Augmenting Multimodal Deep Learning with Attention Mechanisms to Recognize 'Sludge' Videos from Short-Form Content

Marc Olata
BSCSSE Student
FEU Institute of Technology
mmolata@fit.edu.ph

Alpha Romer Coma
BSCSSE Student
FEU Institute of Technology
ancoma@fit.edu.ph

Job Isaac Ong
BSCSSE Student
FEU Institute of Technology
jmong@fit.edu.ph

Kristoffer Ian Sioson
BSCSSE Student
FEU Institute of Technology
ktsioson@fit.edu.ph

Justine Jude Pura
Project Mentor
FEU Institute of Technology
jcpura@feutech.edu.ph

Shaneth Ambat
Course Adviser
FEU Institute of Technology
scambat@feutech.edu.ph

ABSTRACT

"Sludge" content refers to short-form videos that display multiple unrelated streams on a single screen, designed to exploit algorithms and sustain engagement through sensory overstimulation. This format presents significant challenges for content moderation, as it often bypasses conventional detection systems while raising concerns such as reduced attention spans and copyright infringement. This study introduces Visual-Qwen, a multimodal deep learning architecture with attention mechanisms for automated sludge content detection. The model combines EVA-CLIP-G/14 for visual encoding, Whisper V3 Turbo for audio transcription, a Querying Transformer with cross-modal attention for feature fusion, and the Qwen3-4B language model for classification and explanation generation. A two-stage training strategy using transfer learning and Low-Rank Adaptation (LoRA) was employed to maintain computational efficiency. The system was trained and evaluated on a curated dataset of 2,000 TikTok and YouTube Shorts videos. It achieved strong performance on the held-out test set, with 96.67% accuracy, 95.58% precision, 98.86% recall and a 97.19% F1 score. Expert evaluations under ISO/IEC TR 24028 standards rated the model highly for functionality, performance, and explainability, while usability testing indicated strong acceptance among content creators and moderators.

CCS CONCEPTS

• **General and reference** → **Design**; **Performance**; Computing standards, RFCs and guidelines; • **Computer systems organization** → **Special purpose systems**; • **Software and its engineering** → *Model checking*; • **Computing methodologies** → **Neural networks**; **Activity recognition and understanding**; **Speech recognition**; • **Human-centered computing** → Social media; User models.

KEYWORDS

Multimodal deep learning, sludge content, short-form videos, content moderation, video classification, EVA-CLIP-G/14, Q-Former, Qwen3-4B

1 INTRODUCTION AND BACKGROUND

[14]Short-form content has revolutionized digital media consumption, with 72% of social media users preferring to watch the said format for entertainment, trends, products, and services over traditional-length videos. [12] This led content creators to find innovative ways to maximize engagement on short-form content. One trend in particular saw a way to increase user viewership: "sludge" content. [30]This relatively new phenomenon involves the simultaneous playback of two unrelated video clips on a single screen. For example, one clip could be a scene from a cartoon or a person talking on a podcast, and the other could be a satisfying soap-cutting clip - anything that could catch a user's attention (Mattson, 2024). [19]The "sludge" content format is designed to make the user watch a certain short-form video longer by providing more attention points, as users could just switch their focus between the two unrelated videos (González, 2023). [59]The increase in watch time, even if only a few seconds, increases the chances of it being recommended to other users by the platforms' recommendation algorithms, which are heavily based on engagement metrics (Zhang et al., 2023).

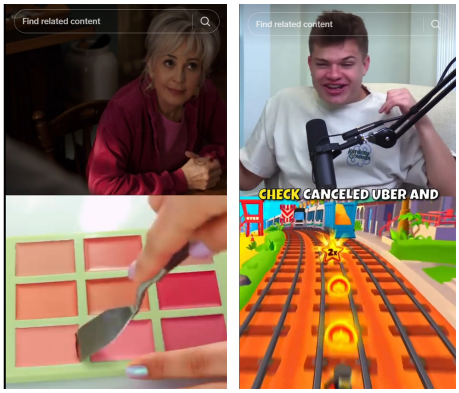

Figure 1: Examples of "sludge content" taken from the social media platform TikTok

[30]However, this type of content comes with many drawbacks. One issue centers on the inherent media multitasking encouraged by sludge content. By delivering viewers various, unrelated

streams of visual and auditory input, it compels an attention split. [9]This type of media engagement is directly correlated with studies showing the adverse effects of multitasking on cognitive abilities (Baumgartner, 2017). Interacting with multiple media sources simultaneously has been associated with lower memory capacity and a reduced capability to switch between tasks effectively. [29]Additionally, media multitasking has been linked to higher levels of stress and anxiety, indicating that the continuous influx of varied stimuli can overstretch the brain's processing capabilities (Li, 2022).

