

PRISM: Perceptual Recognition for Identifying Standout Moments in Human-Centric Keyframe Extraction

Mert Can Cakmak¹, Nitin Agarwal^{1,2}, Diwash Poudel¹

¹COSMOS Research Center, University of Arkansas, Little Rock, USA

²ICSI, University of California, Berkeley, USA

{mccakmak, nxagarwal, dpoudel}@ualr.edu

Abstract

Online videos play a central role in shaping political discourse and amplifying cyber social threats such as misinformation, propaganda, and radicalization. Detecting the most impactful or “standout” moments in video content is crucial for content moderation, summarization, and forensic analysis. In this paper, we introduce PRISM (Perceptual Recognition for Identifying Standout Moments), a lightweight and perceptually-aligned framework for keyframe extraction. PRISM operates in the CIELAB color space and uses perceptual color difference metrics to identify frames that align with human visual sensitivity. Unlike deep learning-based approaches, PRISM is interpretable, training-free, and computationally efficient, making it well suited for real-time and resource-constrained environments. We evaluate PRISM on four benchmark datasets: BBC, TVSum, SumMe, and Clip-Shots, and demonstrate that it achieves strong accuracy and fidelity while maintaining high compression ratios. These results highlight PRISM’s effectiveness in both structured and unstructured video content, and its potential as a scalable tool for analyzing and moderating harmful or politically sensitive media in online platforms.

Introduction

Online platforms like YouTube, TikTok, and X have become powerful engines for shaping public discourse, where video content spreads political narratives, social ideologies, and cyber threats at scale. Within this environment, identifying *standout moments*, which are emotionally or ideologically charged video segments, is crucial for moderating harmful content and understanding its impact. Recent studies have examined how algorithmic bias and emotional content shape viewer exposure and engagement (Okeke et al. 2023; Cakmak et al. 2023a,b).

These segments are often amplified through algorithmic curation and social sharing, and are frequently used to mislead or incite during sensitive events (King and Wang 2023; Abdali, shaham, and Krishnamachari 2024; Kaur et al. 2024). As generative AI further obscures the line between real and synthetic media, detecting perceptually significant keyframes becomes vital for forensic analysis and misinformation detection.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

While NLP and network analysis are common in cyber threat research, visual signals in video remain underexplored, despite their strong emotional and narrative power (Amerini et al. 2025; Seo 2020; Cakmak et al. 2025; Bhattacharya, Agarwal, and Poudel 2024; Shajari and Agarwal 2025). Keyframe extraction helps surface these signals by condensing video into its most semantically relevant components, enabling tasks like moderation, summarization, and content indexing (Gurung, Agarwal, and Al-Taweel 2025).

In this paper, we propose **PRISM** (Perceptual Recognition for Identifying Standout Moments), a lightweight, perceptually-driven framework for extracting visually meaningful frames. PRISM uses color-based perceptual metrics and adaptive filtering to identify standout content without training or heavy computation. As shown in Figure 1, we evaluate PRISM across four diverse datasets and demonstrate its effectiveness in accuracy, fidelity, and compression, highlighting its potential for scalable and interpretable analysis of harmful media online.

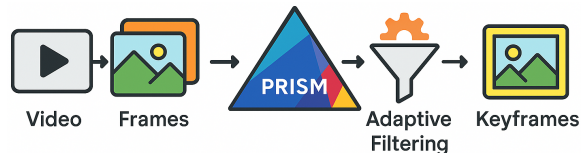


Figure 1: The PRISM framework for perceptually-guided keyframe extraction.

Literature Review

The surge in video content across online platforms has driven demand for effective keyframe extraction methods (Alp et al. 2022), essential for tasks such as summarization, indexing, and content retrieval. A variety of techniques have been proposed in recent years, each balancing different trade-offs between accuracy, complexity, and practical deployment.

Deep learning-based approaches have dominated the field, with frameworks like LMSKE (Tan et al. 2024) combining models such as TransNetV2 and CLIP for shot segmentation and semantic representation. Similarly, reinforcement learning methods (Huang et al. 2022) have been applied in medical imaging scenarios to extract diagnostically

relevant frames. While these methods achieve high performance, they are often computationally intensive, require large annotated datasets, and are difficult to interpret or deploy in real-time systems.

