Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

SNIFR : Boosting Fine-Grained Child Harmful Content Detection Through Audio-Visual Alignment with Cascaded Cross-Transformer

Orchid Chetia Phukan[1], *Mohd Mujtaba Akhtar**[1,2], *Girish** [1,3], *Swarup Ranjan Behera*[4], *Abu Osama Siddiqui*[1], *Sarthak Jain*[1], *Priyabrata Mallick*[4], *Jaya Sai Kiran Patibandla*[4], *Pailla Balakrishna Reddy*[5], *Arun Balaji Buduru*[1], *Rajesh Sharma*[6,7]

[1]IIIT-Delhi, India, [2]V.B.S.P.U, India, [3]UPES, India, [4]Independent Researcher, India, [5]Reliance AI, India, [6]University of Tartu, Estonia, [7]Plaksha University, India

Correspondence: orchidp@iiitd.ac.in

Abstract

As video-sharing platforms have grown over the past decade, child viewership has surged, increasing the need for precise detection of harmful content like violence or explicit scenes. Malicious users exploit moderation systems by embedding unsafe content in minimal frames to evade detection. While prior research has focused on visual cues and advanced such fine-grained detection, audio features remain underexplored. In this study, we embed audio cues with visual for fine-grained child harmful content detection and introduce **SNIFR**, a novel framework for effective alignment. **SNIFR** employs a transformer encoder for intra-modality interaction, followed by a cascaded cross-transformer for inter-modality alignment. Our approach achieves superior performance over unimodal and baseline fusion methods, setting a new state-of-the-art.

Index Terms: Child Unsafe Content, Multimodal Learning, Cross-Transformer

1. Introduction

Warning: The following study includes visualizations of sensitive content. Readers are advised to proceed with discretion.

In today's digital landscape, children are increasingly exposed to a vast array of video content, some of which poses significant risks to their mental and emotional well-being. Harmful content - ranging from violent scenes (Figure 1) to explicit material - can have profound effects on young viewers, leading to heightened aggression, anxiety, and confusion. With platforms like YouTube, TikTok, and Netflix becoming a staple of children's daily media consumption, often without supervision, the likelihood of encountering such content has risen dramatically. These challenges underscore the urgent need for advanced detection systems capable of reliably identifying and mitigating exposure to child-harmful content (CHC). As a remedy, researchers have explored various methods for detecting such child-harmful content (CHC) [1, 2, 3, 4, 5, 6, 7]. Kaushal et al. [8] explored video-level features, user-level features, comment-level and CNN-based modeling with visual-features for detecting CHC. Papadamou et al. [9] made use of title, tags, thumbnail, and style features and so on from the video. Similarly, Binh et al. [10] leveraged video titles, thumbnails, and comments with LSTM and transformer networks for CHC detection. Further, Balat et al. [11] used video pre-trained models such as VideoMAE, TimesFormers, VIVIT for CHC detection from TikTok.

Despite significant advancements in content moderation, the need for fine-grained detection of CHC remains critical as malicious users may insert harmful content within videos, sometimes in as little as a few frames, in an effort to evade detection from content moderators. While some progress been made in addressing this challenge by previous researchers, particularly through visual content analysis [12, 13], critical gaps remain. For instance, Singh et al. [12] introduced a large-scale dataset consisting of 109,835 video clips, sliced from full-length videos, aimed at fine-grained detection of child-harmful content. Their approach, utilizing an autoencoder to extract video representations followed by an LSTM classifier, effectively categorized content into safe, violent, sexual, and other harmful categories. Building on this work, Yousaf et al. [13] employed EfficientNet-B7 with a Bi-LSTM architecture to further improve detection accuracy. However, despite these advances, most methods have been primarily confined to visual cues, with limited exploration of audio signals as complementary information that could significantly enhance detection performance.