In addition to the cognitive effects, sludge content represents a considerable danger to the content moderation procedures on social media platforms. Creators have found ways to manipulate this format to bypass platform regulations and standards. By overlaying seemingly harmless videos, such as ASMR content or gaming streams, with material that violates platform rules (like hate speech, misinformation, or explicit content), creators aim to obscure the violating material by using the innocuous content as a distraction, complicating the task for both automated detection algorithms and human reviewers to swiftly locate and eliminate the underlying harmful material. As a result, this strategy enables restricted content to remain on platforms for longer durations, potentially reaching and influencing a broader audience.

Further compounding these concerns are the ethical issues surrounding copyright infringement and subsequent monetization. [48]Sludge content frequently incorporates copyrighted material, such as clips from movies, television shows, music videos, and video games, without obtaining the required licenses or crediting the copyright holders (Tran, 2023). This practice constitutes a clear violation of intellectual property rights. Moreover, creators often monetize this content, generating revenue through platform-based advertising or other means, thereby directly profiting from the unauthorized use of copyrighted works.

The very structure that makes sludge content engaging also allows it to be exploited to circumvent platform policies. Creators have discovered that layering seemingly innocuous visuals (e.g., ASMR content, game recordings) over content violating platform guidelines (hate speech, misinformation, explicit material) [24]can obscure the objectionable material from both automated detection systems and human moderators (Hua, 2023). This "sludge" format of having two running clips at the same time serves as visual concealment to [57]circumvent content moderation flagging built on social media platforms, increasing the difficulty of rapidly identifying and addressing policy violations, thus allowing prohibited material to persist and potentially reach a wider audience (Zannettou et al., 2018).

The limitations of unimodal analysis, focusing on single elements like text or individual video frames, have spurred a crucial shift toward multimodal deep learning in content moderation research.[37] This approach involves training models on multiple data modalities concurrently – integrating video and textual analysis (Poria et al., 2020).[40] By synthesizing information from diverse sources, multimodal models achieve a more comprehensive understanding, enabling the detection of subtle cues and contextual relationships that single-modality systems might miss (Ramachandram & Taylor, 2017). Within multimodal deep learning, attention mechanisms have emerged as an influential technique.[11] These mechanisms allow models to focus on the most relevant components of the input data instead of treating all input elements equally (Chaudhari et al., 2021). In a multimodal setting, this ability enables the model to [16]prioritize information from various modalities dynamically, depending on its contextual importance for a specific task (Gao et al., 2020).

[43]This selective attention enables the model to place greater emphasis on the semantic relationship between text and video when evaluating potential policy infringements or to focus on specific time segments within a video where inconsistencies between the modalities are most evident (Shi, 2022). The capability to flexibly direct attention across and within different modalities presents valuable and sophisticated content moderation systems. By thoughtfully incorporating attention mechanisms, approaches to tackle the challenges introduced by misleading content formats like sludge, which could result in more effective multimodal deep learning models, exhibit the potential for enhanced identification and flagging of sludge content.

The primary objective of this study is to develop and evaluate a novel multimodal deep learning model, augmented with attention mechanisms, specifically designed for the automated detection of sludge content within short-form videos.

2 RELATED WORK

2.1 Sludge Content and Content Moderation Challenges

Automated moderation systems typically analyze video frames through computer-vision classifiers or process audio transcripts via speech-to-text models using predefined violation patterns. However, sludge formats undermine these unimodal pipelines by fragmenting and masking policy-violating material across separate panels, reducing detection accuracy [17].

Since most moderation algorithms operate in purely visual or audio domains, they fail to capture cross-modal inconsistencies inherent in sludge videos, which is a concern as disallowed content can evade automated moderation systems when made into 'sludge' format.

