Beyond the Click: How YouTube Thumbnails Shape User Interaction and Algorithmic Recommendations

Diwash Poudel, Mert Can Cakmak, and Nitin Agarwal

COSMOS Research Center, University of Arkansas - Little Rock
Little Rock, Arkansas, USA
{dpoudel, mccakmak, nxagarwal}@ualr.edu

Abstract. First impressions are key, especially in the crowded world of digital content on platforms like YouTube, where thumbnails play a vital role in grabbing viewer attention. This paper investigates the impact of thumbnail presentations on user engagement and bias within YouTube's recommendation algorithms. Our study systematically examines the attributes of thumbnails, such as colorfulness, brightness, and image quality, and their potential associations with user engagement metrics. Additionally, we delve into the analysis of thumbnail captions using advanced image captioning models, exploring how accurately these descriptions reflect the content and influence viewer behavior. From these analyses, we infer that attributes like brightness, colorfulness, or the quality of thumbnails do not significantly influence engagement metrics for these narratives. Our findings indicate that thumbnails featuring content with a universal appeal or related to universally engaging themes such as mystery, fantasy, science, and military subjects are more likely to attract higher viewership. This research not only offers a comprehensive assessment of thumbnail effectiveness on viewer engagement but also provides insights into the biases perpetuated by YouTube's algorithms. Through our findings, we aim to enhance understanding of content strategy on digital platforms and propose design practices that can mitigate bias and improve user interaction, thereby contributing to the field of social computing and digital media studies.

Keywords: YouTube Thumbnails · Algorithmic Bias · User Engagement · Recommender Systems · Generative AI · Large Language Models.

1 Introduction

The initial presentation of content plays a crucial role in capturing audience interest, often determining the engagement level even before the content is consumed. While the intrinsic value of content remains undisputed, an ineffective presentation can significantly undermine its perceived value. Essentially, capturing an audience's interest from the start is just as important, perhaps even more importantly, than the content itself. If the presentation fails to capture interest, the content, regardless of its quality, may remain overlooked.

On platforms like YouTube, thumbnails serve as the primary means of presentation. They are not merely previews but pivotal in attracting viewer attention. Thumbnails have the potential to significantly influence user engagement by making a first impression that can either draw viewers in or push them away. This influence extends to YouTube's recommendation algorithms, which may exhibit biases towards certain types of thumbnails that historically generate more views.

Our research investigates these biases and their implications for content visibility and user engagement on YouTube. Specifically, our study addresses the following research questions:

- **RQ1:** How do the characteristic attributes of thumbnails affect user engagement and contribute to biases in YouTube's recommendation algorithm?
- **RQ2:** How does the content described by these thumbnails affect user engagement and perpetuate biases on the platform?

Through this research, we aim to deepen the understanding of how initial impressions through thumbnails influence viewer behavior and algorithmic recommendations, potentially leading to a cycle of biased content promotion.

2 Literature Review

The literature review discusses existing research on the role of thumbnail characteristics in influencing user engagement on digital platforms and explores studies on bias within algorithmic recommender systems.

2.1 Recommender Bias

Bias in YouTube thumbnails and recommender systems influences user perception and engagement, often perpetuating prejudices. Thumbnails may use sensational or misleading visuals to increase views, known as clickbait. This approach manipulates user behavior and prioritizes certain content, contributing to biased information presentation [1, 2]. Research on social networks shows that analyzing public sentiment, such as tweets, can reveal biases in user-generated content and platforms [3]. Similar patterns appear in social movements, where multimedia content fosters connective action, as seen in recent protests in Brazil and Peru [4].

YouTube's recommender systems enhance these biases by favoring videos that keep users engaged longer. They often recommend videos with clickbait thumbnails, which show higher engagement due to their eye-catching nature. The algorithms propagate content similar to what users have previously watched, potentially creating filter bubbles and reinforcing biases [5, 6].

These biases affect the overall content landscape on YouTube, leading to the underrepresentation of minority viewpoints or sensitive topics and skewing public understanding [7–10]. Additionally, content creators' reliance on biased thumbnails and algorithms to maintain visibility exacerbates content homogenization, suppressing diversity and fairness [11].

2.2 Thumbnail Attributes and User Engagement

Thumbnails significantly impact user engagement on digital platforms. Attributes like colorfulness, brightness, and image quality are crucial, with well-designed thumbnails notably increasing view-through rates [12–14]. Customized thumbnails, featuring elements such as celebrity images or high-quality graphics, generally attract more engagement compared to automatically generated ones.

