HEAR ME OUT (& THINK): MAESTRO, A MULTIMODAL AGENTIC MODEL WITH EFFICIENT, SYNERGISTIC TEXT-REASONING OPTIMISATION FRAMEWORK

Members:
Felicia Tan Ee Shan, Low Li Ying Amy
(Raffles Institution)

Mentor:
Kuek Yong Jie Adriel
(DSO National Laboratories)

INTRODUCTION

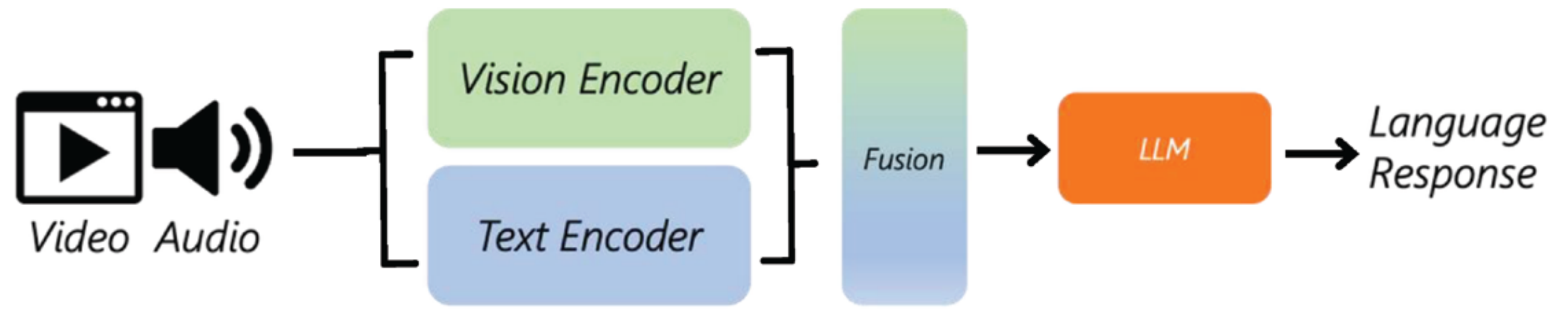
Specific Use-Case: Hateful Video Classification

***500 hours** of videos/min*

***1 Billion** users*

1. *Demands Inference Reasoning*
2. *Multimodal Reasoning*
3. *Temporal Understanding*
4. *Adaptability*

CURRENT VLMs



High Computational Overhead

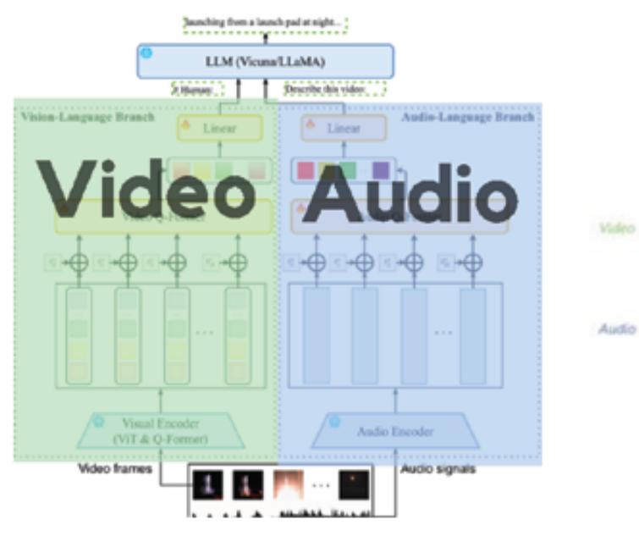
Exhaustive frame-by-frame analysis

Use of Q-formers / transformers

Achieving fine-grained understanding in VLMs often necessitates processing a vast number of video frames, leading to high computational costs that scale significantly with input length