2.2 Multimodal Deep Learning Approaches

2.2.1 Theoretical Foundations. Baltrušaitis et al.[7] established five core challenges in multimodal machine learning: representation, translation, alignment, fusion, and co-learning. These principles provide the theoretical foundation for designing systems that integrate diverse data types. Recent empirical studies have validated these theoretical frameworks, with hybrid fusion models demonstrating 19% increases in classification accuracy compared to single-modality models on datasets including CMU-MOSEI and AVEC [56]

2.2.2 Vision-Language Pre-training Components. The development of robust vision-language models has been crucial for multimodal understanding. Radford et al. [38] introduced CLIP (Contrastive Language-Image Pre-training), which connects visual data with textual descriptions through contrastive learning. Building on this foundation, [28] Li et al. developed BLIP-2, introducing the Querying Transformer (Q-Former) as a lightweight bridge between frozen

vision encoders and language models. The Q-Former learns to extract concise and semantically meaningful visual features using learnable query embeddings, achieving an 8.7 percentage point improvement over the 80-billion-parameter Flamingo model on zero-shot VQAv2 despite using 54× fewer trainable parameters. Recent extensions have further validated the Q-Former's effectiveness. Azad et al. [6] introduced HierarQ, a hierarchical framework for sequential frame processing that overcomes context-length limitations. Chatterjee et al. [13] demonstrated the effectiveness of ProVideLLM in real-time procedural video understanding, achieving state-of-the-art performance in online step detection and forecasting. For video-specific applications, several studies have adapted image-based architectures. [54]Xing et al. (2023) showed that CLIP's visual embeddings achieve competitive performance in few-shot video action recognition. [5]Arnab et al. (2021) established with ViViT (Vision Transformer for Video) that transformer architectures effectively capture spatio-temporal dependencies, while [47]Tong et al. demonstrated that masked autoencoders pre-trained on video datasets generate powerful spatio-temporal representations.

2.2.3 Audio Processing and Integration. Radford et al. [39] introduced Whisper, showcasing remarkable robustness in speech recognition across diverse languages and noisy conditions. [51]Wang et al. developed multimodal approaches for video anomaly detection using deep audio features alongside visual features, demonstrating significant performance improvements in complex video environments.

2.3 Attention Mechanisms for Multimodal Learning

[21]Gorti et al. introduced cross-modal attention networks that dynamically align visual and textual features, achieving 12% improvement in multimodal alignment tasks on the MSR-VTT dataset compared to non-attention-based models. These mechanisms are particularly relevant for identifying misaligned content where visual and audio/textual elements may not correspond.

2.4 Modality Selection and Computational Efficiency

The selection of input modalities significantly impacts both performance and computational cost. [45]Sun et al. (2019) demonstrated with VideoBERT that combining ASR-derived word tokens with visual features yields state-of-the-art action classification. [23]Hessel et al. (2019) showed that multimodal models using ASR word embeddings plus visual frame features outperformed visual-only baselines by over 10% in ROUGE-L and 15% in CIDEr. [31]Miech et al. (2019) released HowTo100M, finding that pretraining on aligned text-video data produces up to 10% accuracy gains over visual-only pretraining. Critically, ASR transcripts generate only tens of tokens per second versus hundreds of spectrogram patches, reducing transformer self-attention operations by roughly 100× [2](Akbari et al., 2021). While tri-modal approaches adding raw mel-spectrogram patches can gain further improvements, they incur 1.5× computational overhead for marginal gains[2].

Table 1: Performance vs. Computational Overhead for Different Modality Combinations Summary

Modalities	Relative Gain in F1-score (%)	Compute Overhead (%)
Visual frames only	0	100
Visual frames + ASR text transcription	10	105
Frames + raw audio spectrogram	5	150
Frames + ASR text + audio spectrogram	12	155

2.5 Synthesis

Advances in multimodal machine learning show that integrating visual, textual, and audio modalities consistently outperforms unimodal approaches. Vision-language pretraining models such as CLIP and BLIP-2 have provided effective mechanisms for aligning visual and textual inputs, while recent extensions have expanded these methods to sequential and real-time contexts. Work on video-specific transformers and robust speech transcription models like Whisper has further enriched the multimodal toolkit. At the same time, cross-modal attention mechanisms and modality selection strategies demonstrate that combining visual features with ASR-derived transcripts offers the best balance of accuracy and efficiency, without the heavy computational cost of tri-modal approaches.