The strategic use of color and objects, as seen in YouTube videos from Sri Lanka, highlights the importance of culturally relevant elements in thumbnail design [15]. Additionally, research on emotion assessment using color theory explores how specific colors in video frames can evoke distinct emotional responses, influencing viewer engagement [16, 17].

Captions in thumbnails also play a vital role. Advanced image processing combined with natural language generation can create engaging captions that complement thumbnail visuals, enhancing appeal [18]. Moreover, advancements in image recognition technology, such as GPT-4V models, show promise in analyzing and categorizing thumbnails to optimize engagement across various contexts, including medical fields and video streaming services [19, 20].

Studies support the effectiveness of specific thumbnail features, like the presence of text or human faces, in boosting user engagement metrics on platforms like YouTube [21, 22].

3 Methodology

This section outlines the research methodologies employed, detailing the systematic approaches used for data collection, analysis, and interpretation of how thumbnail attributes affect viewer engagement on YouTube.

3.1 Data Collection

The study focused on analyzing thumbnail behavior across YouTube videos concerning geopolitical issues. In collaboration with subject matter experts, a meticulous approach was used to develop keywords reflecting significant narratives: the China-Uyghur Conflict, the South China Sea Dispute, and Cheng Ho Propaganda. These keywords spanned topics from human rights to geopolitical tensions and historical narratives.

Using the YouTube Data API v3 with parallelization techniques [23–25], the study queried videos related to these topics, enabling a detailed examination of thumbnail behaviors within specific geopolitical contexts. The keyword selection and querying process was critical to ensuring the relevance of the data to the study's objectives.

Geopolitical topics were prioritized due to their complexity and the need for nuanced exploration. The study collected a substantial dataset of thumbnails: 4,923 for the China-Uyghur Conflict, 4,418 for the South China Sea Dispute, and 4,163 for Cheng Ho Propaganda. This approach provided insights into the design and effectiveness of thumbnails used to attract viewers for content involving sensitive and complex information.

3.2 Investigation of Image Attributes

To gain a deeper insight into the behaviors of thumbnail images, we conducted an examination of the properties defining their visual characteristics, known as image attributes. These attributes can be primarily categorized into three types: colorfulness, brightness, and quality, each contributing distinctively to the overall visual perception of an image.

Colorfulness Colorfulness in an image is quantified by the range of colors present, with higher values indicating a broader color spectrum. This measurement is based on the guidelines from [26] and involves using the Python OpenCV library to analyze the RGB values of thumbnails. By calculating the difference in opposing color values (e.g. red-green and yellow-blue), which simulates human color perception, the colorfulness metric is established, ranging from 0 to 255, where higher numbers represent greater color diversity.

Brightness Brightness pertains to the overall lightness or luminance of an image. It is determined by converting the image to grayscale and then calculating the average pixel value. Images with higher average values are perceived as lighter, while those with lower averages are seen as darker. This evaluation is performed using the OpenCV library, and brightness values also range from 0 to 255, with 0 indicating low brightness and 255 representing high brightness.

Quality Image quality is assessed using the BRISQUE algorithm [27], which analyzes grayscale images to identify distortion like noise and blurriness. The algorithm calculates Mean Subtracted Contrast Normalized (MSCN) coefficients, focusing on texture and structural details, and uses a support vector machine to provide a quality score ranging from 0 (low) to 100 (high quality).

The relationship between image attributes (quality, brightness, colorfulness) and viewer engagement (likes, views, comments) is analyzed through a correlation matrix using the Karl Pearson coefficient. This analysis, performed with the numpy.corrcoef function, helps understand how these attributes impact viewer interactions.

3.3 Exploration of Thumbnail Content Through Advanced Algorithms

To achieve a detailed and contextual understanding of the visual information embedded within images such as objects, scenery, text, etc., we have utilized image captioning algorithms. This approach aims to capture the essence of image content through advanced computational methods. We evaluated three distinct algorithms for this purpose: YOLO, BLIP, and GPT-4 Turbo-Vision, each offering unique insights into image analysis.

YOLO (You Only Look Once) YOLO stands out as a real-time object detection model known for its speed and accuracy. Diverging from traditional models that employ a sliding window or region proposal algorithms, YOLO partitions the image into grids to simultaneously predict bounding boxes and class probabilities for each segment. This global perspective enables it to understand contextual relationships between objects and their characteristics, enhancing its applicability across various domains and unforeseen scenarios [28]. While YOLO primarily focuses on object detection and classification, we adapted its output to enumerate detected objects and their quantities within images.

