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

Felicia Tan Ee Shan[1], Low Li Ying Amy[1], Kuek Yong Jie Adriel[2]
[1]Raffles Institution, 1 Raffles Institution Ln, Singapore 575954
[2]DSO National Laboratories, 12 Science Park Dr, Singapore 118225

Abstract

The video-audio modality remains a critical challenge for current Vision-Language Models (VLMs) due to weak static reasoning, ineffective integration of the auditory modality and high computational overhead, thus compromising performance on complex reasoning tasks like hateful video detection. Moreover, multimodal content's proliferation has significantly increased the complexity of hateful video detection. Our framework, MAESTRO, overcomes these challenges by aligning visual and auditory modalities into a unified space through our proposed MAESTRO–Unified Multimodal Alignment framework, enabling a more holistic understanding of multimodal interactions. It also utilises our proposed MAESTRO– Adaptive Global-Local Reasoning Loop that combines detailed local insights with broader contextual analysis and employs a dynamic iterative reasoning approach, which adapts to tasks without the computational burden of exhaustive frame-by-frame processing. MAESTRO achieves state-of-the-art (SOTA) on MultiHateClip, while achieving localisation ability and improved multimodal reasoning. Additionally, MAESTRO achieves SOTA on industry benchmarks MSRVTT-QA, MSVD-QA and ActivityNet-QA for general video understanding, highlighting its advancement of general VLM video understanding. These results demonstrate its potential as a foundational framework for diverse multimodal reasoning tasks as well as broader applications beyond the hateful video detection use-case like misinformation detection, disaster response, and human rights monitoring

1 Introduction

The proliferation of multimedia content has significantly increased the complexity of multimodal information, particularly for hateful video detection. Platforms like YouTube see over 500 hours of video uploaded every minute [1], while TikTok and Bilibili have reached billions of global users [2]. The sheer volume and diversity of such content exacerbate the difficulty of timely and effective content moderation. Yet, current content moderation systems rely on labour-intensive manual reviews, which are inefficient, limited in scale, and expose moderators to mental distress, resulting in a higher risk of mental disorders such as PTSD [3].

While advancements in Vision-Language Models (VLMs) have facilitated some automation in hateful video detection, their capabilities remain constrained in multimodal reasoning, computational efficiency, and integration of diverse data sources [4]. Evidently, the video modality, often accompanied by complex relationships with audio and text, remains a critical challenge for VLMs. To address these gaps, we propose MAESTRO (Multimodal Agentic model with Efficient, Synergistic Text-Reasoning Optimisation framework), a novel framework that redefines multimodal understanding and reasoning in VLMs by integrating vision, language, and audio into a unified, context-aware system.

Unlike traditional VLMs, MAESTRO

1. Achieves novel and effective integration of the commonly overlooked audio modality through alignment of visual and audio components into a unified space

2. Utilises dynamic, iterative and context-aware reasoning without the need to sample every frame, while still ensuring fine-grained understanding (local) and global understanding
3. Achieves SOTA reasoning and performance on general video Visual Question Answering (VQA) as well as the specific use-case of hateful video detection

2 Literature Review

2.1 Vision Language Models (VLMs)

VLMs have shown significant advances by integrating visual and textual data through multimodal architectures that typically consists of three main components: an image encoder, a text encoder, and a fusion mechanism [4]. While current VLMs like VideoChat and LLaVA predominantly rely on large language models (LLMs) as decoders, this limits their control over cross-modal interactions and reasoning abilities.

In contrast, in MAESTRO, we propose for LLMs to act as control agents so as to leverage an agentic framework for more dynamic and context-aware cross-modal reasoning.

Additionally, MAESTRO overcomes the following limitations of current VLMs:

Limitations of current VLMs	How our model addresses this
1. High computational overhead	Optimised architecture that allows for dynamic reasoning decisions, reducing computational overhead
2. High rate of hallucinations	Reduced hallucinations
3. Weak reasoning capabilities	Synergistic text-reasoning framework that improves multimodal reasoning accuracy and depth
4. Missing/Ineffective integration of audio modality	Advanced audio-text-video alignment for effective multimodal integration through projection into a unified space
5. Lack of grounding capabilities	Ability to localise and ground hateful elements

Table 1: Limitations of current VLMs and how MAESTRO addresses them

2.2 Importance of Audio Modality

The audio modality remains a critically underexplored dimension in video-language modelling, despite its crucial role in providing rich contextual and semantic cues. Contemporary methods like PG-Video-LLaVA [5] treat the audio modality as an auxiliary text source by appending the transcript to a prompt for processing by LLMs, neglecting non-speech elements like environmental sounds, music, and speaker intonation, as well as the temporal alignment with video. Alternatives, like Video-LLaMA [6] and MA-LMM [7], use query transformers (Q-formers) to process audio directly, but introduce significant computational overhead, with costs scaling quadratically with temporal resolution. For both approaches, most models still treat video and audio as entirely separate modalities, processed in parallel streams, limiting their ability to capture complex multimodal relationships.

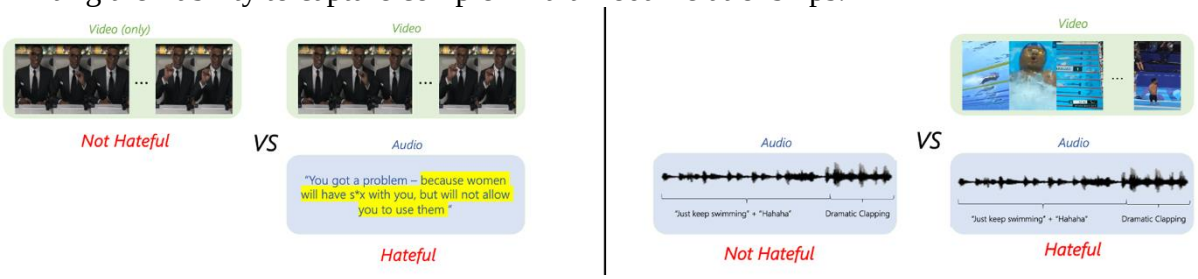


Figure 1: Left: An example of a video where visual content is non-hateful, but the audio introduces hateful elements. Right: An example of a video where audio is non-hateful, but combining both video and audio presents hateful content

As seen in the above example, the audio modality and its temporal alignment with the video modality is essential to understanding the hateful content above.