Clustering-based methods (Kaur, Kaur, and Lal 2024), such as those using fuzzy C-means with metaheuristics, offer a more lightweight alternative, but can suffer from sensitivity to initialization and a lack of perceptual alignment. Perceptual cues, including color-emotion links, have also been explored (Cakmak, Shaik, and Agarwal 2024; Yousefi, Cakmak, and Agarwal 2024). Entropy and information-theoretic techniques (Zhang, Tian, and Li 2017) attempt to select frames with high information content, though they may misfire in noisy or visually subtle transitions. Motion-based approaches (Dong et al. 2022) use temporal dynamics to locate changes but can miss static yet semantically important scenes. Object detection-based strategies (Bharathi, Senthilarasi, and Hari 2023) enhance semantic relevance but are constrained by category-specific detectors. Visual presentation factors, such as thumbnails and their design have also been shown to influence viewer engagement and algorithmic behavior (Poudel, Cakmak, and Agarwal 2024). Bias in Shorts thumbnails and topic drift in recommendations have also been documented (Cakmak et al. 2024; Cakmak and Agarwal 2025; Cakmak, Agarwal, and Oni 2024).

Despite this variety, common challenges persist across existing approaches, including poor interpretability, limited perceptual sensitivity, and high resource demands. Many techniques are ill-suited for real-time or large-scale applications where speed and simplicity are critical. Studies have also highlighted how multimedia plays a key role in shaping mobilization across sociopolitical contexts (Shaik et al. 2024).

In contrast, we propose PRISM, a perceptually-guided, lightweight framework that selects standout frames using color differences in the CIELAB space. Unlike deep or heuristic systems, PRISM is training-free, interpretable, and efficient, making it practical for real-world use in constrained or high-throughput environments.

Methodology

In this research, we propose a perceptually-guided method for extracting standout keyframes from videos, using the CIELAB color space to better align with human visual perception.

Justification of CIELAB Color Space

Standard RGB spaces do not reflect human perception, as they are designed for hardware representation rather than perceptual uniformity. In contrast, CIELAB was specifically developed to align with human color sensitivity, where equal distances correspond to equal perceived differences. Using CIELAB thus enables more perceptually accurate analysis of color changes between frames.

Frame Conversion Procedure

The initial stage involves converting video frames from RGB color space (stored as BGR in OpenCV) to CIELAB

color space. Mathematically, this transformation can be succinctly represented as

$$F_{LAB} = \text{cvtColor}(F_{RGB}, \text{COLOR_BGR2LAB}) \quad (1)$$

This step ensures subsequent frame comparison computations reflect genuine perceptual differences, facilitating a robust and meaningful analysis of standout frames.

Perceptual Color Difference Calculation

We use the CIEDE2000 metric (ΔE_{00}) (Pereira et al. 2019) to measure color differences in a way that aligns with human perception. Unlike basic Euclidean distance, it accounts for variations in hue, chroma, and lightness. To simplify computation, we average the CIELAB values per frame and compute the perceptual difference accordingly.

$$\Delta E_{00}(F_i, F_{i+1}) = \text{CIEDE2000}(\overline{LAB}(F_i), \overline{LAB}(F_{i+1})) \quad (2)$$

Two-Stage Adaptive Thresholding for Keyframe Identification

Our methodology utilizes an intricate two-stage adaptive thresholding mechanism specifically designed to distinguish significant perceptual changes from trivial variations.

Initial Just Noticeable Difference (JND) Thresholding

We first discard perceptual differences below the Just Noticeable Difference (JND), typically defined as a ΔE_{00} value of 1 (CHNSpec 2024). Since CIEDE2000 is designed to reflect human color perception (ViewSonic 2021), this threshold effectively filters out imperceptible variations from noise or lighting, allowing the analysis to focus on meaningful frame changes.

