# Decoding Visual Storytelling in Mental Health Narratives: Aesthetics, Disparity, and Peer Support in YouTube Vlogs

Under review at *Social Media + Society*

**Jiaying "Lizzy" Liu[1], Yunlong Wang[2], Allen Jue[3], Yao Lyu[4], Yiheng Su[1], Shuo Niu[5], Nathan TeBlunthuis[1], Yan Zhang[1]**

[1]School of Information, The University of Texas at Austin, USA

[2]Institute of High Performance Computing (IHPC), A*STAR, Singapore

[3]Computer Science, The University of Texas at Austin, Austin, Texas, USA

[4]School of Information, University of Michigan, Ann Arbor, Michigan, USA

[5]Department of Computer Science, Clark University, Worcester, MA, USA

**Abstract.** Individuals with health concerns increasingly publish videos on video-sharing platforms to present themselves, tell their stories, and connect with supportive peers. Many past studies have analyzed transcripts of such videos, yet, we know little about the role of visual elements in multimodal health support on social media. This study employs mixed methods to analyze 401 YouTube videos by schizophrenia vloggers, examining visual storytelling techniques, the interplay between visual and verbal narration, and their relationship with audience engagement. Guided by Goffman's self-presentation theory and semiotic analysis, our qualitative coding identifies two key dimensions of visual storytelling: (1) visual staging and (2) visual grammar. We then use LLM-assisted annotation to scale these qualitative codes into quantitative variables for regression analysis. As evidence of the importance of understanding visual elements, our regression models incorporating visual features consistently outperformed verbal-only models in predicting engagement metrics. Results show that videos employing richer visual staging and affective grammar received significantly more views, likes, and comments, while verbal narratives were more strongly associated with supportive responses. This study introduces visual staging and grammar as analytical tools for understanding multimodal self-presentation in health communication. The observed visual disparities highlight how platform logics may disadvantage vulnerable creators. The framework offers methodological tools for future research on multimodal communication and raises practical implications for platform design, creator support, and reducing inequalities in the visibility of health-related content.

**Keywords**: Sensitive Disclosure, Video-Sharing Platform, LLM-Assisted Video Analysis, Self-Presentation, Semiotics

## 1 INTRODUCTION

Video-based social media platforms such as YouTube (J. Liu & Zhang, 2024b), TikTok (Schaadhardt et al., 2023), and Instagram (Andalibi, 2017) have become central spaces for digital health communication. Unlike text-based platforms, video environments enable multimodal storytelling that integrates visual, auditory, and performative elements. Individuals with health concerns increasingly use these affordances through vlogs—video blogs that document personal journeys

(Burgess & Green, 2018; Lyu et al., 2024)—to seek peer support and cultivate online communities. Multimodal storytelling can heighten audience immersion and perceived authenticity of illness narratives by combining visual and verbal modes of expression (Huh et al., 2014; J. Liu & Zhang, 2024b).

Despite these affordances, video-mediated health discourse introduces new complexities of visibility that remain underexplored. Visual languages in health communication are entangled with platform logics that may amplify certain narratives while marginalizing others (Li et al., 2024). While health communication research has examined visibility disparities in text-centric contexts (Andalibi, 2016; Joinson, 2001; Y. Wang et al., 2011), the visual dimensions of health discourse have received limited attention. Most existing work on video modality has relied on textual proxies such as transcripts and captions, constraining analyses of visual content to small-scale manual coding (Misoch, 2014; Wu et al., 2025), limiting understanding of how patterns of self-presentation shape audience responses and peer support within video-based health spaces (Milton et al., 2023; Naslund et al., 2014).

Drawing on Goffman's dramaturgical framework (Goffman, 1959), we conceptualize video platforms as performative arenas where narratives of vulnerability, authenticity, and support are enacted in tandem with creation labor (Hu et al., 2023). This study investigates the visual semiotics of mental health discourse on YouTube, examining how multimodal elements structure video communication and shape audience engagement with illness narratives. In this study, we define *visual storytelling* as the crafting of meaning through visual information and *verbal storytelling* as meaning conveyed through spoken language. We focus on schizophrenia vlogs as an empirical case, given schizophrenia's prevalence as a severe mental illness (National Institute of Mental Health, 2023a) and YouTube's prominence as a site of health vlogging and community building (Huh et al., 2014; Milton et al., 2023). This work addresses three interconnected research questions:

- RQ1: What visual storytelling techniques characterize schizophrenia vlogs on YouTube, and how do these techniques function as modes of self-presentation?
- RQ2: How do visual and verbal storytelling elements interact through multimodal narratives?
- RQ3: How do the visual and verbal storytelling influence audience engagement patterns and the generation of supportive community responses?

