and thereby to learn latent feature

npj | Digital Medicine www.nature.com/npjdigitalmed

Nature article

REVIEW ARTICLE OPEN

Check for updates

Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines

Shih-Cheng Huang[1,2,6], Anuj Pareek[2,3,6], Saeed Seyyedi[2,3], Imon Banerjee[2,4,5] and Matthew P. Lungren[1,2,3]

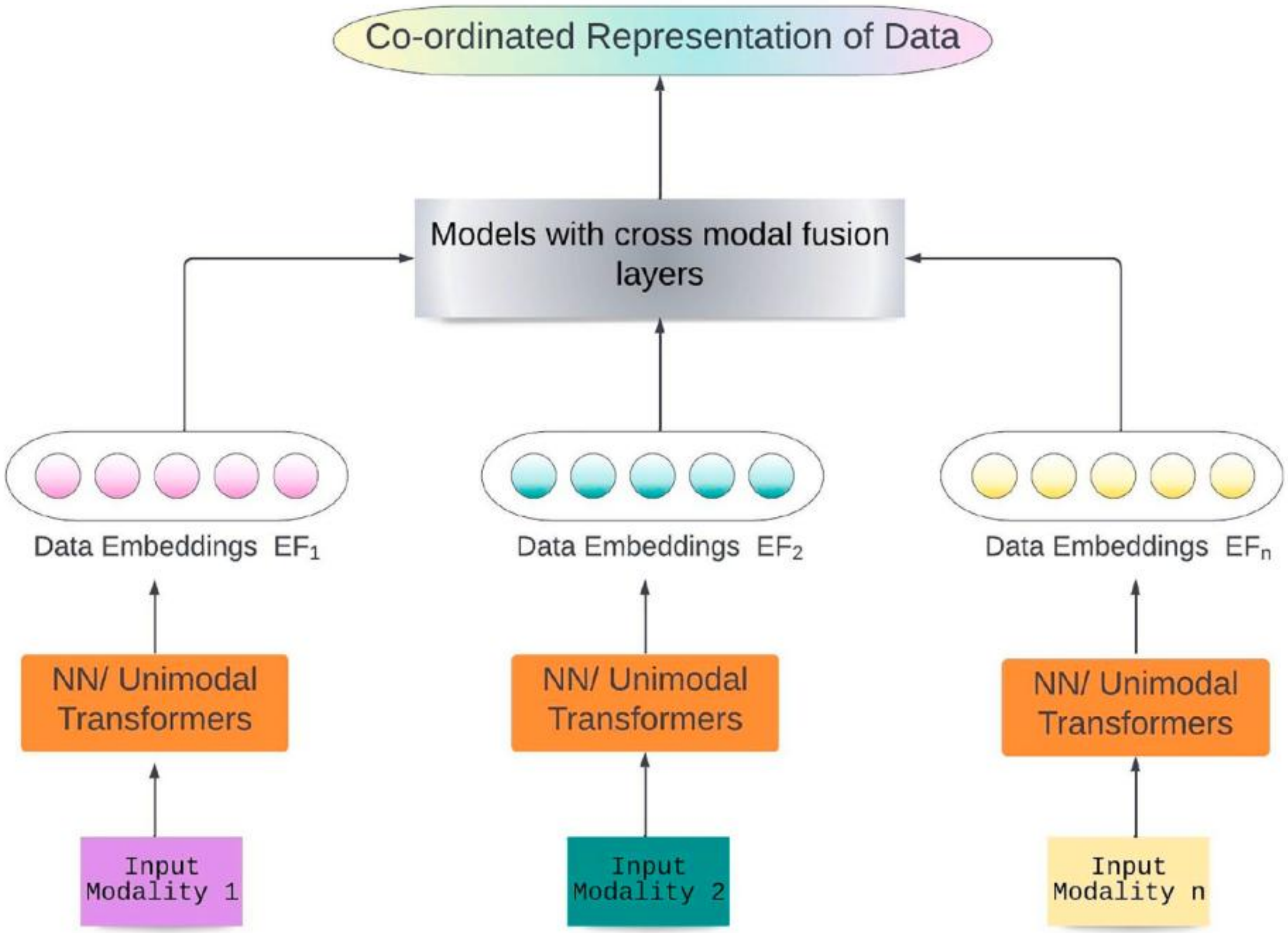


Fig. 19. Hybrid Fusion.

Image source: “Survey on multimodal approaches to emotion recognition”/ A. Aruna Gladys *, V. Vetriselvi / 2023

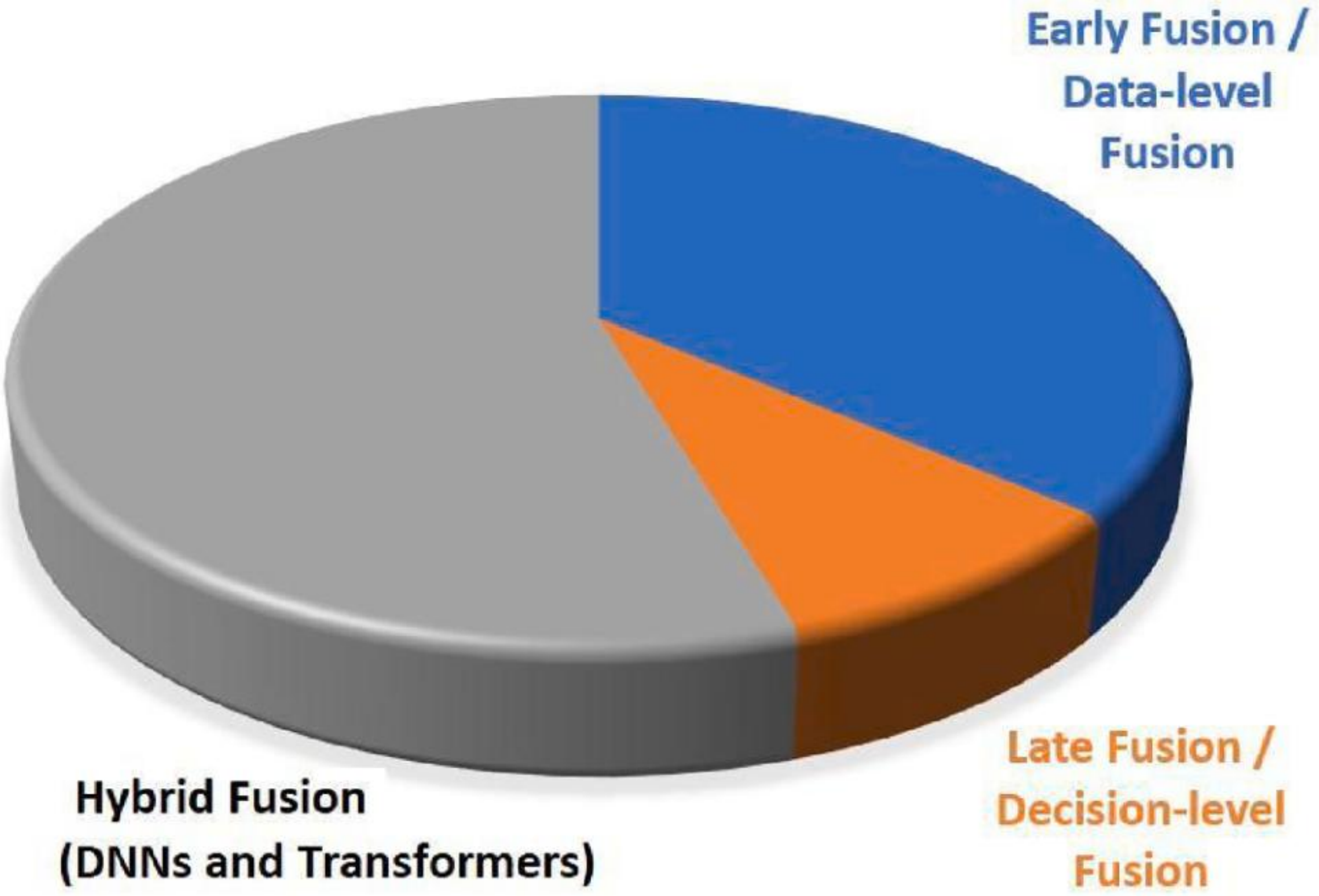


Fig. 20. Popularity of Hybrid Fusion in recent research.

Image source: "Survey on multimodal approaches to emotion recognition"/ A. Aruna Gladys *, V. Vetriselvi / 2023

Multimodal Architectures

Image To Text

Video description DRL & VQA

Image To Text architecture

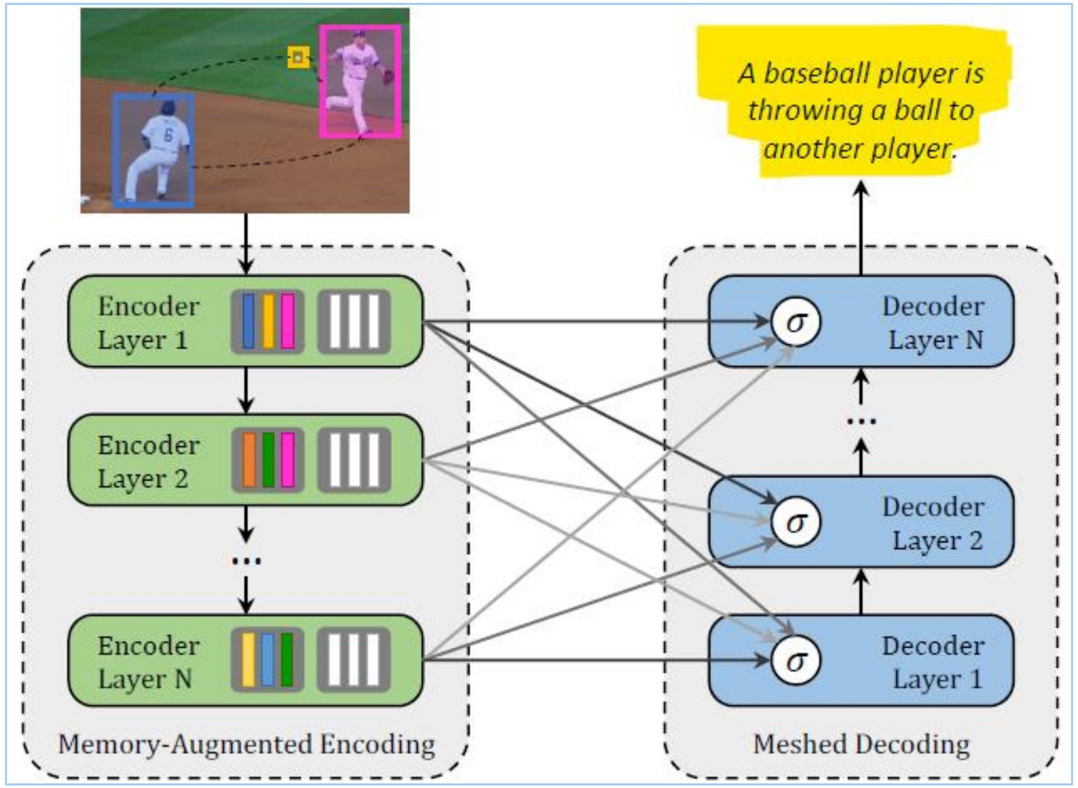
- Image To Text is a core architecture of multimodal learning

Meshed-Memory Transformer for Image Captioning

Marcella Cornia* Matteo Stefanini* Lorenzo Baraldi* Rita Cucchiara
University of Modena and Reggio Emilia
{name.surname}@unimore.it

Meshed-Memory Transformer for Image Captioning –M²

- This work is among the first to apply transformers to multimodal tasks such as image captioning.
- Present a fully-attentive approach
- Has two new novelties:
  - The encoder encodes a multi-level representation of the relationships between image regions with respect to low-level and high-level relations
  - a-priori knowledge can be learned and modeled by using persistent memory vectors



Meshed-Memory Transformer for Image Captioning

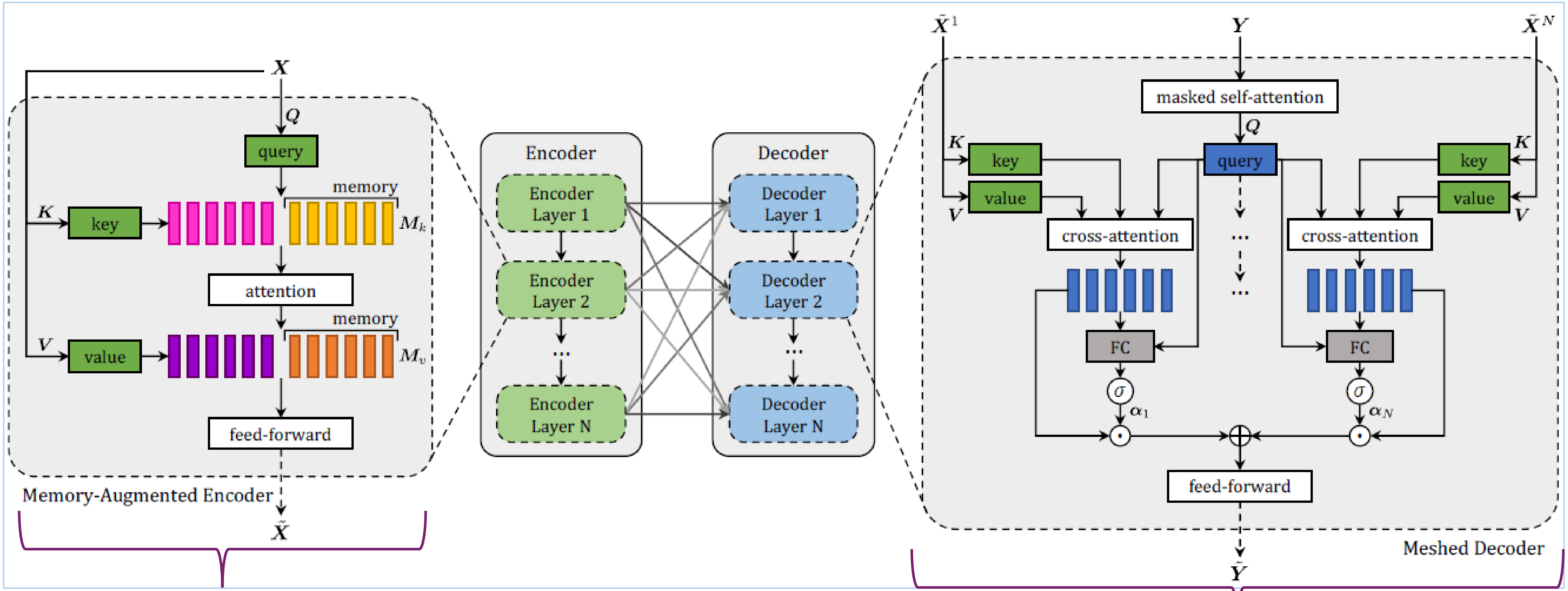
Marcella Cornia* Matteo Stefanini* Lorenzo Baraldi* Rita Cucchiara
University of Modena and Reggio Emilia

