Towards a Robust Framework for Multimodal Hate Detection: A Study on Video *vs.* Image-based Content

Girish A. Koushik, Diptesh Kanojia, Helen Treharne

Nature **I**nspired **C**omputing and **E**ngineering Research (NICE),
Surrey **C**entre for **C**yber **S**ecurity (SCCS)

April 29, 2025

Introduction

- **Hate speech:** Any communication attacking a person/group based on
  - Religion, ethnicity, nationality
  - Race, colour, descent
  - Gender, other identity factors
- Impact
  - Individual: Mental health, self-esteem
  - Community: Social division
  - Economic: Healthcare costs
  - Legal: Balance between “free speech” and “hate speech”

Figure: Source:https://codalab.lisn.upsaclay.fr/competitions/20699

NB This presentation uses hateful example images only for illustrative purposes.

Problem Statement

Task Definition

Given input x from modality space $\mathcal{M}$, where $\mathcal{M} \in \{\text{Text}, \text{Image}, \text{Audio}, \text{Video}\}$. Learn a function f that maps input to binary label space:

$$f : \mathcal{M} \rightarrow \{0, 1\}$$

where 0 represents non-hateful content and 1 represents hateful content.

Challenges

- An effective framework should address three key dimensions of hate speech detection:

$$F_{total} = \sum_{j \in \{text, image, audio\}} \left[\alpha(w_j F^j_{semantic}) + \beta(w_j F^j_{contextual}) + \gamma(F_{temporal}) \right]$$

  where α, β, γ are coefficients, w_j represents the modality inputs. $F_{semantic}$ captures content meaning in each modality, $F_{contextual}$ addresses context across modalities, and $F_{temporal}$ accounts for evolution over time. This is for a dataset containing temporal information in audio or video. For the visio-textual datasets, the $F_{temporal}$ factor would be absent.

From Text to Multimodal Analysis

Evolution of Hate Speech Detection

- Traditional: Text-only analysis
- Modern: Multiple modalities
- Need for joint understanding

Foundation: BERT

- Bidirectional Encoder Representations
- Pre-trained on massive text data
- Contextual understanding

Multimodal Fusion

- Early Fusion:

$$E = f([e_t; e_i; e_a])$$

- Late Fusion:

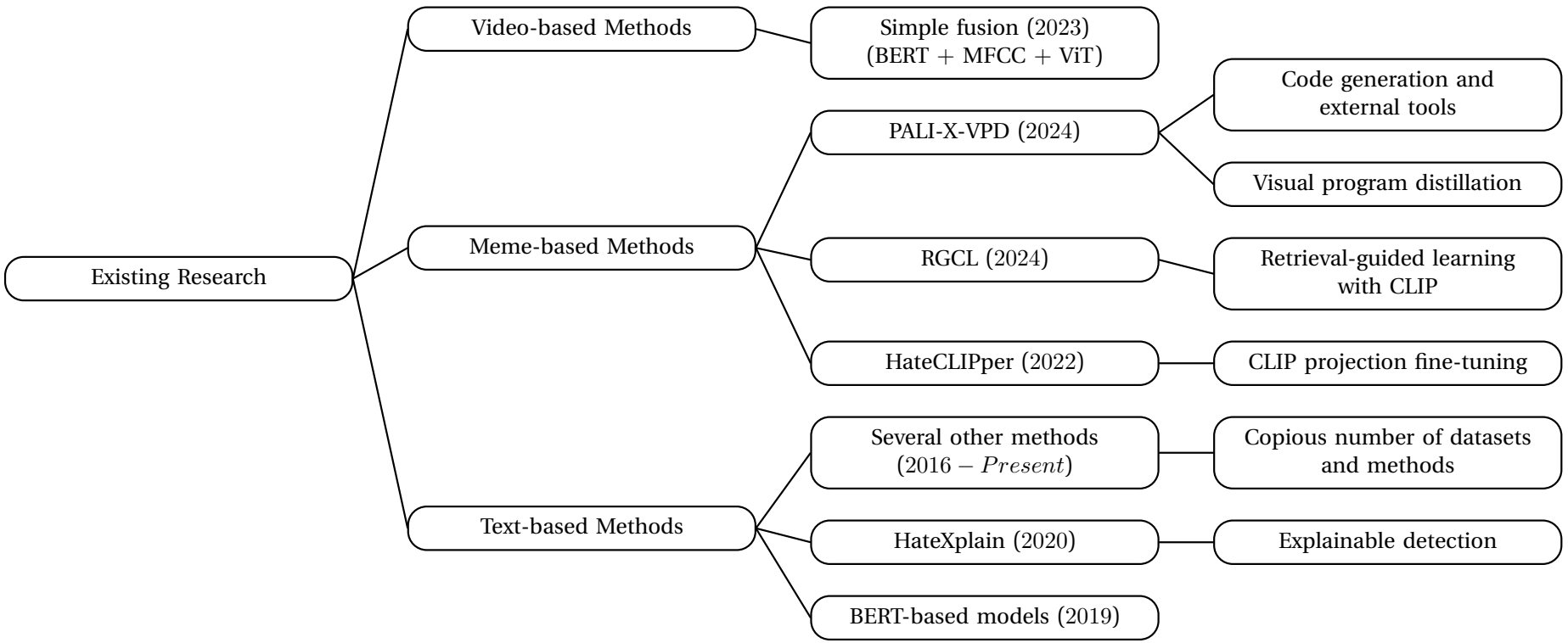
$$L = g(f_t(e_t), f_i(e_i), f_a(e_a))$$

where e_t, e_i, e_a are text, image, and audio embeddings

Key Challenge

How to effectively combine information from multiple modalities while preserving their semantic relationships?

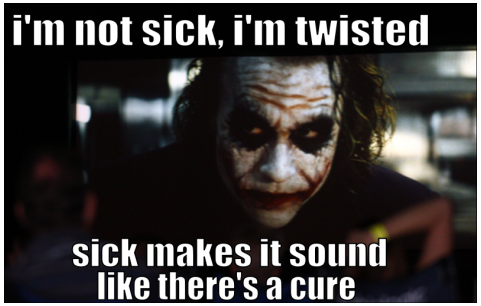
Literature: Text to Multimodal



Datasets Used

Meme Datasets

- Hateful Memes Challenge (HMC): 10k memes



Video Datasets

- HateMM: 1k videos (English only)

Multimodality Challenges

Visio-Textual

- Combined understanding of image and text
- Benign Confounders

Figure: Depiction of benign confounders (absent from the dataset) as presented by Kiela *et al.* (2020): (left) the meme, (middle) the image confounder, (right) the text confounder.

Audio-Visual

- Audio-visual synchronisation (temporal dependencies)
- Misleading features (benign visuals with hateful audio or vice-versa)

Approach (Fusion-based)

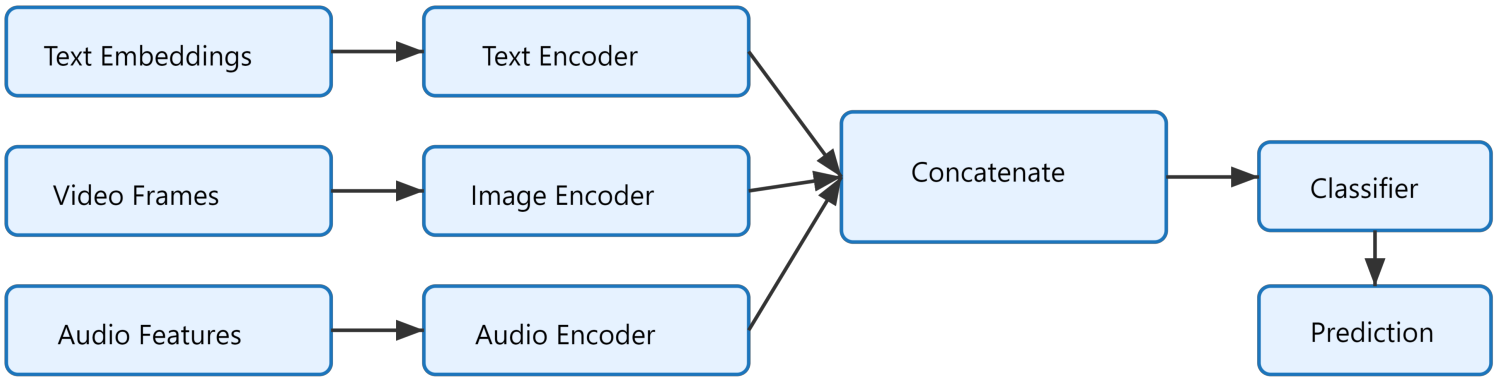
Simple Embedding Fusion

- Text encoders (BERT, HateXplain)
- Audio encoders (Wav2Vec2, CLAP, MFCC, AudioVGG19)
- Vision encoders (ViT, CLIP)

Different Combinations include

- BERT + MFCC + ViT
- CLAP Text + CLAP Audio + CLIP
- HXP + CLAP + CLIP

And so on...



Modality Order-Aware Fusion (MO-Hate)

- BART for text embeddings
- Contextual understanding
- Sequential processing of modalities

Different Combinations include

- BART + CLAP + DINOv2
- BART + Wav2Vec2 + CLIP

And so on...