2.3 Visual Question-Answering (VQA) and Hateful Video Classification

Despite progress in text-based hate speech detection through Natural Language Processing (NLP) techniques, current models face substantial challenges when applied to multimodal content, where both visual and auditory cues shape the overall context. To illustrate, for the general video VQA task, on MSRVTT, the current state-of-the-art (SOTA) performance, achieved by LLaVA-OneVision, is an F1 score of 49.8. While this is the best result reported to date, it remains unsatisfactory for real-world applications. Similarly, on ActivityNet-QA, LLaVA-OneVision achieves a SOTA F1 score of 56.6, which also reflects significant room for improvement.

For the specific task of hateful video classification, empirical experimentation on the MultiHateClip dataset [8] reveals that most existing models perform poorly. Common issues include producing entirely incorrect predictions or demonstrating flawed reasoning (refer to Figure 6 for illustrative examples). Upon error analysis of models' performance on the MultiHateClip dataset, while some models struggle to detect specific hateful elements, the larger issue lies in their inability to reason about the content. Moreover, hateful content often relies on subtle cues like sarcasm and coded language that carries different connotations across cultural contexts, making classification especially challenging. The temporal modality, where hatefulness emerges only when multiple frames or audio segments are considered together, further complicates detection. These challenges underscore the necessity for more advanced architectures that integrate the audio modality, temporal analysis, and dynamic reasoning.

3 Model Architecture

3.1 Pipeline

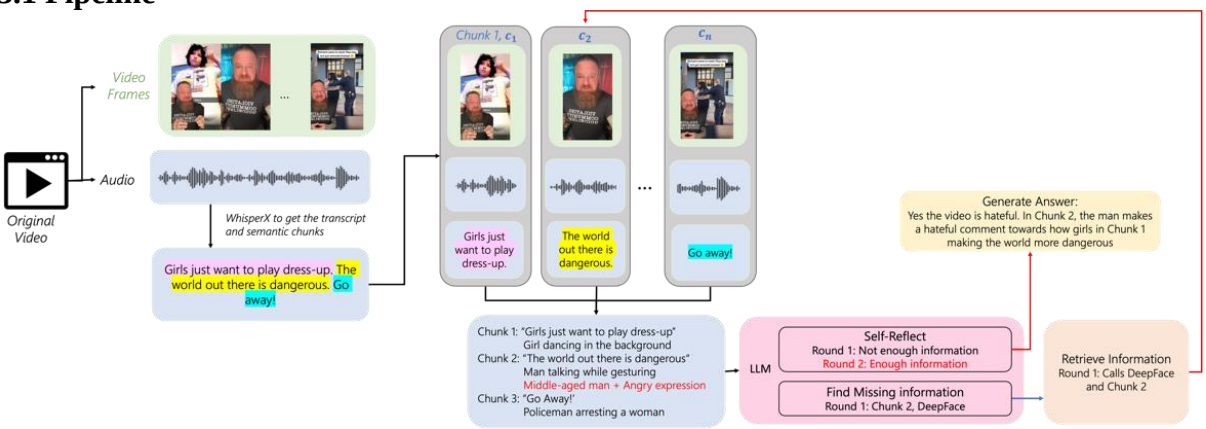


Figure 2: An overview of MAESTRO's pipeline

3.2 Video Modality

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. However, the necessity of processing every single frame to understand a video is questionable. However, processing every frame may not be necessary. Unlike exhaustive frame-by-frame methods, humans first grasp the broader context (global) and then focus selectively on key segments (local), iteratively refining their understanding. Rather than demanding the capacity

to process all visual inputs simultaneously, this approach underscores the greater importance of reasoning capabilities to prioritise and revisit content efficiently. Inspired by this principle, we propose MAESTRO, a system that mimics this selective and iterative process. By dynamically identifying and analysing the most pertinent frames or segments, MAESTRO avoids the computational overhead of exhaustive video analysis while achieving SOTA performance in hateful video detection and general video VQA.

MAESTRO's pipeline begins with:

3.2.1 Transcript Chunking and Concept Representation

Let $T = \{t_1, t_2, \dots, t_n\}$ represent the transcript tokens extracted from the video's accompanying text. Using a BERT-based module [9], the transcript is segmented into $C = \{c_1, c_2, \dots, c_k\}$, where c_i denotes the i-th semantic "chunk," encapsulating a coherent conceptual unit. Formally, this is expressed as $c_i = \text{BERT}(T_i), \quad i \in \{1, \dots, k\}$ where $T_i \subseteq T$ refers to the tokens grouped into the i-th chunk. Each c_i acts as a guiding concept for subsequent stages. For a given chunk c_i, its associated temporal span in the video is denoted as $\Delta t_i = [t^i_{\text{start}}, t^i_{\text{end}}]$. Corresponding video frames $F_{\Delta t_i} = \{f_1, f_2, \dots, f_m\}$ within this interval are selected. These frames are paired with the semantic chunk to form what we term idea-frame pairs $P = \{(c_1, F_{\Delta t_1}), \dots, (c_k, F_{\Delta t_k})\}$. This pairing is represented as $P_i = (c_i, F_{\Delta t_i}), i \in \{1, \dots, k\}$.