Table 1: Summary of ΔE Values and Their Interpretation

ΔE Range	Meaning
0.0 – 0.5	Invisible to slight; acceptable
0.5 – 2.0	Noticeable on inspection; often acceptable
2.0 – 4.0	Clear difference; conditionally acceptable
4.0 – 10.0	Obvious difference; often unacceptable
> 10.0	Strong contrast; unacceptable

Adaptive Statistical Thresholding Following JND filtering, we apply adaptive statistical thresholding. Specifically, we calculate the mean (μ) and standard deviation (σ) of the perceptually significant ΔE_{00} values within each video sequence as follows:

$$\mu = \frac{1}{N} \sum_{i=1}^N \Delta E_{00}(F_i, F_{i+1}), \quad (3)$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (\Delta E_{00}(F_i, F_{i+1}) - \mu)^2} \quad (4)$$

Frames exhibiting perceptual differences exceeding the adaptive threshold ($\mu + \sigma$) are identified as standout frames or keyframes:

$$\Delta E_{00}(F_i, F_{i+1}) > \mu + \sigma \quad (5)$$

This thresholding approach assumes a roughly Gaussian distribution of ΔE_{00} values. In such distributions, values exceeding $\mu + \sigma$ represent statistically significant deviations, commonly associated with meaningful outliers (Montgomery 2007). This enables robust detection of perceptual changes while preserving relevant content.

Our method improves keyframe extraction by mimicking human perception and using adaptive thresholds to detect meaningful changes. It’s a lightweight, efficient alternative to deep models, ideal for summarization and retrieval in resource-limited settings.

Experiments and Evaluation

This section evaluates the proposed PRISM framework using multiple benchmark datasets and a range of performance metrics to assess its accuracy, fidelity, and efficiency.

Evaluation Datasets

We evaluated the effectiveness of the proposed PRISM framework on a diverse set of publicly available video datasets that vary in length, content type, and annotation style. This allowed us to assess the model’s generalizability across different keyframe extraction scenarios, from user-generated clips to professionally edited documentaries.

The TVSum20 dataset (Tan et al. 2024) contained 20 videos across 10 categories, each annotated with frame-level importance scores derived from crowdsourced feedback. SumMe (Gygli et al. 2014) included 25 user-generated videos, each paired with multiple human-generated summaries, enabling evaluation on subjective content. To test PRISM on long-form, high-resolution material, we used 10 selected episodes from the BBC Planet Earth series (Baraldi, Grana, and Cucchiara 2016), which featured rich visuals and complex transitions. We also included a subset of the Clip-Shots dataset (Tang et al. 2018), composed of thousands of short videos with annotated shot boundaries, to evaluate PRISM’s robustness in fast-paced and diverse video content.

These datasets provided a strong benchmark for evaluating PRISM’s scalability and efficiency, supported by prior work on large-scale transcript processing (Cakmak et al. 2023c; Cakmak and Agarwal 2024).

Evaluation Metrics

To fairly evaluate PRISM across videos of varying length and frame rate, we employed a dynamic frame-matching threshold that adapted to each video’s properties. Fixed thresholds often disadvantage longer or high-fps videos, so our method scaled the tolerance window based on frame rate and total frame count.

As described in Algorithm 1, a predicted keyframe is considered correct if it falls within a computed range of any ground-truth keyframe. The final score reflects the percentage of matched predictions.

Algorithm 1: Frame Matching with FPS-Scaled Threshold

Require: Row with actual_frame_number, predicted_frame_number, frame_count, fps
Ensure: Matching percentage