1279 citations

Architectures: Image To Text

Detailed Architecture of M^2

- Given an input image region X, the model applies attention and feed-forward layers to encode relationships between regions using prior knowledge. The decoder then generates the image caption word by word from the encoder outputs



Encoder in details

Decoder in details

Memory-Augmented Attention

Key equation:

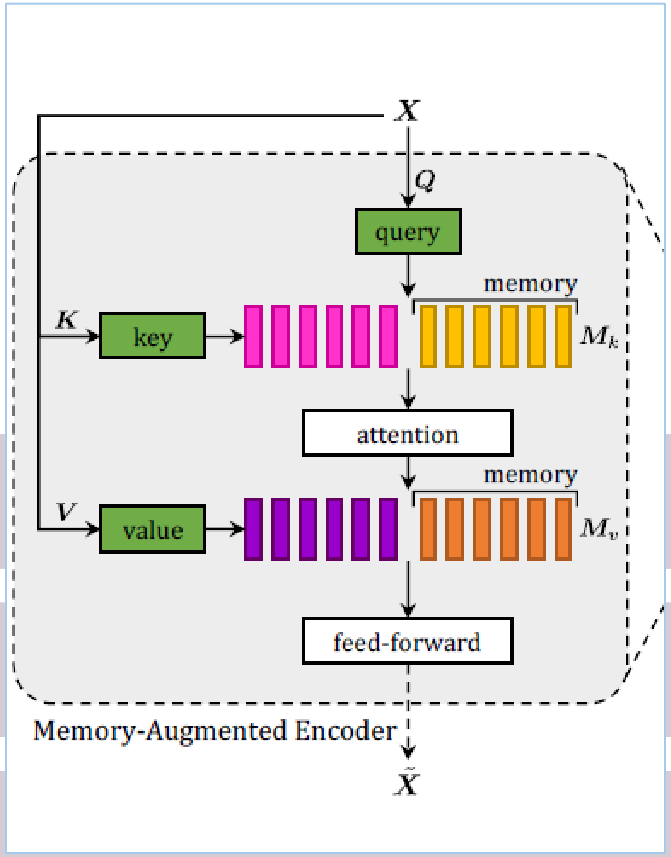
$$\mathrm{M}_{\mathrm{mem}}(X) = \mathbf{Attention}(W_q X,\ [W_k X, M_k],\ [W_v X, M_v])$$

$$M_{mem}(X) = Attention(W_q X, K, V)$$

$$K = [W_k X, M_k]$$

$$V = [W_v X, M_v]$$

$$Attention(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d}})V$$



- M_k, M_v: learnable memory (prior knowledge)
- Retrieved independently of input X
- Improves modelling of external context

Memory-Augmented Transformer Encoder Layer

- Feed-forward:

$$F(X)_i = U\,\mathrm{ReLU}(VX_i + b) + c$$

- Full layer:

$$Z = \mathrm{AddNorm}(M_{\mathrm{mem}}(X)), \quad \tilde{X} = \mathrm{AddNorm}(F(Z))$$

- Residual connections + layer normalization
- Stack of *n* layers refines features across steps

Why is memory augmentation important in multimodal or vision tasks?

- It improves interpretation and tasks as rare object recognition

Encoding Layer with Memory-Augmented Operator

- Memory-augmented operator d is injected into a transformer-like architecture
- Output passes through a position-wise feed-forward layer:

$$F(X)_i = U\sigma(VX_i + b) + c;$$

- Each block includes:

  - Residual connection

  - Layer Normalization

- Encoding Layer computations:

$$Z = AddNorm(M_{mem}(X))$$
$$\tilde{X} = AddNorm(F(Z))$$

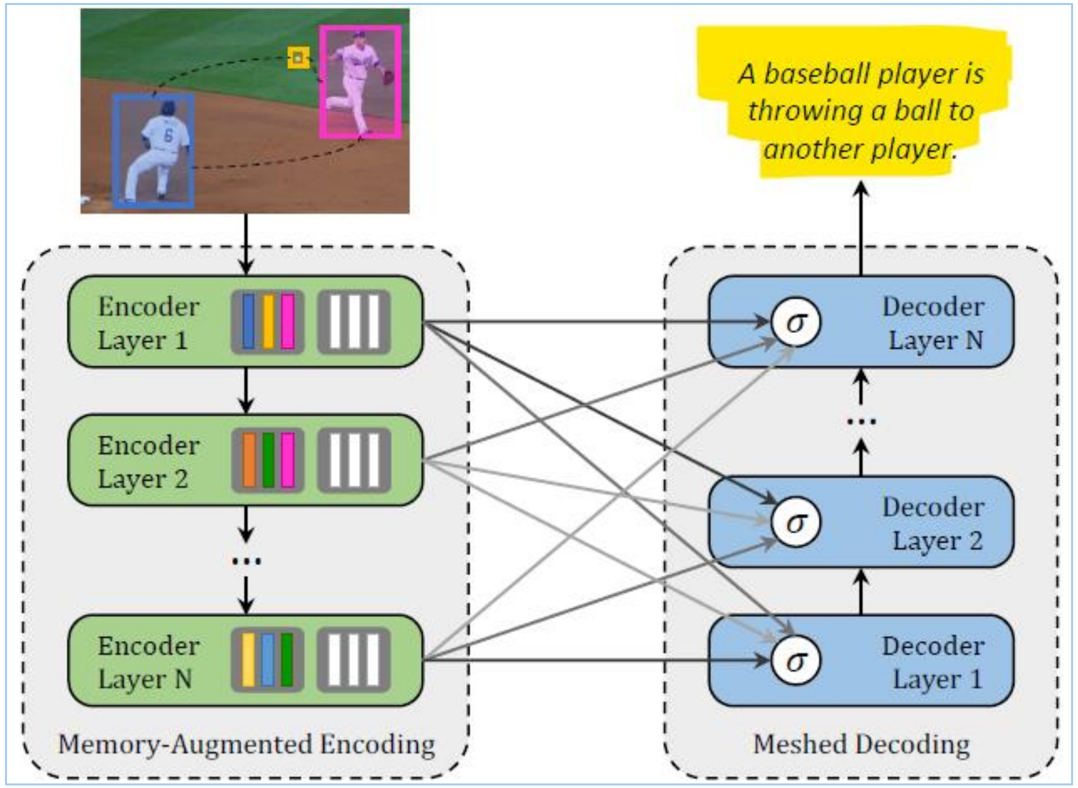
Final output: $\tilde{X} = (\tilde{X}^1 \ldots \tilde{X}^n)$

What Is Mashed Decoder

Takes into account both **previously generated words** and **image region encodings**.

Connects the decoder to **all encoder layers**, not just the last one.

Uses **cross-attention with gating** to integrate multiple layers adaptively.



Architectures: Image To Text

Meshed Cross-Attention Mechanism

Core idea:

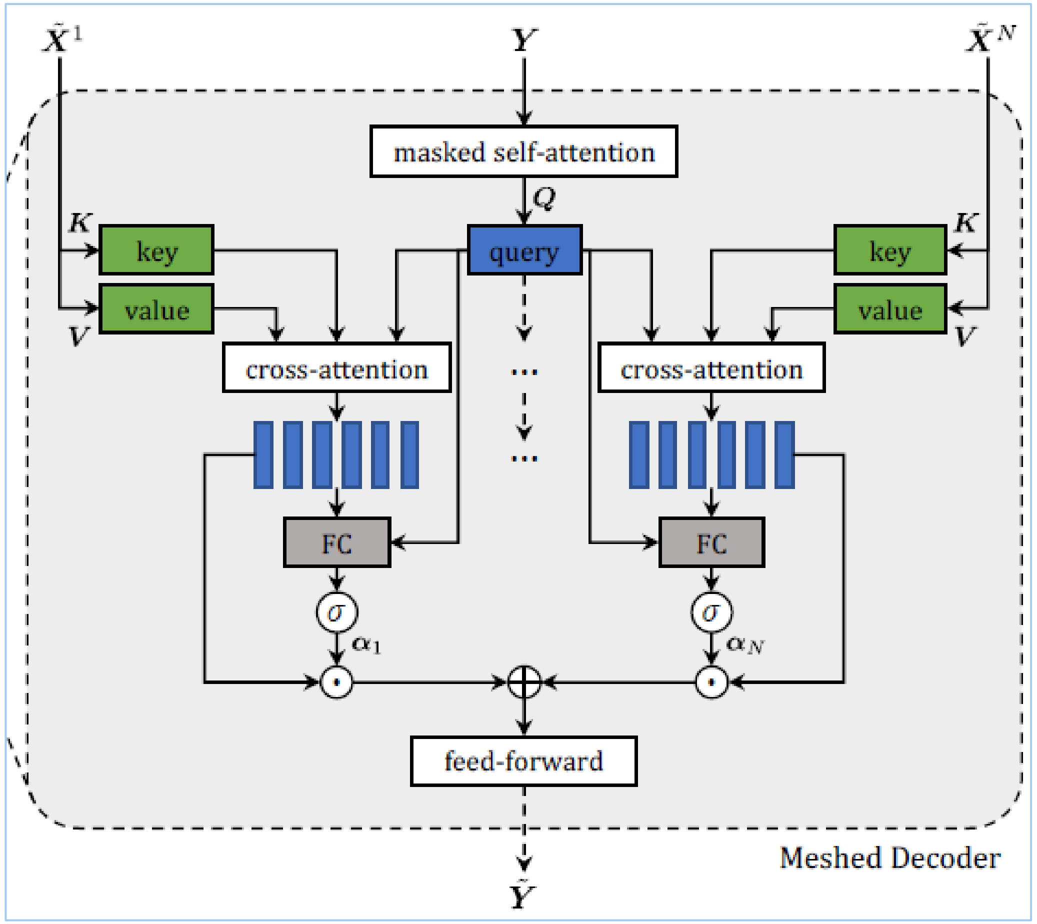
$$M_{\text{mesh}}(\tilde{X}, Y) = \sum_{i=1}^{N} \beta_i \cdot C(\tilde{X}_i, Y)$$

Cross-attention:

$$C(\tilde{X}_i, Y) = \text{Attention}(W_q Y,\ W_k \tilde{X}_i,\ W_v \tilde{X}_i)$$

Gating weights:

$$\beta_i = \sigma(W_i[Y,\ C(\tilde{X}_i, Y)] + b_i)$$

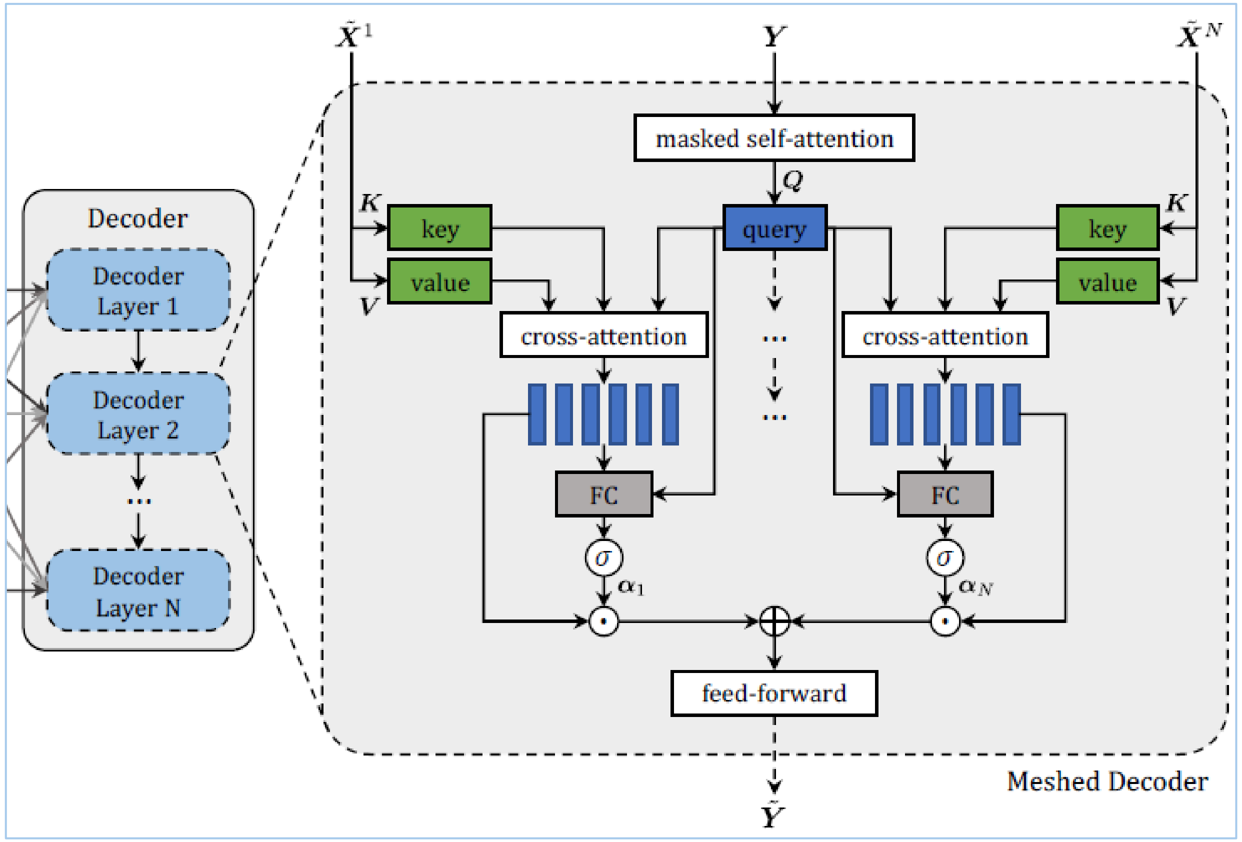


- $C(\tilde{X}_i, Y)$: attends decoder to each coder layer
- β_i: sigmoid gate controlling how much each layer contributes
- Combines **multiple** encoder views of the image **adaptively**

