Multi Modal Learning Emotion Recognition

Focusing on Emotion LLaMA

Team members:
Linoy Halifa, Bezalel Itzhaky, Dr. Ezra Ella.

Supervisor:
Dr. Yehudit Aperstein

Agenda

- Linoy
  - The Dataset - Overview and key characteristics
  - Our Process – End-to-end workflow
  - Unimodal Models – Architectures and results
- Bentzi
  - Multimodal Model – Methods and experiments variants
  - Improvements Implemented – Methods and optimizations
  - Challenges Faced – Issues and solutions
- Ezra
  - Gradio Interface – Demonstration of the interactive tool
  - Future Work – Potential extensions and improvements
  - Q&A- Questions and Answers

The Dataset - Overview and key characteristics

- Data Set Name : MER2025 Track 1 (Multimodal Emotion Recognition 2025).
- Contents:
  - Three modalities: Video, Audio, Text.
  - Pre-extracted features (embeddings) for each modality.
- Goal: Recognize human emotions (Sad, Happy, Angry, Neutral, Fear, Disgust) using all three modalities.
- Number of samples (if available): Indicate how many video clips/samples are in train, and test sets.

Our Process – End-to-end workflow

Visual Feature Extraction

```
[1] visual feature extraction
cd feature_extraction/visual
python -u extract_manet_embedding.py    --dataset=MER2025 --feature_level='UTTERANCE'                                          --g
python -u extract_emonet_embedding.py   --dataset=MER2025 --feature_level='UTTERANCE'                                          --g
python -u extract_ferplus_embedding.py  --dataset=MER2025 --feature_level='UTTERANCE' --model='resnet50_ferplus_dag'          --g
python -u extract_ferplus_embedding.py  --dataset=MER2025 --feature_level='UTTERANCE' --model='senet50_ferplus_dag'           --g
python -u extract_vision_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='clip-vit-base-patch32'   --g
python -u extract_vision_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='clip-vit-large-patch14' --g
python -u extract_vision_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='videomae-base'          --g
python -u extract_vision_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='videomae-large'         --g
python -u extract_vision_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='dinov2-large'           --g
```

Our Process – End-to-end workflow

Visual Feature Extraction

```
[1] visual feature extraction
cd feature_extraction/visual
python -u extract_manet_embedding.py    --dataset=MER2025 --feature_level='UTTERANCE'                                    --g
python -u extract_emonet_embedding.py   --dataset=MER2025 --feature_level='UTTERANCE'                                    --g
python -u extract_ferplus_embedding.py  --dataset=MER2025 --feature_level='UTTERANCE' --model='resnet50_ferplus_dag'    --g
python -u extract_ferplus_embedding.py  --dataset=MER2025 --feature_level='UTTERANCE' --model='senet50_ferplus_dag'     --g
python -u extract_vision_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='clip-vit-base-patch32'  --g
python -u extract_vision_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='clip-vit-large-patch14' --g
python -u extract_vision_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='videomae-base'        --g
python -u extract_vision_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='videomae-large'       --g
python -u extract_vision_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='dinov2-large'        --g
```

CLIP-Vit-Base-Patch32- is a vision-language model developed by OpenAI that learns to connect images with text.
CLIP - (Contrastive Language-Image Pre-training): Trains on millions of image–text pairs to map both modalities into the same embedding space.
ViT-Base - Patch32: Uses a Vision Transformer (ViT) architecture, where each image is divided into 32×32 pixel patches that are processed by the transformer.
Base / large : the size of the mode. Base – medium version , large – large version.

Our Process – End-to-end workflow

Audio Feature Extraction

```
[2] audio feature extraction
cd feature_extraction/audio
python -u extract_audio_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-hubert-base'     --g
python -u extract_audio_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-hubert-large'    --g
python -u extract_audio_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-wav2vec2-base'   --g
python -u extract_audio_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-wav2vec2-large'  --g
python -u extract_audio_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='wavlm-base'              --g
python -u extract_audio_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='whisper-large-v2'       --g
```

Our Process – End-to-end workflow

Audio Feature Extraction

```
[2] audio feature extraction
cd feature_extraction/audio
python -u extract_audio_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-hubert-base'     --g
python -u extract_audio_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-hubert-large'    --g
python -u extract_audio_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-wav2vec2-base'   --g
python -u extract_audio_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-wav2vec2-large'  --g
python -u extract_audio_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='wavlm-base'              --g
python -u extract_audio_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='whisper-large-v2'       --g
```

Chinese – Hubert : designed to extract **high-quality audio embeddings** (features) from raw speech signals, capturing tone, prosody, and acoustic patterns. **HuBERT** is trained with **self-supervised learning**, meaning it learns to represent speech by predicting masked parts of the audio signal without needing labeled data.

Our Process – End-to-end workflow

Text Feature Extraction

```
[3] lexical feature extraction
cd feature_extraction/text
python extract_text_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-roberta-wwm-ext'       -
python extract_text_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-roberta-wwm-ext-large' -
python extract_text_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-macbert-base'          -
python extract_text_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-macbert-large'         -
python extract_text_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='bloom-7b1'                     -
```

Our Process – End-to-end workflow

Text Feature Extraction

```
[3] lexical feature extraction
cd feature_extraction/text
python extract_text_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-roberta-wwm-ext'
python extract_text_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-roberta-wwm-ext-large'
python extract_text_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-macbert-base'
python extract_text_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='chinese-macbert-large'
python extract_text_huggingface.py --dataset=MER2025 --feature_level='UTTERANCE' --model_name='bloom-7b1'
```

Chinese-RoBERTa-wwm-ext – NLP model based on the RoBERTa architecture , an improved variant of BERT.
Layer size (base): 12 transformer layers 12 attention heads and a hidden size of 768
Chinese-RoBERTa-wwm-ext- large - **24 Transformer layers**, 16 attention heads, and a hidden size of 1024.
Provides deeper semantic understanding and more powerful text embeddings

Unimodal Models – Architectures and results

```
ModelsArry = [Audio_Model1, Audio_Model2, Text_Model1, Text_Model2]
for each_model in ModelsArry:
    cmd = (
                f"python -u main-release.py "
                f"--model attention "
                f"--feat_type utt "
                f"--dataset MER2025 "
                f"--audio_feature {each_model} "
                f"--text_feature {each_model}  "
                f"--video_feature {each_model}  "
                f"--epochs {epocs} "
                f"--gpu 0 "
            )
```

Unimodal Models – Architectures and results

```
#Bentzi replace for gpu
    model = get_models(args).to(device)
    reg_loss = MSELoss().to(device)
    cls_loss = CELoss().to(device)
    # model = get_models(args).cuda()
    # reg_loss = MSELoss().cuda()
    # cls_loss = CELoss().cuda()
```

```
loss = interloss + cls_loss(emos_out, emos) + reg_loss(vals_out, vals)
```

Unimodal Models – Architectures and results

Most Common			attention	attention	attention	attention	attention	attention
-			chinese-hubert -base-UTT	chinese-hubert -large-UTT	chinese-roberta -wwm-ext-UTT	chinese-roberta -wwm-ext-large-UTT	clip-vit-base-patch32-UTT	clip-vit-large-patch14-UTT
74.37%			74.37%	73.21%	49.54%	52.92%	59.55%	63.18%
59.19%			59.19%	57.75%	32.93%	35.99%	42.40%	46.18%
Majority Vote	Total aggreed		uni_1	uni_2	uni_3	uni_4	uni_5	uni_6
sad	4	4/6	sad	sad	sad	sad	happy	happy
happy	5	5/6	happy	happy	happy	happy	neutral	happy
sad	3	3/6	neutral	angry	sad	neutral	sad	sad
angry	3	3/6	angry	happy	neutral	neutral	angry	angry
sad	3	3/6	sad	sad	angry	angry	happy	sad
sad	2	2/6	sad	sad	neutral	neutral	happy	happy
sad	3	3/6	sad	happy	sad	sad	fear	fear
happy	5	5/6	happy	happy	happy	happy	neutral	happy
happy	5	5/6	happy	happy	happy	happy	happy	fear

MULTI MODAL : BENTZI

Multi Modal Agenda

- Multi Modal Terminology
- Illustration of MER Inputs and Annotations
- Feature extraction
- Project General overview
- Track 1 Processing Pipeline
- Training results
- Project Challenges

Multi Modal Terminology (1/4)

Term	Meaning
Modality	A type or source of data (e.g., text, audio, video, physiological signals).
Multimodal	Involving multiple modalities (e.g., combining audio and text for emotion recognition).
Feature Extraction	The process of converting raw data into a set of usable features for model input.
Fusion	Combining features or decisions from multiple modalities. Can be **early (feature-level)** or **late (decision-level)** **LLM most common Hybrid**.
Alignment	Ensuring data from different modalities corresponds in time or meaning (e.g., aligning words with audio frames).

Multi Modal Terminology(2/4)

Term	Meaning
Representation Learning	Learning how to encode each modality into a vector format that captures relevant information.
Cross-modal	Involving interactions between different modalities (e.g., using text to enhance understanding of video).
Attention Mechanism	A technique that allows the model to focus on important parts of the input, often used to weigh modalities or time steps differently.
Temporal Modeling	Capturing how data changes over time (especially important in audio and video).
Encoder	A model component that transforms input data into a hidden representation (often per modality).

Multi Modal Terminology (3/4)

Term	Meaning
Decoder	A model component that transforms hidden representations into outputs (e.g., emotion labels).
Multimodal Fusion Network	A neural network that integrates multiple modalities into a shared representation.
Missing Modality	A situation where data for one or more modalities is unavailable; models may need to handle this robustly.
Data Synchronization	Adjusting and aligning multimodal data sources so that they can be jointly processed (e.g., syncing audio with facial expressions).
Late Fusion	Combining predictions from separate unimodal models.

Multi Modal Terminology (4/4)

Term	Meaning
Early Fusion	Combining raw or preprocessed features from different modalities before model input.
Self-Attention	A mechanism for weighting different parts of a single modality's sequence, as used in Transformers.
Cross-Attention	A mechanism where one modality attends to another (e.g., audio attends to text).
Multimodal Transformer	A Transformer model adapted to process and combine multiple modalities.
Emotion Recognition	The task of classifying emotional states based on data (e.g., speech tone, text sentiment, facial expressions).

MER Challenge 2025

Illustration of MER Inputs and Annotations

Audio Tone Description (C_{atd}):

"The woman in the video speaks with an excited voice."

Lexical Subtitle (C_{ls}):

"Oh my!"

Visual Objective Description (C_{vod}):

"The woman in the video is talking to a man, possibly discussing some- thing important or sharing her thoughts and feelings."