Audio signals - such as disturbing language, alarming sound effects, or suggestive tones - can provide crucial contextual information that can significantly augment fine-grained child harmful content detection (FGCHCD). This gap in the integration of audio and visual signals represents a major limitation in current detection frameworks. To address this gap, we explore fusing audio cues with visual modality and hypothesize the such integration will lead to improved performance in FGCHCD in comparison to visual-only approaches. To our end, we propose a novel framework, **SNIFR** (Cros**S**-Modality **IN**teract**I**on Cascaded Trans**F**o**R**mer) for effective alignment of audio and visual modalities. **SNIFR** employs a transformer encoder to capture intra-modality interactions where it efficiently models the dependencies within each modality followed by a cascaded cross-transformer that utilizes sequential cross-attention mechanisms for better inter-modality interaction. **SNIFR** achieves superior performance compared to visual-only, audio-only approaches and baseline fusion techniques. We also report improvement over previous state-of-the-art (SOTA) works for FGCHCD.

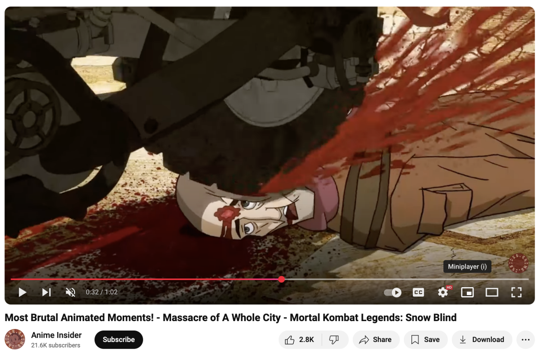


Figure 1: *Violent content in animated media*

* Equal contribution as second authors

10.21437/Interspeech.2025-2251

To summarize, the main contributions of this work are as follows:

- We propose, **SNIFR**, a novel framework that aligns audio and visual modalities for FGCHCD. It ensures effective alignment by first using a transformer encoder to model intra-modality interactions, followed by a cascaded cross-transformer that progressively enhances inter-modality alignment through successive cross-attention mechanisms.
- With **SNIFR**, we report topmost performance compared to unimodal video-only, audio-only as well as baseline fusion methods. We demonstrate, for the first time to the best of our knowledge, that audio serves as a complementary modality for FGCHCD. We also report SOTA in comparison to previous SOTA approaches.

The implementation and trained models for this work are available at: `https://github.com/Helix-IIIT-Delhi/SNIFR-Child_Harmful`

2. Methodology

In this section, we will discuss the feature extraction models followed by baseline models for working with unimodal features, baseline fusion techniques and finally, the proposed framework, **SNIFR**.

2.1. Feature Extraction

Audio Spectrogram Transformer (AST)[1] [14]: It is trained on AudioSet dataset, with weights initialized from ViT, enabling it to effectively capture and encode diverse audio features. AST processes audio by first converting raw waveforms into spectrogram, which are then fed into the transformer to extract meaningful audio representations and shows SOTA behavior in diverse applications.

VideoMAE[2] [15]: It is a video representation learning model and learns to reconstruct masked portions of the input video. VideoMAE is pre-trained on large-scale video datasets, allowing it to produce generalized representations for downstream video tasks and shows SOTA performance in tasks such as video classification.

For feature extraction, we sample video clips at a frame rate of 16 frames per second and resample audio clips to a frequency of 16 kHz. These pre-processed audio and video clips are then passed through AST and VideoMAE, respectively. Each model outputs feature repersentations with a dimension of 768, achieved through average pooling across frames (for video) and time segments (for audio).

2.2. Modeling

2.2.1. Unimodal

Both visual and audio modalities undergo a similar modeling pipeline (Figure 2a). First, feature extraction through VideoMAE or AST and followed by vanilla transformer encoder [16] that models the interaction among the features. The extracted representations are then passed a classifier that contains through a dense layer with 120 neurons and finally the output layer that outputs class probabilities.

2.2.2. Baseline Audio-Visual Fusion Methods

We explore multiple fusion strategies for aligning audio and visual modalities for FGCHCD. In Early Concatenation (EC), the AST and VideoMAE features are passed through transformer encoder then concatenated. Similarly, in Late Concatenation (LC) AST and VideoMAE features are forwarded through transformer encoder and a dense layer of 128 neurons before concatenation. Element-wise Average (EA) fuses the transformer outputs by averaging them element-wise followed by the classifier. In Element-wise Product (EP), the modality-specific transformer outputs undergo element-wise multiplication and finally the classifier. After concatentation in EC, and LC, it is followed by the classifier. Classifier for EC, LC, EA, EP uses the same modeling details as the unimodal modeling above.