Architectures: Image To Text

Complete Decoder Layer

- Masked self attention
- Meshed cross attention
- Feed-forward layer
- Add & norm (residuals)



Equation summary:

$$Z = \text{AddNorm}(M_{\text{mesh}}(X,\ \text{AddNorm}(S_{\text{mask}}(Y))))$$

$$\tilde{Y} = \text{AddNorm}(F(Z))$$

- Final softmax layer turns decoder output into word probabilities.
- Supports context-aware captioning, word by word

Loss Function: Cross Entropy

$$J_{cce} = -\frac{1}{M} \sum_{k=1}^{K} \sum_{m=1}^{M} y_m^k \times \log\left(h_\theta\left(x_m, k\right)\right)$$

where

M	number of training examples
K	number of classes
y_m^k	target label for training example m for class k
x	input for training example m
h_θ	model with neural network weights θ

Equation source: "The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling "/ YAOSHIANG HO AND SAMUEL WOOKEY / 2020

Video description deep reinforcement learning & VQA architectures

- VQA & video description DRL are core architectures of multimodal learning

A Review on Methods and Applications in Multimodal Deep Learning

SUMMAIRA JABEEN and XI LI, College of Computer Science, Zhejiang University, China
MUHAMMAD SHOIB AMIN, School of Software Engineering, East China Normal University, China
OMAR BOURAHLA, SONGYUAN LI, and ABDUL JABBAR, College of Computer Science, Zhejiang University, China

Architectures: Video Description DRL & VQA

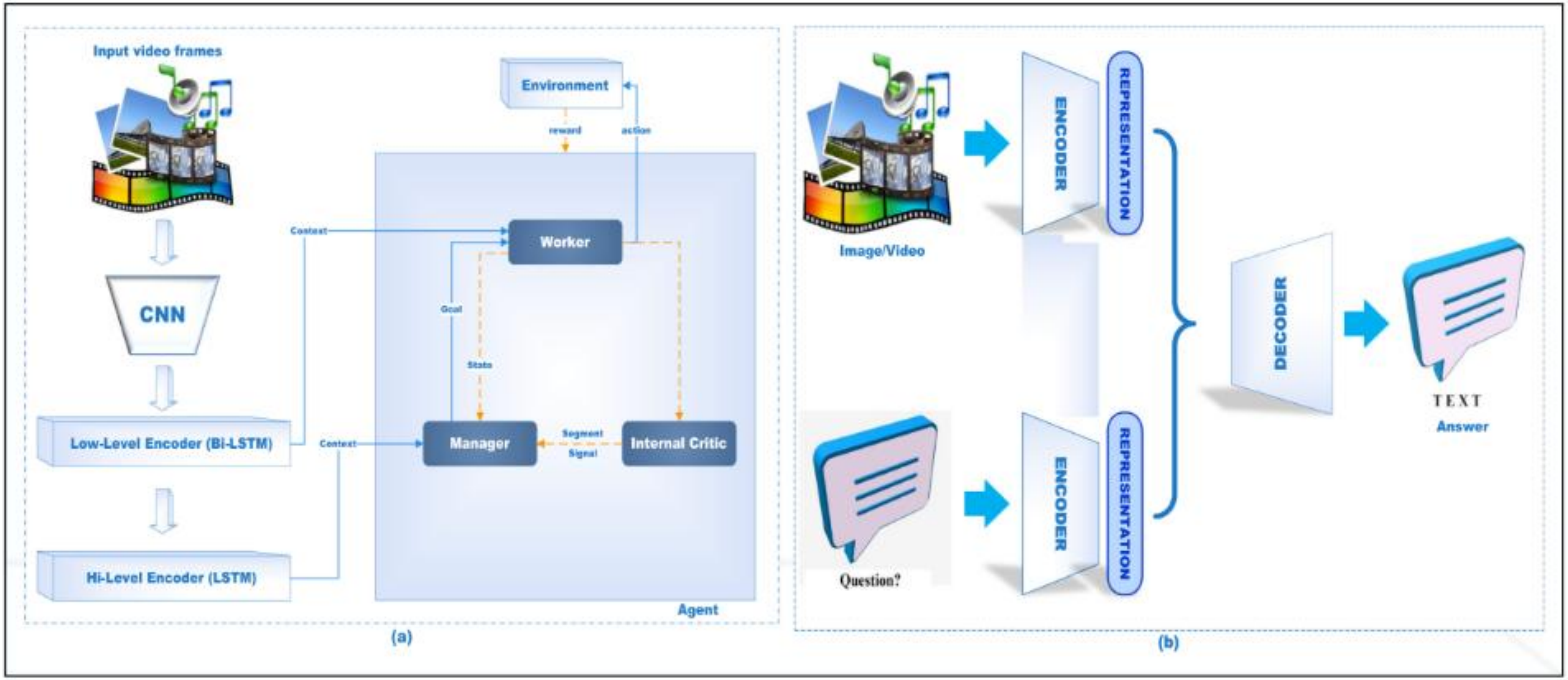


Fig. 5. (a) General structure diagram of Video Description Deep Reinforcement Learning Architectures and (b) General structure diagram of VQA System.

Architectures: Video Description DRL

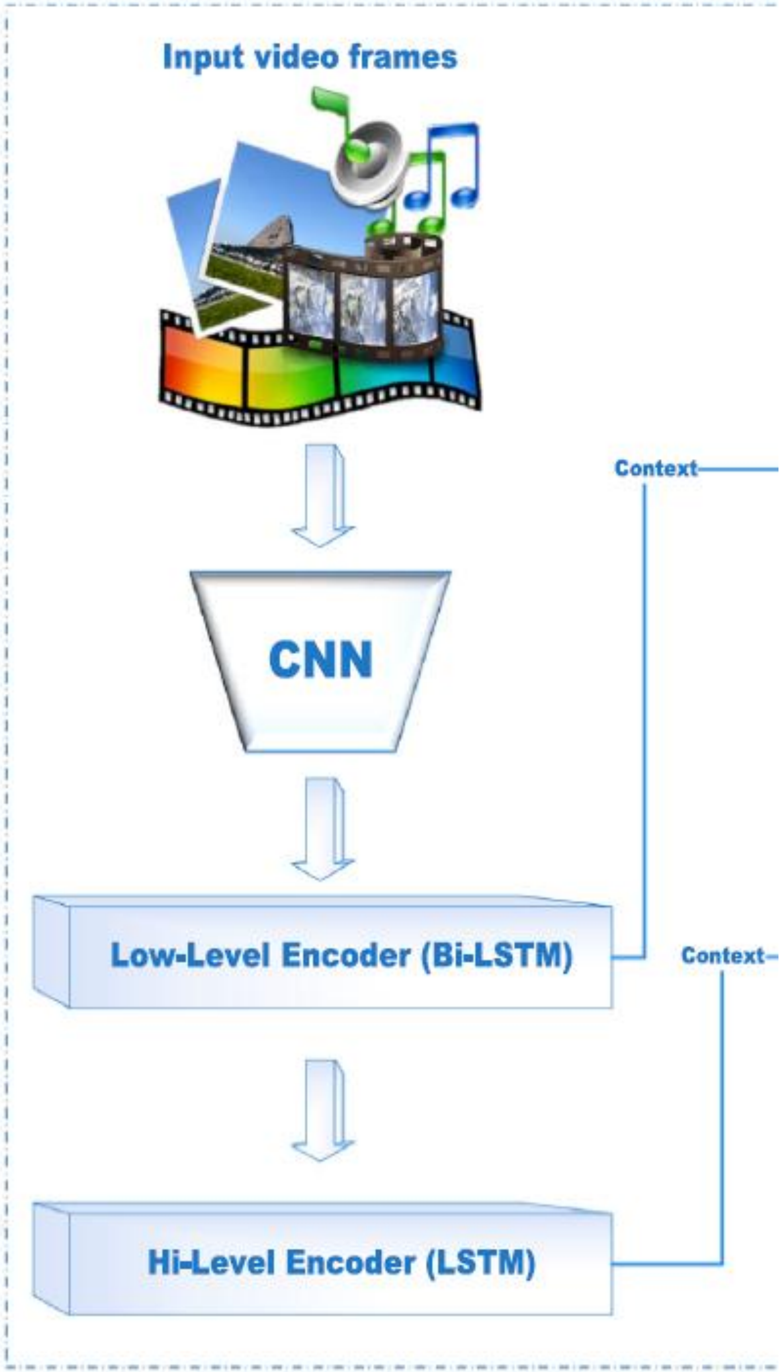
Step 1: Video Processing

The video undergoes three stages:

CNN – Extracts a feature vector from each frame (image).

Bi-LSTM – Takes the sequence of vectors from the CNN and adds bidirectional context (forward and backward in time).

LSTM – Summarizes the entire sequence into a single vector that represents the whole video.



Step 2 : Intelligent Agent

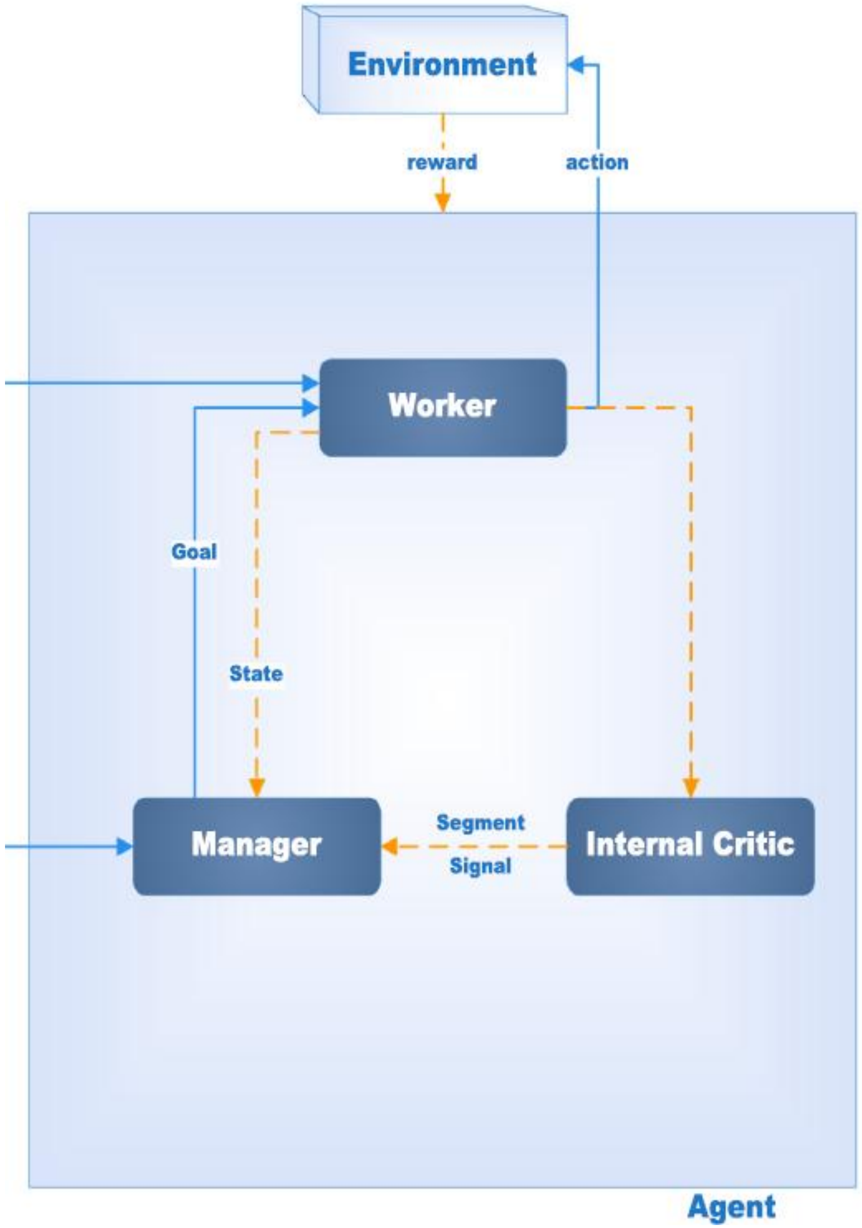
The agent consists of four components:

Manager – Determines which frames to focus on (goal settings) , based on a global representation of the video

Worker - Receives the goal , processes only the embeddings of the selected relevant frames and generates the answer.