Classification Label (C_{cl}):

"Surprise"

Raw Video & Audio

Visual Expression Description (C_{ved}):

AU-05: 0.36

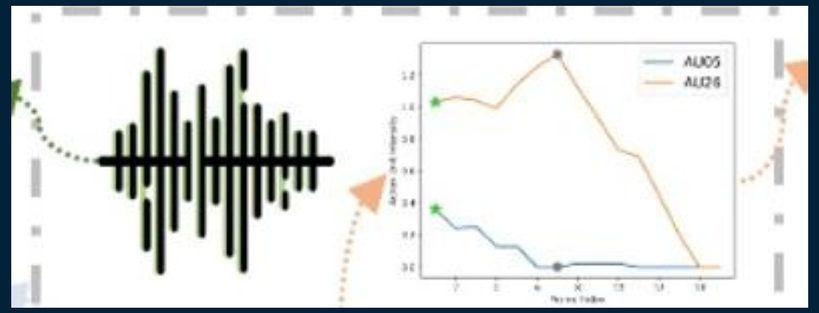
AU-26: 1.03

"Eyes widened, Wide-mouthed."

Multi-modal Description (C_{md}):

"In the video, a woman is conversing with a man. Her facial expressions, with eyes widened and mouth wide open, clearly show surprise. She amplifies this visual cue with an excited tone as she exclaims, "Oh my." This combination of voice and expression indicates that she is experiencing surprise, likely in response to unexpected news shared during the conversation."

Multimodal Emotion Recognition – Track 1: MER Our Pipeline

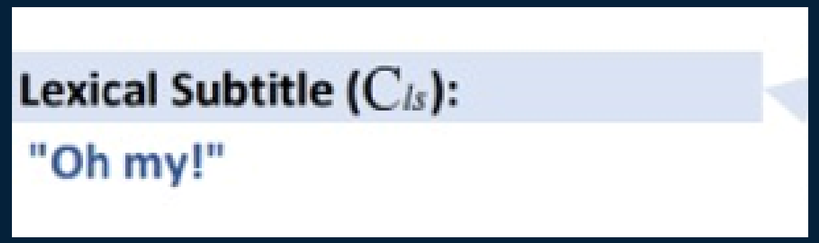


Audio

chinese-hubert-large-UTT ===> dim is (1, 1024)chinese-hubert-base-UTT ===> dim is (1, 768)

Visual

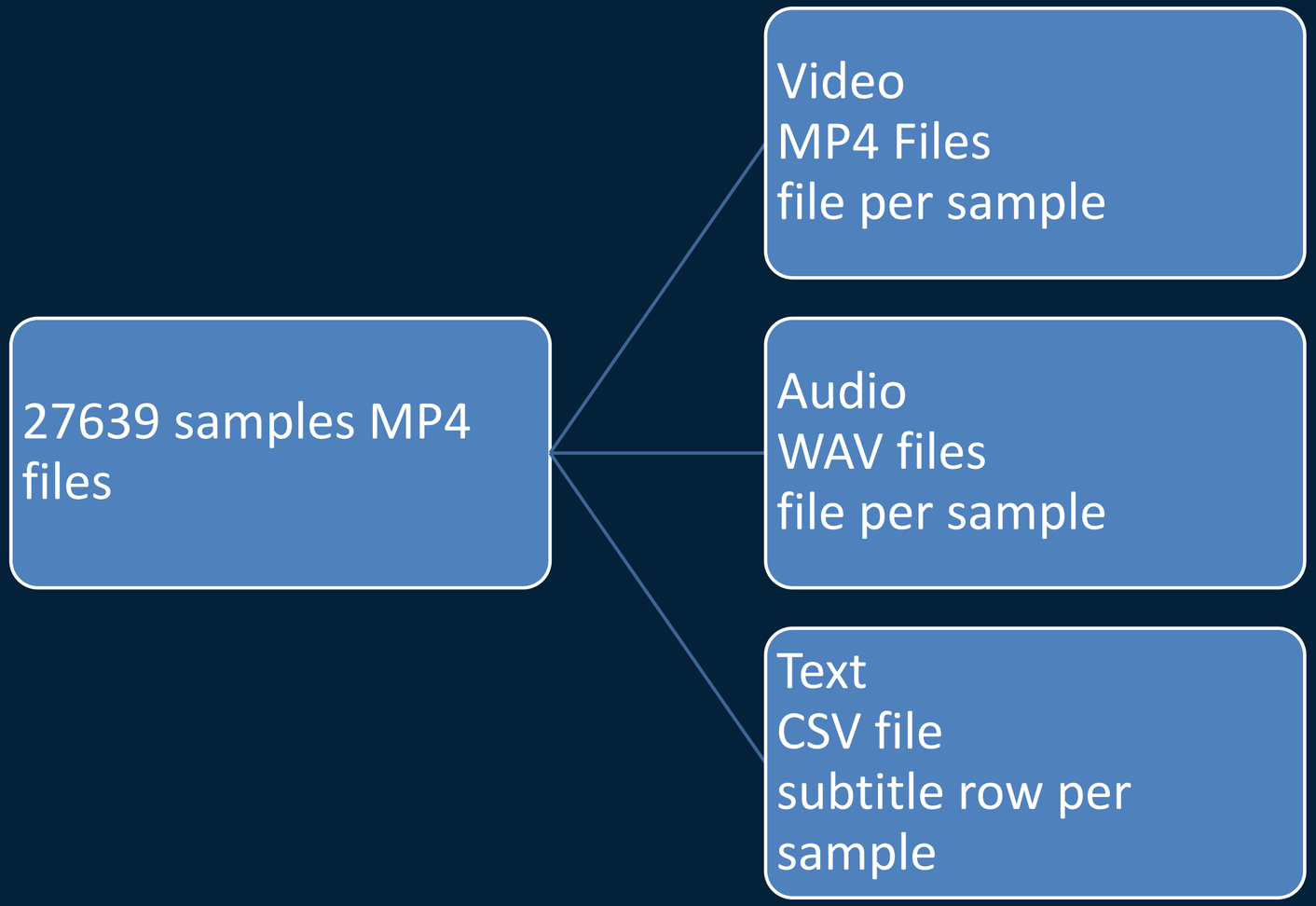
clip-vit-base-patch32-UTT ===> dim is (1, 512)clip-vit-large-patch14-UTT ===> dim is (1, 768)



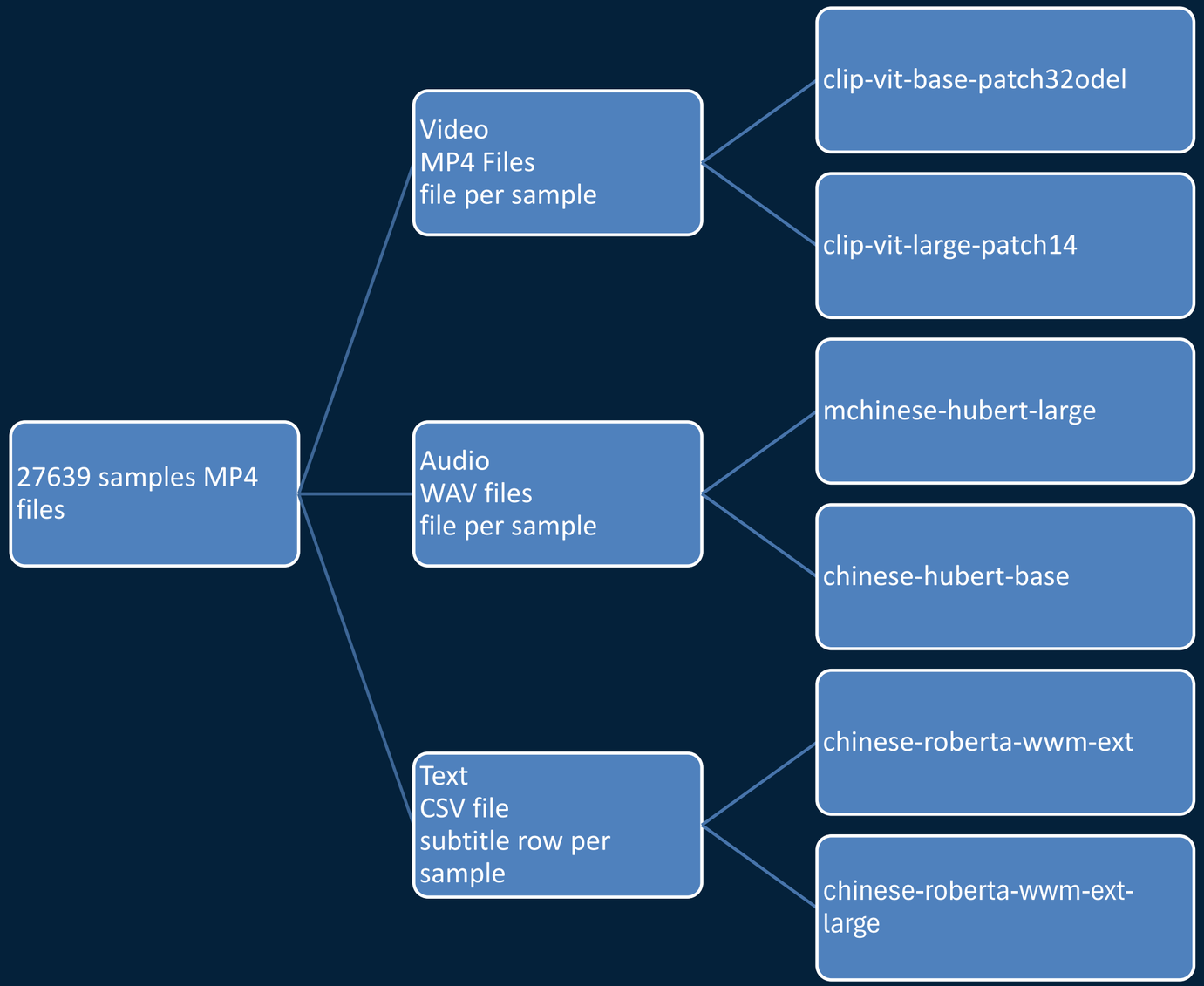
Text

chinese-roberta-wwm-ext-UTT ===> dim is (1, 768)chinese-roberta-wwm-ext-large-UTT ===> dim is (1, 1024)