BLIP (Bootstrapping Language-Image Pre-training) BLIP leverages a transformer-based framework, trained on extensive image-description pairs, to generate descriptive captions from images. It employs the Visual Transformer architecture [29] as an image encoder, which partitions images into patches and encodes them as a sequence of embeddings. This process facilitates the extraction of high-level semantic information, encompassing scenes, objects, and attributes. The end result is a comprehensive textual description that mirrors the visual content of the image [30].

GPT-4 Turbo-Vision This cutting-edge algorithm harnesses a vast neural network, trained on a substantial corpus of text and images, to understand the interplay between visual elements and language. GPT-4 Turbo-Vision is capable of generating precise and contextually relevant responses based on user prompts. It operates through the "GPT-4-Turbo-Vision" API [31], interpreting base64 encoded images and user queries to produce insightful summaries or answers to specific questions. Unlike BLIP, which autonomously generates captions, GPT-4 Turbo-Vision's output significantly depends on both the visual content and the user's input, offering a more interactive approach to image analysis.

3.4 Analysis of Image Captions

To explore the core aspects of images, we leveraged the OpenAI GPT-4-Turbo-Vision, also referred to as the GPT-4V API, from the models we have discussed. This model excels in identifying entities within images most effectively. The identified entities were annotated with captions and stored in a text document, providing a structured methodology for analyzing image content.

All the steps that we took for the analysis are summarized in this figure (refer to Fig. 1) to provide an overview.

Caption Pre-Processing The initial stage involved preparing the collected data for analysis. Each thumbnail image was captioned using the GPT-4V model, followed by a standard pre-processing routine. This routine included tokenizing the captions to isolate individual words, eliminating stop words to focus on content-rich terms, and applying lemmatization to condense words to their base

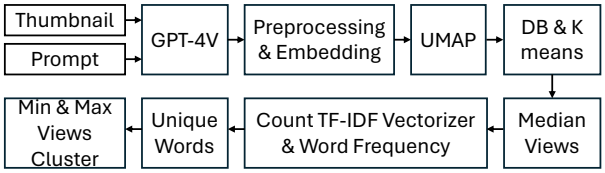


Fig. 1: Flow diagram of analysis.

or root forms. Lemmatization aids in the analytical process by enabling the grouping of similar words, thus enhancing the efficiency of subsequent clustering tasks.

S-BERT Embedding We converted the textual captions into numerical vectors using Sentence-BERT (S-BERT), an adaptation of the original BERT model optimized for sentence embeddings. S-BERT, based on Siamese and triplet network architectures, excels at producing embeddings that accurately reflect sentence similarities. The model is trained on datasets containing sentence pairs with corresponding similarity scores, ensuring that the embeddings accurately capture semantic relationships. S-BERT's efficiency in capturing nuanced textual relations marks a significant improvement over traditional BERT embeddings [32].

UMAP Dimension Reduction We used dimension reduction to tackle issues related to high-dimensional data, such as the curse of dimensionality, improving generalization and accuracy [33, 34]. We selected Uniform Manifold Approximation and Projection (UMAP) for this task, utilizing the Python UMAP library for its ability to preserve global data structure and provide excellent runtime performance [35]. UMAP employs iterative optimization with stochastic gradient descent to create weighted graphs, representing data point proximity in a lower-dimensional space, closely mirroring their relationships in the original high-dimensional context.

DB Score & K-means Clustering We used K-means clustering to group similar words in thumbnail captions based on their features due to its accuracy and speed. To improve data representation, we applied the UMAP embedding technique, simplifying the data structure and enhancing K-means clustering effectiveness. Determining the optimal number of clusters, K, was addressed using Davis & Bouldin scores, aiming for a lower score to indicate well-separated and dense clusters [36].

Median View Analysis per Cluster Upon clustering, we analyzed the data with a preference for the median view within each cluster. This approach, favored for its robustness against outliers, proved particularly valuable given the nature of our dataset.

Count Vectorizer and Word Frequency We applied the Sklearn's Count Vectorizer algorithm to quantify word frequencies within each cluster's captions, simultaneously filtering out singleton words. This step was critical for identifying frequently occurring terms.

TF-IDF for Key Word Extraction Further analysis employed the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer to highlight important words in the captions. By balancing term frequency with the uniqueness of terms, TF-IDF underlines words with the greatest relevance and distinctiveness [37].