2.2.3. SNIFR

We propose a novel framework, **SNIFR**, for effective alignment of audio and visual modalities. The modeling architecture is illustrated in Figure 2 (b). While **SNIFR** shares some architectural similarities with [17], its novelty lies in the cascaded cross-attention mechanisms, which enables more effective inter-modality interaction. The detail architectural flow of **SNIFR** is given as follows: First, audio and visual features are extracted using AST and VideoMAE, respectively. The intra-modality interaction is modeled using a transformer encoder that captures dependencies within each modality. Each input modality undergoes self-attention, followed by a feed-forward network (FFN) and layer normalization within the encoder. This ensures that modality-specific features are effectively learned before engaging in cross-modal interactions. For inter-modality fusion, we design a cascaded cross-transformer, which progressively aligns representations across modalities in two stages. This structured interaction ensures a gradual and effective fusion of features, allowing the model to refine relationships between modalities in a hierarchical manner. The core of the cross-transformer is the cross-attention mechanism where the query ($\mathbf{Q}$) is generated from one modality, while the key ($\mathbf{K}$) and value ($\mathbf{V}$) are generated from the other modality, ensuring effective cross-modal interaction. Given the outputs of the intra-modality transformer encoder, the transformations are defined as follows: $\mathbf{Q}_A = \mathbf{Z}_A\mathbf{W}_Q^A, \quad \mathbf{K}_B = \mathbf{Z}_B\mathbf{W}_K^B, \quad \mathbf{V}_B = \mathbf{Z}_B\mathbf{W}_V^B$ and $\mathbf{Q}_B = \mathbf{Z}_B\mathbf{W}_Q^B, \quad \mathbf{K}_A = \mathbf{Z}_A\mathbf{W}_K^A, \quad \mathbf{V}_A = \mathbf{Z}_A\mathbf{W}_V^A$. $\mathbf{Z}_A, \mathbf{Z}_B$ are the input features for modality A (audio) and modality B (visual) after intra-modality encoding, and $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable weight matrices for each modality. In the first cascade, each modality attends to the other through a cross-attention mechanism, enabling the exchange of complementary features: $\mathbf{Z}_A^{(1)} = \text{LayerNorm}\left(\mathbf{Z}_A + \text{Attention}(\mathbf{Q}_A, \mathbf{K}_B, \mathbf{V}_B)\right)$ and $\mathbf{Z}_B^{(1)} = \text{LayerNorm}\left(\mathbf{Z}_B + \text{Attention}(\mathbf{Q}_B, \mathbf{K}_A, \mathbf{V}_A)\right)$. where the attention scores are computed as:

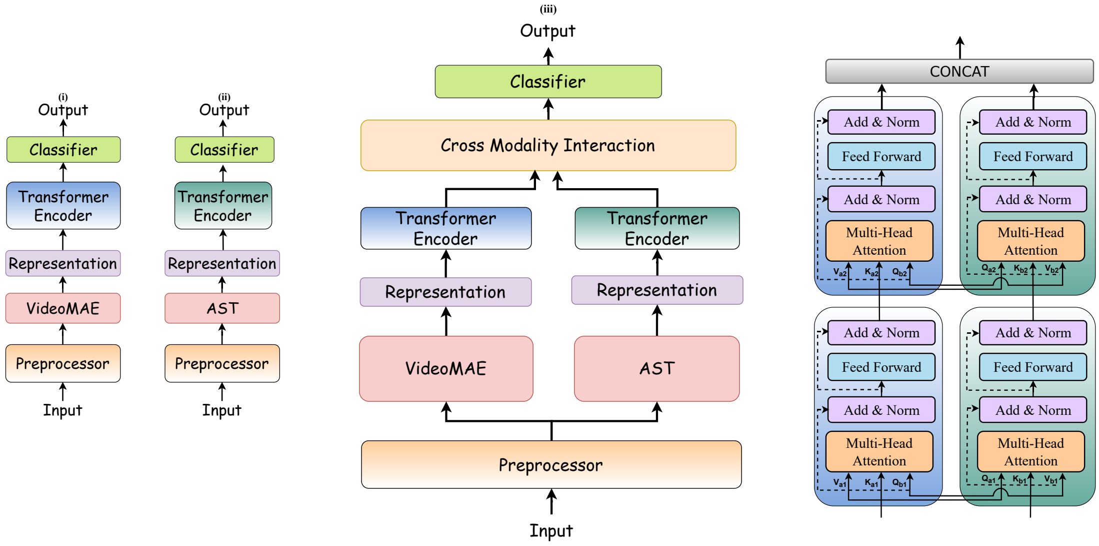
$$\text{Attention}(\mathbf{Q}_A, \mathbf{K}_B, \mathbf{V}_B) = \text{softmax}\left(\frac{\mathbf{Q}_A\mathbf{K}_B^T}{\sqrt{d_k}}\right)\mathbf{V}_B, \tag{1}$$

$$\text{Attention}(\mathbf{Q}_B, \mathbf{K}_A, \mathbf{V}_A) = \text{softmax}\left(\frac{\mathbf{Q}_B\mathbf{K}_A^T}{\sqrt{d_k}}\right)\mathbf{V}_A. \tag{2}$$

where d_k is the dimensionality of the key vectors. After the cross-attention mechanism, the representations pass through a FFN followed by another layer normalization: $\mathbf{Z}_A^{(1)} = \text{LayerNorm}\left(\mathbf{Z}_A^{(1)} + \text{FFN}(\mathbf{Z}_A^{(1)})\right)$ and $\mathbf{Z}_B^{(1)} = \text{LayerNorm}\left(\mathbf{Z}_B^{(1)} + \text{FFN}(\mathbf{Z}_B^{(1)})\right)$. The second cascade further refines this interaction by

[1] `https://huggingface.co/MIT/ast-finetuned-audioset-14-14-0.443`

[2] `https://huggingface.co/MCG-NJU/videomae-base`



(a) (b) (c)

Figure 2: *Modeling Architectures: Subfigure (a) show the individual unimodal modeling pipeline for video and audio, respectively; Subfigure (b) shows the proposed framework,* **SNIFR***; Subfigure (c) provides the detailed illustration of the cross modality interaction through the cascaded cross-transformer*

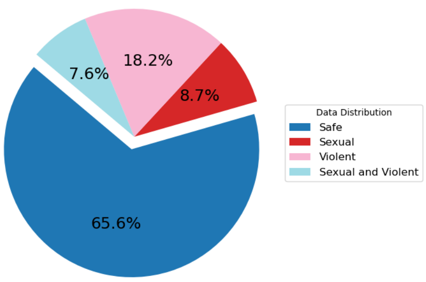


Figure 3: *Percentage Distribution*

reinforcing inter-modal dependencies. The representations from the first cascade are used as input, and another round of cross-attention is applied: $\mathbf{Z}_A^{(2)} = \text{LayerNorm}\left(\mathbf{Z}_A^{(1)} + \text{Attention}(\mathbf{Q}_A^{(1)}, \mathbf{K}_B^{(1)}, \mathbf{V}_B^{(1)})\right)$ and $\mathbf{Z}_B^{(2)} = \text{LayerNorm}\left(\mathbf{Z}_B^{(1)} + \text{Attention}(\mathbf{Q}_B^{(1)}, \mathbf{K}_A^{(1)}, \mathbf{V}_A^{(1)})\right)$.