Our project MER 2025
MER Tools – feature extraction

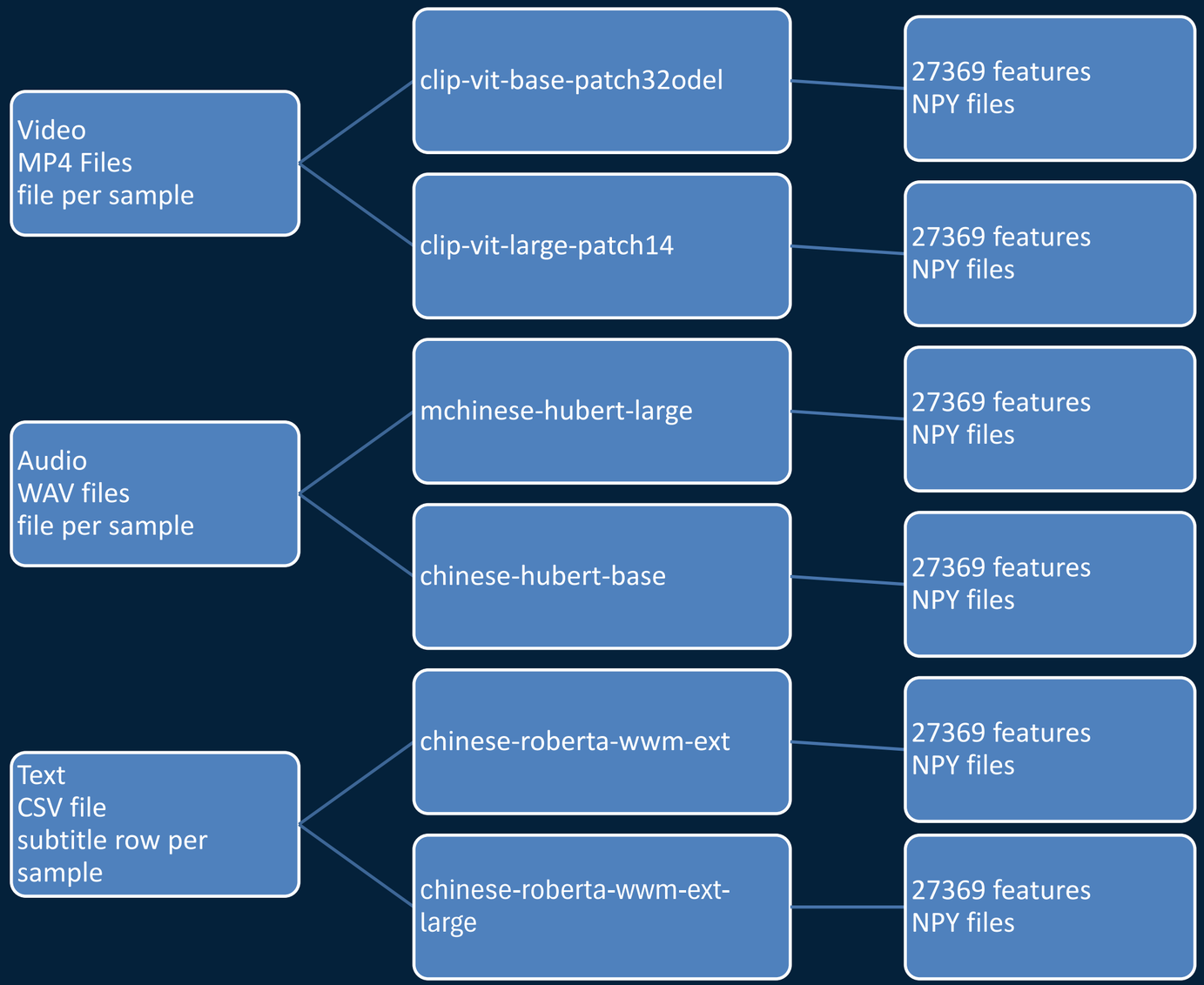


MER Tools – feature extraction
Choose model

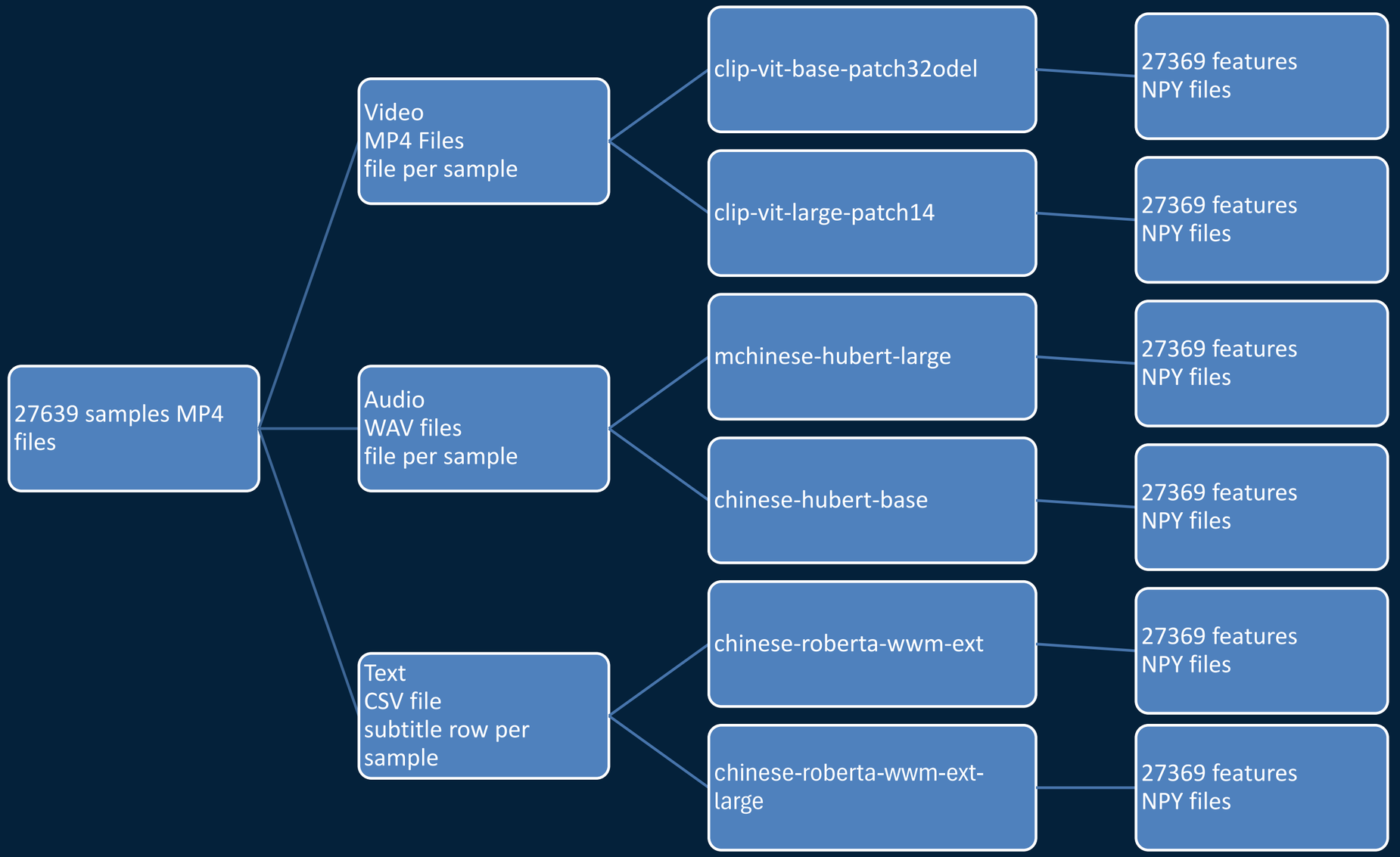


MER Tools – feature extraction

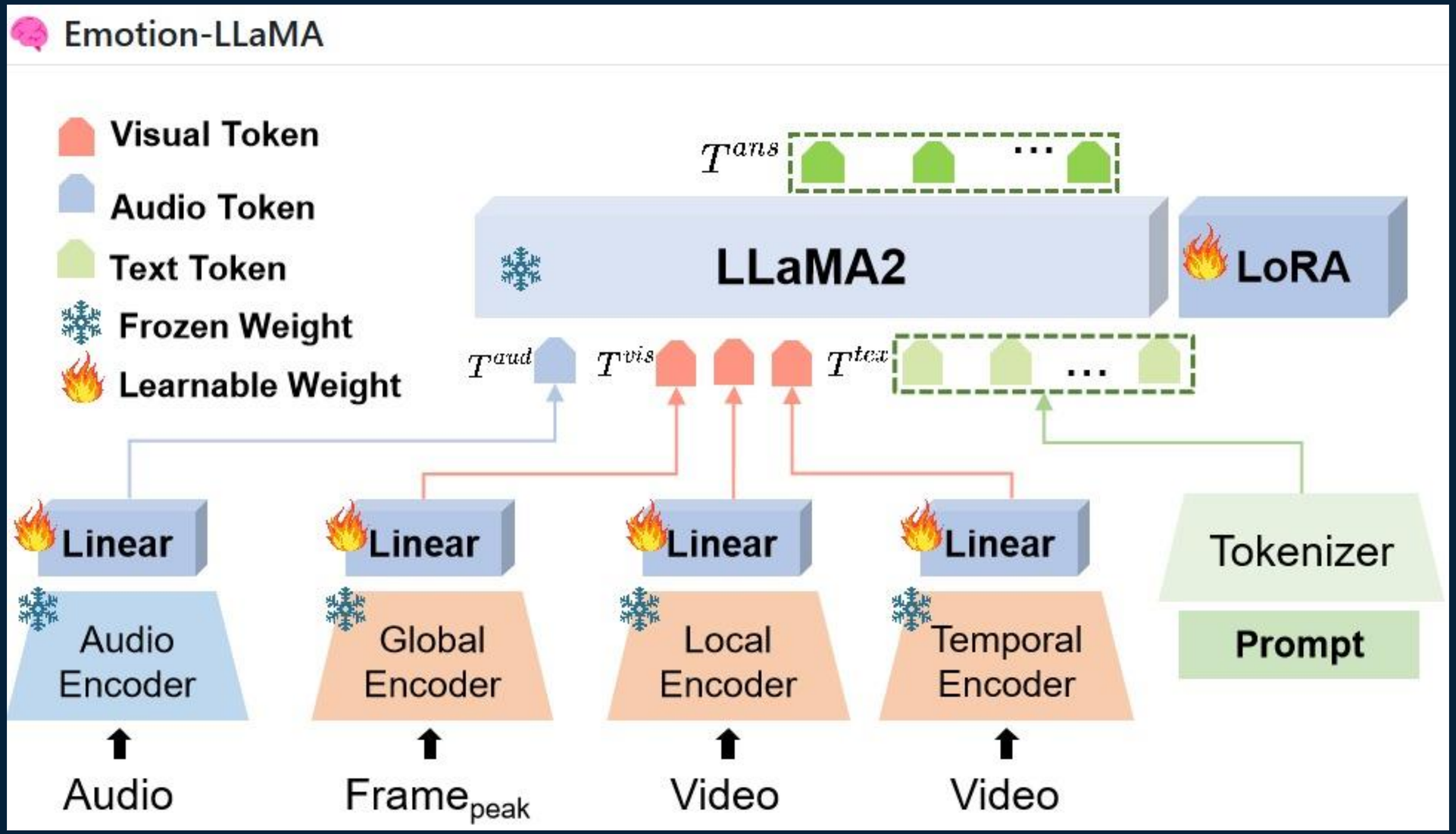
Extract features



Our project MER 2025
MER Tools feature extraction overview

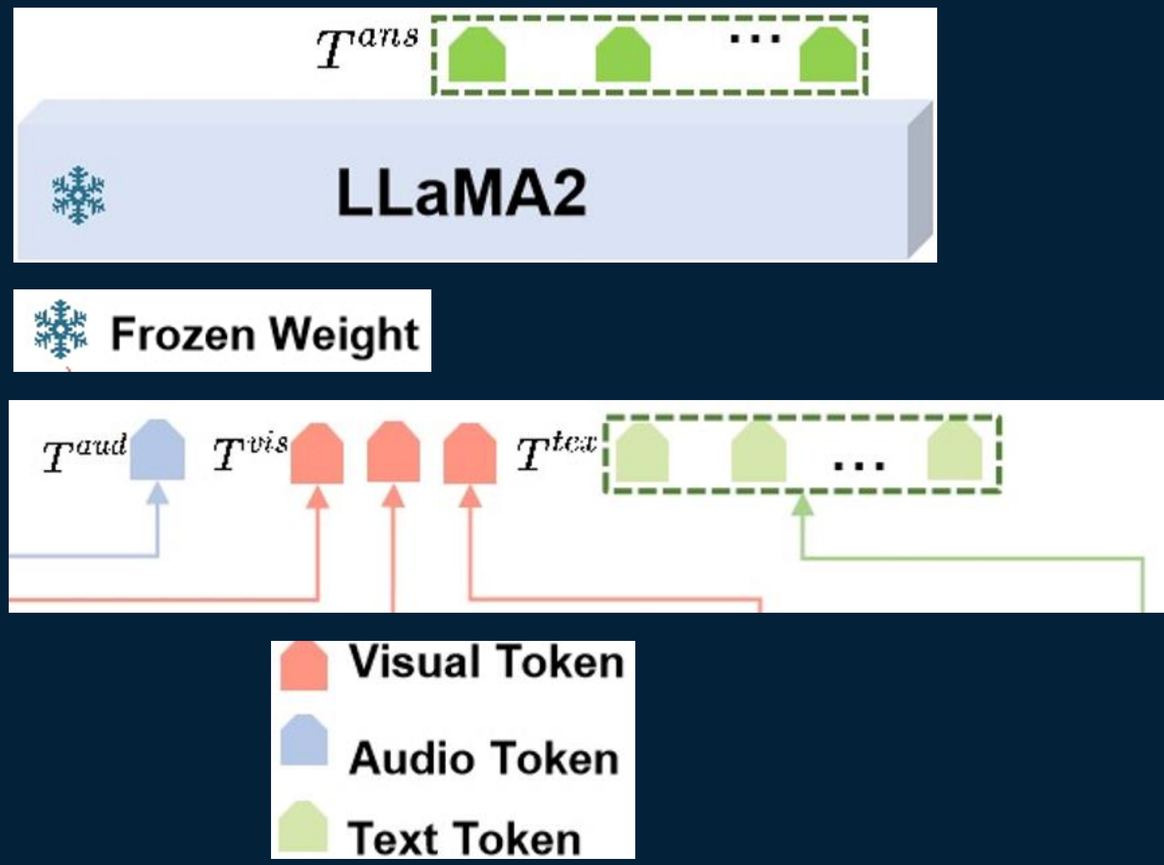


Our project MER 2025 -General overview
MER Challenge Based LLaMA2 Architecture



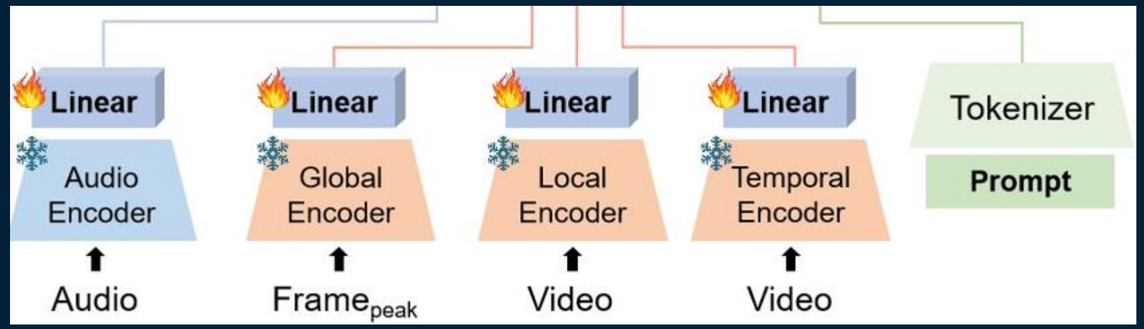
Our project MER 2025
LLaMA2: From Tokens to Answer

- LLaMA2 is:
- large language model
- developed by Meta
- designed to understand and generate human-like text.
- In our project, LLaMA2 acts as a reasoning engine



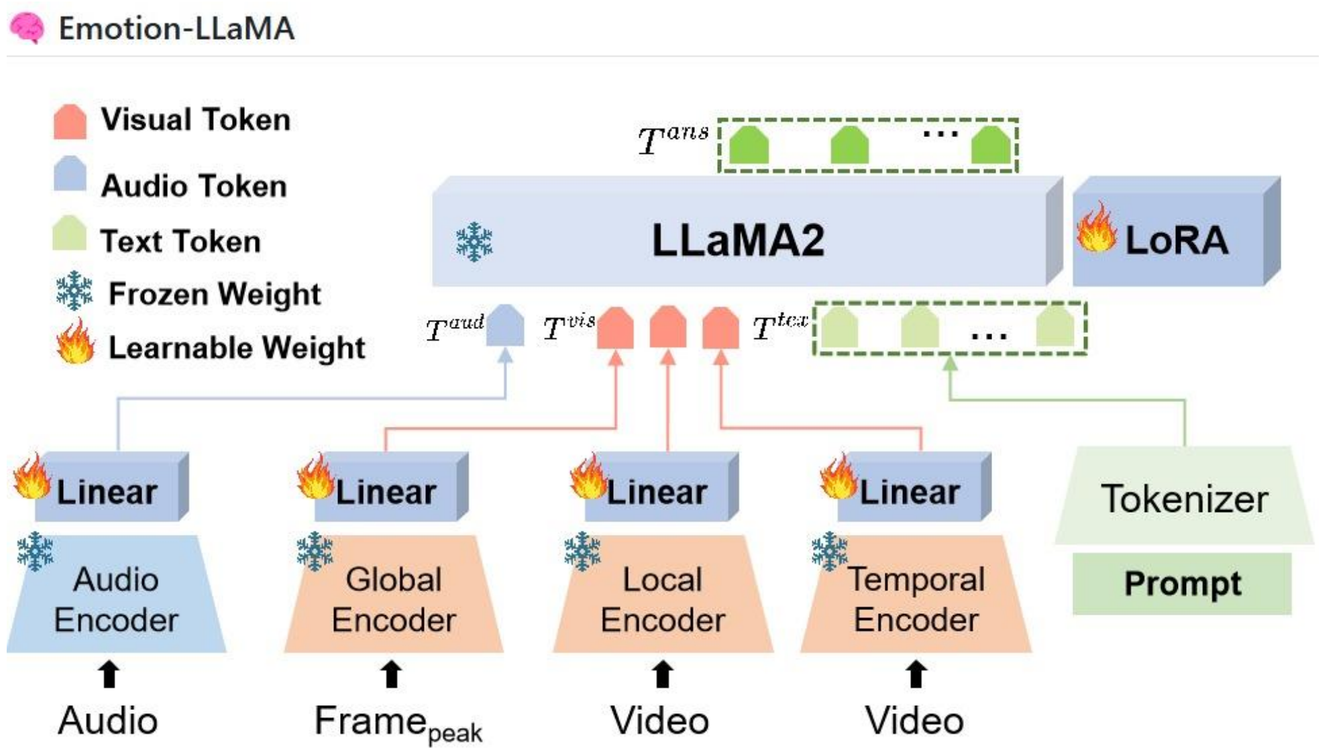
Our project MER 2025
MER Tools prepare Tokens for LLaMA2