Missing integration of audio modality

Separate Streams



Neglects Time Alignment

Video
Audio
"Jump in! Jump in!"
Nuclear bomb explosion Sound Effect

Weak/Static Reasoning

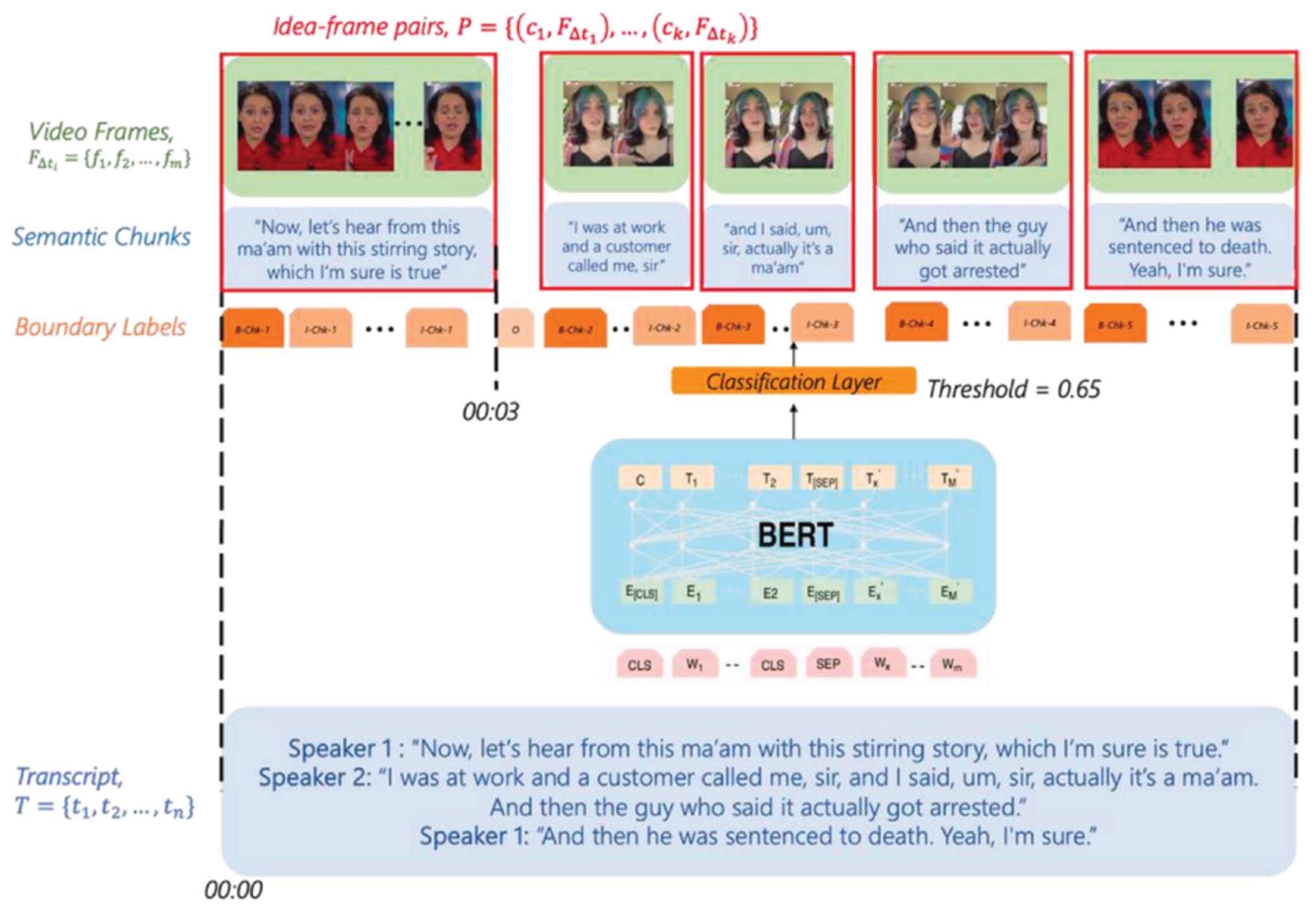
Static Reasoning Pipeline

USER: Is this video hateful? Explain why or why not

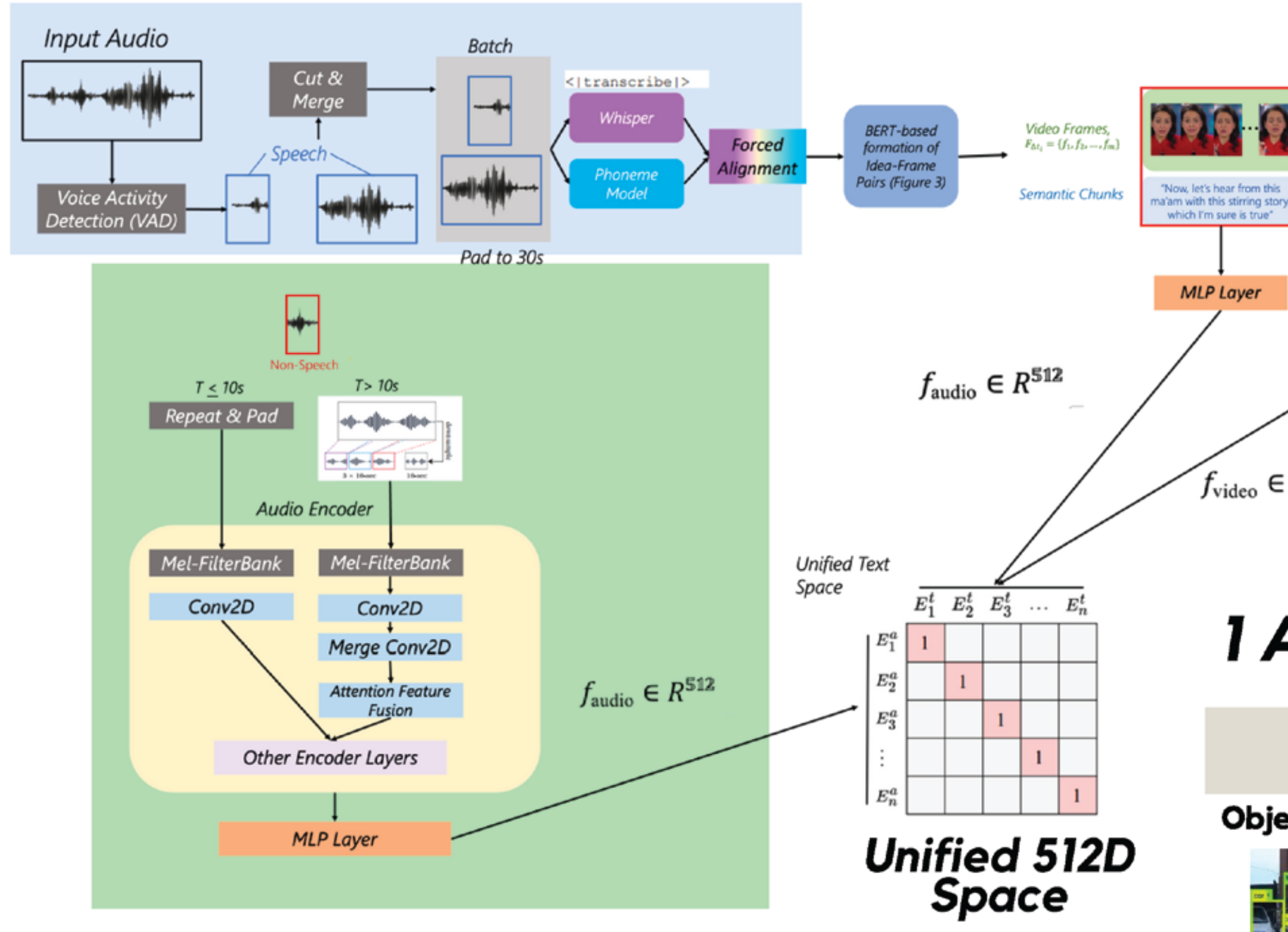
Video-LLaMA 🦙 : Based on the video descriptions, it is not clear whether the video is hateful or not. However, the video shows a large crowd of people at an airport, and some of them appear to be excited or happy. ***The video also shows a person with a phone, which could be a sign of annoyance or frustration***