Internal Critic – Evaluates the quality of the answer and provides a reward, which is fed back to both the manager and the worker.

Goal: To direct the model's attention only to the most relevant parts of the video clip, rather than the entire clip, in order to improve both efficiency and accuracy.



Architectures: VQA

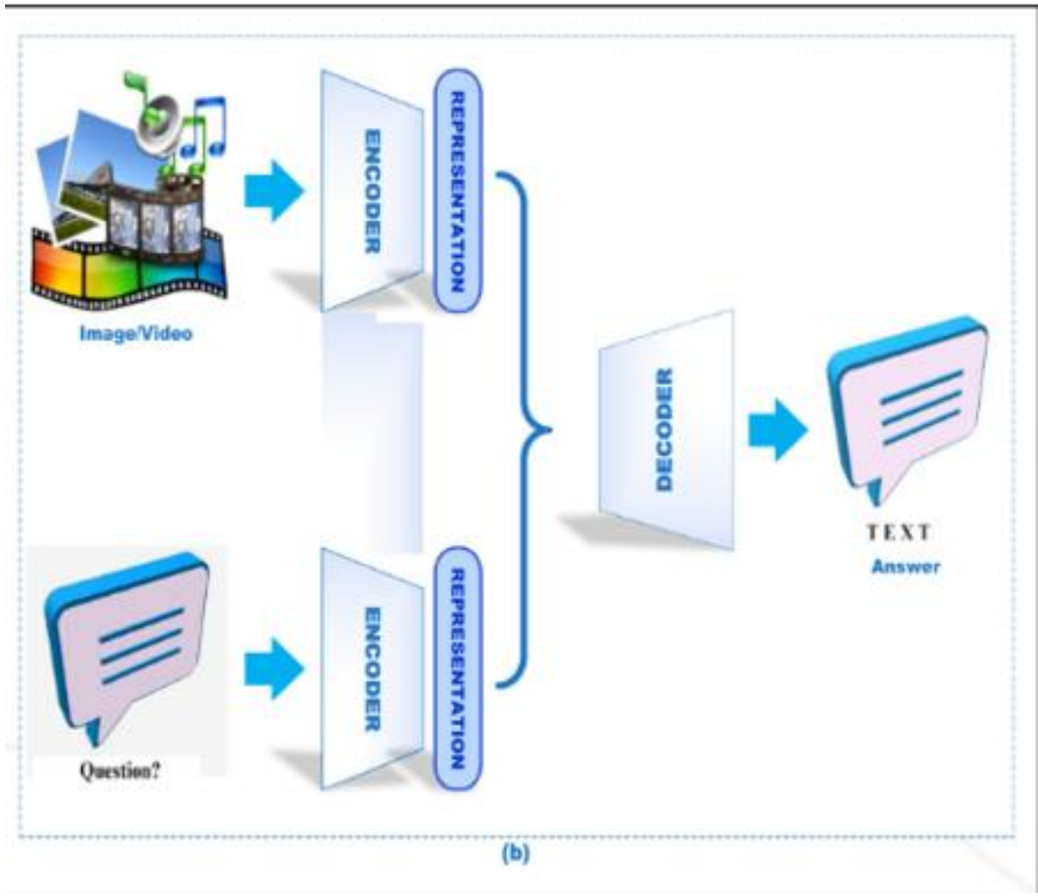
Encoding and fusing the information

this step consists of three components

Encoding the question and video - The video and the question are each processed by a separate encoder, producing a visual representation of the video and a semantic representation of the text.

Fusing and information -The model integrates both representations using an attention mechanism, which identifies the regions of the video that are most relevant to the given question.

Generating and answer - The fused representation is passed into a decoder that generates the answer word by word, taking into account both the question and the visual context from the scene.



Our Project

Multimodal emotion recognition and reasoning with instruction tuning

Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning

Zebang Cheng[1*] **Zhi-Qi Cheng**[2*†] **Jun-Yan He**[3] **Jingdong Sun**[2]
Kai Wang[4] **Yuxiang Lin**[1] **Zheng Lian**[5] **Xiaojiang Peng**[1†] **Alexander G. Hauptmann**[2]
[1]Shenzhen Technology University [2]Carnegie Mellon University [3]Alibaba Group
[4]National University of Singapore [5]Institute of Automation, Chinese Academy of Sciences

Apr-2024

LLaMA: Large Language Model Meta AI

The unmet need for emotion recognition

Traditional single-modality approaches often fail to capture the complexity of real-world emotional expressions, which are inherently multimodal

Human computer interaction

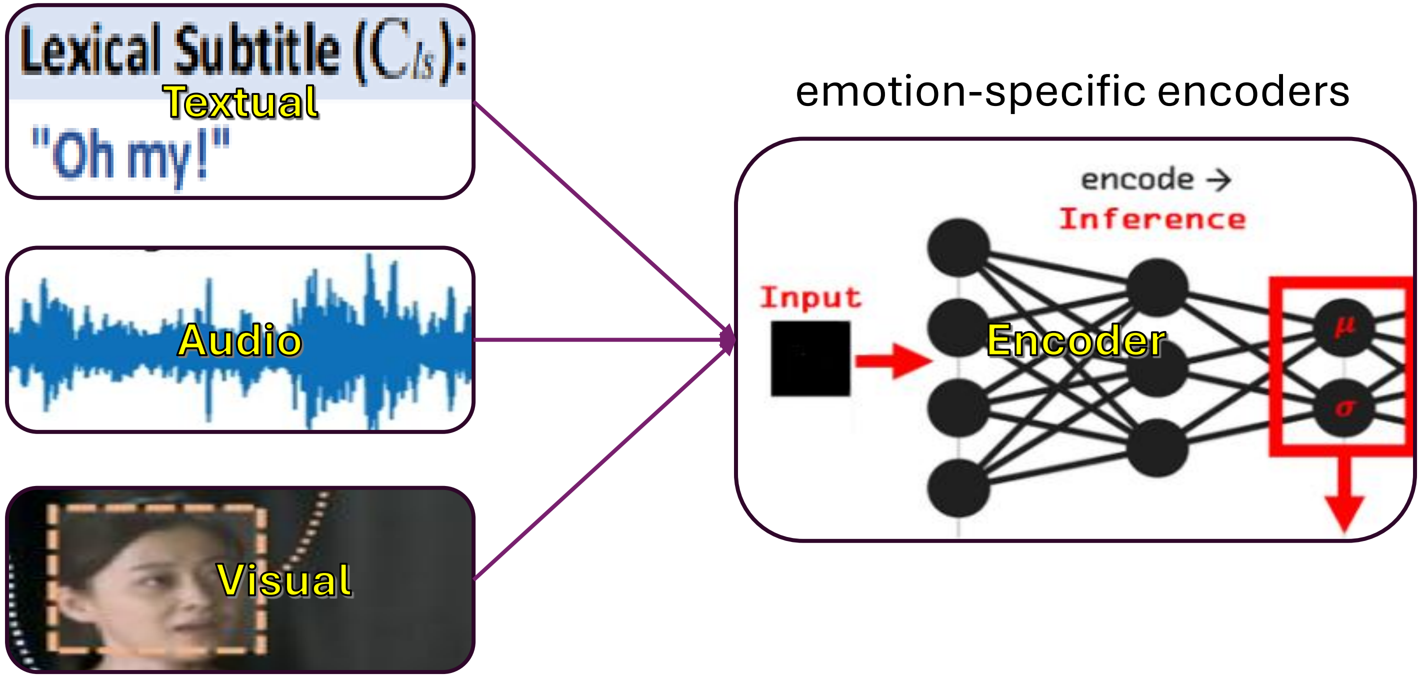
Counseling

mental health, emotional support

Security

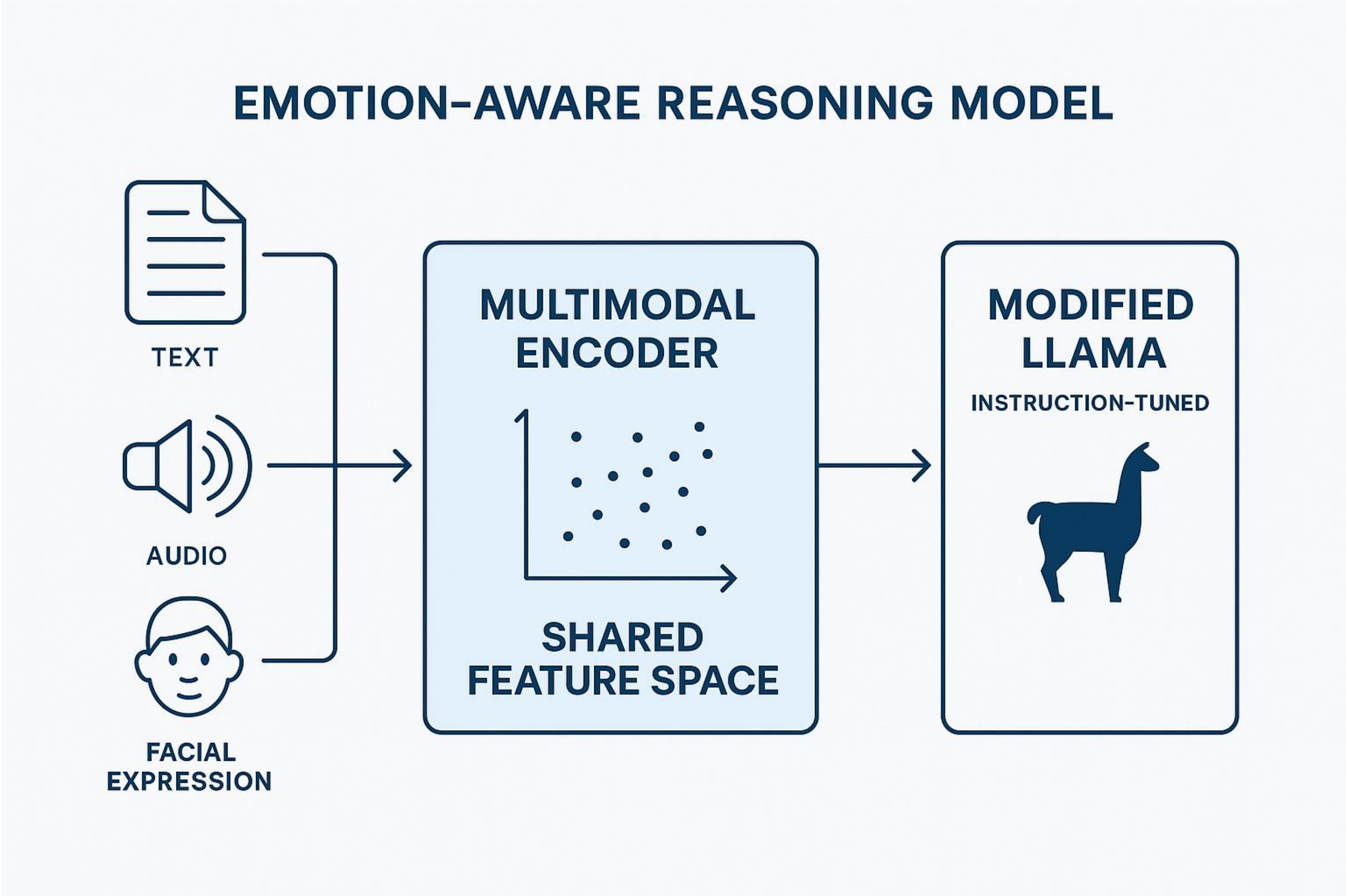
Education

The innovative solution of the article



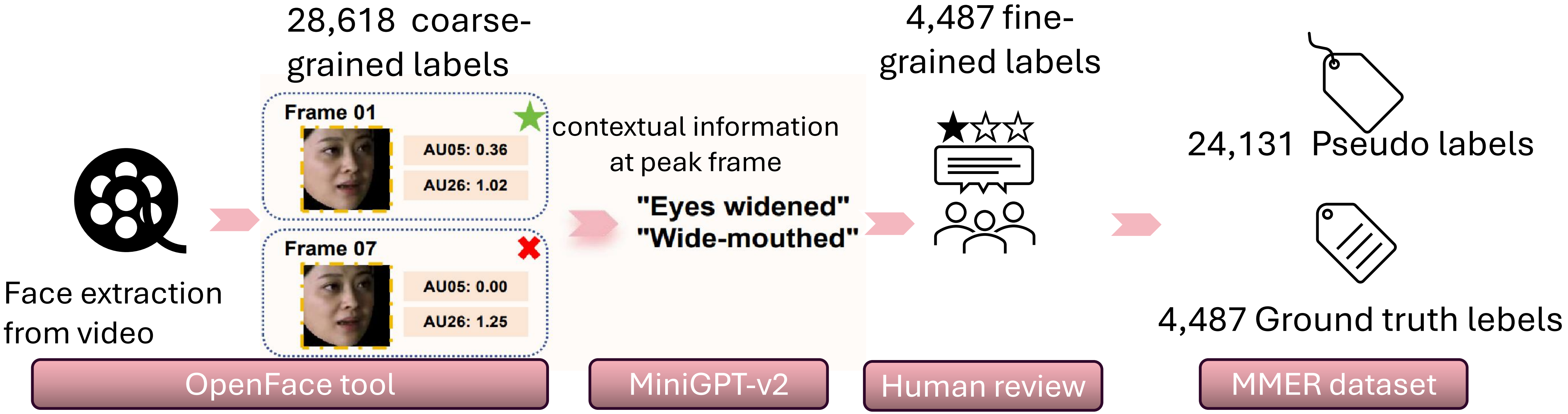
Proposed LLaMA emotion as a model that integrate audio, visual, and textual inputs through **emotion-specific encoders**.