Words in Max & Min View Cluster We removed words appearing in multiple clusters to identify unique words in each cluster. By extracting words from clusters with the highest and lowest views, our approach mirrors the structure of BERTopic, a topic modeling algorithm [38]. Unlike BERTopic, which uses HDBSCAN for clustering, we chose K-means. HDBSCAN, a density-based method that doesn't require specifying the number of clusters in advance [39], produced over 100 clusters, making it impractical for our analysis. Thus, we opted for K-means to manage and analyze our data effectively.

4 Results

This section presents the comprehensive results of our analysis, including the correlation between thumbnail attributes and user engagement, as well as the effectiveness of different captioning models in accurately representing content and influencing viewer behavior.

4.1 Correlation between image attributes and engagement scores

The correlation matrix heatmaps illustrated in Fig. 2 display the correlation coefficients among several variables: number of views, likes, comments, colorfulness, brightness, and quality. This tool helps to quantify the relationships between pairs of variables, where a higher correlation coefficient indicates a stronger relationship.

Our analysis reveals a negligible correlation between the number of views and quality, with a coefficient of -0.02. Similarly, the data shows a lack of significant correlation between the number of views and both brightness and colorfulness across all datasets examined.

However, an interesting pattern emerges among the number of views, likes, and comments. These variables tend to correlate with each other, unlike colorfulness, brightness, and quality, which show minimal correlation among themselves.

From these observations, we infer that attributes like brightness, colorfulness, or the quality of thumbnails do not significantly influence engagement metrics for these narratives. Due to the consistency of these patterns across different

datasets—such as Cheng Ho and South Sea China—only the Uyghur correlation graph is retained to conserve space.

This suggests that factors other than the visual quality of thumbnails may play a more substantial role in attracting viewer attention and interaction.

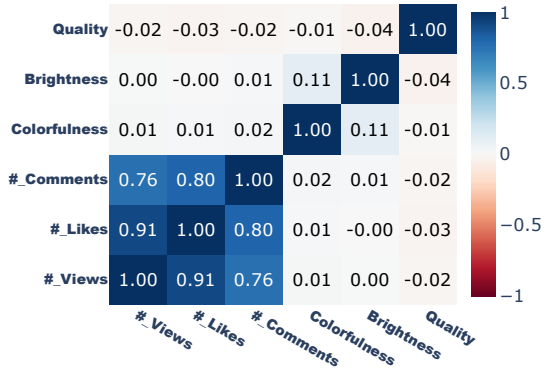


Fig. 2: Correlation Matrix for China-Uyghur Crisis.

4.2 Comparison of Image Captioning Algorithms

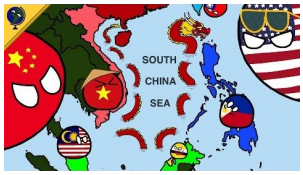
We conducted a comparative analysis of three distinguished image captioning algorithms, as introduced in Section 3.3, using thumbnail images relevant to distinct geopolitical topics. Our goal was to evaluate which algorithm yields the most accurate and detailed descriptions. The significance of selecting the optimal image captioning model is paramount in our study, which is why we picked to examine three models known for their diverse architectural designs.

We provided thumbnail images related to three sensitive geopolitical issues as depicted in the figures below. Each thumbnail was processedto compare the effectiveness and accuracy of the model's captioning capabilities.

1. YOLOv5 model's outputs:
   - For Figure 3a: "5 person."
   - For Figure 3b: "1 person."
   - For Figure 3c: "1 stop sign."
2. BLIP model's outputs:
   - For Figure 3a: "A group of people standing in front of a church."
   - For Figure 3b: "A man in a white hat is praying."
   - For Figure 3c: "A cartoon map of the world with countries and flags."
3. GPT-4 Turbo-Vision Preview model's outputs:

(a) Thumbnail image related to China-Uyghur Crisis.

(b) Thumbnail image related to Cheng Ho Propaganda.



(c) Thumbnail image related to South China Sea Dispute.