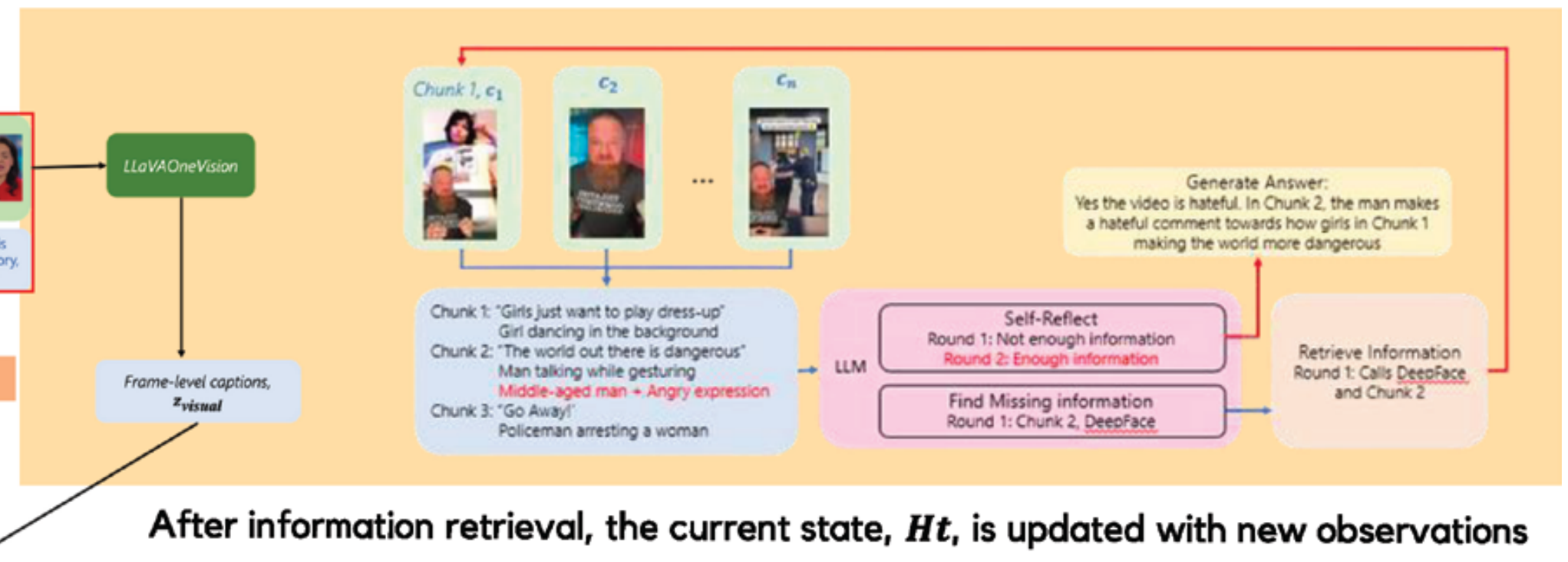
METHODOLOGY

Transcript Chunking → *Unified Modality Alignment* → *Global-Local Reasoning Loop*


Idea-frame pairs, $P = \{(c_1, F_{\Delta t_1}), \dots, (c_k, F_{\Delta t_k})\}$
Video Frames, $F_{\Delta t_i} = \{f_1, f_2, \dots, f_m\}$
Semantic Chunks
"Now, let's hear from this ma'am with this stirring story, which I'm sure is true"
"I was at work and a customer called me, sir"
"and I said, um, sir, actually it's a ma'am"
"And then the guy who said it actually got arrested"
"And then he was sentenced to death. Yeah, I'm sure."
Boundary Labels
B-Chk-1
I-Chk-1
O
B-Chk-2
I-Chk-2
B-Chk-3
I-Chk-3
B-Chk-4
I-Chk-4
B-Chk-5
I-Chk-5
00:03
Classification Layer
Threshold = 0.65
BERT
CLS
SEP
Transcript, $T = \{t_1, t_2, \dots, t_n\}$
00:00

Speaker 1 : "Now, let's hear from this ma'am with this stirring story, which I'm sure is true."
Speaker 2: "I was at work and a customer called me, sir, and I said, um, sir, actually it's a ma'am. And then the guy who said it actually got arrested."
Speaker 1: "And then he was sentenced to death. Yeah, I'm sure."




LLaVAOneVision
Frame-level captions, z_{visual}
Chunk 1, c_1
c_2
c_n
Chunk 1: "Girls just want to play dress-up" Girl dancing in the background
Chunk 2: "The world out there is dangerous" Man talking while gesturing Middle-aged man + Angry expression
Chunk 3: "Go Away!" Policeman arresting a woman
LLM
Self-Reflect
Round 1: Not enough information
Round 2: Enough information
Find Missing information
Round 1: Chunk 2, DeepFace
Retrieve Information
Round 1: Calls DeepFace and Chunk 2
Generate Answer:
Yes the video is hateful. In Chunk 2, the man makes a hateful comment towards how girls in Chunk 1 making the world more dangerous

After information retrieval, the current state, Ht, is updated with new observations and iteratively refined, ensuring the model adapts its reasoning to the specific task. After each update, the system re-evaluates the confidence score (ct) to determine whether the model is ready to classify the video (Action 1) or if further context is needed, thus continuing the cycle of state updates (Action 2)