- 1: actual \leftarrow row.actual_frame_number
- 2: predicted \leftarrow row.predicted_frame_number
- 3: frame_count \leftarrow row.frame_count
- 4: fps \leftarrow row.fps
- 5: **if** predicted is empty OR fps is 0 OR frame_count is 0 **then**
- 6: **return** 0.0
- 7: **end if**
- 8: max_time_window \leftarrow 10.0
- 9: alpha \leftarrow 10
- 10: time_scaling \leftarrow max_time_window $\times \frac{fps}{fps+alpha}$
- 11: threshold \leftarrow int(fps \times time_scaling)
- 12: min_threshold \leftarrow 30
- 13: max_threshold \leftarrow int(frame_count \times 0.03)
- 14: threshold \leftarrow clamp(threshold, min_threshold, max_threshold)
- 15: match_count \leftarrow 0
- 16: **for all** pred in predicted **do**
- 17: **if any** |pred - act| \leq threshold for act in actual **then**
- 18: match_count \leftarrow match_count + 1
- 19: **end if**
- 20: **end for**
- 21: **return** round($\frac{match_count}{len(predicted)} \times 100, 2$)

This adaptive strategy provided a balanced comparison across diverse datasets and video styles, from fast-paced clips to slow, documentary footage.

In addition to frame-matching accuracy (Algorithm 1), we evaluated PRISM using two complementary metrics: *fidelity* and *compression ratio*, which capture representativeness and efficiency, respectively.

Fidelity measures how well the selected keyframes preserved the visual content of the original video. We computed it using cosine similarity between normalized color histograms of predicted and ground-truth keyframes, following the approach in (Tan et al. 2024). For each predicted frame, we found the most similar ground-truth frame, and defined the final score as

$$\text{Fidelity} = 1 - \max_i \left(\min_j \text{CosSim}(k_i, g_j) \right) \quad (6)$$

Compression ratio is defined as

$$\text{Compression Ratio} = \frac{\text{Total frames}}{\text{Selected keyframes}} \quad (7)$$

Higher fidelity indicates better content preservation, while a higher compression ratio reflects greater summarization efficiency. These metrics naturally trade off: fewer keyframes yield better compression but may reduce fidelity. PRISM aims to balance this trade-off by selecting a minimal yet perceptually meaningful set of frames.

Evaluation Results

Table 2 presents PRISM’s performance across four benchmark video datasets, evaluated in terms of accuracy, fidelity, and compression ratio (CR). PRISM consistently achieves

high compression ratios (above 98.8%) while maintaining strong accuracy and perceptual fidelity.

Table 2: Evaluation results of PRISM across four benchmark datasets. (CR: Compression Ratio)

Dataset	Accuracy (%)	Fidelity (%)	CR (%)
BBC	99.50	75.70	99.50
TVSum	78.40	74.85	99.20
ClipShots	66.24	68.31	99.42
SumMe	96.16	64.35	98.80

The highest accuracy (99.50%) and fidelity (75.70%) were observed on the BBC dataset, which contains structured, long-form documentary footage, well-suited to PRISM’s perceptual approach. On TVSum, PRISM showed balanced performance (78.40% accuracy, 74.85% fidelity), reflecting its robustness across varied genres. In contrast, SumMe’s subjective, user-generated content led to a higher accuracy (96.16%) but lower fidelity (64.35%), revealing limitations when semantic importance is weakly correlated with visual cues.

ClipShots posed the greatest challenge due to fast editing and noise, yet PRISM still achieved reasonable fidelity (68.31%) and the highest compression ratio (99.42%). These results underscore PRISM’s effectiveness for both structured and unstructured content, validating its utility as a lightweight and generalizable keyframe extraction framework.

Table 3: Average evaluation metrics for PRISM and other keyframe extraction models for the four datasets.

Model	Accuracy (%)	Fidelity (%)	CR (%)
PRISM (Ours)	85.58	70.30	99.23
LiveLight	72.30	80.00	90.00
DSVS	66.00	75.00	95.00
MKFE	58.00	—	—
VSLS	41.40	—	98.60

To assess the overall effectiveness of PRISM, we compare its average performance across four benchmark datasets (TVSum, SumMe, ClipShots, and BBC Planet Earth) using three key metrics: accuracy, fidelity, and compression ratio (CR). As shown in Table 3, PRISM achieves the highest average accuracy at 85.58%, surpassing established methods such as LiveLight (Zhao and Xing 2014) (72.30%) and DSVS (Cong, Yuan, and Luo 2011) (66.00%). It also offers an exceptional compression ratio of 99.23%, significantly reducing video length while retaining critical content. Although PRISM’s fidelity (70.30%) is slightly lower than that of LiveLight (80.00%), this reflects a deliberate trade-off favoring correctness and compactness over broader visual coverage. In scenarios where frame accuracy and compression are prioritized, such as indexing, summarization, or constrained environments, PRISM offers a highly effective solution.