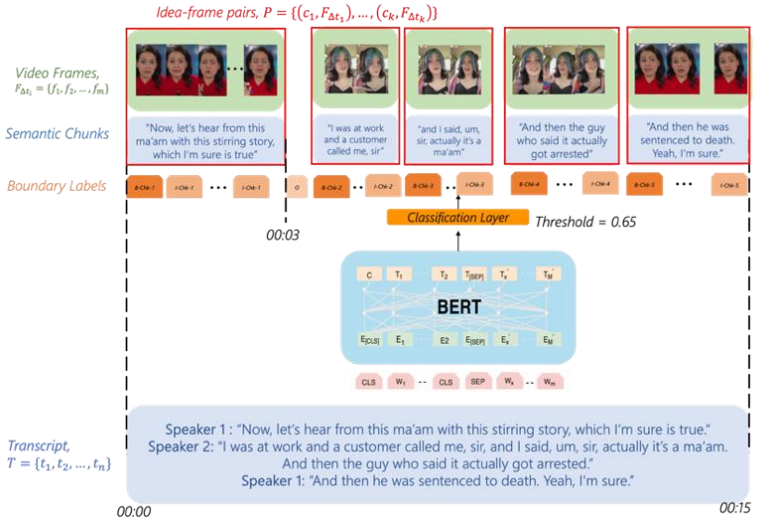


Figure 3: An overview of how Idea-Frame Pairs are formed through semantic chunking and grouping of video frames

3.3 Audio Modality

As aforementioned, the missing/ineffective integration of the audio modality into current VLMs critically limits performance, as they often miss out on important context present in the audio only.

3.3.1 MAESTRO-Unified Multimodal Alignment

To address this gap of audio integration, we propose the MAESTRO framework, a novel approach that unifies visual, textual, and audio modalities into a shared feature space. This integration enables rich multimodal interactions, enhancing comprehensive understanding while reducing computational inefficiencies and preserving contextual richness. MAESTRO first decomposes audio input into two parts (1) Speech Audio (2) Non-Speech Audio.

(1) Speech Audio

The audio input is segmented into speech intervals using Voice Activity Detection (VAD), which maps an acoustic feature sequence $A = \{a_1, \dots, a_T\}$ to binary labels $y = \{y_1, \dots, y_T\}$ where $y_t = 1$ indicates speech presence

Post-processing converts y into active speech segments $S = \{s_1, \dots, s_N\}$, defined by timestamps (t_i^0, t_i^1). To handle variable segment lengths, we employ a min-cut strategy for overly long segments and split at the point of minimum voice activation score. Conversely, short segments are merged with neighbours if their combined duration is below a threshold $\tau = |A_{\text{train}}|$, maintaining temporal consistency. Processed segments are transcribed in parallel using Whisper [10] yielding independent text outputs $T = \{T_1, \dots, T_N\}$. This approach avoids context-dependent errors like repetition. For precise word timing, forced phoneme alignment is applied. Each segment's phonemes C_{T_i} are classified, producing logits $L_i \in R^{|C_{T_i}| \times T}$. Dynamic Time Warping (DTW) aligns phonemes temporally, and word boundaries are derived from their start and end times. These high-quality text embeddings are then aligned into the unified text feature space, ensuring compatibility with other modalities and preserving semantic alignment.

(2) Non-Speech Audio

However, non-speech audio components, such as environmental sounds, music, and speaker intonations, carry significant semantic value that contributes to the overall meaning of the video. Despite this, they are often underrepresented or neglected in current audio modelling approaches, which primarily focus on speech content.

To capture non-speech audio, our custom MAESTRO-CLAP [11] module employ two distinct audio encoders: PANN [12] and HTSAT [13] (Refer to Appendix 1.1 for detailed architecture)

- **PANN**: A CNN-based audio classification model with 7 downsampling and 7 upsampling blocks. Its penultimate layer produces a 2048-dimensional feature vector, where $z_{\mathbf{PANN}} \in R^{2048}$
- **HTSAT**: A state-of-the-art transformer model featuring 4 groups of Swin Transformer blocks, outputting a 768-dimensional feature vector, where $\boldsymbol{z}_{\mathbf{HTSAT}} \in R^{768}$

These audio encodings are passed through a two-layer MLP with ReLU activation to project them into a unified 512-dimensional space, with $f_{\text{audio}} = \text{MLP}_{\text{audio}}(z_{\text{audio}}), \quad f_{\text{audio}} \in R^{512}$. For text, three encoders are utilised: CLIP [14], BERT [9], and RoBERTa [15]. Their respective outputs are $z_{\mathbf{CLIP}} \in R^{512}, \quad \boldsymbol{z}_{\mathbf{BERT}}\ and\ \boldsymbol{z}_{\mathbf{RoBERTa}} \in R^{768}$. These are also projected to the shared 512-dimensional space through a Multilayer Perceptron (MLP), with $f_{\text{text}} = \text{MLP}_{\text{text}}(z_{\text{text}}), \quad f_{\text{text}} \in R^{512}$.

Unified Representation with Contrastive Learning

Given audio features f_{audio} and text features f_{text}, encoders are trained using a contrastive loss $L_{contrastive} = -\frac{1}{N}\sum_{i=1}^{N} log\ \frac{exp(f_{\text{audio}}^i \cdot f_{\text{text}}^i/\tau)}{\sum_{j=1}^{N} exp(f_{\text{audio}}^i \cdot f_{\text{text}}^i/\tau)}$ to align audio and text embeddings in the shared space.

Num. of Training Examples	8822 (out of 11028)
Num. of Epochs	20
Instantaneous batch size per device	24
Total train batch size (w. parallel, distributed & accumulation)	96
Gradient Accumulation steps	4
Total optimisation steps	2000
Num. of trainable parameters	153,507,530

Table 2: Details of training configuration of our custom MAESTRO-CLAP module

MAESTRO-CLAP has been found to demonstrate strong zero-shot performance (F1-score of 0.84, Refer to Appendix 1.2 for detailed results) compared to other non-speech audio classifiers when evaluated on our custom SocialSound-11k dataset containing 11028 audio-text pairs of common sound effects found in social media videos. Its zero-shot capability enables it to adapt to new and unseen sounds, making it highly scalable for dynamic environments like social media, where viral audio trends and memes emerge daily.