1 Agent, 3 Tools

ChatGPT-3.5

YOLO

Object & Symbol Detection

LLaVAOneVision

Action Recognition

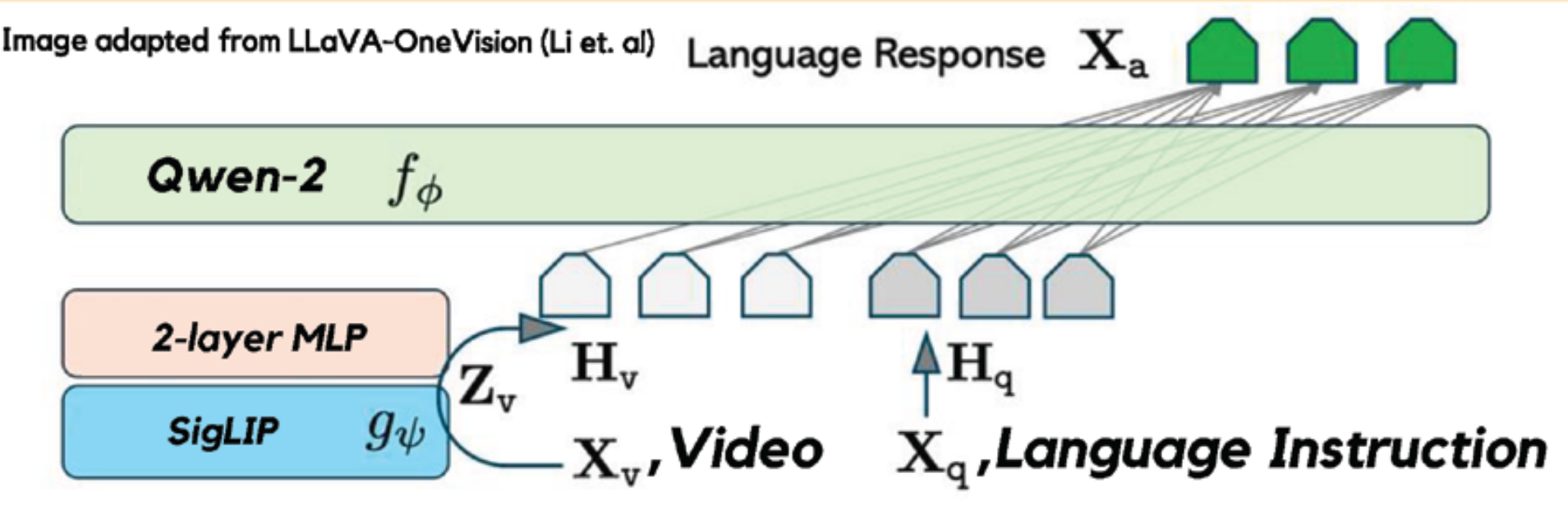
- Accurate Descriptions
- Weak Reasoning

videoDeepFace

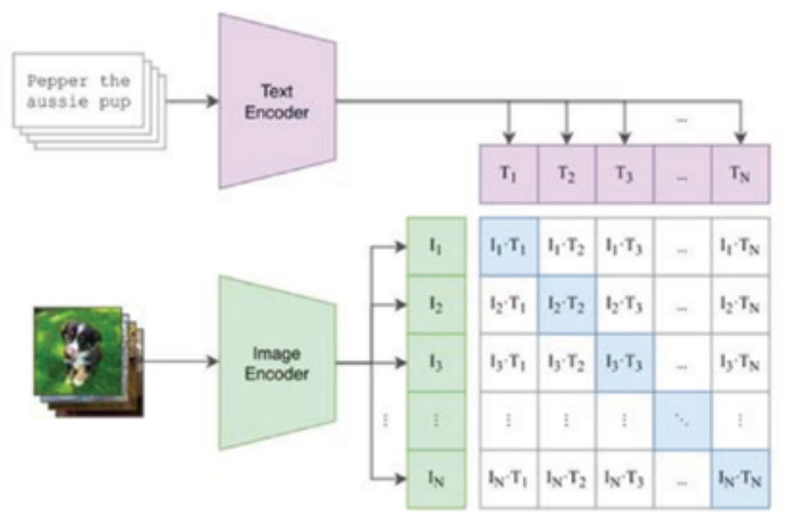
Race, Age, Gender, Emotion

Video

Image adapted from LLaVA-OneVision (Li et. al)



SigLIP Vision Encoder



Qwen-2

offers various model size and exhibits strong language capabilities to date among publicly available checkpoints

Speech Audio

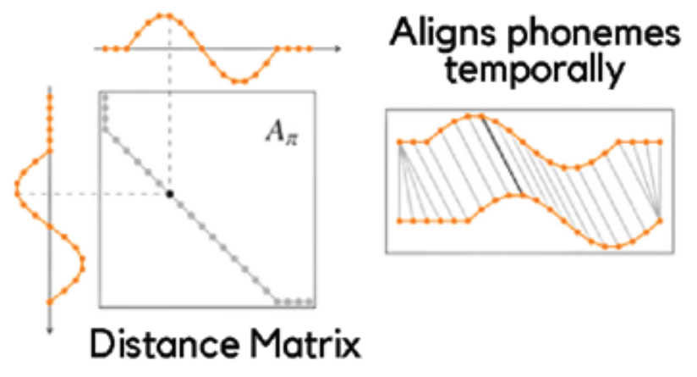
Voice Activity Detection (VAD)