The innovative solution of the article



Managed to significantly **enhance** both **emotional recognition** and **reasoning capabilities** by aligning features into a shared space and employing a modified LLaMA model with instruction tuning.

Methodology: construction of MERR* dataset



$$I_{peak} = \arg\max_{k} \left(\sum_{i} S^{k}_{au_i} \right)$$

Generated 28,618 pseudo-labels based on facial muscle movements combinations (AU), miniGPT generated a contextual text to the frames with highest I_{peak} value, after a human review 4,487 frames were labeled as ground truth with accurate emotion.

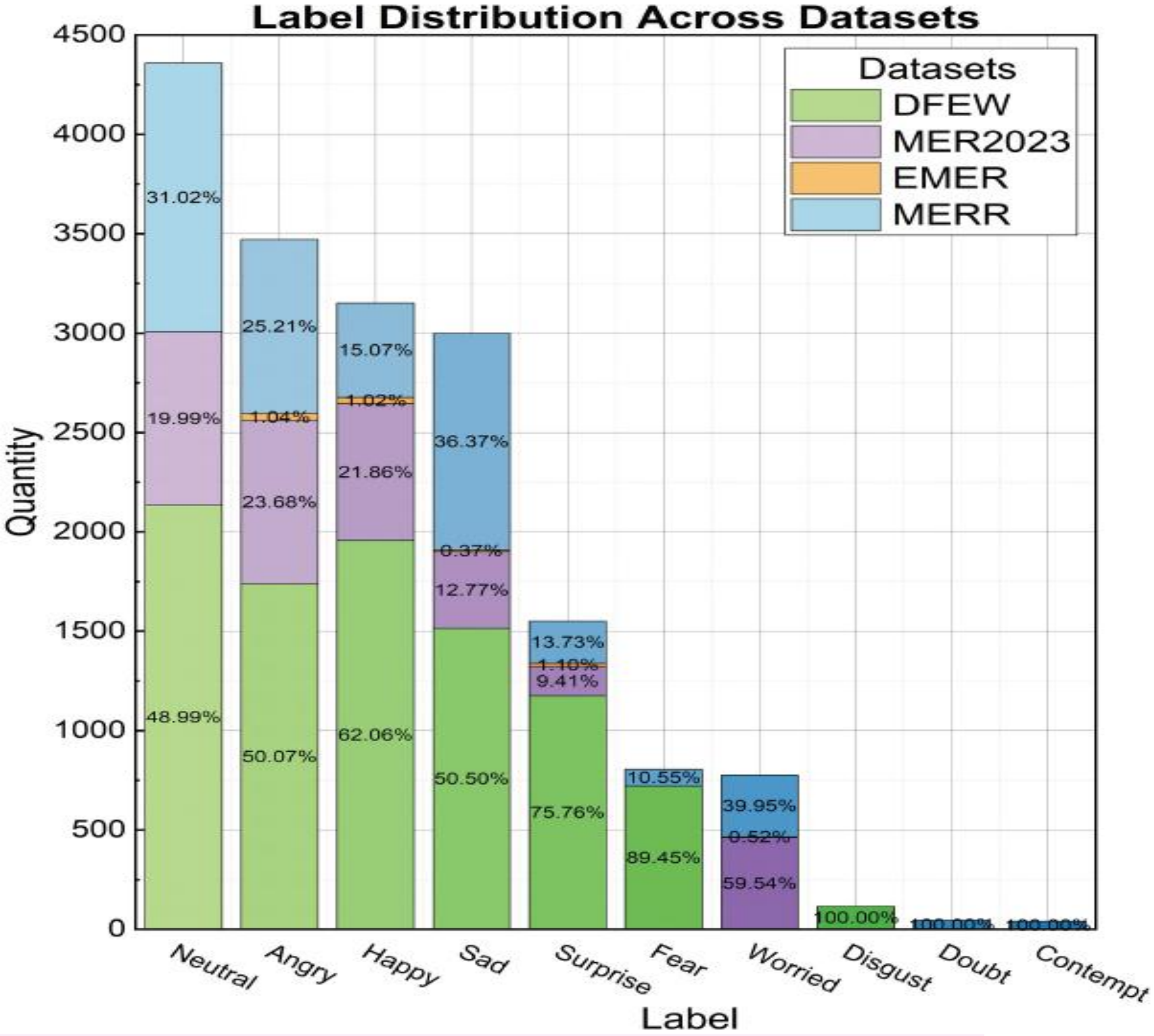
*MERR = multimodal emotion recognition and reasoning

The MERR dataset extends the range of emotional categories and annotations beyond those found in existing datasets

	Sufficient Quantity	Audio Description	Visual Objective Description	Visual Expression Description	Classification Label	**Multimodal Description**
EmoSet [93]	✔	✗	✔	✓✗	✓✗	✗
EmoVIT [92]	✔	✗	✔	✓✗	✓✗	✓✗
DFEW [45]	✔	✗	✗	✗	✔	✗
MER2023 [59]	✔	✗	✗	✗	✔	✗
EMER [62]	✗	✔	✔	✔	✔	✔
MERR (ours)	✔	✔	✔	✔	✔	✔

*MERR = multimodal emotion recognition and reasoning

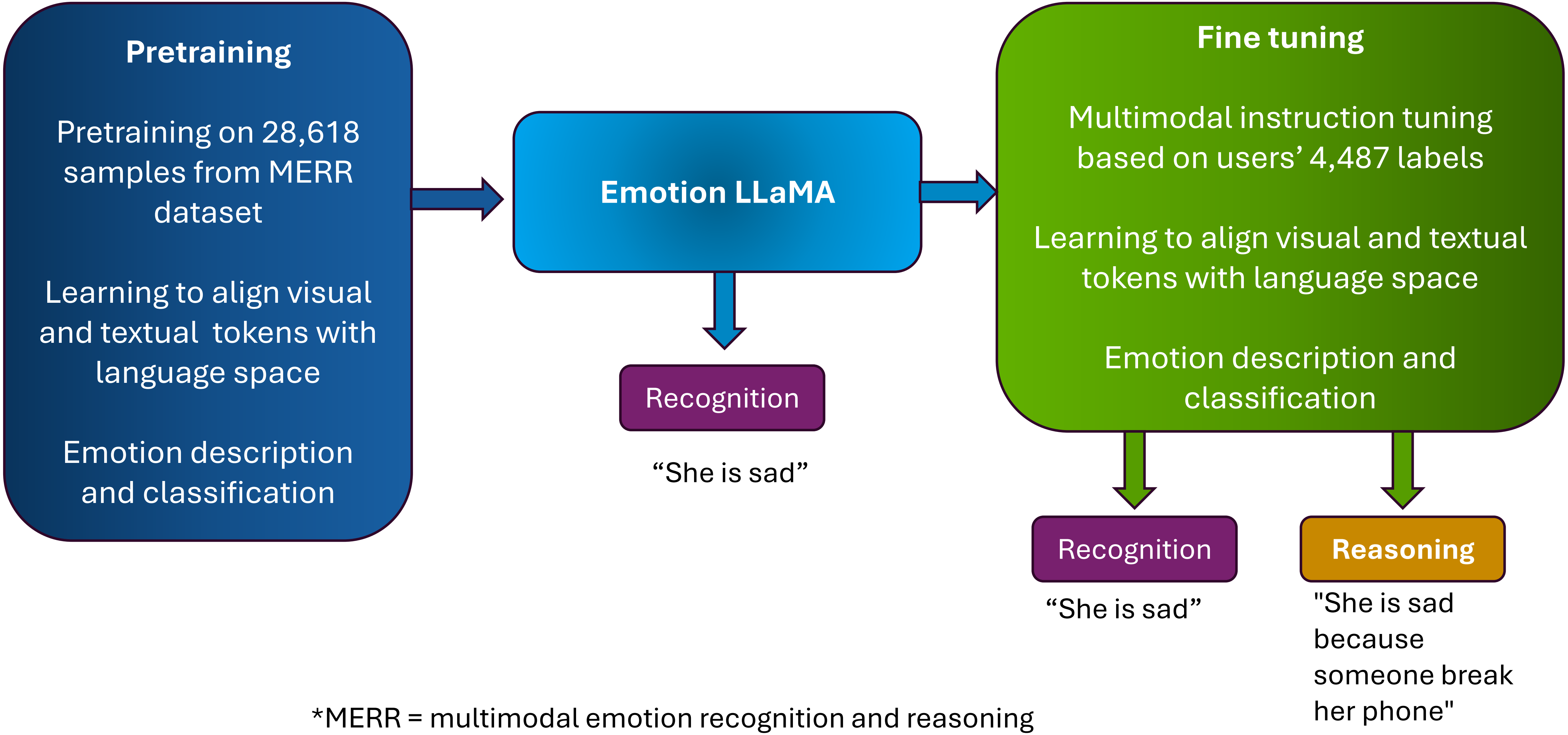
The MERR dataset extends the range of emotional categories and annotations beyond those found in existing datasets



	Su Q	lassification Label	Multimodal Description
EmoSet [93]		✓	✗
EmoVIT [92]		✓	✓
DFEW [45]		✓	✗
MER2023 [59]		✓	✗
EMER [62]		✓	✓
MERR (ours)		✓	✓

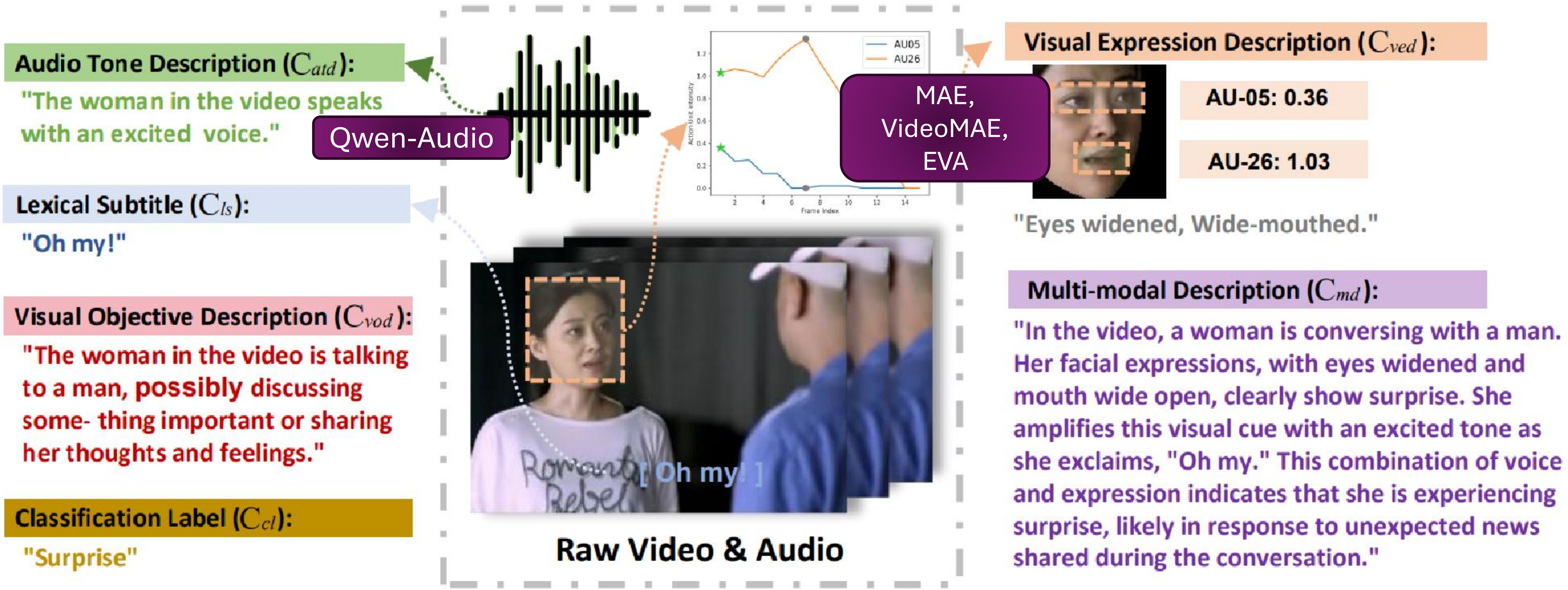
*MERR = multimodal emotion recognition and reasoning