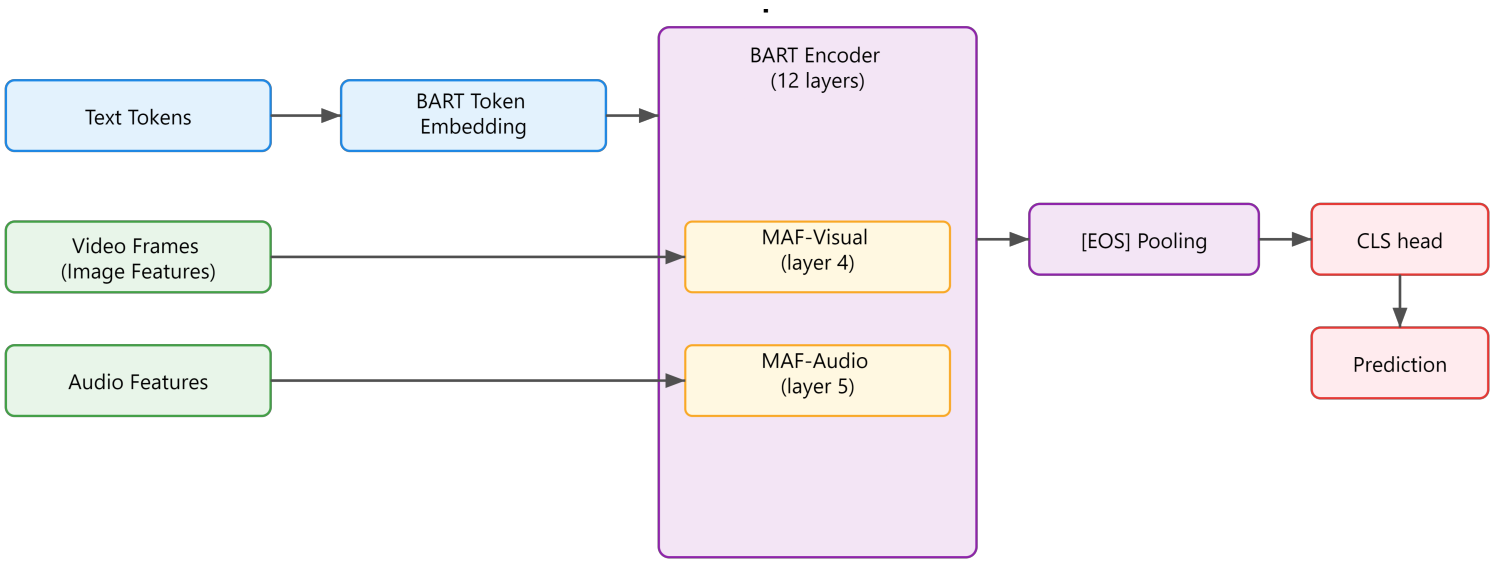


Figure: Modality Order-Aware Fusion architecture adapted from Tomar *et al.* (2023).

Results: Video Content

Performance on HateMM Dataset:

- BA1 (HXP + CLAP + CLIP) achieves best F1 of 0.848 (9.9% points improvement) with concatenation
- MA1 (BART → CLAP → DINOv2) shows strong performance in modality-ordered fusion
- BA1 performs better than MA1 owing to the use of HXP which underscores the importance of having a good text encoder

	Models	F1 (M)	P (M)	R (M)	Acc
Existing	BERT + MFCC + ViT	0.749	0.742	0.758	0.798
	HXP + MFCC + ViT	0.720	0.718	0.726	0.777
	BERT + AVGG19 + ViT	0.718	0.723	0.719	0.755
	HXP + AVGG19 + ViT	0.707	0.714	0.712	0.767
Sim. Fusion	HXP + CLAP + ViT (Concat)	0.823	0.803	0.765	0.832
	CLAP Text + CLAP Audio + CLIP (Concat)	0.802	0.788	0.741	0.811
	HXP + CLAP + CLIP (Concat)	**0.848**	**0.840**	0.800	**0.854**
MO-H	BART → Wav2Vec2 → DINOv2	0.821	0.822	**0.820**	0.820
	Wav2Vec2 → ViT → Text (BART)	0.794	0.794	0.794	0.794
	BART → CLAP → DINOv2	<u>0.821</u>	0.821	0.820	0.820

Ablation Studies

Impact of Individual Modalities:

- Text + Video combination shows strongest performance (F1: 0.819)
- Text alone achieves competitive results (F1: 0.786)
- MO-Hate benefits most from Text + Audio fusion (F1: 0.831)