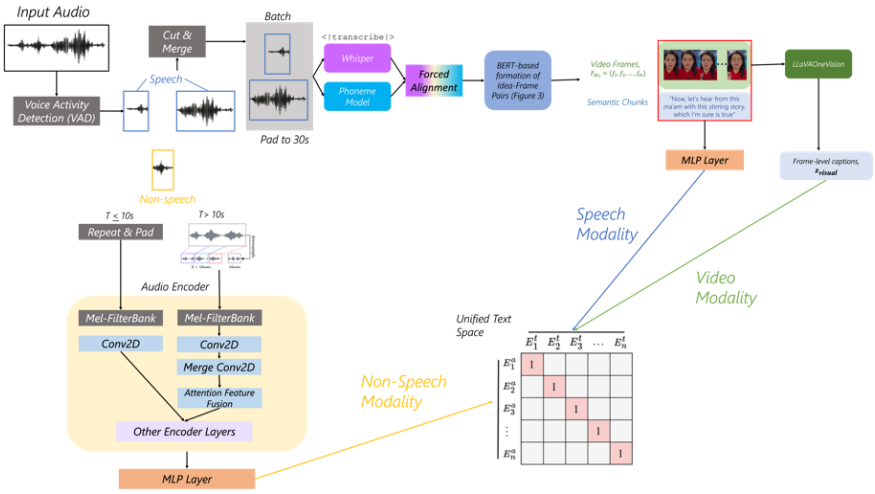


Figure 4: MAESTRO projects 1) Speech Audio 2) Non-speech Audio 3) Visual Frames into one unified text space for multimodal alignment

3.4 MAESTRO– Adaptive Global-Local Reasoning Loop

We propose MAESTRO– Adaptive Global-Local Reasoning Loop to ensure comprehensive reasoning across temporal and fine-grained dimensions, motivated by the reasoning in Appendix 1.3. Inspired by the iterative reasoning approach for long form video understanding of VideoAgent [16], we extend this method for enhancing the depth and localisation of reasoning in our framework. This enables context-aware and temporally coherent video understanding that combines local insights with broader contextual analysis.

Global Reasoning

To achieve global reasoning, we construct a Global Representation by summarizing the video across its temporal dimension. Each video V is divided into N chunks $\{C_1, C_2, \ldots, C_N\}$. From each chunk, we select a pair of keyframes $\{F_i^1, F_i^2\}$ to serve as representative frames $\mathbf{G} = \mathcal{F}_{\text{global}}(V) = \bigcup_{i=1}^{N}\{g(F_i^1), g(F_i^2)\}$, where $g(\cdot)$ represents the feature extraction function of the vision encoder. The aggregated global representation $\mathbf{G} \in R^{N\times d}$ captures temporal dependencies and broader context.

Local Reasoning

For localised understanding, each chunk C_i undergoes fine-grained analysis to generate Local Representations, focusing on textual and visual features:

1. **Speech Audio**: Text embeddings $\mathbf{z}_{\mathbf{speech}}$ are derived from the speech component.
2. **Non-Speech Audio**: Non-verbal sound components $\mathbf{z}_{\mathbf{non\text{-}speech}}$ are processed as explained.
3. **Visual Frames**: Frame-level fine-grained captions $\mathbf{z}_{\mathbf{visual}}$ are generated using LLaVA-OneVision.

The local representation $\mathbf{L_i}$ for chunk C_i is then expressed as $\mathbf{L_i} = \{p_\theta(z_{\mathbf{speech}}), p_\theta(z_{\mathbf{non\text{-}speech}}), p_\theta(z_{\mathbf{visual}})\}$, where $p_\theta(\cdot)$ aligns each modality into the unified text space. Both global ($\mathbf{G}$) and local ($\mathbf{L_i}$) representations are aligned within a shared text space to enable cross-modal interaction.

Initial State Representation

The initial state of the video H_{init} is formed by combining the Global Representation G with the set of initial Local Representations $\{L_1, L_2, \dots, L_N\}$, providing a comprehensive starting point $H_{\mathbf{init}} = G \cup \bigcup_{i=1}^{N} L_i$

Deciding the Subsequent Action

At each round, the current state, $\boldsymbol{H_t}$, aggregating all observed frames and context, is input into ChatGPT-3.5 to decide the next action via a structured reasoning processing:

- **Action 1: Generate an Answer** if $\boldsymbol{H_t}$ contains sufficient information
- **Action 2: Gather Additional Context** if $\boldsymbol{H_t}$ lacks critical details

This decision is facilitated through a structured three-step reasoning approach:

1. **Hypothesis Generation:** Proposes a preliminary answer with chain-of-thought reasoning.
2. **Confidence Evaluation:** Assigns a confidence score $c_t \in \{1,2,3\}$ to H_t
   - $c_t = 1$: The information is insufficient.
   - $c_t = 2$: Partial information is available but incomplete.
   - $c_t = 3$: The information is complete.
3. **Action Decision:** Executes Action 1 for $c_t = 3$; otherwise, performs Action 2 to gather more information

The initial state, $\boldsymbol{H_{init}}$, comprises the **Global Representation G** with the set of **initial Local Representations** $\{\boldsymbol{L_1}, \boldsymbol{L_2}, \dots, \boldsymbol{L_N}\}$, and the process progressively refines $\boldsymbol{H_t}$ to achieve a comprehensive understanding before generating the final output for reasoning.

Agents for Gathering New Observations

The recent emergence of agentic systems, which are advanced AI models with autonomous decision-making and adaptability for complex tasks have gained significant traction [17]. Using a modular architecture, they assign specialised tools to different data features, enhancing efficiency and performance. This approach enables adaptive information retrieval, making them ideal for automating complex workflows across various domains.