Some values for models like MKFE (Asim et al. 2018)

and VSLS (Guo et al. 2025) are not available, as the original publications did not report fidelity or CR metrics. Nonetheless, the inclusion of these models provides historical and methodological context for comparing PRISM’s performance against both traditional and modern approaches.

Time Complexity

To assess the computational efficiency of PRISM, we compare its frame processing speed (FPS) with prior models across multiple datasets. Here, FPS refers to the number of video frames processed per second by the model, not the frame rate of the video itself. A higher FPS indicates faster analysis and better suitability for real-time or large-scale applications.

As shown in Table 4, PRISM outperforms all baselines, achieving up to 454 FPS on ClipShots and over 130 FPS on other datasets, confirming its lightweight and efficient design.

Table 4: FPS performance comparison of PRISM and prior models across benchmark datasets.

Dataset	Model	FPS	Dataset	Model	FPS
TVSum	PRISM (Ours)	167	SumMe	PRISM (Ours)	131
	FIVT	56		FIVT	58
	LiveLight	25–30		LiveLight	25–30
	D-KTS	15		D-KTS	15
ClipShots	PRISM (Ours)	454	BBC	PRISM (Ours)	185
	DeepSBD	382		TransNet V2	25–30
	FastShot	25		MFSBD	31
	TransNet V2	30		DASBD	125
	DASBD	125		FFmpeg	165

Compared models include FIVT (Hsu, Liao, and Huang 2021), LiveLight (Zhao and Xing 2014), and D-KTS (Ke et al. 2022), as well as recent and competitive baselines like DeepSBD (Tang et al. 2018), TransNet V2 (Souček, Moravec, and Lokoč 2019), FastShot (Apostolidis and Mezaris 2014), DASBD (Esteve Brotons et al. 2023), and MFSBD (Gushchin, Antsiferova, and Vatolin 2021). Many of these methods rely on deep learning or handcrafted features, which often introduce latency. In contrast, PRISM remains training-free and perceptually grounded, enabling real-time speeds without compromising performance.

These findings highlight PRISM’s practicality for deployment in systems requiring fast, interpretable, and scalable video analysis.

Conclusion

We introduced PRISM, a lightweight and perceptually-driven framework for keyframe extraction that operates without training or labeled data. By leveraging perceptual color differences in the CIELAB space, PRISM selects visually meaningful frames aligned with human perception.

Experiments across four diverse datasets show that PRISM achieves strong accuracy and fidelity while maintaining high compression, making it suitable for real-time

and resource-constrained applications. PRISM also significantly outperforms prior methods in processing speed, confirming its practicality for large-scale and time-sensitive video analysis.

Future work will explore integrating semantic cues and temporal dynamics to further enhance performance in more complex or subjective video scenarios.

Acknowledgments

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-23-1-0011, W911NF-24-1-0078), U.S. Office of Naval Research (N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency, the Australian Department of Defense Strategic Policy Grants Program, Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment, and the Donaghey Foundation at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