	T	A	V	F1 (M)	P (M)	R (M)	Acc
Simple Fusion	✓	✗	✗	0.786	0.733	0.776	0.791
	✗	✓	✗	0.730	0.739	0.600	0.747
	✗	✗	✓	0.738	0.698	0.682	0.747
	✓	✓	✗	0.778	0.791	0.670	0.791
	✓	✗	✓	**0.819**	0.795	0.776	0.825
	✗	✓	✓	0.803	0.780	0.753	0.810
MO-Hate	✓	✗	✗	0.785	0.784	0.785	0.791
	✗	✓	✗	0.614	0.631	0.614	0.645
	✗	✗	✓	0.717	0.719	0.716	0.728
	✓	✓	✗	**0.831**	0.829	0.833	0.835
	✓	✗	✓	0.806	0.832	0.798	0.820
	✗	✓	✓	0.802	0.807	0.798	0.810

Ablation Studies

- Model performance is most reliable when all modalities are present and aligned
- Both MA1 and BA1 struggle with misleading visuals, particularly when hateful audio/text is incomplete
- MO-Hate (MA1) handles visual-only cases better than Simple Fusion (BA1), showing better contextual understanding

Video Description	**Modality**	**MO-Hate** → (MA1) **BART + Wav2Vec2 + DINOv2**	**Sim Fusion** → (BA1) **HXP + CLAP + CLIP (Concat)**
Video contains anime and subtitles that do not match the audio. Audio contains hate speech; subtitles and visuals are misleading.	Text + Audio + Video	Pred Label: 1, True Label: 1 Correctly classified the video utilizing audio input.	Pred Label: 1, True Label: 1 Correctly classified video utilizing audio input.
Video contains a cartoon while the audio contains hate speech repeating the word *n*gg**.	Audio + Text	Pred Label: 0, True Label: 1 Unable to correctly classify video due to the partial utterance of slur.	Pred Label: 0, True Label: 1 Incorrectly classified video due to partial utterance of slur.
Video contains hateful symbols displayed throughout the video along with some sound.	Audio + Video	Pred Label: 0, True Label: 1 Without text data, the model can only learn from the video frames.	Pred Label: 0, True Label: 1 Without text data, model can only learn from the video frames.
Video shows violence and physical altercation, but, there is no hate speech.	Text + Audio + Video	Pred Label: 0, True Label: 0 Correctly classified video as non-hateful.	Pred Label: 0, True Label: 0 Correctly classified video as non-hateful.
Video shows picture of a cop restraining a person; no explicit signs of hate speech.	Video	Pred Label: 0, True Label: 0 Correctly classified as non-hateful even though picture looks aggressive.	Pred Label: 1, True Label: 0 Incorrectly classified as hate speech due to aggressive-looking visuals.

Results: Meme Content

Performance on HMC Dataset:

- PALI-X-VPD remains SoTA with AUROC of 0.892
- Our MBD1 (BART → DINOv2) achieves AUROC of 0.628
- Struggling with memes since they require a combined understanding of image and text
- Challenges with benign confounders and the need for sophisticated architectures

	Models	AUROC	F1 (M)	P (M)	R (M)	Acc
VLM 0-shot	LLaVA-1.5 (13B)	0.618	-	-	-	0.614
	InstructBLIP (13B)	0.596	-	-	-	0.601
	Evolver (13B)	0.603	-	-	-	0.604
Existing	**PALI-X-VPD**	**0.892**	-	-	-	-
	RGCL-HateCLIPper	0.867	-	-	-	0.788
	Hate-CLIPper - Align	0.858	-	-	-	-
Sim Fusion	HXP + CLIP (Concat)	0.615	0.557	0.643	0.492	0.617
	CLIP Text + CLIP Image (Concat)	0.606	0.531	0.644	0.451	0.609
	CLIP Text + CLIP Image (EW Product)	0.591	0.467	0.660	0.361	0.596
MO-H	BART → CLIP	0.618	0.608	0.637	0.622	0.622
	BART → DINOv2	<u>0.628</u>	0.619	0.645	0.631	0.631

Ablation Study

- MBD1 struggles with textual confounders, often misclassifying examples.
- Better performance in cases with more explicit hateful content.
- Conservative bias: More false negatives than false positives.