Acoustic Features — Binary Labels

$$f_\theta : A = \{a_1, \dots, a_T\} \rightarrow y = \{y_1, \dots, y_T\}$$

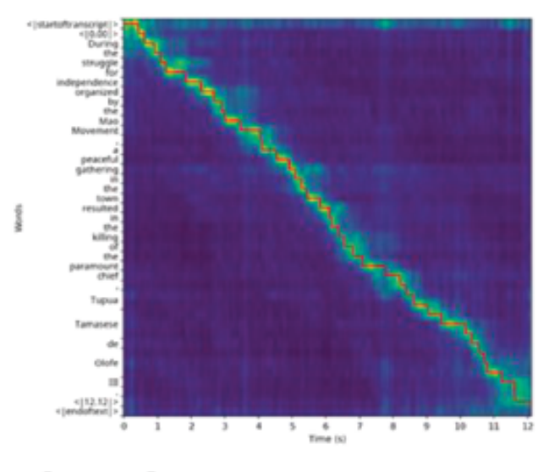
$$y_t \in \{0, \underline{1}\}$$ Speech present

Dynamic Time Warping



Word boundaries derived from t_{start} and t_{end}

Forced Phoneme Alignment



$$L_i \in R^{|c_{T_i}| \times T}$$

Min-Cut Strategy

Overly long segments

Split at the point of **minimum voice activation score**

Neighbour Merging

Overly short segments

Below $\tau = |A_{\text{train}}|$

Maintains Temporal Consistency

MAESTRO-CLAP (Non-Speech Audio)

2 Audio Encoders

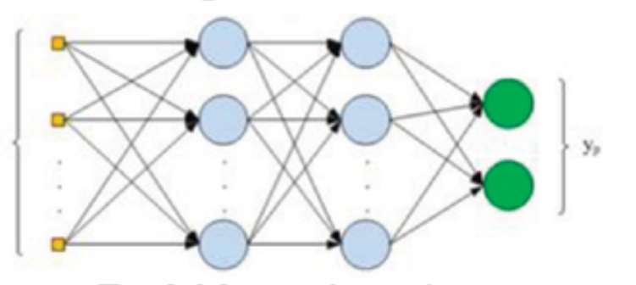
PANN
- CNN-based audio classification model
- 7 downsampling and 7 upsampling blocks
- Penultimate layer : $z_{\text{PANN}} \in R^{2048}$

HTSAT
- SOTA Transformer model featuring 4 groups of Swin Transformer blocks,
- Output: $z_{\text{HTSAT}} \in R^{768}$

Contrastive Loss

$$-\frac{1}{N}\sum_{i=1}^{N} log \frac{exp(f_{\text{audio}}^i \cdot f_{\text{text}}^i / \tau)}{\sum_{j=1}^{N} exp(f_{\text{audio}}^i \cdot f_{\text{text}}^i / \tau)}$$

2 layer MLP



ReLU activation

$$f_{\text{non-speech}} = \text{MLP}_{\text{non-speech}}(z_{\text{non-speech}}), \quad f_{\text{non-speech}} \in R^{512}$$