Despite these advances, research has largely concentrated on tasks such as captioning, sentiment analysis, and action recognition, leaving unexplored the challenge of detecting intentional audio-visual misalignments. This study addresses that gap by adapting proven components such as EVA-CLIP-G/14 for visual encoding, Whisper for transcription, and Q-Former with cross-modal attention for fusion into an architecture tailored for sludge content detection. In doing so, it extends multimodal learning into a novel application where deliberate inconsistencies between modalities are the defining signal.

3 METHODOLOGY

3.1 Model Architecture

The proposed Visual-Qwen architecture comprises four modular components processing visual and audio inputs into language-interpretable features.

3.1.1 EVA-CLIP-G/14 Vision Encoder. Video frames are preprocessed to 224×224 pixels and encoded using frozen EVA-CLIP-G/14 [46], producing 257×1408-dimensional embeddings per frame. The frozen encoder preserves robust pre-trained visual features while avoiding overfitting during multimodal training.

3.1.2 Q-Former and Linear Projection. A lightweight Q-Former with 32 learnable query embeddings attends to EVA-CLIP-G/14 features through cross-attention layers, condensing visual information into 32×768-dimensional representations. A linear projection layer maps these to 2560 dimensions, matching the Qwen3-4B input space.

3.1.3 Audio Processing. Audio streams are separated using MoviePy and transcribed via Whisper V3 Turbo, producing timestamped text segments converted to tokens, preserving temporal alignment with visual content.

3.1.4 Language Model Integration. Qwen3-4B processes projected visual tokens, transcription tokens, and instruction tokens through

transformer layers enhanced by attention mechanisms, generating sludge/non-sludge classifications with natural language explanations.

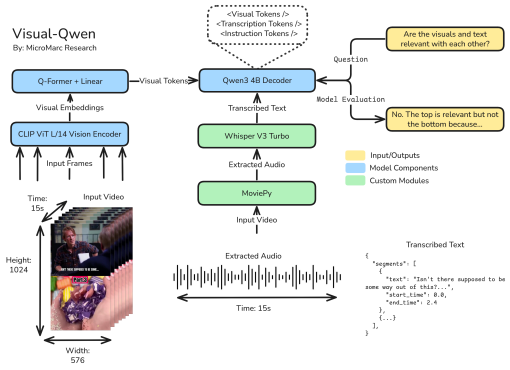

Figure 2: Visual-Qwen model architecture.

3.2 Training Strategy

Training proceeds through two stages, leveraging transfer learning principles:

Stage 1 - Pre-training: EVA-CLIP-G/14, Q-Former, and Qwen remain frozen while training only the linear projection layer on the LLAVA pre-training dataset containing image-caption pairs. This preserves broad visual-language capabilities while learning multimodal alignment.

Stage 2 - Fine-tuning: The projection layer freezes, and Low-Rank Adaptation (LoRA) modules are injected into Qwen layers, fine-tuning only these adapters on the custom sludge dataset. This approach confines learning to minimal parameters while specializing for sludge detection.

3.3 Dataset Construction

A balanced dataset[1] of 2,000 videos was assembled through ethical scraping from public TikTok and YouTube Shorts feeds, searching sludge-related hashtags. The collection process involved:

(1) Automated scraping using official platform APIs
(2) Manual screening by trained reviewers
(3) Synthetic feature generation using Gemini 2.5 Flash
(4) Human verification of generated features
(5) Dataset validation by external experts

The final dataset contains 1,000 sludge and 1,000 non-sludge videos, split into 70% training, 15% validation, and 15% test sets with stratified sampling ensuring balanced representation.

3.4 Inference Pipeline

During inference, the system processes uploaded videos through a multi-step pipeline. The video frames and audio are extracted using MoviePy. The video is sampled at 1 frame per second (fps) across the video's duration, then resized to 224×224 pixels for EVA-CLIP-G/14 encoding. Simultaneously, the audio stream is transcribed via Whisper V3 Turbo to generate timestamped text segments.