Leveraging these benefits, we deploy a modular agentic system to facilitate adaptive information retrieval when the system determines that further context is required (Action 2). The retrieval process is guided by a structured prompt, Prompt 1 (Appendix 1.4).

Tool 1: LLaVAOneVision – Action Recognition

LLaVAOneVision [18] is employed as the action recognition tool. It excels at generating video descriptions with minimal hallucination, making it reliable for tasks demanding accurate interpretation of actions and object interactions. However, it lacks advanced reasoning capabilities, making it less suitable for inference-heavy tasks.

Tool 2: YOLO – Symbol Recognition

YOLO (You Only Look Once) [19] is integrated for its robust ability to recognize symbols, such as logos or hate symbols (e.g., swastikas), making it valuable for scenarios requiring nuanced symbol detection such as our example use case of hateful video detection.

Tool 3: videoDeepFace – Identity Tracking

Building on the image-based DeepFace model [20], we introduce videoDeepFace, a novel extension that extends DeepFace across the temporal dimension to handle video sequences for face identification while ensuring continuous identity tracking for multiple faces across frames, preventing identity swaps. videoDeepface additionally integrates features, such as hair colour, to improve demographic classification (e.g. race).

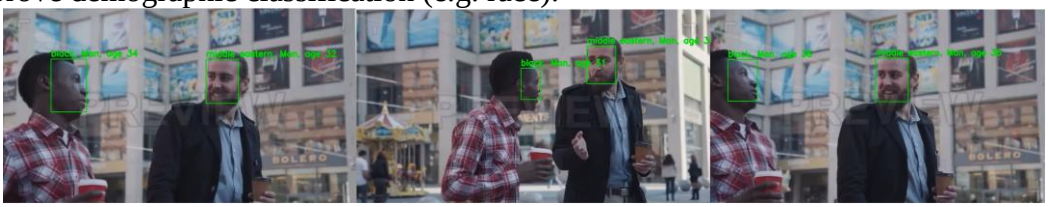


Figure 5: Demonstration of videoDeepFace's capabilities for continuous identity tracking and demographic classification (race, age, gender, emotion) in video sequences

Updating the current state

After information retrieval, the current state, $\boldsymbol{H_t}$, 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 ($\boldsymbol{c_t}$) 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)

Final Decision

The model makes its final decision based on Prompt 2 (Appendix 1.5) that decomposes the task of identifying whether a video is hateful into two specific criteria. Refer to Appendix 1.6 for a diagrammatic overview of MAESTRO's reasoning loop.

4 Results

4.1 Hateful Video Detection (MultiHateClip Dataset)

		Binary				
	Model	**Acc**	**M-F1**	**F1(O)**	**R(O)**	**P(O)**
T1	mBERT	0.57	0.57	0.52	0.42	0.68
T2	GPT-4	<u>0.81</u>	<u>0.79</u>	<u>0.73</u>	0.69	0.78
T3	Qwen	0.72	0.71	0.65	0.57	0.75
A1	MFCC	0.54	0.50	0.36	0.33	0.40
A2	Wav2Vec	0.53	0.48	0.64	0.50	<u>0.90</u>
V1	ViViT	0.73	0.73	0.68	0.57	0.86
V2	Vit	0.63	0.58	0.44	0.46	0.45
VL1	VLM	0.70	0.64	0.48	0.59	0.41
VL2	GPT-4V	<u>0.81</u>	<u>0.79</u>	<u>0.73</u>	<u>0.72</u>	0.73
VL3	Qwen-VL	0.62	0.61	0.56	0.46	0.72
M1	T1 ⊙ A1 ⊙ V1	0.75	0.74	0.67	0.61	0.77
M2	MAESTRO (Ours)	**0.96**	**0.95**	**0.93**	**0.87**	**1.0**

Table 2: Model performance for English YouTube hateful video classification. Metrics: H = Hateful, O = Offensive, Acc = Accuracy, M-F1 = Macro F1, R = Recall, P = Precision. Best results are **bolded**, second-best are *underlined*. Models include Multimodal (M), Video-only (V), Text-only (T), Audio-only (A), and Vision-Language (VL).

Experimental results show our model outperforming other current VLMs across all metrics on the MultiHateClip dataset [8]. Notably, unlike other models that were specifically fine-tuned on the dataset before evaluation, our framework operates without requiring dataset-specific training. Despite this, it achieves SOTA by a considerable margin in both quantitative and qualitative evaluations. Remarkably, MAESTRO surpasses GPT-4V, which is widely regarded as the gold standard for vision-language tasks. It is worth noting that GPT-4V was trained using an immense amount of computational resources by OpenAI – an estimated 25,000 Nvidia A100 GPUs running continuously for approximately 90 to 100 days [21]. In contrast, our framework matches GPT-4V's performance with only a significantly smaller fraction of computation resources (2 Nvidia T4 GPUs), showcasing the comparative computational efficiency and scalability of our approach. Refer to Appendix 1.7 for the breakdown of the usage of the various agents.

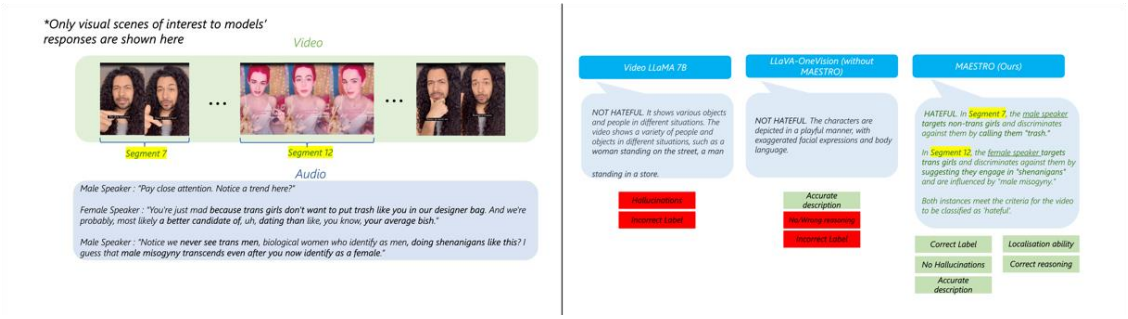


Figure 6: A comparative analysis of MAESTRO v.s. SOTA models, evaluated on a challenging example.