Training of Emotion LLaMA Model procedure

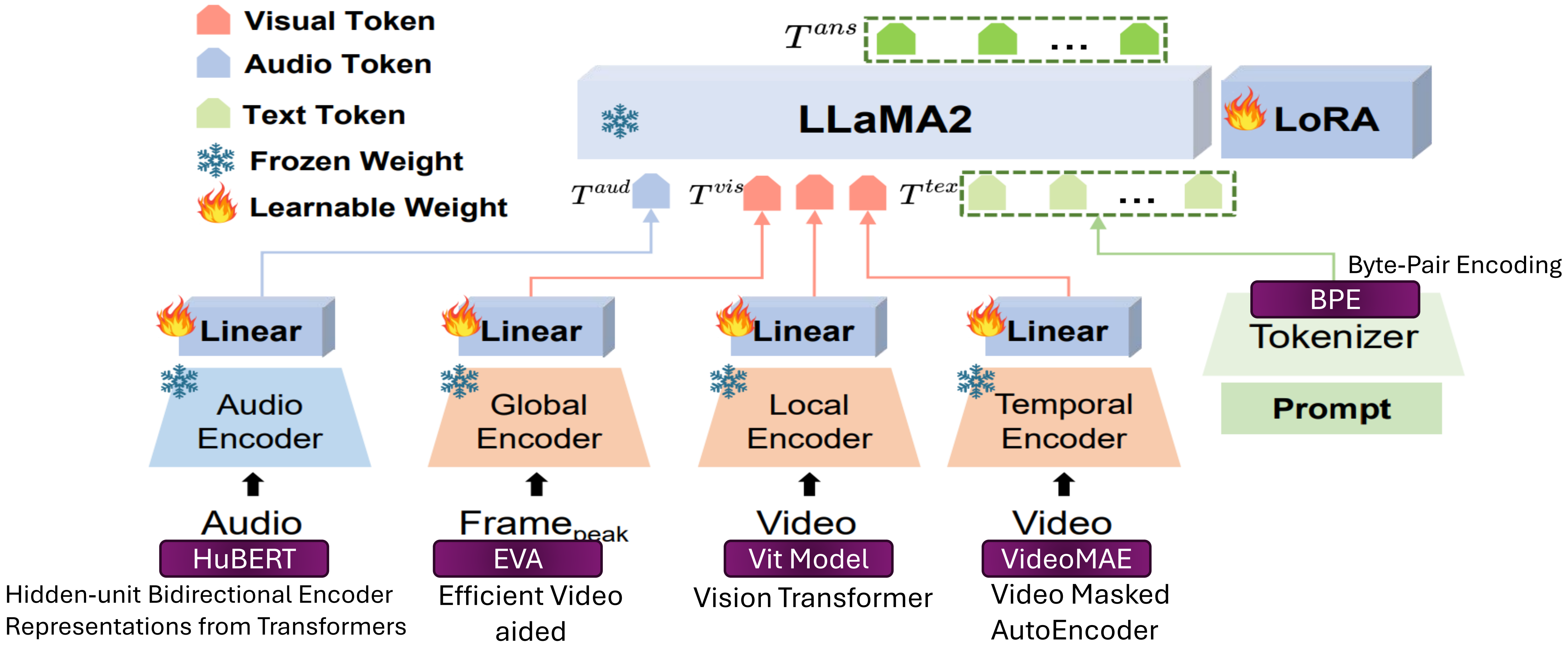


*MERR = multimodal emotion recognition and reasoning

Methodology: single frame example

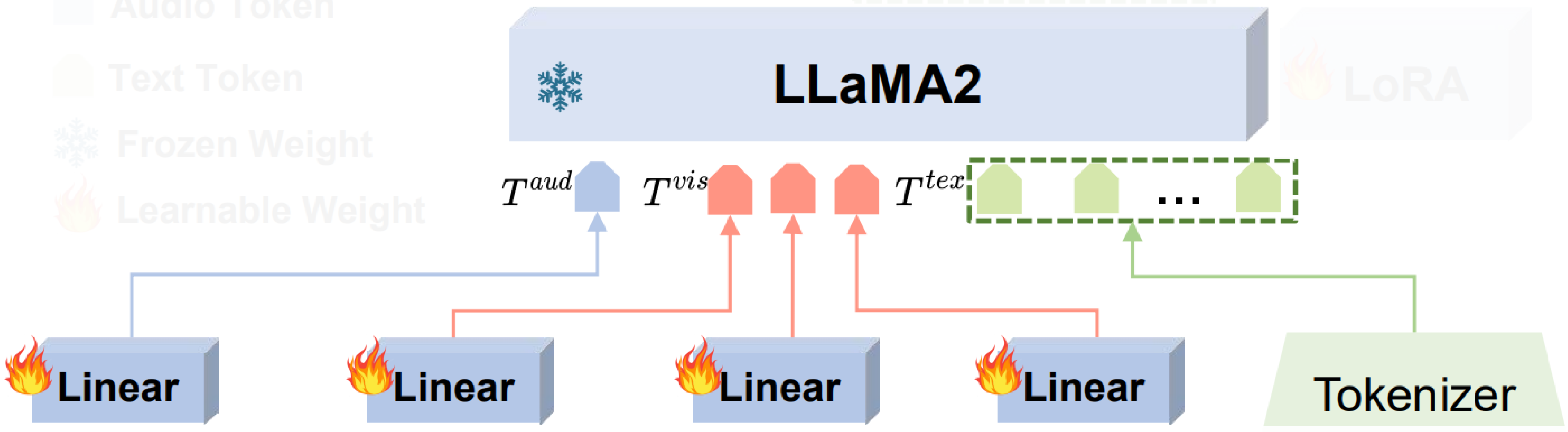


Emotion LLaMA's Architecture



Architecture of Emotion-LLaMA, which integrates audio, visual, and text inputs for multimodal emotional recognition and reasoning

Emotion-LLaMA Training Objective & Optimization



Uses **Language Modeling Loss**

- Measures how well the model predicts ground-truth tokens based on input multimodal data.
- Supports both **emotion recognition** and **emotion reasoning** tasks.
- Encourages the model to generate **contextually relevant and coherent outputs**.

Optimization Strategy

- Employs **Adam Optimizer**
  → Adapts learning rate per parameter based on past gradients.
  → Ensures faster convergence & better generalization.

Efficient fine tunning with Low Rank Adaptation LoRA

- Efficient method to train large language models with **fewer parameters**.
- Only low-rank matrices are adapted, preserving core model knowledge.

Freezes the original model weights

Adds small trainable "adapter" layers

Trains only those adapters
0.5% of model size (34M)

Benefits for Emotion-LLaMA:

- Retains general language understanding.
- Gains emotion-specific knowledge:
  - Tone of speech
  - Facial expression cues
- Achieves efficient specialization without full retraining.

LLaMA performance

Method	Emotion Hap	Sad	Neu	Ang	Sur	Dis	Fea	UAR	WAR
Zero-Shot									
Qwen-Audio [22]	25.97	12.93	67.04	29.20	6.12	0.00	35.36	25.23	31.74
LLaVA-NEXT [64]	57.46	**79.42**	38.95	0.00	0.00	0.00	0.00	25.12	33.75
MiniGPT-v2 [10]	**84.25**	47.23	22.28	20.69	2.04	0.00	0.55	25.29	34.47
Video-LLaVA(image) [63]	37.09	27.18	26.97	58.85	12.97	0.00	3.31	20.78	31.10
Video-LLaVA(video) [63]	51.94	39.84	29.78	58.85	0.00	0.00	2.76	26.17	35.24
Video-Llama [97]	20.25	67.55	**80.15**	5.29	4.76	0.00	9.39	26.77	35.75
GPT-4V [61]	62.35	70.45	56.18	50.69	32.19	10.34	**51.11**	**47.69**	54.85
Emotion-LLaMA (ours)	71.98	76.25	61.99	**71.95**	**33.67**	0.00	3.31	45.59	**59.37**
Fine-tuning									
EC-STFl [45]	79.18	49.05	57.85	60.98	46.15	2.76	21.51	45.35	56.51
Former-DFER [102]	84.05	62.57	67.52	70.03	56.43	3.45	31.78	53.69	65.70
IAL [52]	87.95	67.21	70.10	76.06	62.22	0.00	26.44	55.71	69.24
MAE-DFER [82]	92.92	77.46	74.56	76.94	60.99	**18.62**	42.35	63.41	74.43
VideoMAE [84]	93.09	78.78	71.75	78.74	63.44	17.93	41.46	63.60	74.60
S2D [12]	**93.62**	**80.25**	**77.14**	81.09	64.53	1.38	34.71	61.82	76.03
Emotion-LLaMA (ours)	93.05	79.42	72.47	**84.14**	**72.79**	3.45	**44.20**	**64.21**	**77.06**

Current Challenges in Emotion-LLaMA

- Dataset Annotation Quality:
  – MERR includes coarse- and fine-grained samples, but some mismatches remain
  – Fine-grained subset (4,487 samples) still limited relative to real-world emotion diversity.

- Multimodal Fusion Limitations:
  – Linear projection merges audio/visual features into token space.
  – Lacks deeper cross-modal interactions for subtle cues and temporal dynamics.

- Instruction Tuning Dependence:
  – Relies on MERR instruction datasets; quality of automatic pseudo-labels varies.
  – Model performance sensitive to annotation granularity and instruction design.

Current Challenges in Emotion-LLaMA

Dataset Annotation Quality:

- MERR includes coarse- and fine-grained samples, but some mismatches remain
- Fine-grained subset (4,487 samples) still limited relative to real-world emotion diversity.

• Multimodal Fusion Limitations:

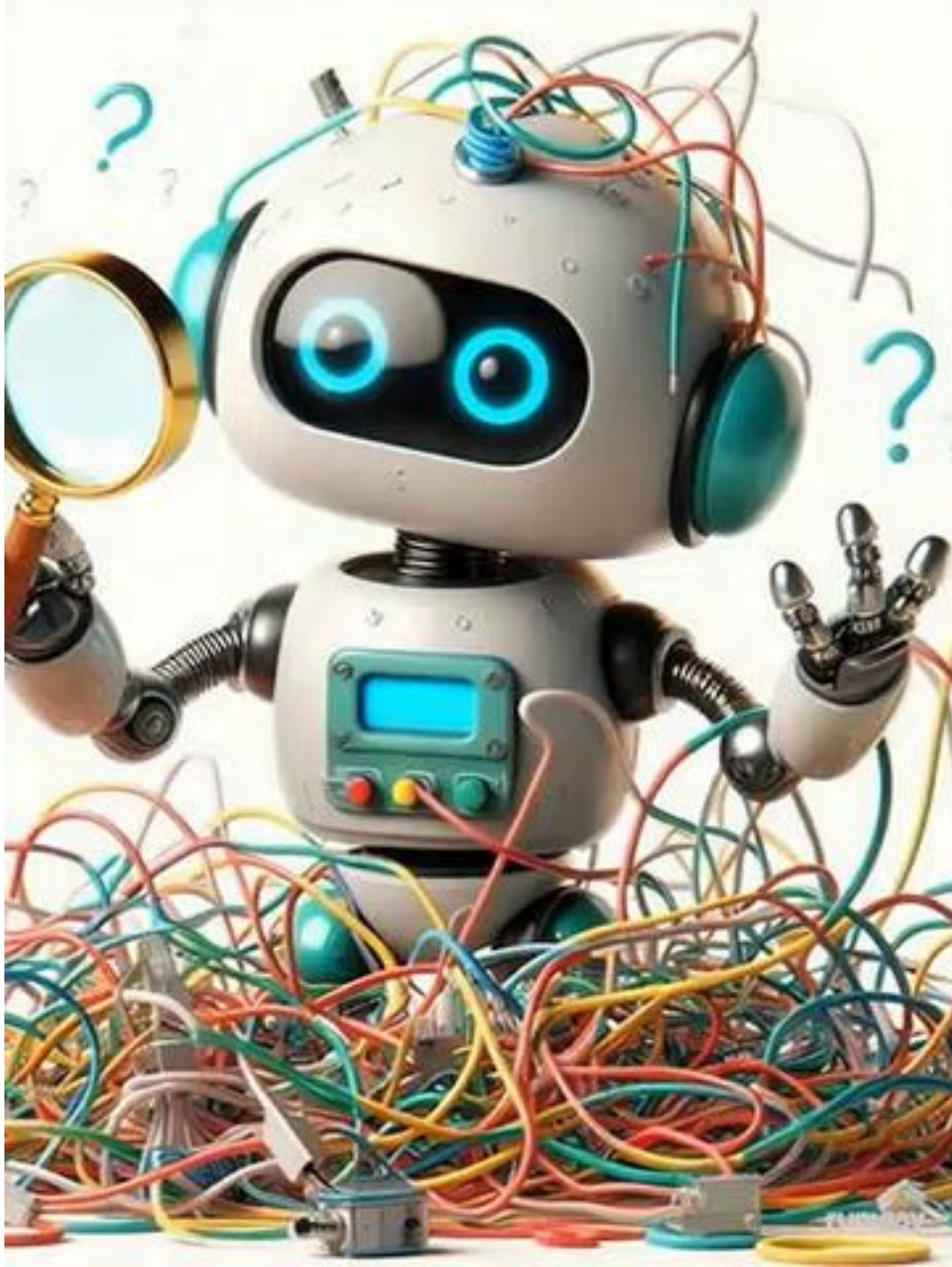
– Linear projection merges audio/visual features into token space.

– Lacks deeper cross-modal interactions for subtle cues and temporal dynamics.

• Instruction Tuning Dependence:

– Relies on MERR instruction datasets; quality of automatic pseudo-labels varies.

– Model performance sensitive to annotation granularity and instruction design.



Current Challenges in Emotion-LLaMA

• Dataset Annotation Quality:

– MERR includes coarse- and fine-grained samples, but some mismatches remain

– Fine-grained subset (4,487 samples) still limited relative to real-world emotion diversity.