- Abdali, S.; shaham, S.; and Krishnamachari, B. 2024. Multimodal Misinformation Detection: Approaches, Challenges and Opportunities. arXiv:2203.13883.
- Alp, E.; Gergin, B.; Eraslan, Y. A.; Çakmak, M. C.; and Alhajj, R. 2022. *Covid-19 and Vaccine Tweet Analysis*, 213–229. Cham: Springer International Publishing. ISBN 978-3-031-08242-9.
- Amerini, I.; Barni, M.; Battiato, S.; Bestagini, P.; Boato, G.; Bruni, V.; Caldelli, R.; De Natale, F.; De Nicola, R.; Guarniera, L.; et al. 2025. Deepfake Media Forensics: Status and Future Challenges. *Journal of Imaging*, 11(3): 73.
- Apostolidis, E.; and Mezaris, V. 2014. Fast shot segmentation combining global and local visual descriptors. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6583–6587. IEEE.
- Asim, M.; Almaadeed, N.; Al-Máadeed, S.; Bouridane, A.; and Beghdadi, A. 2018. A key frame based video summarization using color features. In *2018 Colour and visual computing symposium (CVCS)*, 1–6. IEEE.
- Baraldi, L.; Grana, C.; and Cucchiara, R. 2016. Recognizing and presenting the storytelling video structure with deep multimodal networks. *IEEE Transactions on Multimedia*, 19(5): 955–968.
- Bharathi, S.; Senthilarasi, M.; and Hari, K. 2023. Key frame extraction based on real-time person availability using YOLO. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 14(2): 31–40.
- Bhattacharya, S.; Agarwal, N.; and Poudel, D. 2024. Analyzing the impact of symbols in Taiwan’s election-related anti-disinformation campaign on TikTok. *Social Network Analysis and Mining*, 14(1): 1–20.
- Cakmak, M. C.; and Agarwal, N. 2024. High-Speed Transcript Collection on Multimedia Platforms: Advancing Social Media Research through Parallel Processing. In *2024 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 857–860.
- Cakmak, M. C.; and Agarwal, N. 2025. Unpacking Algorithmic Bias in YouTube Shorts by Analyzing Thumbnails. In *Proceedings of the 58th Hawaii International Conference on System Sciences*.
- Cakmak, M. C.; Agarwal, N.; Dagtas, S.; and Poudel, D. 2024. Unveiling Bias in YouTube Shorts: Analyzing Thumbnail Recommendations and Topic Dynamics. In Thomson, R.; Hariharan, A.; Renshaw, S.; Al-khateeb, S.; Burger, A.; Park, P.; and Pyke, A., eds., *Social, Cultural, and Behavioral Modeling*, 205–215. Cham: Springer Nature Switzerland. ISBN 978-3-031-72241-7.
- Cakmak, M. C.; Agarwal, N.; and Oni, R. 2024. The bias beneath: analyzing drift in YouTube’s algorithmic recommendations. *Social Network Analysis and Mining*, 14(1): 171.
- Cakmak, M. C.; Agarwal, N.; Poudel, D.; and Bhattacharya, S. 2025. Examining the Impact of Symbolic Content on YouTube’s Recommendation System. In Cherifi, H.; Donduran, M.; Rocha, L. M.; Cherifi, C.; and Varol, O., eds., *Complex Networks & Their Applications XIII*, 421–432. Cham: Springer Nature Switzerland. ISBN 978-3-031-82431-9.
- Cakmak, M. C.; Okeke, O.; Onyepunuka, U.; Spann, B.; and Agarwal, N. 2023a. Analyzing Bias in Recommender Systems: A Comprehensive Evaluation of YouTube’s Recommendation Algorithm. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 753–760.
- Cakmak, M. C.; Okeke, O.; Onyepunuka, U.; Spann, B.; and Agarwal, N. 2023b. Investigating bias in YouTube recommendations: emotion, morality, and network dynamics in China-Uyghur content. In *International Conference on Complex Networks and Their Applications*, 351–362. Springer.
- Cakmak, M. C.; Okeke, O.; Spann, B.; and Agarwal, N. 2023c. Adopting Parallel Processing for Rapid Generation of Transcripts in Multimedia-rich Online Information Environment. In *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 832–837.
- Cakmak, M. C.; Shaik, M.; and Agarwal, N. 2024. Emotion assessment of youtube videos using color theory. In *Proceedings of the 2024 9th International Conference on Multimedia and Image Processing*, 6–14.
- CHNSpec. 2024. Why Use CIE Lab Color Metrics. <https://www.chnspec.net/Why-Use-CIE-Lab-Color-Metrics.html>. Accessed: 2025-03-05.
- Cong, Y.; Yuan, J.; and Luo, J. 2011. Towards scalable summarization of consumer videos via sparse dictionary selection. *IEEE Transactions on Multimedia*, 14(1): 66–75.