- Reshape extracted features for compatibility with LLaMA2 token format

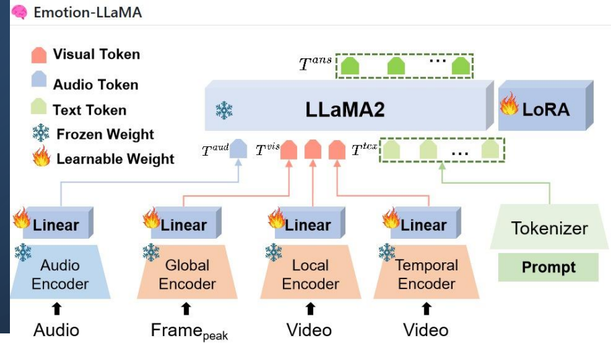


Structure of the Track 1 Processing Pipeline

- Reading Data
  - Read all relevant train features(7369 per feature)
  - Read all relevant test features(20,000 per feature)
  - Top 1 3 features
  - Top 2 6 features
- train&val 5 folder (CV1-CV5)
  - Per folder
    - Build the model
    - Prepare the CV train and eval
    - Train number of defined epochs
    - Save the results
    - Choose the Best epoch result
  - End of folder training
- Prediction and saving
  - Save cv in as NPZ file
  - Save test as NPZ file
- In the file name a lot of info including:
  - Model attention topn
  - Fusion (e.g. AVT Top2)
  - F1 and Acc score



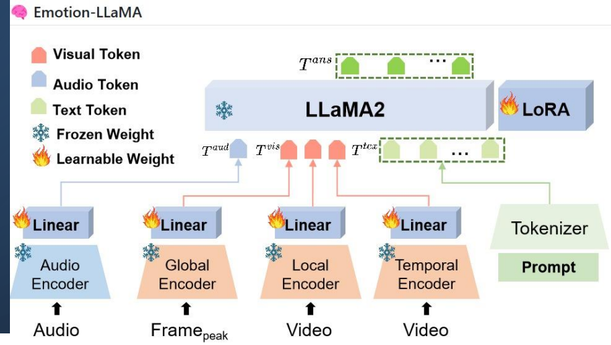
MER Our Pipeline : Cross validation Splits