[1]Dataset in Kaggle: https://doi.org/10.34740/kaggle/dsv/12104583

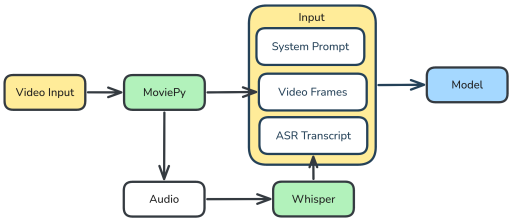

Figure 3: Model inference pipeline.

The system combines the projected visual tokens with the ASR transcript and task-specific instruction prompts (e.g., "Are the visuals and text relevant to each other?"). Qwen3-4B processes this multimodal input sequence to assess the relationships between visual and textual content. The model outputs both a binary sludge/non-sludge classification and a natural language explanation detailing the reasoning behind its decision.

4 RESULTS AND DISCUSSION

4.1 Model Performance

The Visual-Qwen model achieved strong performance on the held-out test set of 300 videos:

- Accuracy: 96.67% (95% CI 94.33–98.67)
- Precision: 95.58% (95% CI 92.61–98.31)
- Recall: 98.86% (95% CI 97.06–100.00)
- F1-Score: 97.19%

The confusion matrix (Figure 4) demonstrates effective classification with minimal false negatives, crucial for content moderation applications.

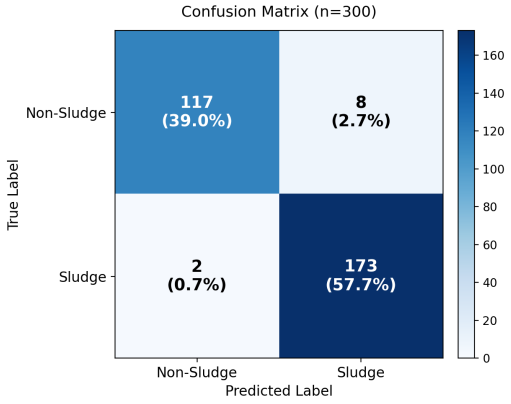

Figure 4: Confusion matrix for sludge content classification

4.2 Training Efficiency

Training in Google Cloud TPU v4-64 demonstrated computational efficiency:

- Stage 1: 177 minutes for 4 epochs on image-caption pairs
- Stage 2: 9.6 minutes for 6 epochs on sludge dataset
- Total training time: 3 hours

This efficiency comes from freezing large components of the backbone while training only lightweight adapters, making the approach practical for real-world deployment.

4.3 User Evaluation

Three stakeholder groups evaluated the system: - 20 content creators rated perceived usefulness (4.73/5), ease of use (4.72/5), and behavioral intention (4.55/5) - 20 content moderators provided similar positive ratings across Technology Acceptance Model dimensions - 10 machine learning experts assessed the system against ISO/IEC TR 24028 standards, rating functionality (4.57/5), performance (4.50/5), and transparency (4.40/5) highly

All survey sections exceeded Cronbach's alpha threshold of 0.70, confirming reliable measurement instruments.

4.4 Multimodal Advantage

The combination of visual embeddings and ASR-transcribed text delivered optimal accuracy-to-compute trade-offs. Related studies consistently show 10% F1-score improvements when pairing visual features with transcribed text, while adding only 5% computational overhead compared to vision-only approaches [23, 31].

This efficiency advantage stems from transcripts generating tens of tokens per second versus hundreds of visual patches, significantly reducing transformer self-attention operations while capturing complementary semantic information.

5 CONCLUSION AND FUTURE WORK

This study successfully developed a multimodal deep learning architecture augmented with attention mechanisms for automated sludge content detection in short-form videos. The Visual-Qwen model achieved 96.67% accuracy by effectively leveraging visual and textual modalities through vision-language training.

Key contributions include:

(1) Novel application of Q-Former architecture for video content moderation
(2) Balanced dataset of 2,000 annotated sludge/non-sludge videos
(3) Comprehensive evaluation following ISO/IEC TR 24028 standards

Future work should explore richer temporal modeling, knowledge distillation for edge deployment, and active learning feedback loops. Pilot integrations with live platforms will address practical deployment challenges and threshold calibration needs.