- Dong, Y.; Zhang, Y.; Zhang, J.; and Zhang, X. 2022. Video key frame extraction based on scale and direction analysis. *The Journal of Engineering*, 2022(9): 910–918.
- Esteve Brotons, M. J.; Lucendo, F. J.; Javier, R.-J.; and Garcia-Rodriguez, J. 2023. Shot Boundary Detection with 3D Depthwise Convolutions and Visual Attention. *Sensors*, 23(16): 7022.
- FORCE11. 2020. The FAIR Data principles. <https://force11.org/info/the-fair-data-principles/>.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.
- Guo, W.; Chen, Z.; Wang, S.; He, J.; Xu, Y.; Ye, J.; Sun, Y.; and Xiong, H. 2025. Logic-in-Frames: Dynamic Keyframe Search via Visual Semantic-Logical Verification for Long Video Understanding. *arXiv preprint arXiv:2503.13139*.
- Gurung, M. I.; Agarwal, N.; and Al-Taweel, A. 2025. Are Narratives Contagious? Modeling Narrative Diffusion Using Epidemiological Theories. In Aiello, L. M.; Chakraborty, T.; and Gaito, S., eds., *Social Networks Analysis and Mining*, 303–318. Cham: Springer Nature Switzerland. ISBN 978-3-031-78554-2.
- Gushchin, A.; Antsiferova, A.; and Vatolin, D. 2021. Shot boundary detection method based on a new extensive dataset and mixed features. *arXiv preprint arXiv:2109.01057*.
- Gygli, M.; Grabner, H.; Riemenschneider, H.; and Van Gool, L. 2014. Creating summaries from user videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VII 13*, 505–520. Springer.
- Hsu, T.-C.; Liao, Y.-S.; and Huang, C.-R. 2021. Video summarization with frame index vision transformer. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, 1–5. IEEE.
- Huang, R.; Ying, Q.; Lin, Z.; Zheng, Z.; Tan, L.; Tang, G.; Zhang, Q.; Luo, M.; Yi, X.; Liu, P.; et al. 2022. Extracting keyframes of breast ultrasound video using deep reinforcement learning. *Medical image analysis*, 80: 102490.
- Kaur, A.; Noori Hoshyar, A.; Saikrishna, V.; Firmin, S.; and Xia, F. 2024. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*, 57(6): 159.
- Kaur, S.; Kaur, L.; and Lal, M. 2024. An effective Key Frame Extraction technique based on Feature Fusion and Fuzzy-C means clustering with Artificial Hummingbird. *Scientific Reports*, 14(1): 26651.
- Ke, X.; Chang, B.; Wu, H.; Xu, F.; and Zhong, S. 2022. Towards practical and efficient long video summary. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1770–1774. IEEE.
- King, K. K.; and Wang, B. 2023. Diffusion of real versus misinformation during a crisis event: A big data-driven approach. *International Journal of Information Management*, 71: 102390.
- Montgomery, D. C. 2007. *Introduction to Statistical Quality Control*. John Wiley & Sons.
- Okeke, O.; Cakmak, M. C.; Spann, B.; and Agarwal, N. 2023. Examining content and emotion bias in youtube’s recommendation algorithm. In *the Ninth International Conference on Human and Social Analytics, Barcelona, Spain*.
- Pereira, A.; Carvalho, P.; Coelho, G.; and Côte-Real, L. 2019. Efficient CIEDE2000-based color similarity decision for computer vision. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7): 2141–2154.
- Poudel, D.; Cakmak, M. C.; and Agarwal, N. 2024. Beyond the click: How youtube thumbnails shape user interaction and algorithmic recommendations. In *The 16th International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, accepted for presentation.
- Seo, H. 2020. Visual propaganda and social media. *Handbook of Propaganda*, 126–137.
- Shaik, M.; Cakmak, M. C.; Spann, B.; and Agarwal, N. 2024. Characterizing Multimedia Adoption and its Role on Mobilization in Social Movements. In *Proceedings of the 57th Hawaii International Conference on System Sciences*.
- Shajari, S.; and Agarwal, N. 2025. Developing a network-centric approach for anomalous behavior detection on youtube. *Social Network Analysis and Mining*, 15(1): 3.
- Souček, T.; Moravec, J.; and Lokoč, J. 2019. Transnet: A deep network for fast detection of common shot transitions. *arXiv preprint arXiv:1906.03363*.
- Tan, K.; Zhou, Y.; Xia, Q.; Liu, R.; and Chen, Y. 2024. Large model based sequential keyframe extraction for video summarization. In *Proceedings of the International Conference on Computing, Machine Learning and Data Science*, 1–5.
- Tang, S.; Feng, L.; Kuang, Z.; Chen, Y.; and Zhang, W. 2018. Fast video shot transition localization with deep structured models. In *Asian Conference on Computer Vision*, 577–592. Springer.
- ViewSonic. 2021. What is Delta E? And Why Is It Important for Color Accuracy? <https://www.viewsonic.com/library/creative-work/what-is-delta-e-and-why-is-it-important-for-color-accuracy/>. Accessed: 2025-03-05.
- Yousefi, N.; Cakmak, M. C.; and Agarwal, N. 2024. Examining Multimodal Emotion Assessment and Resonance with Audience on YouTube. In *Proceedings of the 2024 9th International Conference on Multimedia and Image Processing, ICMIP '24*, 85–93. New York, NY, USA: Association for Computing Machinery. ISBN 9798400716164.
- Zhang, M.; Tian, L.; and Li, C. 2017. Key frame extraction based on entropy difference and perceptual hash. In *2017 IEEE International Symposium on Multimedia (ISM)*, 557–560. IEEE.
- Zhao, B.; and Xing, E. P. 2014. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2513–2520.