We conducted a mixed-methods analysis of 401 YouTube videos to address these questions. Drawing on self-presentation theory and semiotics, this study identifies key visual storytelling techniques in mental health vlogs and examines their relationships with verbal narratives and audience engagement. Through novel application of large language models (LLMs) to video content analysis, we advance computational methodologies for multimodal research and critically examine how platform architectures shape the visibility and circulation of mental health narratives in digital spaces.

## 2 BACKGROUND

### 2.1 Promises and Challenges of Video-Based Social Media as Spaces of Peer Support

A growing body of interdisciplinary social science research seeks to understand how individuals use online spaces to disclose personal health information and to seek peer support, with the aim of informing the design of more supportive digital communities. For individuals experiencing health conditions, online platforms provide opportunities to share personal stories that may be difficult to express in offline contexts (Reavley & Jorm, 2011). These acts of disclosure create

digital narratives that not only document lived experiences but also function as pathways for social support (Sangeorzan et al., 2019). Through such narratives, individuals can encounter others who validate their emotions and demonstrate shared understanding of similar struggles (Aldao et al., 2010; Cohen & Wills, 1985).

Vlogging has become a prominent form of video-mediated health communication, offering multimodal affordances for self-disclosure that extend beyond text-based platforms. Through video diaries, individuals document personal health journeys—sharing daily struggles, medication experiences, and coping strategies (Huh et al., 2014; Naslund et al., 2014; Poquet et al., 2018). The integration of visual and auditory cues conveys affective nuances that text alone cannot capture (Misoch, 2014), making vlogs especially powerful for emotion disclosure, a key process in care-seeking and social support (Chaudoir & Fisher, 2010; J. Liu & Zhang, 2024b; Song et al., 2021). By enabling the expression of raw, unfiltered emotions across diverse mental health experiences (L. S. Liu et al., 2013; Woloshyn & Savage, 2020)—including anxiety (Pyle et al., 2021) and depression (Andalibi et al., 2017; De Choudhury et al., 2014)—vlogs foreground visceral vulnerability. Such emotional authenticity fosters empathetic connections and, through interactions with viewers and commenters, contributes to enhanced well-being and sustained peer support (Andalibi et al., 2016; Chaudoir & Fisher, 2010; Green et al., 2015; J. Kim & Lee, 2011).

However, self-presentation on video platforms does not guarantee engagement. Researchers have observed that individuals seeking care and connections may not receive the attention and engagement they need (Postigo, 2016). Lacking audience and peer engagement can exacerbate distress, undermine help-seeking efforts, and limit access to peer support. Thus to to achieve engagement and peer support, online disclosure requires substantial labor, as individuals must invest effort to navigate audience expectations, algorithmic moderation, and platform conventions (Yi & Xian, 2024). Vloggers engage in creative labor (Duffy et al., 2019), involving the ongoing work of crafting visual narratives and managing audience relationships that platforms benefit from while providing minimal support infrastructure.

Video platforms impose additional layers of performative labor beyond the linguistic craft required on text-based social media (Feuston & Piper, 2018, 2019). Even when disclosing sensitive health and identity topics, creators must navigate the tension between authentic self-disclosure and presentation strategies to generate audience engagement and support (Pinch, Birnholtz, et al., 2024; Pinch, Fiers, et al., 2024). The visual and emotional work required to maintain audience attention may not align with creators' actual support needs or recovery processes. Such pressures create hierarchies within digital health communities, where certain visual presentations of mental illness become more successful at generating support than others, potentially marginalizing experiences that do not conform to audience expectations or platform conventions.

However, the potential mediators that influence video platforms remain critically underexplored. To extend health communication studies into video-based social media, this research examines the visual storytelling techniques in mental health vlogs and their influence on audience support behaviors. The findings will inform how video platforms mediate care-seeking and peer support for individuals with severe mental illness, with broader implications for digital health equity in algorithmic environments.

### 2.2 Theoretical foundations of visual storytelling in self-presentation through vlogs

Goffman's dramaturgical framework (Goffman, 1959) conceptualizes social interaction as theatrical performance and has been widely applied to understand video-mediated communication (Chancellor et al., 2017; Wan & Lu, 2024). In this view, social life is a performance where individuals act as "actors" managing impressions—they present and stage themselves to manage social expectations and sustain desired identities or roles. Extending this framework to digital contexts, Hogan (2010) interpreted online platforms as stages for self-presentation, a perspective that has become foundational in studies of social media and increasingly in video-based platforms such as TikTok (Ditchfield & Vicari, 2025) and YouTube (Xiao et al., 2020). We adopt this perspective and view video platforms as performative stages where mental health identities are enacted through the deliberate orchestration of self-presentation.