The approach demonstrates multimodal deep learning's potential for content moderation, providing foundations for safer online environments while maintaining computational efficiency suitable for large-scale deployment.

6 ACKNOWLEDGMENTS

The researchers wish to express their sincere gratitude to their thesis mentor, Mr. Justine Jude Pura, for his unwavering dedication, insightful guidance, and patient support throughout every stage of this study. His expertise and encouragement were instrumental in shaping the research direction and in overcoming technical and academic challenges.

The researchers also acknowledge Google Research's TPU Research Cloud (TRC) for providing the compute resources that made this thesis possible; access to the TRC program exponentially accelerated experimentation and model development at no cost.

The researchers would also like to thank the entire faculty of the Computer Science Department at FEU Institute of Technology for their collective advice and encouragement throughout their undergraduate journey. Finally, a special thanks is extended to the domain experts who generously shared their time and knowledge to evaluate this thesis.

7 REFERENCES

[1] Abdu, S. A., Yousef, A. H., and Salem, A. 2021. Multimodal video sentiment analysis using deep learning approaches, a survey. *Information Fusion* 76 (2021), 204–226. https://doi.org/10.1016/j.inffus.2021.06.003
[2] Akbari, H., Yuan, L., Qian, R., Chuang, W., Chang, S., Cui, Y., and Gong, B. 2021. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. *ArXiv* (2021). https://arxiv.org/abs/2104.11178
[3] Alayrac, J. B., Recasens, A., Bottenberg, E., Rueda, A., De Fauw, J., Smaira, L., and Carreira, J. 2022. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems* 35 (2022), 19276–19290.
[4] AlDahoul, N., Tan, M. J. T., Kasireddy, H. R., and Zaki, Y. 2024. Advancing content moderation: Evaluating large language models for detecting sensitive content across text, images, and videos. *arXiv preprint* (Nov. 2024). https://arxiv.org/abs/2411.17123
[5] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. 2021. ViViT: a video vision transformer. *arXiv preprint* (Mar. 2021). https://arxiv.org/abs/2103.15691
[6] Azad, S., Vineet, V., and Rawat, Y. S. 2025. HierarQ: Task-Aware Hierarchical Q-Former for Enhanced Video Understanding. *arXiv preprint* arXiv:2503.08585v1 (2025).
[7] Baltrušaitis, T., Ahuja, C., and Morency, L.-P. 2017. Multimodal machine learning: A survey and taxonomy. *arXiv preprint* (2017). https://arxiv.org/abs/1705.09406
[8] Barnett, S. 2023. The newest threat to your attention span? TikTok 'dual' videos. *Wired* (Aug. 2023). https://www.wired.com/story/tiktok-dual-videos-attention-spans/
[9] Baumgartner, S. E., van der Schuur, W. A., Lemmens, J. S., and te Poel, F. 2017. The relationship between media multitasking and attention problems in adolescents: Results of two longitudinal studies. *Human Communication Research* 44, 1 (2017), 3–30. https://doi.org/10.1093/hcre.12111
[10] Castello, J. 2023. TikTok's sludge content isn't just for short attention spans. *Polygon* (Mar. 2023). https://www.polygon.com/23649389/tiktok-sludge-content-subway-surfers-attention-span-hasanabi
[11] Chaudhari, S., Pandey, P., Agrawal, A., and Sohi, B. S. 2021. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology* 12, 5 (2021), 1–32.
[12] Chaves, L. 2025. 20+ interesting short form video trends & statistics (2025). *Vidico* (Aug. 2025). https://vidico.com/news/short-form-video-statistics/
[13] Chatterjee, D., Remelli, E., Song, Y., Tekin, B., Mittal, A., Bhatnagar, B., Camgöz, N. C., Hampali, S., Sauser, E., Ma, S., Yao, A., and Sener, F. 2025. Memory-efficient Streaming VideoLLMs for Real-time Procedural Video Understanding. *arXiv preprint* arXiv:2504.13915 (2025).
[14] Eastdon-Smith, C. 2025. 40+ short form video statistics: The jaw-dropping numbers you must know in 2024. *Firework* (Jun. 2025). https://firework.com/blog/short-form-video-statistics
[15] Feichtenhofer, C., Fan, H., Malik, J., and He, K. 2019. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 6202–6211.
[16] Gao, M., Li, X., Tao, D., and Ji, R. 2020. Multi-modal fusion based on attention mechanism for rumor detection. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI Press, 797–804.
[17] Golemanova, R. 2025. Automated content moderation | What is it? Benefits, tools and more. *Imagga Blog* (Jan. 2025). https://imagga.com/blog/automated-content-moderation/
[18] Gongane, V. U., Munot, M. V., and Anuse, A. D. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining* 12, 1 (2022). https://doi.org/10.1007/s13278-022-00951-3
[19] González, R. 2023. The rise of 'sludge' videos: Gen Z are watching multiple clips at once. *PetaPixel* (2023).
[20] Gorwa, R., Binns, R., and Katzenbach, C. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020). https://journals.sagepub.com/doi/full/10.1177/