Projected into unified space

Output, $f_{\text{output}} \in R^{512}$

RESULTS

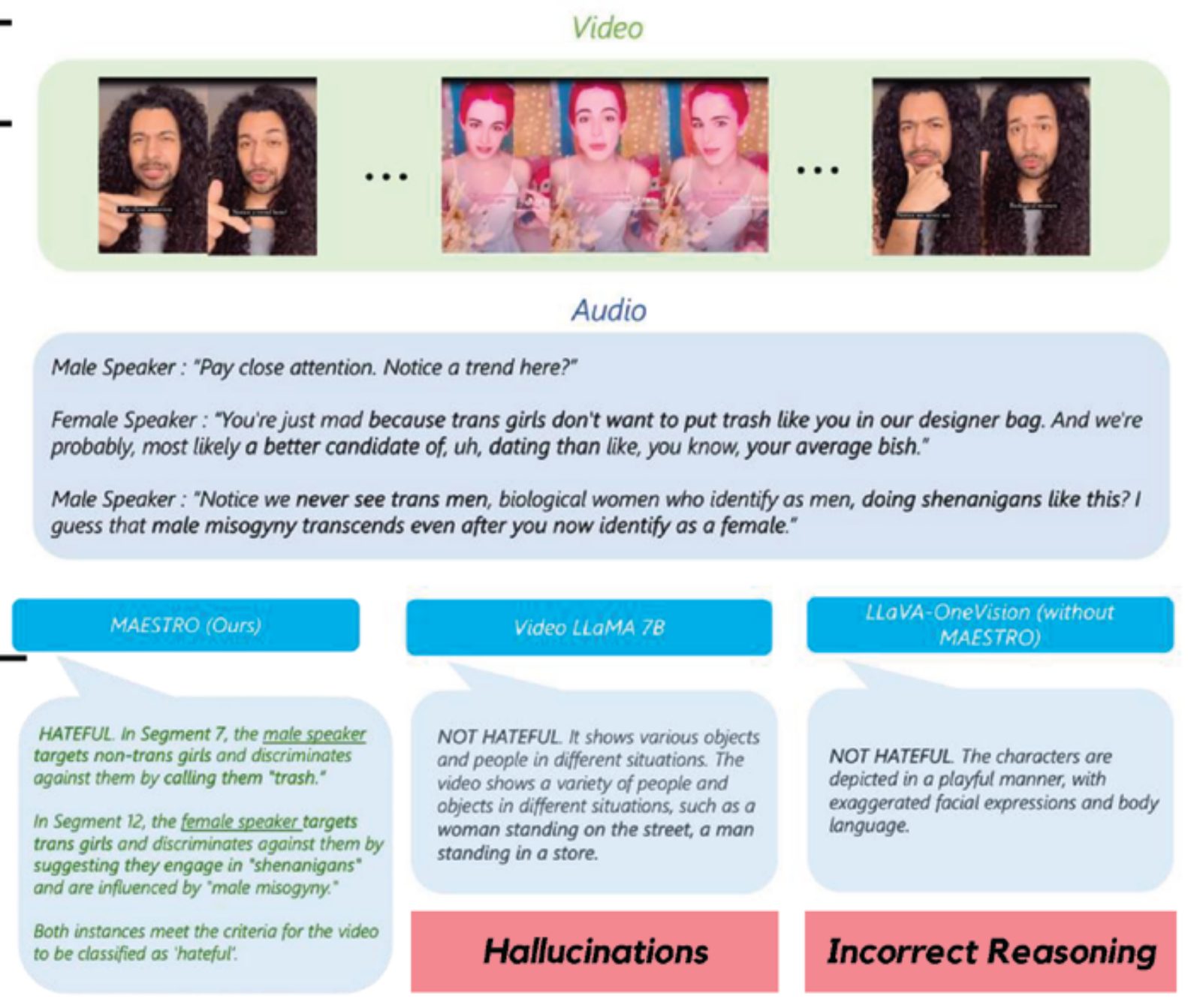
MultiHateClip (Hateful Videos)

Model	Acc	Binary M-F1	F1(O)	R(O)	P(O)
mBERT	0.57	0.57	0.52	0.42	0.68
GPT-4	<u>0.81</u>	<u>0.79</u>	<u>0.73</u>	0.69	0.78
Qwen	0.72	0.71	0.65	0.57	0.75
MFCC	0.54	0.50	0.36	0.33	0.40
Wav2Vec	0.53	0.48	0.64	0.50	<u>0.90</u>
ViViT	0.73	0.73	0.68	0.57	0.86
Vit	0.63	0.58	0.44	0.46	0.45
VLM	0.70	0.64	0.48	0.59	0.41
GPT-4V	<u>0.81</u>	<u>0.79</u>	<u>0.73</u>	<u>0.72</u>	0.73
Qwen-VL	0.62	0.61	0.56	0.46	0.72
T1 ⊙ A1 ⊙ V1	0.75	0.74	0.67	0.61	0.77
MAESTRO (Ours)	**0.96**	**0.95**	**0.93**	**0.87**	**1.0**

Achieves **SOTA** by ≥ 15%

Correct Reasoning — Localisation Ability

Accurate Description — Correct Label



Male Speaker : "Pay close attention. Notice a trend here?"

Female Speaker : "You're just mad because trans girls don't want to put trash like you in our designer bag. And we're probably, most likely a better candidate of, uh, dating than like, you know, your average bish."