- Train samples: 7369
  - CV Train 80%
  - CV Eval 20 %
- Test examples : 20,000
  - Calculate F1 and acc

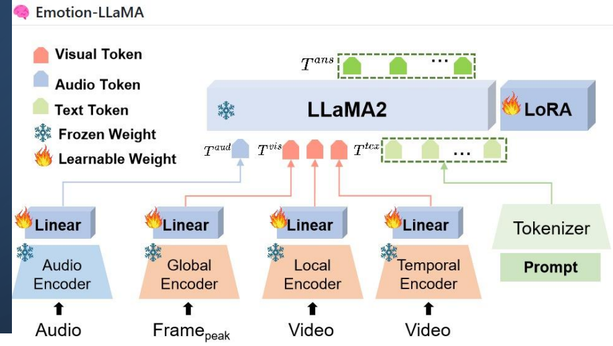
Folder	Train	Evaluate	Total
CV0	5896	1473	7369
CV1	5896	1473	7369
CV2	5896	1473	7369
CV3	5896	1473	7369
CV4	5892	1477	7369

MER Our Pipeline :
Models extracted Feature dimension



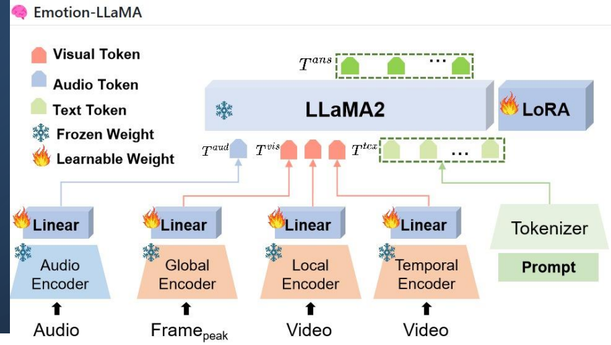
Feature extraction	Model	Dimensions
Audio 1	chinese-hubert-large	(1,1024)
Text 1	chinese-roberta-wwm-ext	(1,768)
Video 1	clip-vit-large-patch14	(1,768)
Audio 2	chinese-hubert-base	(1,768)
Text 2	chinese-roberta-wwm-ext-large	(1,1024)
Video 2	clip-vit-base-patch32	(1,512)

Results: Train Uni Modal
common hyper parameter



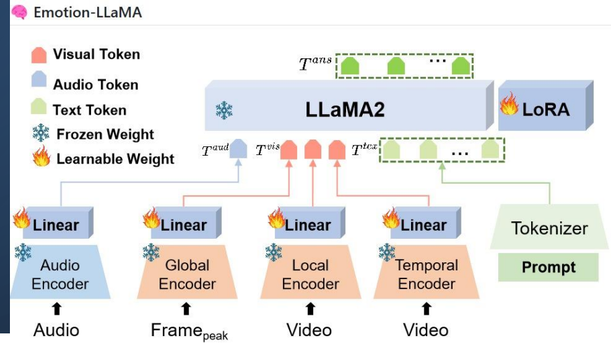
Hyper Parameter	Value	
topn	None	UNI No Topn
fusion modality	AVT	All same model
hidden_dim	64	Dimension of the hidden layers in the model. Determines the internal feature vector size.
feat	'utt'	Feature type – 'utt' stands for : **utterance-level features**, meaning **one** feature vector **per full** sentence or spoken segment
model	attention	Specifies the model type – likely a model with **attention mechanism** focusing on the **Top-N most relevant elements**.
lr	0.001	Controls how quickly the model updates during training. 0.001 is a common starting value.
batch size	128	Number of samples processed together in each training step. A larger batch size can speed up training (depending on available GPU memory).
grad clip	-1	Gradient clipping – -1.0 means **no clipping applied**. If set to a positive value, it limits the maximum gradient size to stabilize training.

Results :Train Uni Modal : Uni Models Train Scores



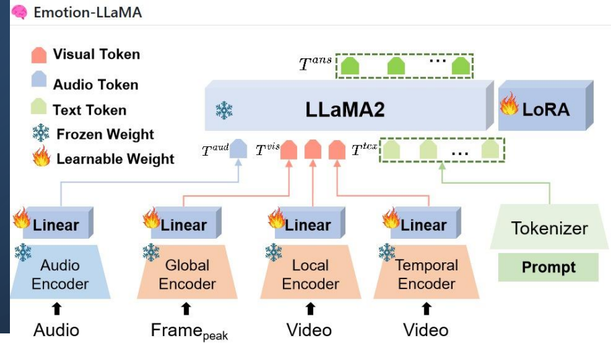
Feature extraction	Model	Duration	f1	acc
Audio 1	chinese-hubert-large	1:15:21	74.41%	59.25%
Text 1	chinese-roberta-wwm-ext	1:15:46	45.64%	59.25%
Video 1	clip-vit-large-patch14	1:14:52	58.83%	41.68%
Audio 2	chinese-hubert-base	1:16:12	74.11%	59.25%
Text 2	chinese-roberta-wwm-ext-large	1:18:19	53.64%	36.64%
Video 2	clip-vit-base-patch32	1:14:43	62.5%	45.46%

Train Multi Modal
common hyper parameter



Parameter	Value	Description
feat_type	'utt'	Feature type – 'utt' stands for **utterance-level features**, meaning one feature vector per full sentence or spoken segment.
model	'attention_topn'	Specifies the model type – likely a model with **attention mechanism** focusing on the **Top-N most relevant elements**.
batch_size	128	Number of samples processed together in each training step. A larger batch size can speed up training (depending on available GPU memory).
epochs	10	Number of complete passes over the training dataset. In this case, the model will train for 10 epochs.
lr (Learning Rate)	0.001	Controls how quickly the model updates during training. 0.001 is a common starting value.
hidden_dim	64	Dimension of the hidden layers in the model. Determines the internal feature vector size.
dropout	0.2	Dropout rate – 20% of neurons are randomly dropped during training to reduce overfitting.
grad_clip	-1.0	Gradient clipping – -1.0 means **no clipping applied**. If set to a positive value, it limits the maximum gradient size to stabilize training.

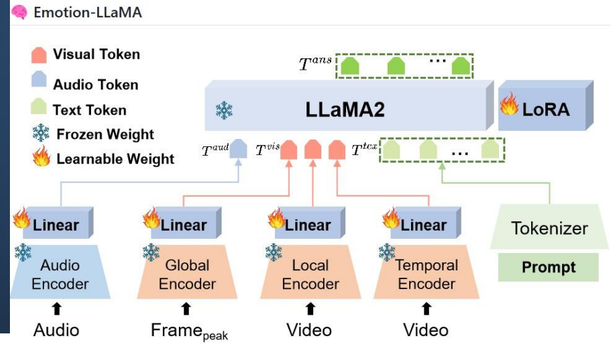
MER Our Pipeline : Key Findings from Experiment



Parameter	Fusion 1	Fusion 2	Fusion 3	Fusion 4	Fusion 5	Fusion 6	Fusion 7	Fusion 8
Fusion type	AVT	AVT	AV	AV	AT	AT	VT	VT
topn	1	2	1	2	1	2	1	2
epocs	10	10	10	10	10	10	10	10
run duration	1:19:36	1:31:49	1:16:44	1:28:23	1:17:09	1:33:56	1:18:19	1:29:23
f1	79.27%	79.96%	73.39%	74.71%	79.25%	77.77%	69.78%	70.88%
acc	65.66%	66.61%	57.97%	59.62%	65.63%	63.63%	53.59%	54.89%
Running hours	1:19:36	1:31:49	1:16:44	1:28:23	1:17:09	1:33:56	1:18:19	1:29:23

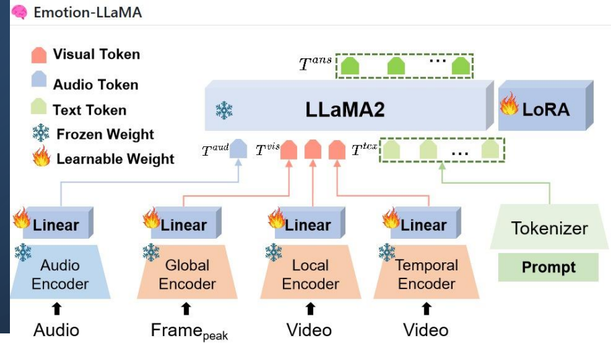
- **Best overall performance** achieved with **AVT 2**, reaching an **F1-score of 79.96**.
- **Audio** was identified as the **most effective modality**.
- Including audio features consistently **boosted the F1-score above 73%**.
- In contrast, **models without audio input produced the lowest F1-scores**, highlighting its critical role in emotion recognition

MER Our Pipeline :
Fusion AVT 2 -Hyper parameters affect



hyper	Base Value	Run Value	F1 score	Acc score	Running time	remarks
epochs	10	20	76.92%	62.50%	2:49:35	Time cost
Learning rate	0.001	0.0001	78.04%	63.99%	1:31:00	improved F1 and ACC
Drop out	0.2	0.5	77.79%	63.65%	1:28:04	improved F1 and ACC
Hidden dimension	64	128	78.46%	78.97%	01:35:15	improved F1 and ACC
Hidden dimension	64	256	64.56%	65.25%	01:29:18	Just Run time affect

MER Our Pipeline : Fusion Vs majority

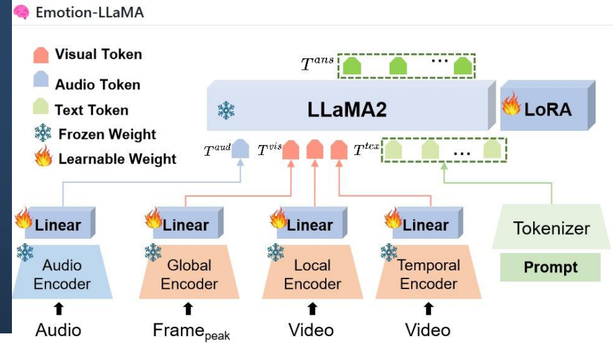


8 Fusion Combinations
A total of 20,000 samples were evaluated.
Predictions are determined using a **majority voting** strategy:

% majority models	Count of samples	% samples
2/8	22	0%
3/8	394	1%
4/8	1481	4%
5/8	1983	7%
6/8	3932	17%
7/8	2251	12%
8/8	9937	58%
Grand Total	**20000**	**100%**

MER Our Pipeline :

Uni Vs Multi



Training model	Model/features	Duration	f1	acc
Uni Audio 1	Most effective Uni model Audio chinese-hubert-large	1:15:21	74.41%	59.25%
Uni Text 1	Least effective Uni modelt Text chinese-roberta-wwm-ext	1:15:46	45.64%	59.25%

Training model	Model/Fusions	Duration	f1	acc
Multi Fusion 1	Most effective Fusion model AVT 2 which utilizes All 6 features	1:31:49	79.96%	66.61%
Multi Fusion 7	Least effective Fusion model is AV 1, which utilizes four features but excludes audio information.	1:18:19	69.78%	53.59%

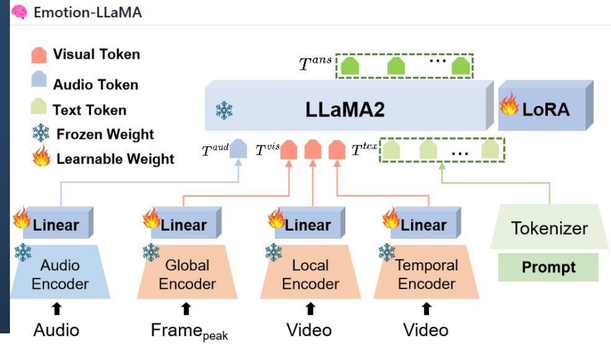
Impact of Audio Modality on Emotion Recognition

Audio modality demonstrates strong predictive power in emotion recognition.