Fig. 3: Thumbnail Images

- For Figure 3a: "The image appears to be a selfie of four individuals inside a mosque. There is a woman in the foreground wearing a blue hijab, with a smiling expression, holding the camera. Behind her, there is a man with glasses holding a young child, and to their right, there is another woman wearing a white hijab. The group seems to be a family. They are posed in front of an ornate interior that suggests they are in the prayer hall of a mosque, as evidenced by the characteristic arches, stained glass windows, and chandeliers. The text overlay on the image says, "1000 NEW MUSLIMS EVERY MONTH" and "110 NEW MASJID." This text suggests that the image may be associated with a message or campaign about the growth of the Muslim community, possibly indicating the number of converts to Islam and the construction of new mosques."
- For Figure 3b: "The image shows an individual with Asian features in a praying posture with their hands raised close to their face, palms facing towards themselves - a typical gesture in certain types of prayer. The person appears to be indoors, and the setting suggests a place of worship or ceremonial space, given the ornate design in the background. The text on the image indicates that it is related to "Rona Ramadan" and mentions "Takmir Tionghoa Masjid Cheng Ho" alongside the location "Surabaya, Jawa Timur," which translates to Surabaya, East Java. This alludes to an association with the Muslim community in Surabaya, specifically the Chinese Muslim community with reference to the Cheng Ho Mosque, known for its cultural and religious significance amongst Chinese Indonesian Muslims."
- For Figure 3c: "The image depicts a stylized and cartoonish interpretation of a map, focusing on the South China Sea and surrounding countries. The countries are represented by circular characters with faces, flags, and personalities that seem to align with political or cultural stereotypes of those nations. These characters are likely referencing the countries' relations with respect to the territorial disputes in the South China Sea. Such representations are often found in internet memes and political cartoons and are sometimes associated with "countryball" or "Polandball," a user-generated internet meme which originated on the

board of Krautchan.net in the latter half of 2009. In this image, you can see characters representing China, Vietnam, Malaysia, the Philippines, and the USA among others, depicted in a way that personifies the countries with expressions and accessories that suggest their attitudes or roles in regard to the South China Sea."

The YOLOv5 model is primarily effective for basic object detection, providing straightforward outputs like "5 person," "1 person," and "1 stop sign," without additional context or detail. In contrast, the BLIP model produces more descriptive captions that attempt to contextualize objects within a broader narrative. However, it struggles with accuracy, as seen in its misidentification of a mosque as a church in an image related to the Uyghur crisis.

The GPT-4 Turbo-Vision Preview model excels beyond these by offering detailed, context-aware, and culturally sensitive descriptions. It effectively analyzes and connects images to broader cultural and historical narratives. For instance, it detailed the cultural significance in a family's selfie at a mosque and linked imagery in other thumbnails to relevant cultural and geopolitical narratives.

Overall, the GPT-4 Turbo-Vision Preview model is superior for providing accurate, detailed, and culturally informed image descriptions, outperforming both the YOLOv5's basic object detection and the BLIP model's descriptive but occasionally inaccurate captions.

4.3 Cluster Results

In our analysis, we first determined the optimal number of clusters for the narratives by applying UMAP embedding and calculating the Davies-Bouldin (DB) Score across the three distinct datasets. The DB Score helps in identifying the most effective cluster count by favoring lower scores, which indicate better separation and compactness of clusters. For the Cheng Ho and the South China Sea datasets, the lowest DB Score was observed with four clusters, signifying that four is the optimal number of clusters for these topics. Conversely, for the Uyghur dataset, the optimal cluster count was identified as six, based on achieving the lowest DB Score at this number. Among these clusters, the highest median views were recorded as follows: the China-Uyghur topic led with 1.6 million views, followed by the Cheng Ho Propaganda and the South China Sea topics, both around 800k median views.

An analysis of the content within these clusters revealed consistent use of the word "text" across all narratives, a finding consistent with the expectation that thumbnails related to these geopolitical topics predominantly originate from news sources. Notably, variations between clusters were observed; for example, the Cheng Ho dataset's high-view clusters frequently featured words like "dress" and "attire", suggesting a visual focus on clothing, while lower-view clusters commonly included words like "group" and "standing". Similar subtle differences were noted in the clusters for the South China Sea and Uyghur datasets. A more detailed examination was conducted by identifying words unique to each high and low view cluster, excluding common terms across clusters.

Word clouds generated from this analysis, represented by Fig. 4, showcase words based on their TF-IDF scores, indicating the relevance of a word within a cluster. The visualization emphasizes words unique to clusters with the highest and lowest views, omitting words prevalent in multiple clusters.

For clusters with the highest views that are displayed in Fig. 4a, Fig. 4c, and Fig. 4e, terms like "supernatural", "space", "planet", and "solar" were prominent. These words, associated with themes of mystery, fantasy, and science, possess the potential to captivate a broader audience. Additionally, terms like "military", "uniform", and "religious", which relate to specific interests or cultural identities, were found to enhance viewership due to their resonance with viewers' personal or cultural affiliations. Visually engaging terms such as "portrait", "painting", "vehicle", "truck", and "tornado" likely contribute to higher click-through rates on YouTube by drawing immediate viewer attention.