Paper Checklist

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes, we use public video data without involving private or sensitive information.**
 - (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? **Yes, they clearly reflect PRISM’s contributions and scope.**
 - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes, Sections 3 and 4 justify the method and evaluation.**
 - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **No, the datasets are public benchmarks without demographic or sensitive data.**
 - (e) Did you describe the limitations of your work? **Yes, discussed in Section 6.**
 - (f) Did you discuss any potential negative societal impacts of your work? **Yes, discussed in the Introduction and Conclusion.**
 - (g) Did you discuss any potential misuse of your work? **Yes, we address potential misuse and responsible deployment in the Conclusion.**
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **We use public data, no training, and detail our reproducible setup.**
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes, the paper follows ethical guidelines and uses no private data.**
2. Additionally, if your study involves hypotheses testing...
- (a) Did you clearly state the assumptions underlying all theoretical results? **NA**
 - (b) Have you provided justifications for all theoretical results? **NA**
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **NA**
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **NA**
 - (e) Did you address potential biases or limitations in your theoretical framework? **NA**
 - (f) Have you related your theoretical results to the existing literature in social science? **NA**
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **NA**
3. Additionally, if you are including theoretical proofs...
- (a) Did you state the full set of assumptions of all theoretical results? **NA**
 - (b) Did you include complete proofs of all theoretical results? **NA**
4. Additionally, if you ran machine learning experiments...
- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **NA**
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **NA**
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **NA**
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **NA**
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **NA**
 - (f) Do you discuss what is “the cost” of misclassification and fault (in)tolerance? **NA**
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
- (a) If your work uses existing assets, did you cite the creators? **Yes, all datasets are cited in the Evaluation section.**
 - (b) Did you mention the license of the assets? **No, but we only use publicly available academic datasets.**
 - (c) Did you include any new assets in the supplemental material or as a URL? **No, we included no new assets or links.**
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **NA – no personal or user data is used.**
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes, the public datasets contain no PII or offensive content.**
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? **NA**
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? **NA**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
- (a) Did you include the full text of instructions given to participants and screenshots? **NA**
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **NA**
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **NA**
 - (d) Did you discuss how data is stored, shared, and de-identified? **NA**