(a) True Label: Not Hateful, Prediction: Not Hateful

(b) True Label: Hateful, Prediction: Not Hateful

(c) True Label: Not Hateful, Prediction: Not Hateful

(d) True Label: Hateful, Prediction: Hateful

Conclusion

Insights from this Work:

- Importance of modality-specific preprocessing for HateMM
- Benign confounders in memes
- Trade-off between performance and efficiency

Key Contributions:

- Extensive literature review on multimodal hate speech datasets and methods
- State-of-the-art performance on HateMM (F1: 0.848)
- Thorough qualitative and quantitative analyses for both datasets using Simple Embedding Fusion and MO-Hate methods

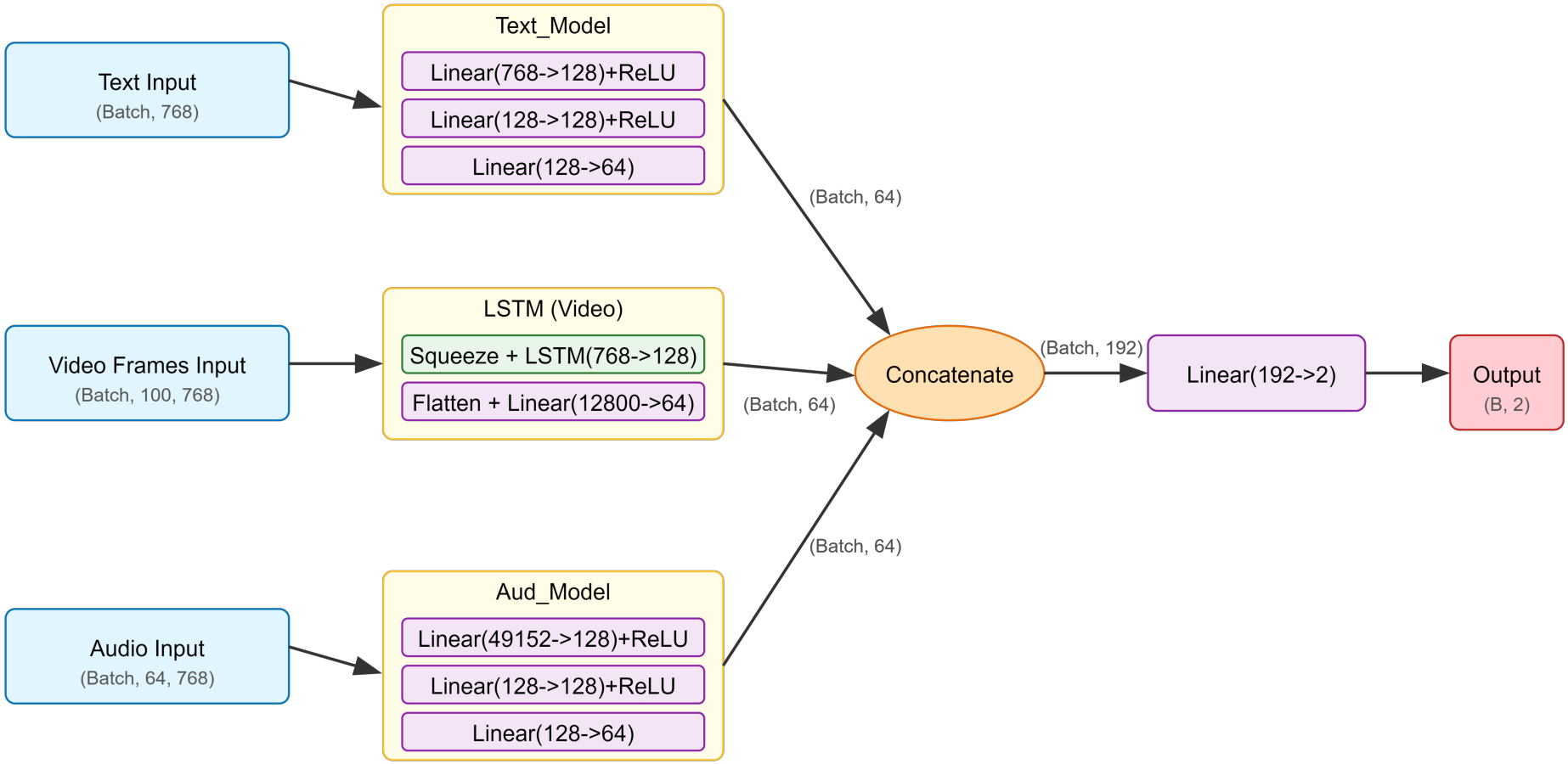
Acknowledgements

Funding Acknowledgement

This work has been supported by the UK EPSRC project AP4L: Adaptive PETs to Protect & emPower People during Life Transitions (Grant number: EP/W032473/1) funded through the UKRI Strategic Priority Fund as part of the wider Protecting Citizens Online programme.

Thank You

Simple Fusion Implementation



Original MO-Sarcasm Architecture