Localisation and Segmentation Capability

As evident from Figure 4, MAESTRO can localise both distinct segments with hateful content present —one involving the male speaker and another involving the female speaker—and provide contextually grounded explanations for both. This capability is in stark contrast to other models like Video-LLaMA-7B and LLaVA-OneVision, which fail to detect either of these segments. Such fine-grained localisation is crucial in multimodal tasks where important content may be sparsely distributed across temporal and visual dimensions.

Improved Reasoning and Labelling Accuracy

MAESTRO demonstrates significant improvements in reasoning ability compared to other models. While Video-LLaMA-7B and LLaVA-OneVision incorrectly classify the video as "not hateful", MAESTRO correctly labels the video as "HATEFUL" and provides the correct explanation. This nuanced reasoning underscores MAESTRO's capacity to align multimodal inputs (audio and video) with contextual semantics to arrive at well-justified conclusions. Refer to Appendix 1.8 for further analysis of results.

4.2 Advancing General VQA in VLMs (Industry Benchmarks)

While MAESTRO demonstrates exceptional performance in hateful video detection, its design as a unified, context-aware multimodal reasoning framework extends far beyond this specific use-case.

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	65.4	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	49.8	51.7	56.6
MacawLLM	V,A	25.5	42.1	14.5
MAESTRO (Ours)	V,A	**82.0**	**86.9**	**87.2**

Table 3: Model performance (F1-score) on Industry Benchmarks

As shown in Table 3, MAESTRO significantly outperforms other competing VLMs across the MSRVTT-QA [22], MSVD-QA [22], and ActivityNet-QA [23] benchmarks, achieving SOTA F1-scores of 82.0%, 86.9%, and 87.2%, respectively. These benchmarks test video question-answering capabilities, evaluating understanding of actions, objects, events, and high-level activities in videos, thus demonstrating how **MAESTRO advances multimodal reasoning in vision-language models (VLMs) broadly, beyond the specific application of hate speech detection**.

MAESTRO's architecture addresses fundamental limitations in existing VLMs for VQA by:

1. **Enabling Rich Multimodal Interactions –** Unlike existing VLMs, MAESTRO fully integrates the audio modality (speech and non-speech) and aligns them temporally with information from video frames in a unified space
2. **Adaptive Global-Local Reasoning Loop–** This iterative reasoning process enables the model to focus selectively on relevant video segments, which adapts the model's understanding to complex queries that require fine-grained temporal and multimodal analysis.

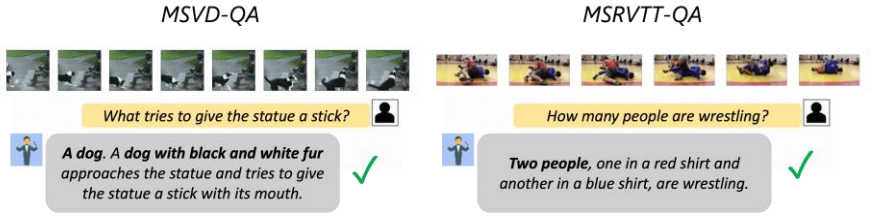


Figure 7: Examples of MAESTRO's responses for MSVD-QA, MSRVTT-QA

5 Discussion

5.1 Automated content moderation

By integrating visual, audio, and textual modalities into a unified system, MAESTRO enables more accurate identification of harmful content, even when hateful messages are subtle or distributed across modalities. Its ability to process multimodal inputs dynamically and contextually ensures higher precision and recall, making it a valuable tool for social media moderation, content review platforms, and regulatory frameworks seeking to curb the spread of harmful multimedia content.

6 Conclusion

Our proposed MAESTRO framework redefines multimodal reasoning in Vision-Language Models (VLMs) by integrating vision, audio, and text into a unified feature space through MAESTRO–Unified Multimodal Alignment and adopts dynamic and iterative reasoning through our MAESTRO– Adaptive Global-Local Reasoning Loop. Beyond achieving state-of-the-art performance in hateful video detection, MAESTRO fundamentally advances VLM capabilities in Video Question Answering (VQA), setting new benchmarks (MSVD-QA, MSRVTT-QA, and ActivityNet-QA) for understanding actions, objects, and high-level events in videos. These results demonstrate its potential as a foundational framework for diverse multimodal reasoning tasks as well as broader applications like misinformation detection, disaster response, and human rights monitoring.

7 References

[1] SOAX, “How many hours of video are uploaded to youtube every minute?” *SOAX*, Oct. 08, 2024. https://soax.com/research/how-many-hours-of-video-are-uploaded-to-youtube-every-minute

[2] M. Dellatto, “TikTok Hits 1 Billion Monthly Active Users,” *Forbes*, Sep. 27, 2021. https://www.forbes.com/sites/marisadellatto/2021/09/27/tiktok-hits-1-billion-monthly-active-users/

[3] R. Spence, A. Bifulco, P. Bradbury, E. Martellozzo, and J. DeMarco, “Content Moderator Mental Health, Secondary Trauma, and Well-being: A Cross-Sectional Study,” *Cyberpsychology, Behavior, and Social Networking*, Dec. 2023, doi: https://doi.org/10.1089/cyber.2023.0298.