Unimodal audio model outperforms multimodal (Audio + Visual) configuration.

Audio-only model achieves higher accuracy, indicating its dominant role.

MER Our Pipeline :
8 Fusion types varios Results



- When the fusion method is AVT:
  - The extracted features are aligned and compatible across all modalities.
- If only two modalities are available
  - we still provide three inputs to LLaMA by replicating one of the existing modality features.

Fusion	Top	Audio	Text	Video	10Epocs hours Running time	F1	Acc	2Epocs hours Running time	F1	Acc	100Epocs hours Running time	F1	Acc
AVT	1	A1	T1	V1	01:20	74.73%	59.66%				12:20	78.68%	64.86%
AVT	2	A1,A2	T1,T2	V1,V2	01:28	79.67%	66.22%	00:30	78.66%	64.33%	13:04	79.30%	65.70%
AV	1	A1	V1	V1	01:13	76.52%	61.97%				12:05	77.27%	62.96%
AV	2	A1,A2	V1,V2	V1,V2	01:21	73.72%	58.38%				12:55	76.33%	61.72%
AT	1	A1	T1	T1	01:14	73.92%	58.64%				11:55	74.25%	59.05%
AT	2	A1,A2	T1,T2	T1,T2	01:26	75.49%	60.63%				15:00	78.68%	64.86%
VT	1	T1	T1	V1	01:15	65.35%	48.53%						
VT	2	T1,T2	T1,T2	V1,V2	01:23	65.81%	49.04%	00:47	49.91%	33.25%			

Project Challenges

- **Project Environment and Dataset Challenges**
- The full project dataset is approximately **2 TB** in size.
- The code expects a **specific directory structure** on disk for correct execution.
- **Afeka's storage limitation**: only **500 GB** available — workaround uses only **Track 1 raw data** (~27,639 samples / instead of full data set with more than 132K samples).
- **Environment Adaptation – Linux to Windows**
- **MER Tools** were originally built for **Linux**.
  Porting to Windows required:
  - Updating **file save paths** to match Windows conventions.
  - Replacing Linux-only **Python packages** with Windows-compatible versions *(involved many hours of searching and installing)*.
  - Rewriting **.sh scripts** into **Python equivalents** for single and multi-run execution.
- **System Complexity & Support**
- The MER project includes many scripts, tools, and dependencies — designed with **flexibility**, but also **steep learning curve**.
- While **MER Tools** are part of the **Emotion LLaMA ecosystem**, the two are **separate systems**.
- **Project support is strong**:
  Initially, we contacted **customer support via email**.
  Now we have **direct email communication with Dr. Zheng Lian**, the project lead.

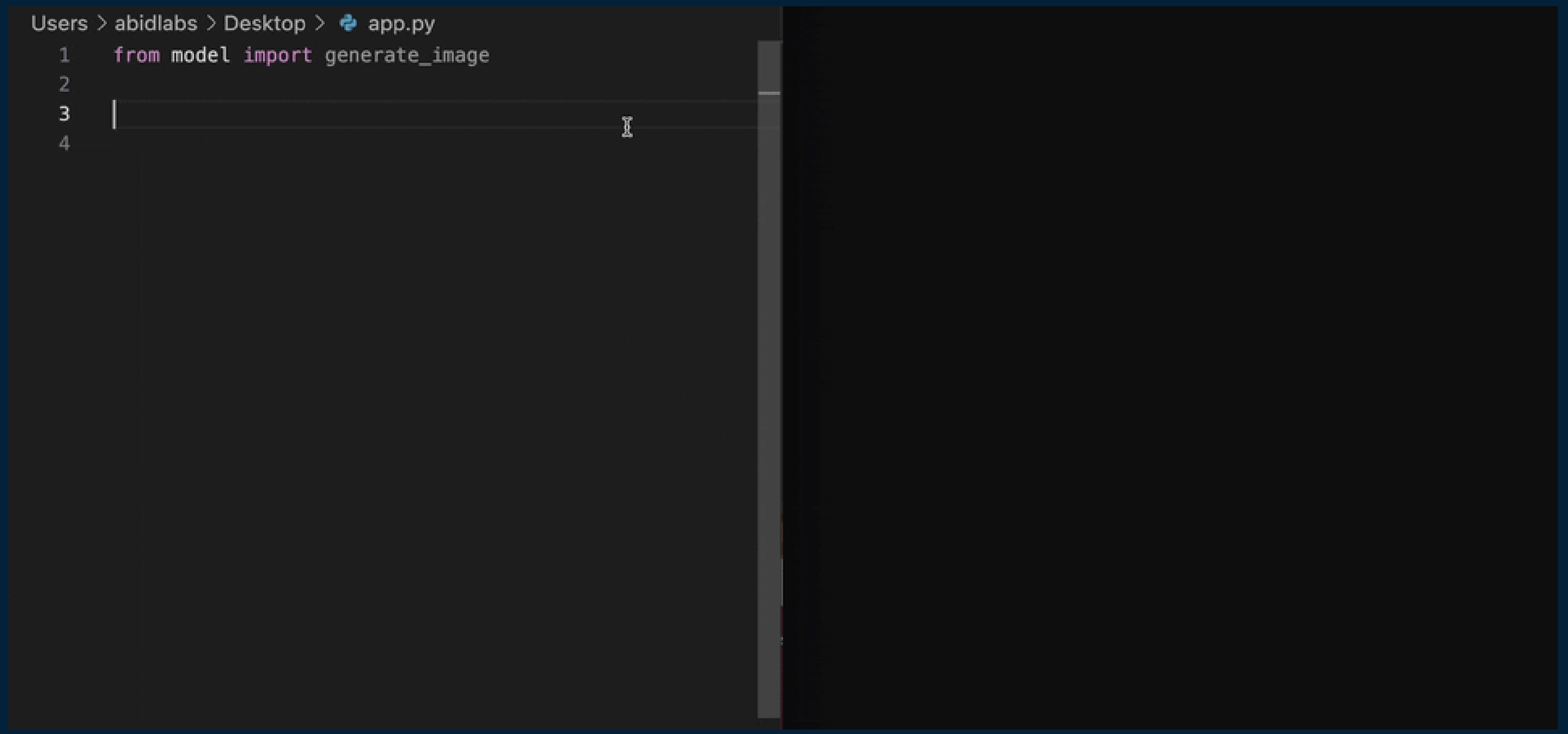
Project Challenges

- **Afeka VM – Technical & Collaboration Challenges**
- The virtual machine (VM) forces use of a **predefined student username**, making **user-specific installations** the default and **project sharing difficult**.
- The VM environment is essentially **bare/minimal ("naked")**, requiring students to invest significant time in setting up packages and tools.
- **Collaboration limitations**:
  - No shared user support — our workaround was to **share a single user account**, including **sharing credentials** within the team.
  - This approach allowed smoother teamwork but raises concerns about security and scalability.
- **Resource allocation**:
  - Initially limited to **12 GB GPU memory**.
    After request, it was increased to **24 GB**, which was sufficient for our needs.
- **Lack of local access tools**:
  - No **Microsoft Office** or document tools installed.
  - No ability to **copy-paste** between the VM and personal PCs —
    we used **WhatsApp as a makeshift clipboard** for transferring code and results.
- The current setup **assumes prior IT and DevOps skills**, which may not match all students' backgrounds.

MULTI MODAL : EZRA

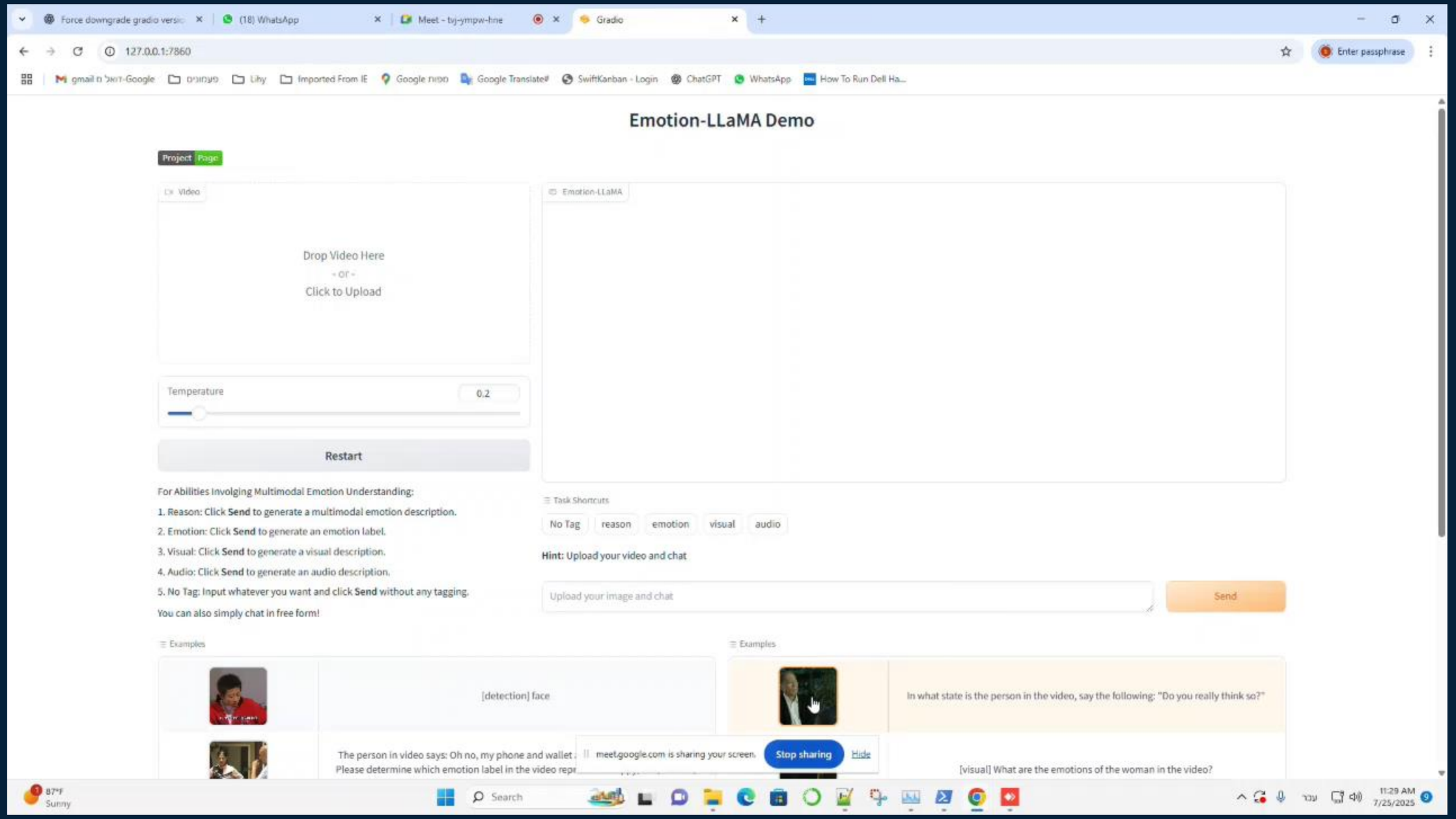
Gradio Interface – an interactive tool for ai developers