To further enhance feature refinement, another FFN is applied: $\mathbf{Z}_A^{(2)} = \text{LayerNorm}\left(\mathbf{Z}_A^{(2)} + \text{FFN}(\mathbf{Z}_A^{(2)})\right)$ and $\mathbf{Z}_B^{(2)} = \text{LayerNorm}\left(\mathbf{Z}_B^{(2)} + \text{FFN}(\mathbf{Z}_B^{(2)})\right)$. This iterative refinement ensures that higher-order dependencies between modalities are captured, leading to a deeper integration of complementary features. After two stages of cascaded cross-transformer processing, the final fused representation is obtained by concatenating the refined outputs from both modalities: $\mathbf{Z}_{\text{fused}} = \text{Concat}(\mathbf{Z}_A^{(2)}, \mathbf{Z}_B^{(2)})$. The fused representation is then fed to the classifier with the same modeling details as used in unimodal modeling above. **SNIFR** constitutes of 8.9M trainable parameters.

3. Experiments

3.1. Dataset

We conduct our experiments using the dataset presented by Singh et al. [12], which is specifically curated for FGCHCD. We got hold of the dataset by signing a consent form given by the corresponding author. To the best of our knowledge, this is the only accessible dataset for FGCHCD with both audio and visual information present. This dataset comprises 107907 one-second video clips, divided into four categories: 70741 safe clips, 9335 containing sexual content, 19658 containing violent content and 8173 containing both sexual and violent content. The percentage distribution plot is given in Figure 3. We use FFmpeg to extract corresponding audio from each video segment.

3.2. Training Details

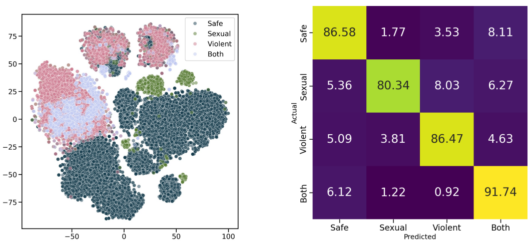
We use cross-entropy loss function, AdamW as the optimizer and setting the learning rate to 1e-4. We train the models for 25 epochs with batch size of 16. We apply a weight decay of 1e-5, dropout, and early stopping to prevent overfitting. We use 5-fold cross-validation for training and evaluation of our models where four folds are used for training and one fold for evaluation.

3.3. Experimental Results

Table 1 presents the class-wise evaluation results for models trained using single-modality and multimodal approaches for FGCHCD. Among unimodal models, visual-based models consistently outperform audio-based models, highlighting the dominance of visual cues in content classification. However, while weaker in isolation, audio models capture critical cues such as alarming sound effects, background music, and suggestive dialogue, which contribute significantly to classification. Despite this, unimodal audio models struggle with subtle harmful content when visual cues are dominant, reinforcing the necessity of multimodal fusion. However, we hypothesize that integrating audio with visual

Modality	Safe			Sexual			Violent			Both		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
V	85.45	89.48	90.55	90.70	73.68	95.37	66.18	56.96	87.63	64.71	64.08	91.49
A	71.27	80.29	77.92	83.33	58.82	82.87	46.48	35.11	74.81	50.00	27.91	78.58
AV (EC)	87.19	87.85	88.93	72.73	71.91	96.58	58.88	61.17	87.01	48.48	40.51	89.10
AV (LC)	82.29	84.40	85.62	85.71	75.00	91.06	58.25	56.34	83.85	60.00	57.14	92.95
AV (EA)	79.91	84.99	86.67	81.48	60.27	88.13	53.42	46.43	85.38	53.85	46.67	89.09
AV (EP)	78.33	81.23	83.60	70.59	58.54	95.28	45.45	44.55	81.15	51.66	45.36	88.29
AV (CT)	84.65	87.92	90.62	81.25	71.23	96.83	66.02	65.38	89.83	79.07	64.76	95.34
AV (SNIFR)	**88.24**	**91.49**	**95.28**	**93.33**	**82.11**	**98.72**	**84.15**	**77.09**	**96.19**	**79.59**	**75.73**	**97.82**
SOTA [12]	-	-	88.00	-	-	95.00	-	-	90.00	-	-	91.00