[4] J. Zhang, J. Huang, S. Jin, and S. Lu, “Vision-Language Models for Vision Tasks: A Survey,” *arXiv.org*, Apr. 02, 2023. https://arxiv.org/abs/2304.00685

[5] S. Munasinghe *et al.*, “PG-Video-LLaVA: Pixel Grounding Large Video-Language Models,” *arXiv.org*, 2023. https://arxiv.org/abs/2311.13435 (accessed Jan. 23, 2025).

[6] H. Zhang, X. Li, and L. Bing, “Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding,” *arXiv (Cornell University)*, Jun. 2023, doi: https://doi.org/10.48550/arxiv.2306.02858.

[7] B. He *et al.*, “MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding,” *arXiv.org*, 2024. https://arxiv.org/abs/2404.05726 (accessed Jan. 23, 2025).

[8] H. Wang, T. R. Yang, U. Naseem, and R. K.-W. Lee, “MultiHateClip: A Multilingual Benchmark Dataset for Hateful Video Detection on YouTube and Bilibili,” *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7493–7502, Oct. 2024, doi: https://doi.org/10.1145/3664647.3681521.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv.org*, Oct. 11, 2018. https://arxiv.org/abs/1810.04805

[10] A. Radford, J. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” Dec. 2022. Available: https://cdn.openai.com/papers/whisper.pdf

[11] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and Shlomo Dubnov, “Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation,” *arXiv (Cornell University)*, Nov. 2022, doi: https://doi.org/10.48550/arxiv.2211.06687.

[12] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, “PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition,” *arXiv:1912.10211 [cs, eess]*, Aug. 2020, Available: https://arxiv.org/abs/1912.10211

[13] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, “HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection,” *arXiv.org*, 2022. https://arxiv.org/abs/2202.00874

[14] A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision,” *arXiv:2103.00020 [cs]*, Feb. 2021, Available: https://arxiv.org/abs/2103.00020

[15] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” *arXiv.org*, Jul. 26, 2019. https://arxiv.org/abs/1907.11692

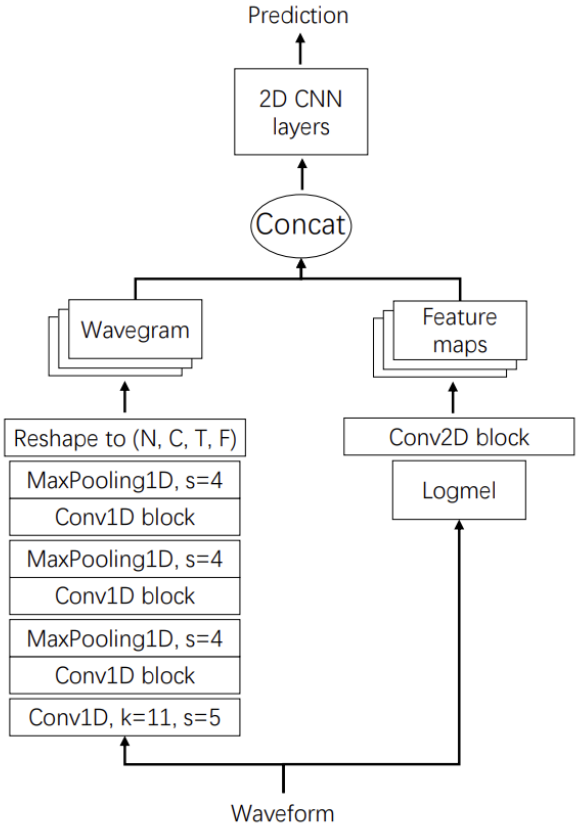
[16] X. Wang, Y. Zhang, O. Zohar, and S. Yeung-Levy, “VideoAgent: Long-form Video Understanding with Large Language Model as Agent,” *arXiv.org*, 2024. https://arxiv.org/abs/2403.10517 (accessed Jan. 23, 2025).

[17] S. Hu, C. Lu, and J. Clune, “Automated Design of Agentic Systems,” *arXiv.org*, 2024. https://arxiv.org/abs/2408.08435

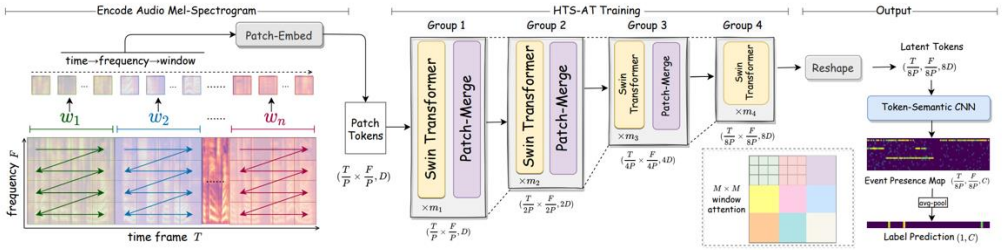
[18] B. Li *et al.*, “LLaVA-OneVision: Easy Visual Task Transfer,” *arXiv.org*, 2024. https://arxiv.org/abs/2408.03326 (accessed Jan. 23, 2025).
[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” *arXiv.org*, Jun. 08, 2015. https://arxiv.org/abs/1506.02640
[20] S. I. Serengil, “serengil/deepface,” *GitHub*, Aug. 30, 2020. https://github.com/serengil/deepface
[21] OpenAI, “GPT-4V(ision) System Card OpenAI,” 2023. Available: https://cdn.openai.com/papers/GPTV_System_Card.pdf
[22] D. Xu *et al.*, “Video Question Answering via Gradually Refined Attention over Appearance and Motion,” Oct. 2017, doi: https://doi.org/10.1145/3123266.3123427.
[23] Z. Yu *et al.*, “ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering,” *arXiv.org*, 2019. https://arxiv.org/abs/1906.02467 (accessed Jan. 23, 2025).