An open-source Python package that allows to build a demo or web application for machine learning model, API, or any arbitrary Python function.
Users can share a link to demo or web application in just a few seconds.
No JavaScript, CSS, or web hosting experience needed.



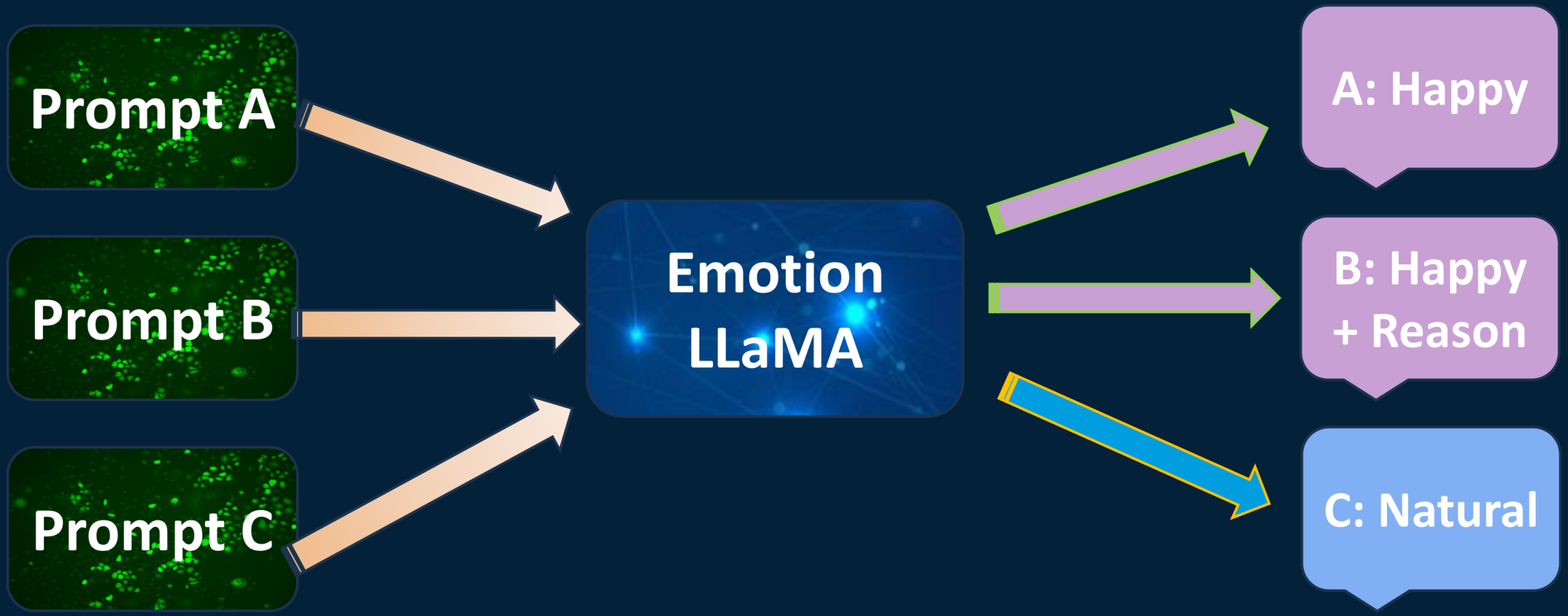
https://www.gradio.app/

Gradio inference – An interactive emotional LLaMA



Blow a kiss / fire a gun – what modality we should lean on ?

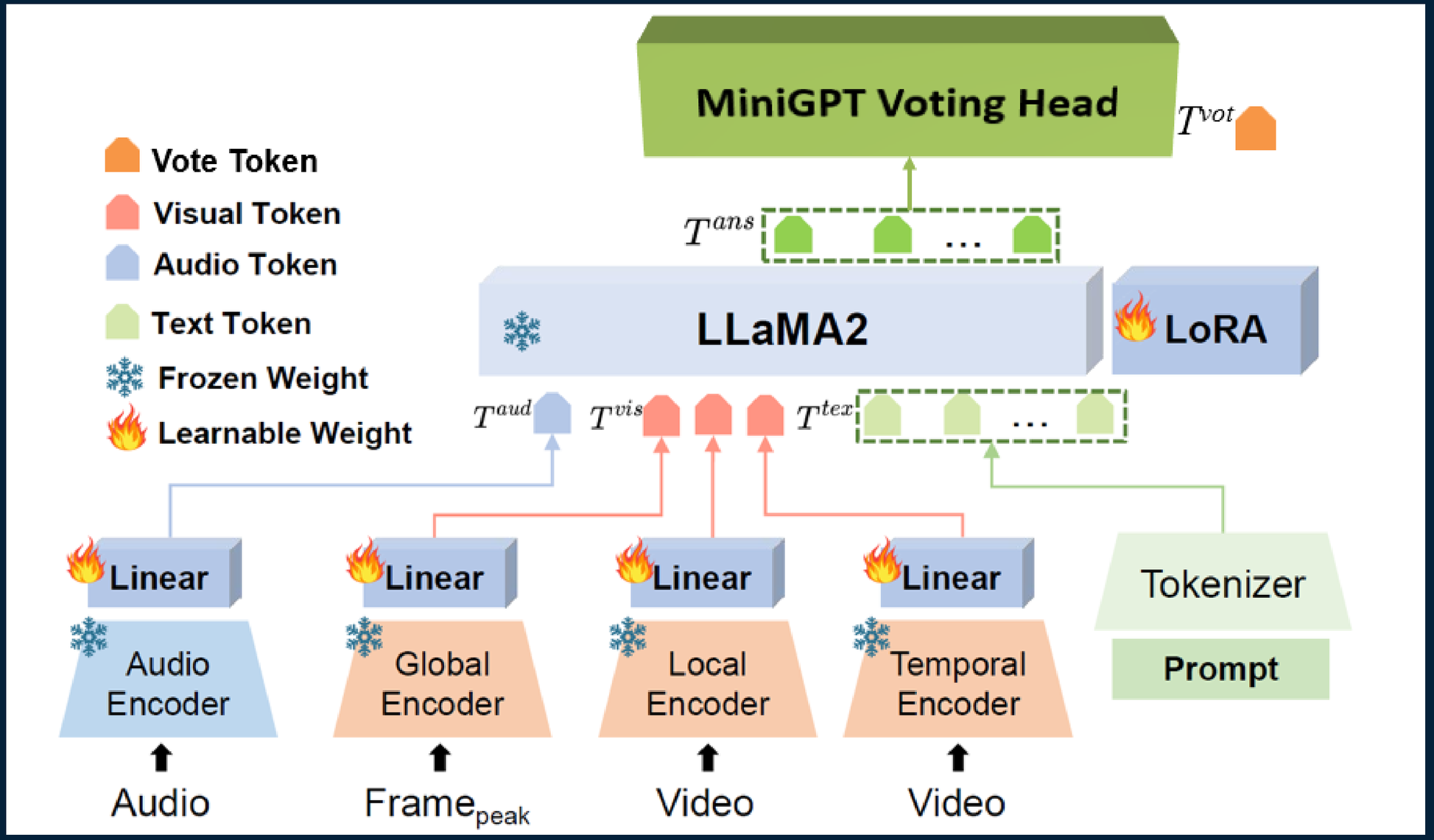
- Model’s output is highly dependent on prompt’s focus
- For better reasoning, the model should know where to focus



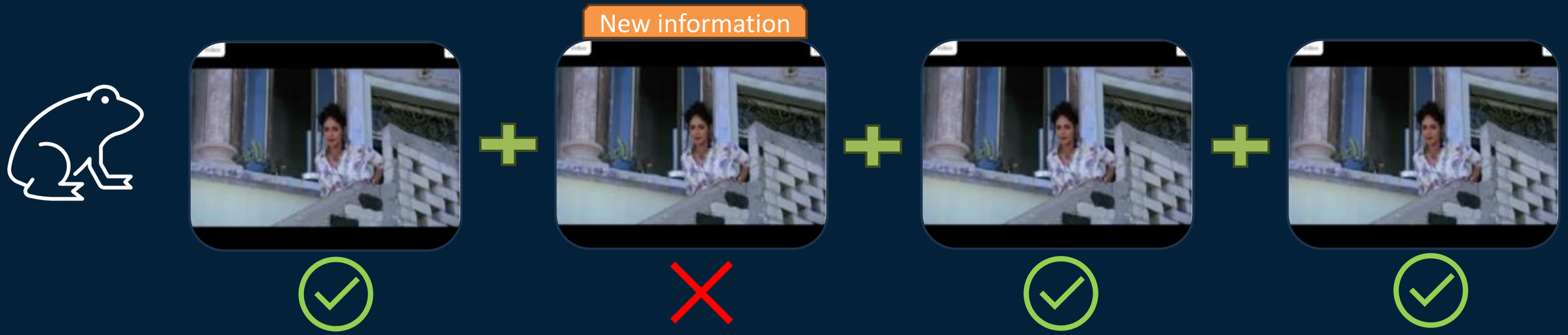
Gradio inference – An interactive emotional LLaMA