Table 1: *Evaluation results showing Accuracy (ACC) in %, Macro-average F1-scores (F1), and AUC-scores (AUC) in % across different classes; 'Both' represents video clips containing both violent and sexual content; V: Visual, A: Audio, AV: Audio-Visual; Scores are presented in average of five-folds; AV (EC), AV (LC), AV (EA), AV (EP), AV (CT), AV (SNIFR) presents the results for audio-visual modeling with Early Concatenation, Late Concatenation, Element-wise Average, Element-wise Product, Cross-Transformer, and proposed novel framework,* **SNIFR**



(a) (b)

Figure 4: *The subfigures (a) and (b) represents the t-SNE plot visualizatoin and confusion matrix of* **SNIFR** *respectively*

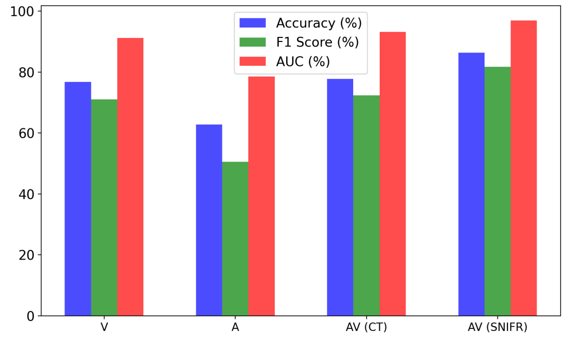


Figure 5: *Total Accuracy, F1 Score, AUC; V: Visual, A: Audio, AV (CT): Cross-Transformer, AV (SNIFR): Proposed framework,* **SNIFR**

modalities will improve FGCHCD. Our experiments with the proposed novel framework, **SNIFR** validates our hypothesis in terms of highest accuracy, F1-score, and AUC across all the classes. We also present the total accruacy, F1-score, and AUC in Figure 5 that is obtained through average of the class-wise scores. We observe the top performance of **SNIFR** with audio-visual alignment. To benchmark **SNIFR** effectiveness, we also compare it against various baseline fusion techniques, including EC, LC, EA, and EP. The results reveal that these baseline approaches fail to effectively integrate audio-visual cues, often underperforming even compared to unimodal visual models. In contrast, **SNIFR** shows top performance as the structured integration of modalities in a cascaded manner through the cascaded cross-transformer blocks allows **SNIFR** to dynamically capture fine-grained dependencies between audio and visual cues, leading to substantial improvements. We present an ablation study of **SNIFR**, if instead of cascaded cross-transformers (AV (CT) in Table 1), we only use a single level of cross-transformer and that make use of cross-attention mechanism that has demonstrated effectiveness for multimodal fusion [18, 19]. We see that it is not able to keep up the results attained by cascaded cross-transformers. We also plot the t-SNE plot visualization of the features from the penultimate layer of **SNIFR** in Figure 4 (a) followed by the confusion matrix of **SNIFR** Figure 4 (b).

3.4. Comparison to SOTA

In this section, we present the comparison of the proposed framework, **SNIFR** with previous SOTA work. As FGCHCD is a underexplored task, so we are not able to find any other studies that worked on the same dataset, to the best of our knowledge, as used in our experiments [12]. So, we compare our results with Singh et al. [12] as the SOTA study where they used only-visual modality for their experiments. We observe sufficient gain in the performance across all the scores with **SNIFR** (See Table 1). These results mirrors our hypothesis that audio cues will act as complementary modality to visual modality for FGCHCD and also showed the superior cross-modality alignment capability of **SNIFR**. We acknowledge that few works have worked on similar topics, however, they haven't made their dataset public [13] and haven't labelled their dataset exactly for FGCHCD [10].

4. Conclusion

In this study, we explore for the first time, aligning audio and visual modalities for FGCHCD and propose **SNIFR**, a novel framework for effective alignment. By leveraging a transformer encoder for intra-modality interactions and a cascaded cross-transformer for inter-modality alignment, **SNIFR** effectively captures dependencies between audio and visual cues, outperforming unimodal and baseline fusion models. This work highlights the crucial role of audio in enhancing visual-only models, setting new SOTA performance for FGCHCD and demonstrating the potential of multimodal integration in content moderation.