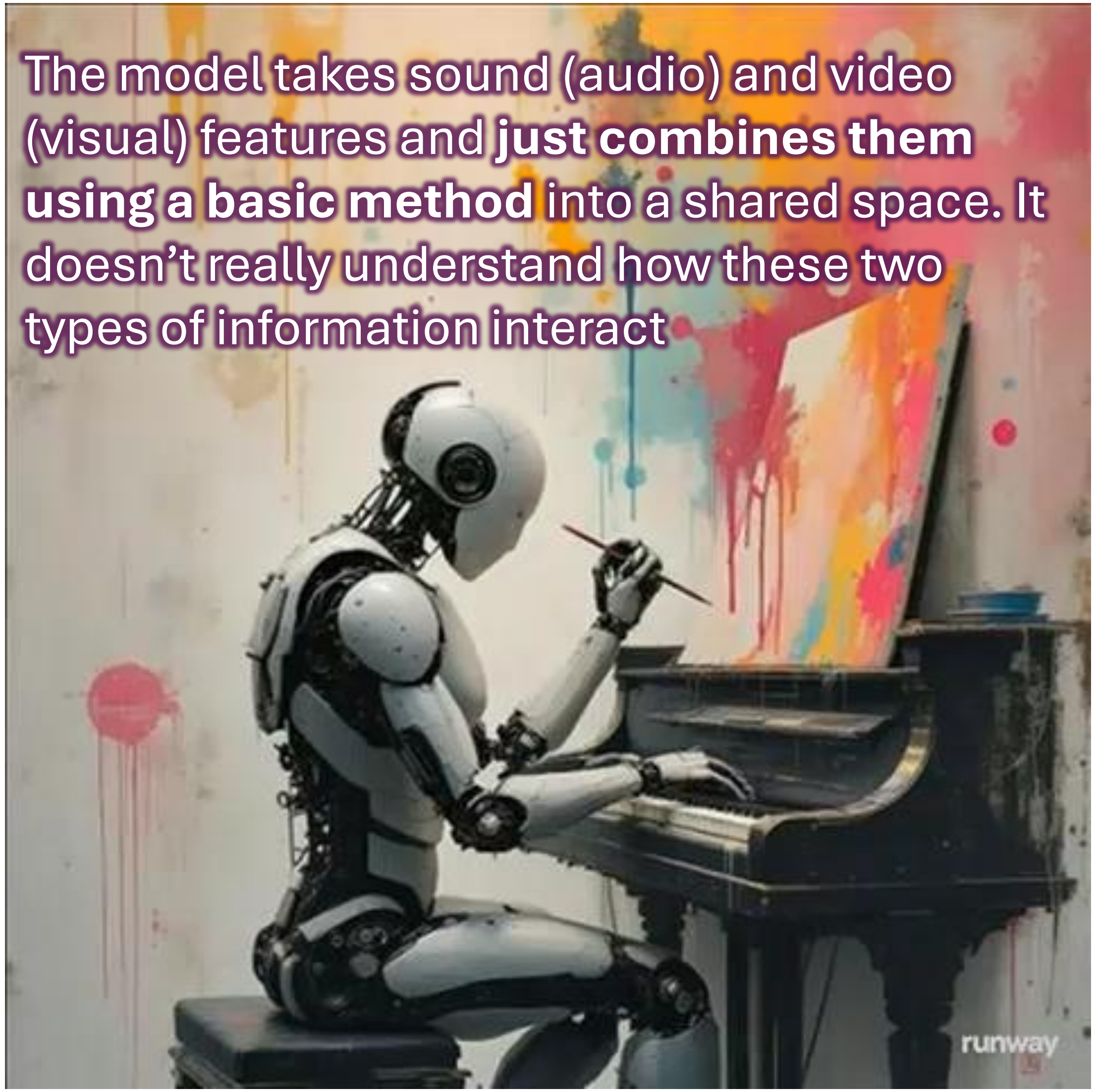
Multimodal Fusion Limitations:

- Linear projection merges audio/visual features into token space.
- Lacks deeper cross-modal interactions for subtle cues and temporal dynamics.

• Instruction Tuning Dependence:

– Relies on MERR instruction datasets; quality of automatic pseudo-labels varies.

– Model performance sensitive to annotation granularity and instruction design.



Current Challenges in Emotion-LLaMA

Pseudo - labels

Pseudo-labels generated automatically by algorithms may lack consistency and accuracy

Instruction design

Unclear instructions as “Describe emotion” vs. “Identify the speaker’s emotional tone from facial and vocal cues” leads to worse performance

Annotation granularity

If the emotion labels in the training data are **too broad or too specific**, the model's accuracy can change:

Labeling something just as "happy" vs. "excited, proud, or relaxed"

Instruction Tuning Dependence:

- Relies on MERR instruction datasets; quality of automatic **pseudo-labels** varies.
- Model performance sensitive to **annotation granularity** and **instruction design**.

Our project aims

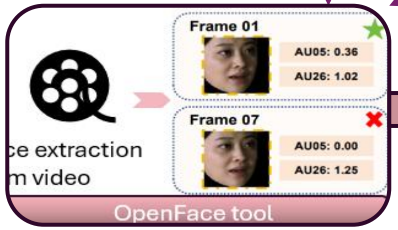
Explore where can we improve the model

Before and after training on masked

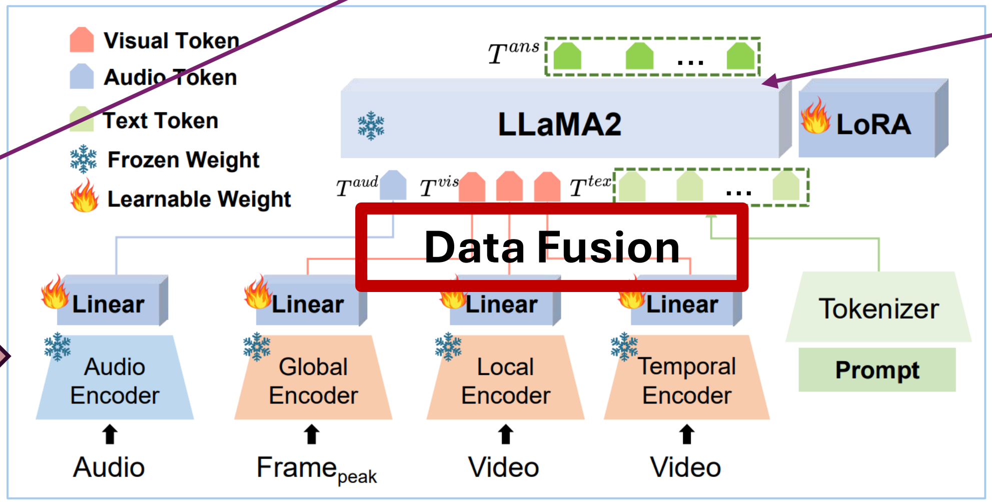
Increasing training dataset

Using newer / better LLM

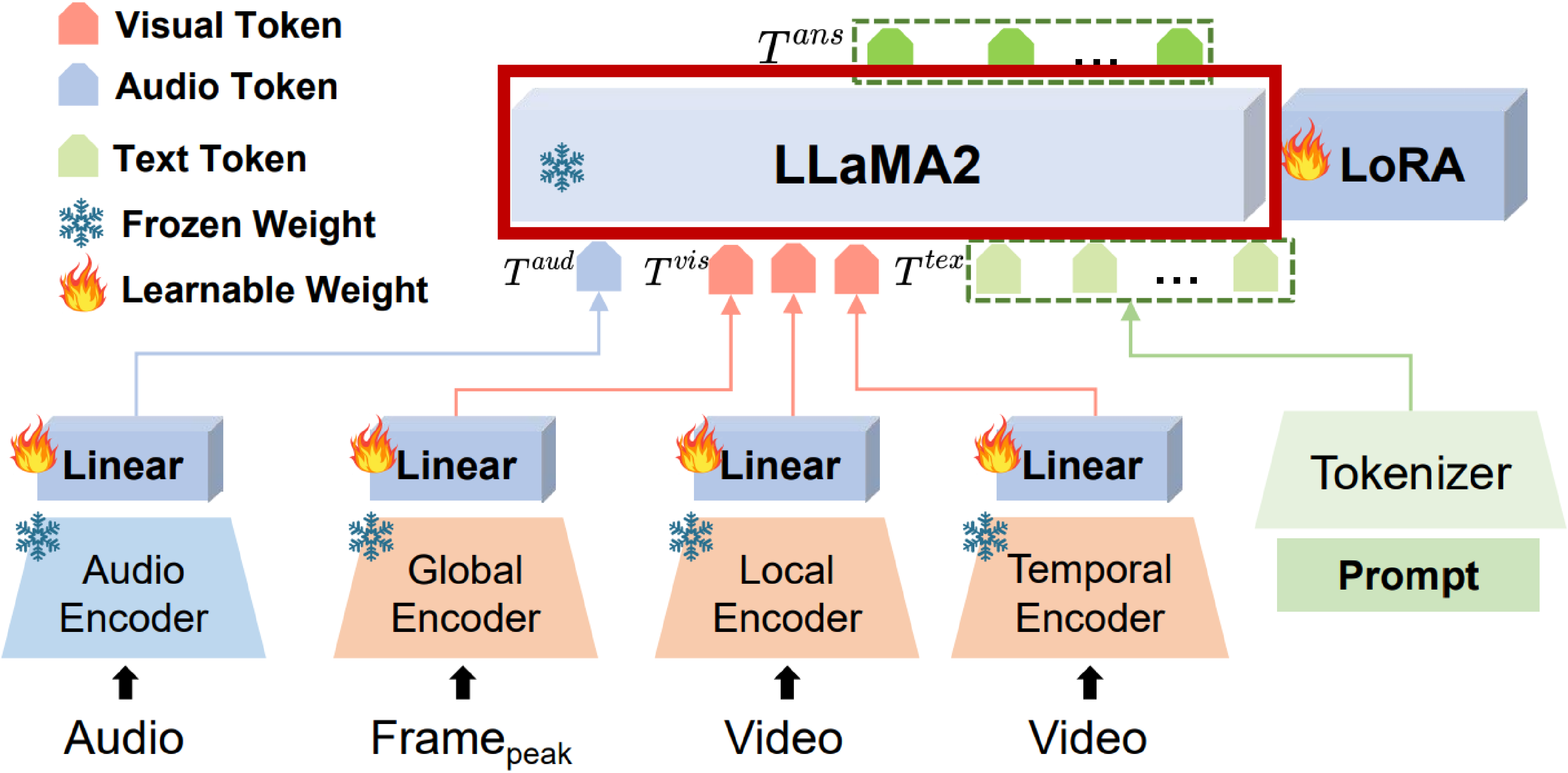
Model Name	Developer	Size / Parameters
Llama 3.1 8B-Instruct	Meta	8B
Mistral 7B Instruct v0.3	Mistral AI	7B
Phi-3 Mini Instruct (3.8B)	Microsoft	3.8B
Gemma 2 9B IT	Google	9B
GPT-4o mini	OpenAI	"Mini" (smaller than GPT-4o)
Claude 3.5 Haiku	Anthropic	"Smallest" in Claude 3.5 family
Gemini 1.5 Flash / 2.5 Flash	Google	"Lightweight"



Data pre-processing



LLaMA’s version as a glass ceiling



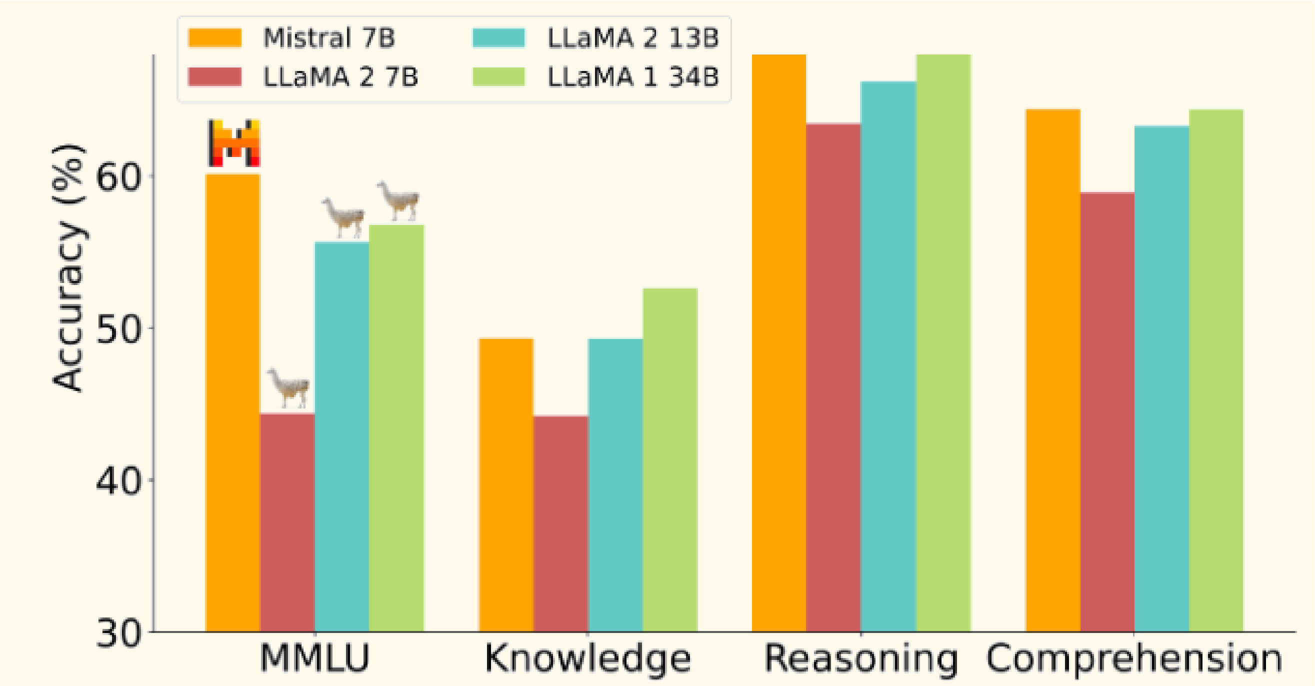
Mistral 7B as a candidate to replace LLaMA2.0 (7B)

A relatively small model:
7.3B Parameters

Performance: Outperforms most of LLaMAs

Open source: allowing for free and unrestricted use by researchers and developers

Fine-tuning: Used with Parameter-Efficient Fine-Tuning (PEFT) techniques like **LORA** for further customization.