Conversely, clusters with lower views that are presented in Fig. 4b, Fig.4d, and Fig. 4f featured words such as "specific", "red", "poster", "performance", "gathering", "meal", "dance", "festival", "speech", and "interview". These terms, often pertaining to specialized, specific events or subjects, tend to have a more limited appeal, potentially restricting their reach to broader audiences [40].

Our findings indicate that thumbnails featuring content with a universal appeal or related to universally engaging themes such as mystery, fantasy, science, and military subjects are more likely to attract higher viewership. In contrast, content that focuses on targeted topics or events tends to garner less attention. This underscores the strategic importance of thumbnail content selection in maximizing engagement on platforms like YouTube.

5 Conclusion and Discussion

This study seeks to explain the complex interplay between the content and attributes of video thumbnails and their impact on viewer engagement across various content categories. Our investigation sheds light on the intricate dynamics of user interactions within digital environments.

Our research, utilizing a correlation matrix, underscores a minimal correlation between certain image attributes such as colorfulness, brightness, and quality and the number of views a video receives. This suggests that such attributes may not be critical in drawing viewer attention, highlighting the need to explore additional characteristics to discern more subtle drivers of viewer engagement.

In the evaluation of tools for generating image descriptions, the GPT-4 Turbo-Vision Preview model proved to be supremely effective, outperforming both the YOLOv5 and BLIP models in extracting rich and varied content. This underscores the paramount importance of selecting the appropriate model for specific tasks, particularly those necessitating a refined understanding and detailed descriptions of visual content.

However, it's crucial to note the constraints imposed by OpenAI's content policy on the GPT-4 Turbo-Vision Preview model, particularly regarding images that depict violence, sensitive content, or recognizable human faces. This

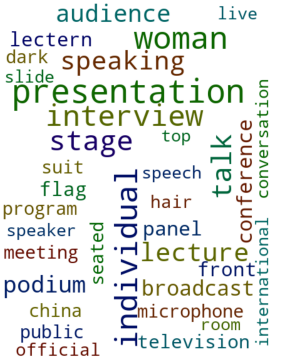
(a) Words with high views for China-Uyghur Conflict

(b) Words with low views for China-Uyghur Conflict

(c) Words with high views for Cheng Ho Propaganda

(d) Words with low views for Cheng Ho Propaganda

(e) Words with high views for South China Sea Dispute



(f) Words with low views for South China Sea Dispute

Fig. 4: Word Clouds for High and Low Views in Clusters

policy restricts the model's ability to fully analyze or describe certain images, potentially omitting critical context or subjects from our analysis. Despite its attempts to contextualize images within its knowledge cut-off of April 2023, future research may benefit from leveraging Large Language Models (LLMs) that offer comparable accuracy without such content restrictions.

Our comparative analysis between thumbnails of high and low median views revealed significant distinctions. Themes such as fantasy and science, as well as visually engaging elements like portraits and natural phenomena, were found to more effectively captivate viewers. This finding underlines the strategic importance of utilizing specific thematic and visual content to spark viewer interest and engagement.

Looking ahead, investigating the relationship between thumbnails and video previews could yield further insights, especially as YouTube Shorts gain popularity. Expanding the dataset to include a wider variety of thumbnails could also enrich our understanding of effective thumbnail strategies in the future.

Ultimately, our study offers valuable insights into optimizing video appeal through thumbnail content, providing guidance for content creators aiming to enhance viewer engagement. Moreover, it lays the groundwork for future research in this field, offering a comprehensive resource for navigating the competitive landscape of digital content.

Acknowledgements

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189, W911NF-23-1-0011, W911NF-24-1-0078), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, 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 (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, 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

1. S. Zannettou, S. Chatzis, K. Papadamou, and M. Sirivianos, "The Good, the Bad and the Bait: Detecting and Characterizing Clickbait on YouTube," in 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 2018, pp. 63-69. doi: 10.1109/SPW.2018.00018.
2. J. Qu, A. M. Hißbach, T. Gollub, and M. Potthast, "Towards Crowdsourcing Clickbait Labels for YouTube Videos," in HCOMP (WIP & Demo), 2018