2053951719897945

[21] Gorti, S. K., Vouitsis, N., Ma, J., Golestan, K., Volkovs, M., Garg, A., and Yu, G. 2022. X-Pool: Cross-Modal Language-Video Attention. *arXiv preprint* arXiv:2203.15086 (2022). https://arxiv.org/abs/2203.15086

[22] Hasan, M. K. 2024. Digital multitasking and hyperactivity: Unveiling the hidden cognitive costs. *Frontiers in Psychology* 15 (2024), Article 1456789. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11543232/

[23] Hessel, J., Pang, B., Zhu, Z., and Soricut, R. 2019. A case study on combining ASR and visual features for generating instructional video captions. *arXiv preprint* (Oct. 2019). https://arxiv.org/abs/1910.02930

[24] Hua, J. 2023. TikTok sludge trend: What parents need to know. *Movieguide* (2023).

[25] Huertas-García, Á., Martín, A., Huertas-Tato, J., and Camacho, D. Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage.

[26] ISO/IEC TR 24028:2020. 2020. Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence. International Organization for Standardization.

[27] Li, J., Hasegawa-Johnson, M., and McElwain, N. L. 2023. Towards robust family-infant audio analysis based on unsupervised pretraining of Wav2vec 2.0 on large-scale unlabeled family audio. *arXiv preprint* (2023). https://doi.org/10.48550/arXiv.2305.12530

[28] Li, J., Li, D., Savarese, S., and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *ArXiv* (2023). https://arxiv.org/abs/2301.12597

[29] Li, S. and Fan, L. 2022. Media multitasking, depression, and anxiety of college students: Serial mediating effects of attention control and negative information attentional bias. *Frontiers in Psychiatry* 13 (2022), Article 989201. https://doi.org/10.3389/fpsyt.2022.989201

[30] Mattson, A. 2024. Sludge videos are taking over TikTok—and people's minds. *Scientific American* (Jan. 2024). https://www.scientificamerican.com/article/sludge-videos-are-taking-over-tiktok-and-peoples-mind1

[31] Miech, A., Zhukov, D., Alayrac, J., Tapaswi, M., Laptev, I., and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. *ArXiv* (2019).

[32] Montag, C. and Elhai, J. D. 2020. Discussing digital technology overuse in children and adolescents during the COVID-19 pandemic and beyond: On the importance of considering affective neuroscience theory. *Addictive Behaviors Reports* 12 (2020), Article 100313. https://doi.org/10.1016/j.abrep.2020.100313

[33] Nansen, B. and Balanzategui, J. 2022. Visual tactility: 'Oddly satisfying' videos, sensory genres and ambiguities in children's YouTube. *Convergence: The International Journal of Research into New Media Technologies* 28, 6 (2022), 1555–1576. https://doi.org/10.1177/13548565221105196

[34] Ophir, E., Nass, C., and Wagner, A. D. 2009. Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences* 106, 37 (2009), 15583–15587. https://doi.org/10.1073/pnas.0903620106

[35] OpenAI. 2022. Whisper: Robust speech recognition model. https://openai.com/research/whisper

[36] Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A. B., Hecker, D., Houben, S., Mock, M., Rosenzweig, J., Sicking, J., Schulz, E., Voss, A., and Wrobel, S. 2023. Guideline for Trustworthy Artificial Intelligence – AI Assessment Catalog. *ArXiv* (2023). https://arxiv.org/abs/2307.03681

[37] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L.-P. 2020. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems* 35, 6 (2020), 17–25.