LINK

Gradio inference – An interactive emotional LLaMA



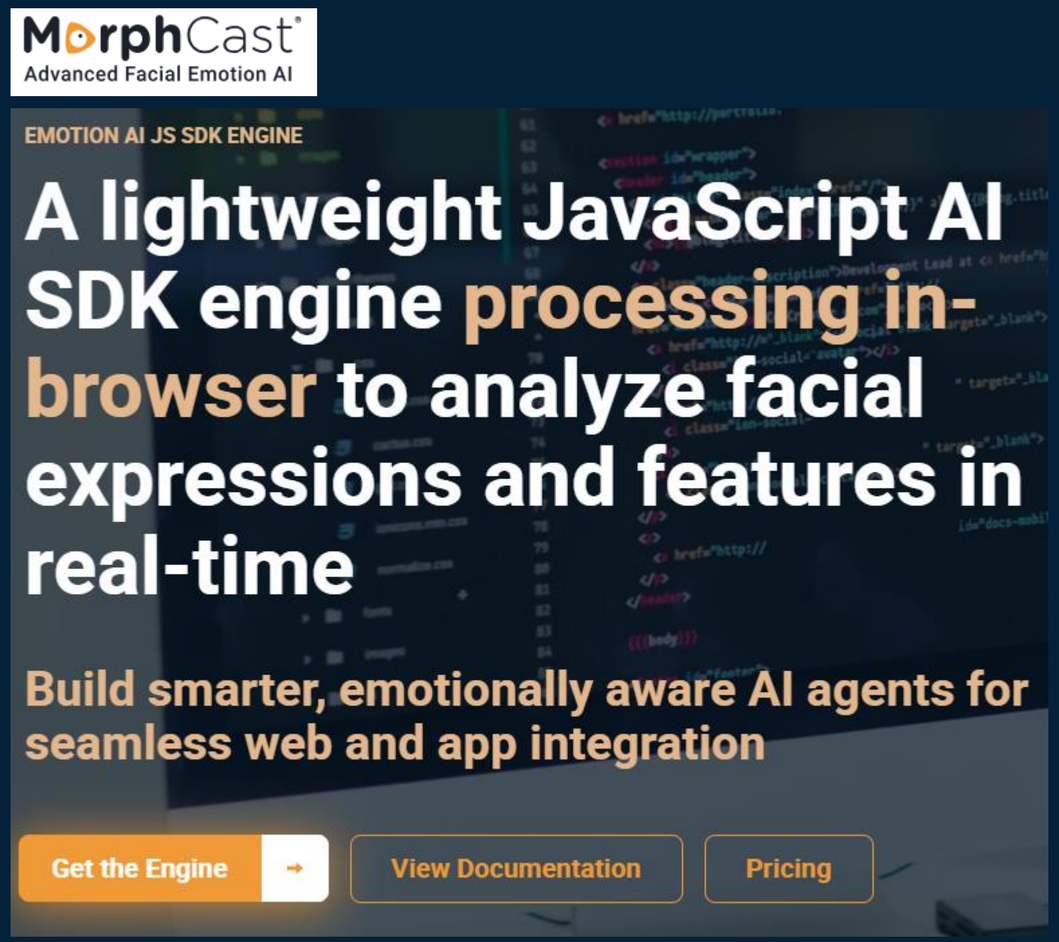
Gradio inference – An interactive emotional LLaMA



Tvot (Token Vote) Output:

- **Emotion**: Angry
- **Situation**: The woman appears to be addressing a stressful situation involving children, potentially in a caregiving or educational context. Her stern facial expression, sharp tone, and abrupt questioning suggest she is experiencing frustration or irritation—possibly due to challenging behavior or lack of support. The setting (balcony/staircase) implies she might be confronting someone from a physical or emotional distance, adding to the tension.
- **Confidence**: 9/10

The future is here: available emotion products

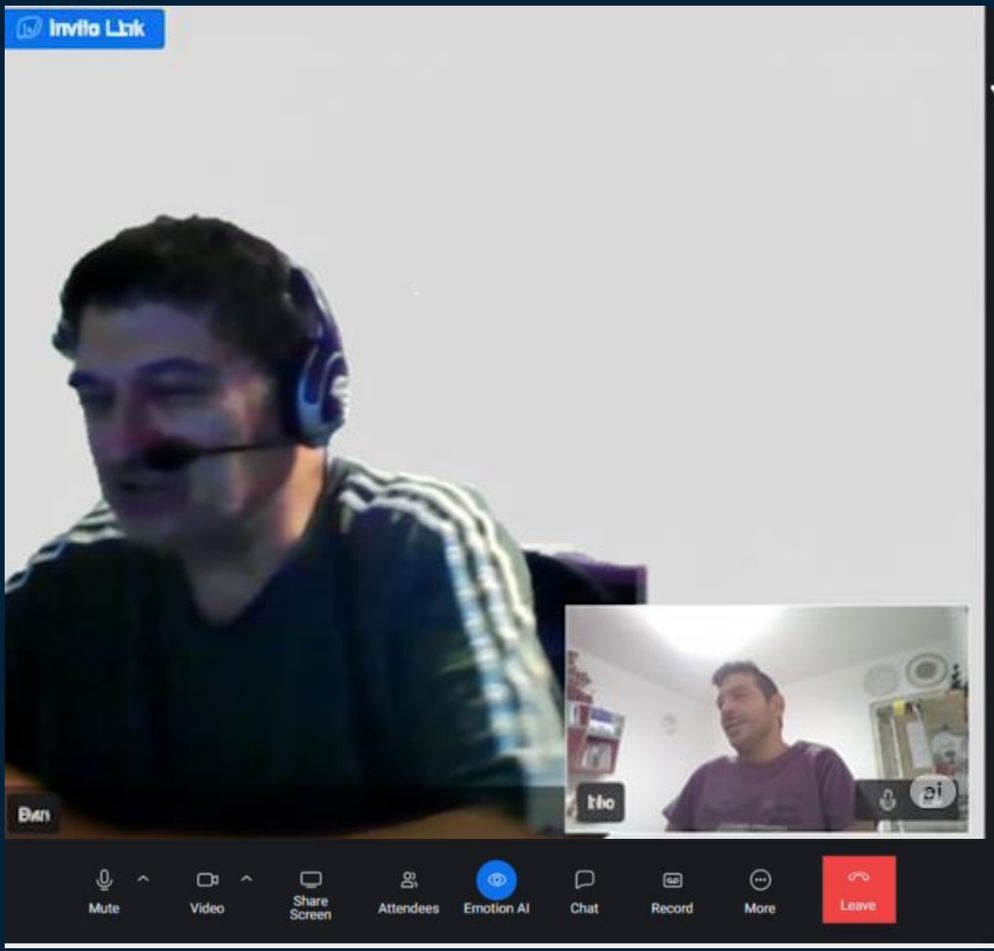


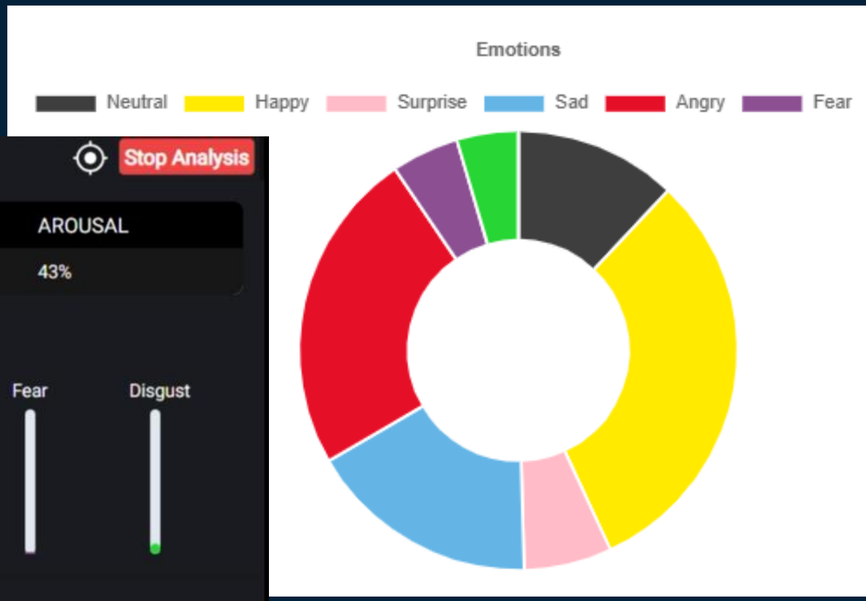
Crafted a Frugal Emotion AI solutions for precise facial emotion analysis—directly in the browser, with zero server processing. Ensuring privacy, sustainability, and smooth integration into empathetic, data-driven digital experiences.

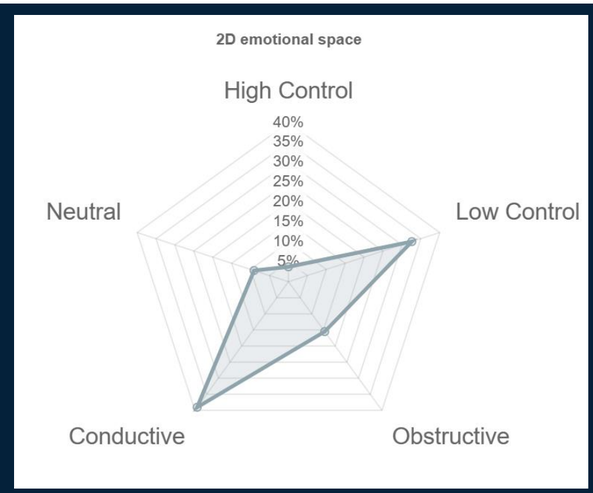
https://www.morphcast.com/

The future is here: online meetings

MorphCast®
Advanced Facial Emotion AI







NEXT STEPS

- Integrating a voting + reasoning head to improve emotion LLaMA2
- Adding an encoder to identify cultural context (race, time geography)
- Improving model by multiple entries, not ignoring minority report
- Expending model to further features from MERR2025:
  - **Track1**: MER-SEMI: Predict one label from six categories: worried, happy, neutral, angry, surprise, sad.
  - **Track2**: MER-FG: predict any emotion labels, capturing more nuanced emotional expressions
  - **Track3**: MER-DES: submit both multimodal evidence and emotion labels, enhancing interpretability
  - **Track4**: determine whether better emotion recognition improves personality prediction

VISION

"There are two pivotal inflection points in the human soul: the quiet ascent of growth through education, and the fragile instant just before we break."

E.E.2025

Leverage emotion recognition mainly to improve:

- Education, character-building, discipline, and moral guidance.
- Improved mental health care and the rise of new profession of specialists for detection and treatment of persons in stress.

Q&A

- Open floor for feedback