3. E. Alp, B. Gergin, Y. A. Eraslan, M. C. Çakmak, and R. Alhajj, "Covid-19 and Vaccine Tweet Analysis," in *Social Media Analysis for Event Detection*, T. Özyer, Ed. Lecture Notes in Social Networks, Cham, Switzerland: Springer, 2022, pp. 169-184. doi: 10.1007/978-3-031-08242-9_9.
4. M. Shaik, M. C. Cakmak, B. Spann, and N. Agarwal, "Characterizing Multimedia Adoption and its Role on Mobilization in Social Movements," in *Proc. 57th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Hilton Hawaiian Village Waikiki Beach Resort, Hawaii, USA, Jan. 2024, pp. 146-155. [Online]. Available: https://hdl.handle.net/10125/106393
5. A. Vitadhani, K. Ramli, and P. D. Purnamasari, "Detection of Clickbait Thumbnails on YouTube Using Tesseract-OCR, Face Recognition, and Text Alteration," in 2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST), Yogyakarta, Indonesia, 2021, pp. 56-61. doi: 10.1109/ICAICST53116.2021.9497811.
6. M. C. Cakmak, O. Okeke, U. Onyepunuka, B. Spann, and N. Agarwal, "Analyzing Bias in Recommender Systems: A Comprehensive Evaluation of YouTube's Recommendation Algorithm," in Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '23), Kusadasi, Turkiye, 2024, pp. 753-760. doi: 10.1145/3625007.3627300
7. O. Okeke, M. C. Cakmak, B. Spann, and N. Agarwal, "Examining Content and Emotion Bias in YouTube's Recommendation Algorithm," in The Proceedings of the Ninth International Conference on Human and Social Analytics (HUSO 2023), Barcelona, Spain, Mar. 2023, pp. 15-20. Available: https://www.thinkmind.org/index.php?view=article&articleid=huso_2023_1_40_80032
8. M. C. Cakmak, O. Okeke, U. Onyepunuka, B. Spann, and N. Agarwal, "Investigating Bias in YouTube Recommendations: Emotion, Morality, and Network Dynamics in China-Uyghur Content," in Complex Networks & Their Applications XII, H. Cherifi, L. M. Rocha, C. Cherifi, and M. Donduran, Eds., Cham: Springer Nature Switzerland, 2024, pp. 351-362. doi: 10.1007/978-3-031-53468-3_30
9. M. I. Gurung, M. M. I. Bhuiyan, A. Al-Taweel, and N. Agarwal, "Decoding YouTube's Recommendation System: A Comparative Study of Metadata and GPT-4 Extracted Narratives," in *Companion Proceedings of the ACM on Web Conference 2024*, New York, NY, USA: Association for Computing Machinery, 2024, pp. 1468–1472, doi: 10.1145/3589335.3651913.
10. M. C. Cakmak, N. Agarwal, S. Dagtas, and D. Poudel, "Unveiling Bias in YouTube Shorts: Analyzing Thumbnail Recommendations and Topic Dynamics," in *Proc. 17th Int. Conf. Soc. Comput., Behav.-Cultural Model. & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS)*, 2024, Accepted for presentation.
11. H. Kim et al., "Towards Visualization Thumbnail Designs That Entice Reading Data-Driven Articles," IEEE Transactions on Visualization and Computer Graphics, doi: 10.1109/TVCG.2023.3278304
12. B. Koh and F. Cui, "An exploration of the relation between the visual attributes of thumbnails and the view-through of videos: The case of branded video content," Decision Support Systems, vol. 160, pp. 113820, 2022. doi: https://doi.org/10.1016/j.dss.2022.113820.
13. J. Park, "The Impact of YouTube's Thumbnail Images and View Counts on Users' Selection of Video Clip, Memory Recall, and Sharing Intentions of Thumbnail Images," The Florida State University, 2022.