[38] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *ArXiv* (2021). https://arxiv.org/abs/2103.00020

[39] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *ArXiv* (2022). https://arxiv.org/abs/2212.04356

[40] Ramachandram, D. and Taylor, G. W. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine* 34, 6 (2017), 96–108.

[41] Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., and Meira, W. 2023. Auditing radicalization pathways on YouTube. In *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–35.

[42] Sankaran, S. 2025. Enhancing Trust Through Standards: A Comparative Risk-Impact Framework for Aligning ISO AI Standards with Global Ethical and Regulatory Contexts. *ArXiv* (2025). https://arxiv.org/abs/2504.16139

[43] Shi, Y., Li, F., Zhao, L., and Huang, F. 2022. Cross-modal contrastive learning for multimodal emotion recognition. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*. AAAI Press, 2047–2055.

[44] Shorten, C. and Khoshgoftaar, T. M. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data* 6, 1 (2019), Article 60. https://doi.org/10.1186/s40537-019-0197-0

[45] Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. *ArXiv* (2019). https://arxiv.org/abs/1904.01766

[46] Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. 2023. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv preprint* arXiv:2303.15389 (2023). https://arxiv.org/abs/2303.15389

[47] Tong, Z., Song, Y., Wang, J., and Wang, L. 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. *ArXiv* (2022). https://arxiv.org/abs/2203.12602

[48] Tran, D. 2023. Sludge content and media multitasking: A deep dive.

[49] Tsai, Y.-H. H., Bai, S., Liang, P. P., Zadeh, A., Morency, L.-P., and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 6558–6569.

[50] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems* 30. Curran Associates.

[51] Wang, Y., Zhao, Y., Huo, Y., and Lu, Y. 2025. Multimodal anomaly detection in complex environments using video and audio fusion. *Scientific Reports* 15, 1 (2025). https://doi.org/10.1038/s41598-025-01146-4

[52] Winslow, C. 2023. Sludge content and media multitasking: A deep dive.

[53] Wohlert, I. K., Vega, D., Magnani, M., and Sergerberg, A. 2025. Detecting Coordination on Short-Video Platforms: The challenge of Multimodality and Complex Similarity on TikTok. *arXiv preprint* (Jun. 2025). https://arxiv.org/abs/2506.05868

[54] Xing, J., Xu, C., Wang, M., Dai, G., Sun, B., Liu, Y., Wang, J., and Zhao, J. 2023. MA-FSAR: Multimodal Adaptation of CLIP for Few-Shot Action Recognition. *ArXiv preprint* (2023). https://arxiv.org/abs/2308.01532

[55] Yuan, J., Yu, Y., Mittal, G., Hall, M., Sajeev, S., and Chen, M. 2023. Rethinking Multimodal Content Moderation from an Asymmetric Angle with Mixed-modality. *ArXiv preprint* (2023). https://arxiv.org/abs/2305.10547

[56] Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. 2018. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1103–1114.

[57] Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Suarez-Tangil, G. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*. ACM, 188–202.

[58] Zhang, X., Wu, Y., and Liu, S. 2019. Exploring short-form video application addiction: socio-technical and attachment perspectives. *Telematrics and Informatics* 42 (2019), 101243. https://doi.org/10.1016/j.tele.2019.101243

[59] Zhang, Y., Bai, Y., Chang, J., Zang, X., Lu, S., Lu, J., Feng, F., Niu, Y., and Song, Y. 2023. Leveraging Watch-time Feedback for Short-Video Recommendations: A Causal Labeling Framework. *ArXiv (Cornell University)* (2023), 4952–4959. https://doi.org/10.1145/3583780.3615483

[60] Zou, S., Xiong, J., Fan, C., Yu, S., and Tang, J. 2023. A Multi-Stage Adaptive Feature Fusion Neural Network for Multimodal Gait Recognition. *ArXiv* (2023). https://arxiv.org/abs/2312.14410