8 Appendix

1.1: Architecture of PANN (top) and HTSAT (bottom)



(Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2019). PANNS: Large-Scale pretrained audio neural networks for audio pattern recognition. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1912.10211**)**



(*HTS-AT: a hierarchical Token-Semantic audio transformer for sound classification and detection*. (2022, May 23). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9746312)

1.2: Table 1: Model performance for zero-shot multi-class labelling of SocialSound-11k test set

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

1.3 : Importance of Time Alignment in Video Understanding

Some previous approaches to incorporating the audio modality, such as PG-Video-LLaVA, typically involve simply appending the entire transcript to the input. However, this approach fails to account for the temporal dependencies inherent in audio data, leading models to overlook crucial time-dependent information.

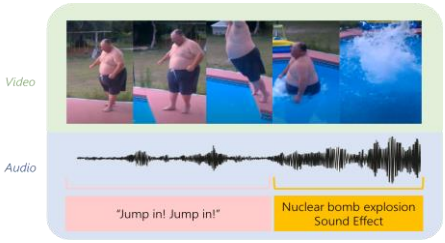


Figure 3: An example where the visual content is non-hateful, but the audio, when synchronised appropriately, introduces hateful elements.

1.4: Prompt 1

"Video Content: {video_content}

Assess whether the video meets the following two criteria: Criteria 1: The video targets a certain individual or group of individuals based on a characteristic. Criteria 2: The video discriminates against, blames, or encourages harm/fear/hatred toward this individual/group, or threatens societal peace/harmony.

Your task is to evaluate whether these criteria are met based on the content provided. You will then choose a **TOOL** and a **CHUNK ID** to apply the TOOL to that CHUNK and gather more information about the video. The available **AGENTS** are:

1. **'YOLO'** to analyse the objects present in the video segment.
2. **'LLOV'** to examine the actions of the people in the segment.
3. **'DEEPFACE'** to analyse the race, age, gender, and emotion of the people in the segment.

If **SUFFICIENT INFORMATION** is available to determine whether the video is hateful, respond in the following format: **[STOP, Explanation on whether the video is hateful or not]**

If **INSUFFICIENT INFORMATION** is available, select a **TOOL** and a **CHUNK ID** to explore further. Respond in the following format: [**CONTINUE, [chunk_id], [tool]]"**

1.5: Prompt 2

Prompt:

"The two criteria for a video being classified as 'hateful' are: Criteria 1: The video targets a certain individual or group of individuals based on a characteristic. Criteria 2: The video discriminates against, blames, or encourages harm/fear/hatred toward this individual/group, or threatens societal peace/harmony. Video Content: {video_content} Based on the provided information, is the video hateful according to these criteria? Answer format: [HATEFUL/NOT HATEFUL], [Explanation]"

1.6: Iterative Reasoning Loop

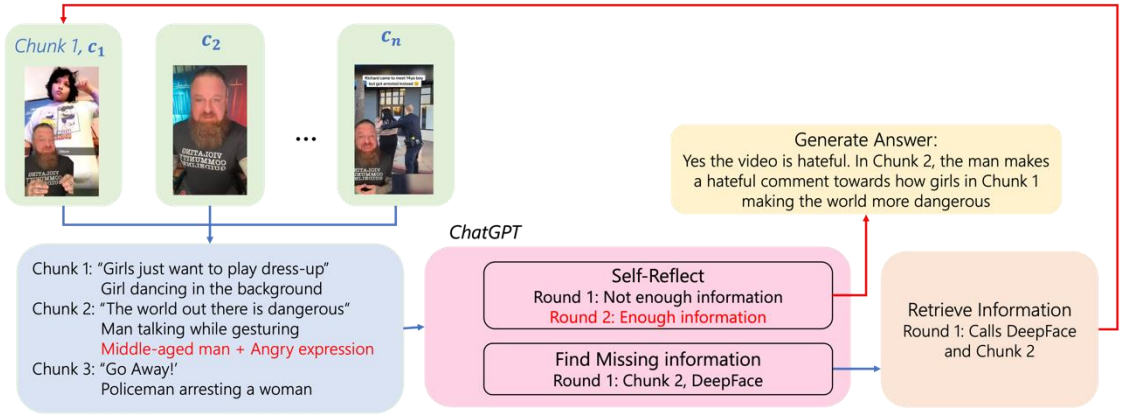
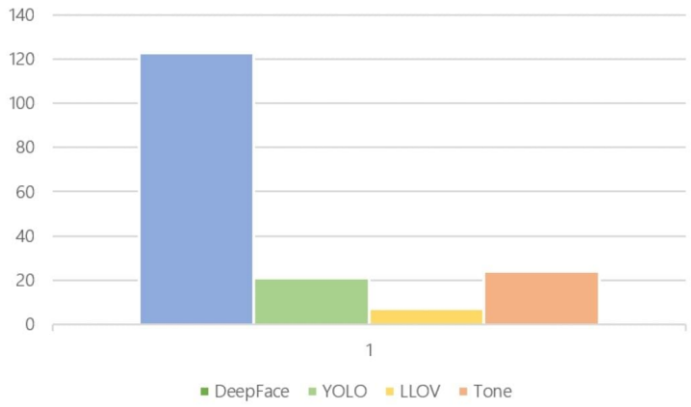


Figure 4: An overview of MAESTRO's dynamic iterative reasoning loop

1.7 : Breakdown of Agent Usage



1.8 : Further Analysis of Results of MAESTRO on MultiHateClip

Incorporation of LLaVA-OneVision Strengths

While MAESTRO inherits the descriptive accuracy of LLaVA-OneVision, it substantially mitigates the latter's weaknesses, such as incorrect reasoning and labelling.

Addressing Hallucinations

Video-LLaMA-7B exhibits hallucinations in its descriptions, introducing irrelevant or fabricated details such as “a man and woman standing in a store”, which are absent from the actual video. In contrast, MAESTRO’s outputs are free from hallucinations, ensuring fidelity to the input data and delivering accurate, context-aware assessments.