14. S. Shajari, R. Amure, and N. Agarwal, "Analyzing Anomalous Engagement and Commenter Behavior on YouTube," in *AMCIS 2024 Proceedings*, 6, 2024. [Online]. Available: https://aisel.aisnet.org/amcis2024/social_comp/social_comput/6
15. S. I. A. P. Diddeniya, H. N. Gunasinghe, and C. Premachandra, "YouTube Trending Video Analysis in Sri Lanka," in 2022 2nd International Conference on Image Processing and Robotics (ICIPRob), Colombo, Sri Lanka, 2022, pp. 1-6. doi: 10.1109/ICIPRob54042.2022.9798745.
16. M. C. Cakmak, M. Shaik, and N. Agarwal, "Emotion Assessment of YouTube Videos using Color Theory," in Proceedings of the 9th International Conference on Multimedia and Image Processing (ICMIP), 2024. doi: 10.1145/3665026.3665028.
17. N. Yousefi, M. C. Cakmak, and N. Agarwal, "Examining Multimodal Emotion Assessment and Resonance with Audience on YouTube," in Proceedings of the 9th International Conference on Multimedia and Image Processing (ICMIP), 2024. doi: 10.1145/3665026.3665039.
18. D. R. Beddiar, M. Oussalah, and T. Seppänen, "Automatic captioning for medical imaging (MIC): a rapid review of literature," Artificial Intelligence Review, vol. 56, pp. 4019–4076, 2023. doi: 10.1007/s10462-022-10270-w.
19. R. Chen, T. Xiong, Y. Wu, G. Liu, Z. Hu, L. Chen, Y. Chen, C. Liu, and H. Huang, "GPT-4 Vision on Medical Image Classification – A Case Study on COVID-19 Dataset," 2023, arXiv:2310.18498 [eess.IV].
20. J. Deng, K. Heybati, and M. Shammas-Toma, "When vision meets reality: Exploring the clinical applicability of GPT-4 with vision," Clinical Imaging, vol. 108, pp. 110101, 2024. doi: https://doi.org/10.1016/j.clinimag.2024.110101.
21. S. Lee, "A Study on Visual Expression Elements and User Satisfaction in Video Streaming Services on the Web: Focusing on Video Thumbnails," Journal of Web Engineering, vol. 22, no. 1, pp. 27-40, January 2023. doi: 10.13052/jwe1540-9589.2212.
22. H. E. Jang, S. H. Kim, J. S. Jeon, and J. H. Oh, "Visual Attributes of Thumbnails in Predicting YouTube Brand Channel Views in the Marketing Digitalization Era," IEEE Transactions on Computational Social Systems. doi: 10.1109/TCSS.2023.3289410.
23. "YouTube Data API," YouTube. Available: https://developers.google.com/youtube/v3/docs. Accessed on: Nov. 15, 2023.
24. M. C. Cakmak and N. Agarwal, "High-Speed Transcript Collection on Multimedia Platforms: Advancing Social Media Research through Parallel Processing," in *Proc. 2024 IEEE Int. Parallel Distrib. Process. Symp. Workshops (IPDPSW)*, San Francisco, CA, USA, 2024, pp. 857-860, doi: 10.1109/IPDPSW63119.2024.00153.
25. M. C. Cakmak, O. Okeke, B. Spann, and N. Agarwal, "Adopting Parallel Processing for Rapid Generation of Transcripts in Multimedia-rich Online Information Environment," 2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), St. Petersburg, FL, USA, 2023, pp. 832-837, doi: 10.1109/IPDPSW59300.2023.00139.
26. D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in Proc. SPIE 5007, Human Vision and Electronic Imaging VIII, June 17, 2003, doi: 10.1117/12.477378.
27. A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," in IEEE Transactions on Image Processing, vol. 21, no. 12, pp. 4695-4708, Dec. 2012, doi: 10.1109/TIP.2012.2214050.
28. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

29. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2021, arXiv:2010.11929 [cs.CV].
30. J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in Proceedings of the 39th International Conference on Machine Learning, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162, pp. 12888–12900, PMLR, 17–23 Jul. 2022. Available: https://proceedings.mlr.press/v162/li22n.html
31. J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.
32. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," 2019, arXiv:1908.10084 [cs.CL].
33. L. J. P. Van der Maaten, "An introduction to dimensionality reduction using MATLAB," Report, vol. 1201, no. 07-07, pp. 62, 2007.
34. S. Shajari, M. Alassad, and N. Agarwal, "Characterizing suspicious commenter behaviors," in Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, 2023, pp. 631–635, doi: https://doi.org/10.1145/3625007.3627309
35. L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," 2020, arXiv:1802.03426 [stat.ML].
36. D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224-227, April 1979. doi: 10.1109/TPAMI.1979.4766909.
37. K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," Journal of Documentation, vol. 28, no. 1, pp. 11–21, 1972, MCB UP Ltd.
38. M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022, arXiv:2203.05794 [cs.CL].
39. R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," in Advances in Knowledge Discovery and Data Mining. PAKDD 2013, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds., vol. 7819, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013. doi: 10.1007/978-3-642-37456-2_14
40. S. S. Sundar, "The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility," Cambridge, MA: MacArthur Foundation Digital Media and Learning Initiative, 2008.