Male Speaker : "Notice we never see trans men, biological women who identify as men, doing shenanigans like this? I guess that male misogyny transcends even after you now identify as a female."

HATEFUL. In Segment 7, the male speaker targets non-trans girls and discriminates against them by calling them "trash."

In Segment 12, the female speaker targets trans girls and discriminates against them by suggesting they engage in "shenanigans" and are influenced by "male misogyny."

Both instances meet the criteria for the video to be classified as 'hateful'.

NOT HATEFUL. It shows various objects and people in different situations. The video shows a variety of people and objects in different situations, such as a woman standing on the street, a man standing in a store.

NOT HATEFUL. The characters are depicted in a playful manner, with exaggerated facial expressions and body language.

Industry Benchmarks (General VQA)

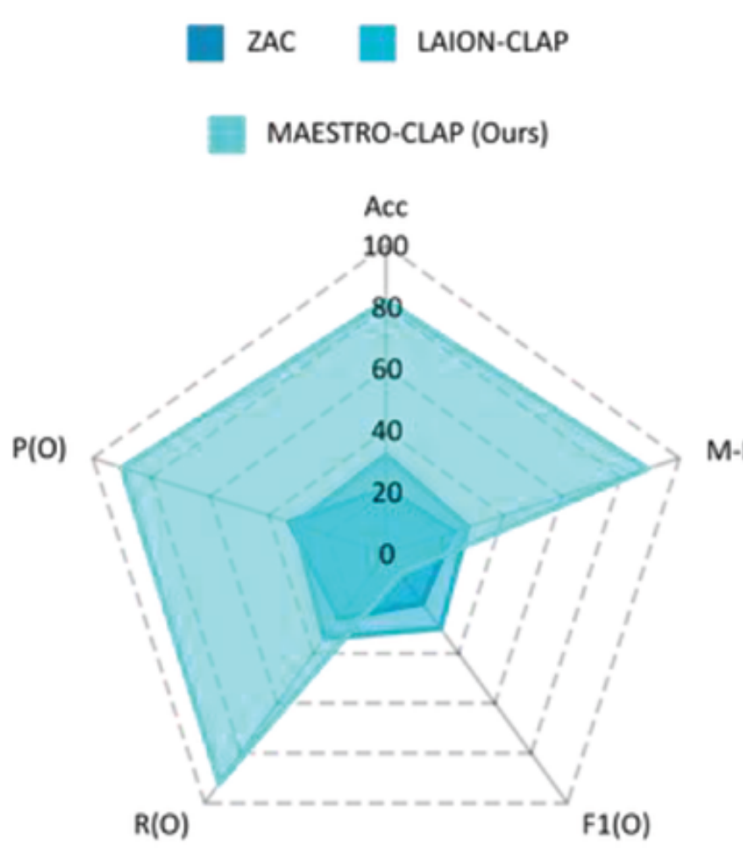
Model	Modality	MSRVTT-QA	MSVD-QA	ActivityNet-QA
QueST	V	34.6	34.6	-
ClipBERT	V	37.4	-	-
JustAsk	V	41.5	46.3	38.9
GIT	V	42.7	55.1	-
MERLOT	V	43.1	-	41.4
Singularity	V	43.5	-	43.1
Clover	V	43.9	51.9	-
VideoChat	V	45.0	56.3	26.5
Video-ChatGPT	V	49.3	64.9	35.2
VALOR	V,A	46.7	56.4	44.8
FrozenBiLM	V	47.0	54.8	43.2
Valley	V	45.7	<u>65.4</u>	42.9
Video-LLaMA	V,A	29.6	51.6	12.4
PandaGPT	V,A	25.5	42.1	14.5
LLaVA-OneVision-7B	V	<u>49.8</u>	51.7	<u>56.6</u>
MacawLLM	V,A	25.5	42.1	14.5
MAESTRO (Ours)	V,A	**82.0**	**86.9**	**87.2**

Achieves **SOTA** by 32.2%, 21.5%, 30.6%

Advances general multimodal understanding of VLMs

MAESTRO-CLAP

Model	Acc	Multiclass (Zero-shot) M-F1	F1(O)	R(O)	P(O)
ZAC	0.21	0.23	0.20	0.26	0.33
LAION-CLAP	<u>0.32</u>	<u>0.28</u>	<u>0.30</u>	<u>0.34</u>	<u>0.31</u>
MAESTRO–CLAP (Ours)	**0.82**	**0.89**	**0.84**	**0.92**	**0.90**



Strong Zero-Shot Abilities

Suited for performance in the wild