MMLU (Massive Multitask Language Understanding) is a benchmark designed to evaluate the multitask capabilities of language models across diverse subjects. It covers 57 tasks spanning topics like humanities, STEM, social sciences, and more, with questions ranging from elementary to professional levels.)

Replacing LLaMA with a More Emotionally Tuned Core

Recommended Core:

Mistral-7B or GPT-NeoX-20B*:

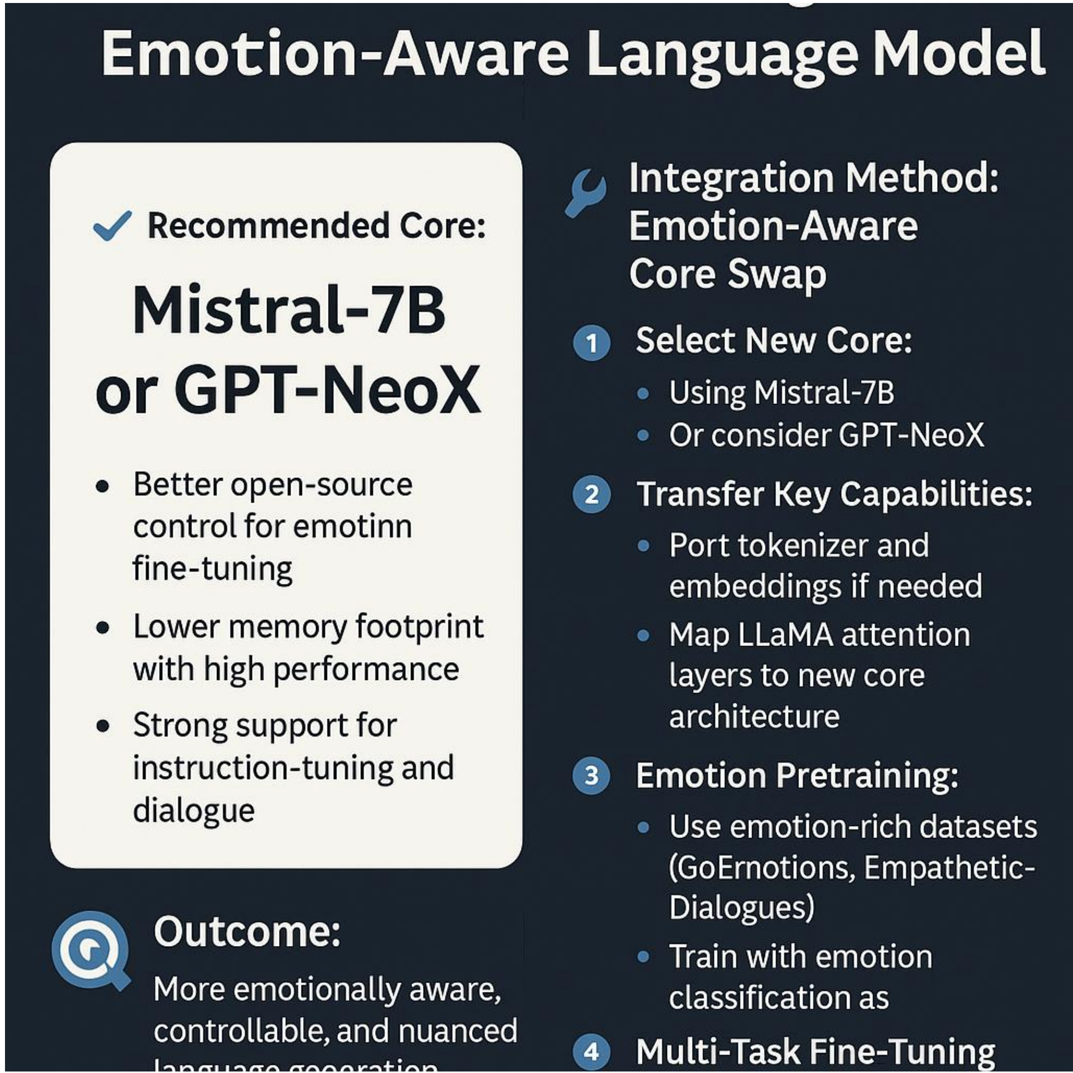
- Better open-source control for emotion fine-tuning
- Lower memory footprint with high performance
- Strong support for instruction-tuning and dialogue

Outcome:

More emotionally aware, controllable, and nuanced language generation

Better performance in dialogue, therapy bots, storytelling, and character design

*** AffectGPT is considered to have the top score in emotion understanding**



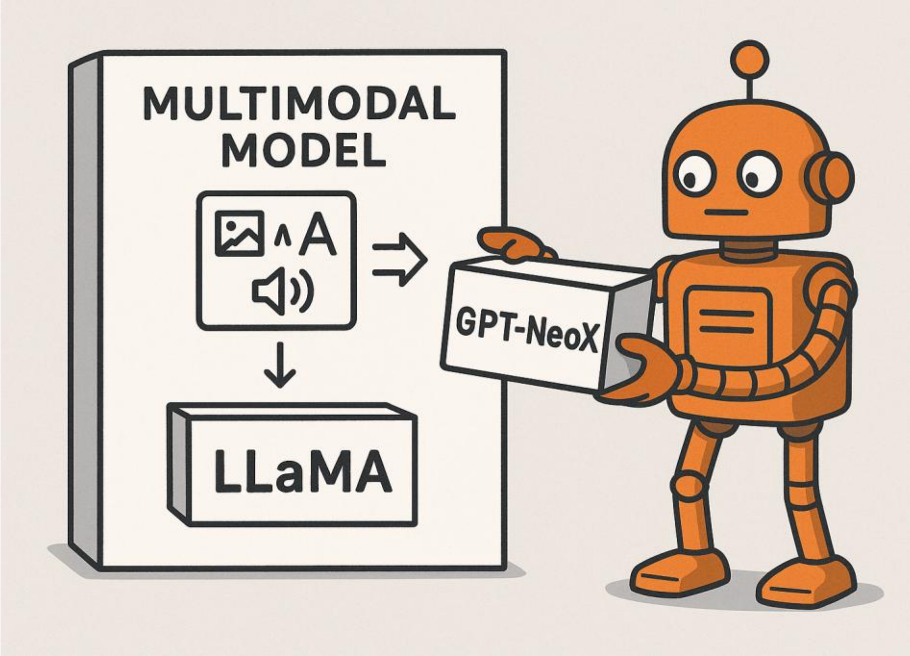
Integration Method: Emotion-Aware Core Swap

The Requirements to Replace LLaMA Core

Transfer Key Capabilities:

Port tokenizer and embeddings

Map LLaMA attention layers to the new core architecture



Emotion Pretraining:

Use emotion-rich datasets: GoEmotions, Empathetic Dialogues

Train with emotion classification as auxiliary task

Multi-Task Fine-Tuning:

Blend standard language modeling with emotion conditioning

Inject "emotion tags" as soft prompts or embeddings

Evaluation:

Benchmark vs. LLaMA using:

- EMER (Emotion Merging Reasoning)
- MER2023 Challenge (F1 score)
- DFEW (Dynamic Facial Expression in the Wild

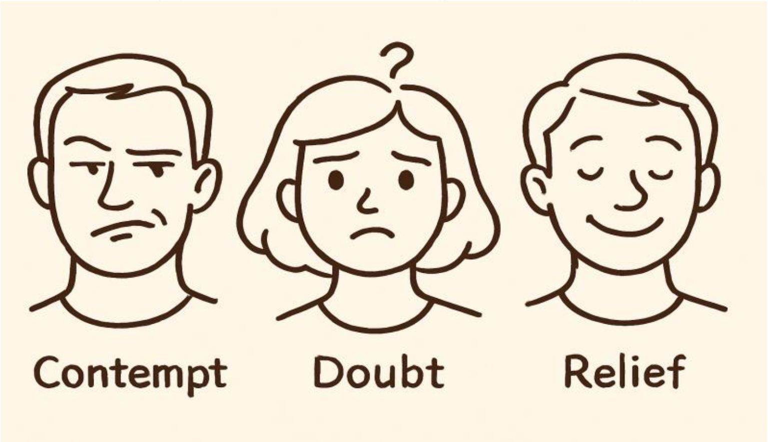
Alternative 1:Enhanced Data Quality & Diversity

Iterative Pseudo-Label Refinement:

- Use active human-in-the-loop verification on low-confidence MERR samples.
- Incrementally correct mismatches identified in coarse-to-fine pipeline or add more samples from this category

Expand Fine-Grained Annotations:

- Increase fine-grained subset beyond 4,487 by sampling diverse contexts from MER2023.
- Incorporate more nuanced emotion categories as “contempt,” “doubt,” “relief” etc.



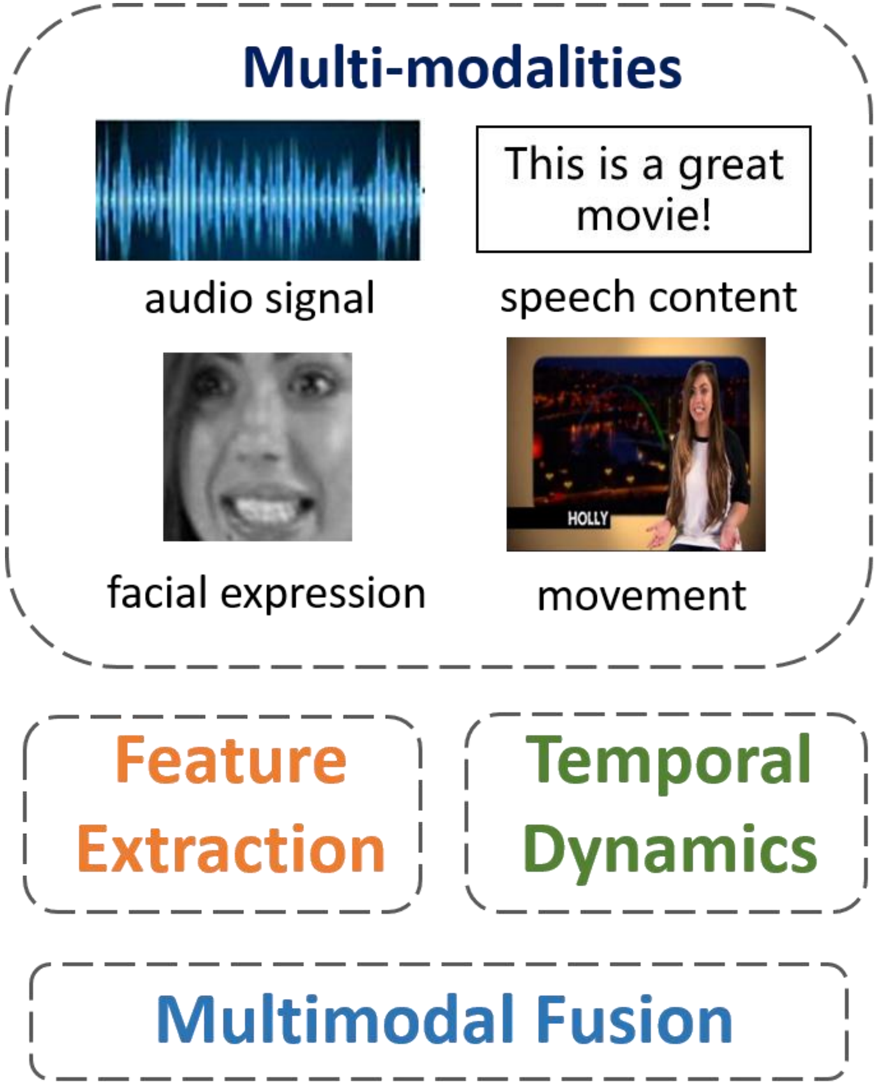
Alternative 2: Advanced Multimodal Fusion

Cross-Modal Transformer Layers:

- Integrate audio, visual, and text features via cross-attention at multiple depths instead of 3 linear.
- Enables fine-grained interactions, improving micro-expression and tone recognition.

Modality-Specific Prompt Tokens:

- Introduce learned special tokens per modality to guide attention
- Improve instruction tuning by signaling whether to focus on text, audio nuance or facial detail.



Summary & Next Steps

Upgraded Backbone:

Replace LLaMA2-chat with Mistral, GPT, LLaMA3 or like for better reasoning, fluency, and emotion alignment.

Enhanced Data:

Refine MERR annotations, expand fine-grained labels, apply augmentation.

Improved Fusion:

Implement cross-modal transformers, deeper temporal modeling, modality tokens.

Next Steps:

1. Evaluate newer core (Mistral/LLaMA3) on EMER, MER2023, and DFEW tasks.
2. Collect and validate augmented fine-grained samples with human review.
3. Prototype cross-modal transformer modules and benchmark on MER2023/DFEW.

Milestones

Installing the model

Understanding the barriers

Replacing LLM core / improve fusion

Alternatively Increase data size and variability

Alternatively Training on masked faces

Conclusions

Thanks