5. References

[1] S. Alshamrani, A. Abusnaina, and D. Mohaisen, "Hiding in plain sight: A measurement and analysis of kids' exposure to malicious urls on youtube," in *2020 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 2020, pp. 321–326.

[2] N. Aldahoul, H. A. Karim, M. H. L. Abdullah, A. S. B. Wazir,

M. F. A. Fauzi, M. J. T. Tan, S. Mansor, and H. S. Lyn, "An evaluation of traditional and cnn-based feature descriptors for cartoon pornography detection," *IEEE Access*, vol. 9, pp. 39 910–39 925, 2021.

[3] M. Y. Chuttur and A. Nazurally, "A multi-modal approach to detect inappropriate cartoon video contents using deep learning networks," *Multimedia Tools and Applications*, vol. 81, no. 12, pp. 16 881–16 900, 2022.

[4] M. Gkolemi, P. Papadopoulos, E. Markatos, and N. Kourtellis, "Youtubers not madeforkids: Detecting channels sharing inappropriate videos targeting children," in *Proceedings of the 14th ACM Web Science Conference 2022*, 2022, pp. 370–381.

[5] K. Ramesh, A. R. KhudaBukhsh, and S. Kumar, "'beach'to 'bitch': Inadvertent unsafe transcription of kids' content on youtube," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 108–12 118.

[6] S. Aggarwal and D. K. Vishwakarma, "Protecting our children from the dark corners of youtube: A cutting-edge analysis," in *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*. IEEE, 2023, pp. 1–5.

[7] S. I. Alqahtani, W. M. Yafooz, A. Alsaeedi, L. Syed, and R. Alluhaibi, "Children's safety on youtube: A systematic review," *Applied Sciences*, vol. 13, no. 6, p. 4044, 2023.

[8] R. Kaushal, S. Saha, P. Bajaj, and P. Kumaraguru, "Kidstube: Detection, characterization and analysis of child unsafe content promoters on youtube," in *2016 14th Annual Conference on Privacy, Security and Trust (PST)*, 2016, pp. 157–164.

[9] K. Papadamou, A. Papasavva, S. Zannettou, J. Blackburn, N. Kourtellis, I. Leontiadis, G. Stringhini, and M. Sirivianos, "Disturbed youtube for kids: Characterizing and detecting inappropriate videos targeting young children," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 522–533.

[10] L. Binh, R. Tandon, C. Oinar, J. Liu, U. Durairaj, J. Guo, S. Zahabizadeh, S. Ilango, J. Tang, F. Morstatter *et al.*, "Samba: Identifying inappropriate videos for young children on youtube," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 88–97.

[11] M. Balat, M. Gabr, H. Bakr, and A. B. Zaky, "Tikguard: A deep learning transformer-based solution for detecting unsuitable tiktok content for kids," in *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*. IEEE, 2024, pp. 337–340.

[12] S. Singh, R. Kaushal, A. B. Buduru, and P. Kumaraguru, "Kidsguard: fine grained approach for child unsafe video representation and detection," in *Proceedings of the 34th ACM/SIGAPP symposium on applied computing*, 2019, pp. 2104–2111.

[13] K. Yousaf and T. Nawaz, "A deep learning-based approach for inappropriate content detection and classification of youtube videos," *IEEE Access*, vol. 10, pp. 16 283–16 298, 2022.

[14] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[15] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10 078–10 093, 2022.

[16] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[17] Y. Gong, A. H. Liu, A. Rouditchenko, and J. Glass, "Uavm: Towards unifying audio and visual models," *IEEE Signal Processing Letters*, vol. 29, pp. 2437–2441, 2022.

[18] J. Lin, X. Cai, H. Dinkel, J. Chen, Z. Yan, Y. Wang, J. Zhang, Z. Wu, Y. Wang, and H. Meng, "Av-sepformer: Cross-attention sepformer for audio-visual target speaker extraction," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[19] Z. Xu, X. Fan, and M. Hasegawa-Johnson, "Dual-path cross-modal attention for better audio-visual speech extraction," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.