Videos, as products of self-presentation, thus involve careful alignment of a vlogger's appearance, manner, and setting to construct identities for particular audiences (Goffman, 1959). Scholars have applied the notion of the front stage to analyze online videos and streaming platforms. For example, Wan and Lu (2024) showed how VTuber streamers strategically construct identities through avatar design and performance techniques. "In mental health contexts, there exists a distinctive tension between the desire for authentic self-disclosure and the risks associated with stigmatized experiences (Boyd, 2010). Chancellor et al. (2017) showed how individuals with eating disorders use Instagram images—such as selfies and food photos—to curate narratives of recovery or resistance.

However, Goffman's framework requires extension to address videos' multimodal specificities, such as how creators articulate their verbal narratives and visually stage themselves. In communication research, semiotics provides a framework for analyzing how visual signifiers—symbols, gestures, staging choices, and aesthetic elements—construct meaning beyond verbal narration in video content (Barthes, 1968). These meanings operate on multiple levels of signification: denotative, or the literal visual elements; connotative, or the cultural associations and emotional resonances; and mythological, or the ideological meanings embedded in representations. This multilayered approach is particularly useful for analyzing video content, where meaning emerges through the interplay of visual, verbal, and contextual cues. One recent study applied the semiotic framework to show how mental health stigma is contested through multimodal discourses on Douyin, Chinese TikTok, short videos (Wu et al., 2025). By examining how visual, textual, and auditory elements coalesce to encode stigma or promote anti-stigma narratives, the study reconceptualizes mental health stigma as a visually mediated and morally framed phenomenon. This underscores the value of semiotic analysis for video-mediated health communication and informs our methodological approach.

This study aims to address a methodological limitation in existing research, which has predominantly relied on textual analysis and treated video transcripts as proxies for multimodal content (Doyle & Campbell, 2020), thereby overlooking the rich visual dimensions that distinguish video-based platforms from text-only environments. While recent work has begun addressing these gaps (Wan & Lu, 2024; Wu et al., 2025), the complex relationship between visual self-presentation strategies and their effects on audience engagement and support provision in health contexts remains underexplored. Building on these foundations, the present analysis examines visual storytelling techniques to investigate how visual signifiers and chromatic choices shape meaning-making in mental health vlogs. By disentangling the visual languages embedded in video content, this study illuminates how visual elements mediate self-presentation and audience interpretation in video-mediated communication.

### 2.3 Decoding versatile visual storytelling techniques in videos

While health communication research has extensively documented how linguistic features influence viewers' perception of online messages, emerging work reveals distinct mechanisms through which visual elements shape audience responses in video content (Anjani et al., 2020; Niu et al., 2021).

The video medium affords unique social presence and multimodal self-presentation strategies that extend beyond verbal narration, fostering parasocial interactions that create immediate, visceral connections between creators and viewers (Horton & Richard Wohl, 1956; J. Liu & Zhang, 2024b; Lu, 2019). Creators strategically deploy visual elements as expressive tools: filters and contrast adjustments enable emotional curation—softening harsh realities or intensifying dramatic moments—while camera angles establish relational dynamics with audiences (Ferwerda et al., 2016; Hong et al., 2020). Manikonda and De Choudhury (2017) identified distinct visual patterns across disclosure types, with emotional distress conveyed through muted tones and close-up framing, while raw vulnerability employed stark lighting and direct eye contact (Andalibi, 2017). Creators also innovate storytelling techniques that leverage video's protective affordances. Misoch (2014) documented the "card story" method employed by YouTubers, where creators display handwritten cards containing personal mental health experiences directly to the camera rather than speaking aloud, creating emotional distance while enabling intimate disclosure (Poria et al., 2017).

Specifically, lighting and color serve as prominent visual signifiers in video content, with communication research demonstrating their psychological and emotional effects across traditional television and advertising domains (Elliot & Maier, 2012; Metallinos, 2013; Seckler et al., 2015). Emerging research demonstrates how visual elements carry specific narrative meanings in social media videos. Z. Wang et al. (2025) found that Douyin conspiracy videos exhibit distinct visual signatures—longer durations, lower brightness and entropy, and greater presence of human faces—that differentiate them from mainstream content and signal alternative narratives. Similarly, Wu et al. (2025) identified systematic visual patterns in mental health content: stigmatizing videos employ dark color schemes and high-angle shots that diminish creators, while anti-stigma content features warmer aesthetics and eye-level framing that establishes equality with viewers. These findings highlight how color and visual framing function as key components of meaning-making in video-mediated communication, where technical choices become rhetorical strategies that shape audience interpretation and engagement.

Together, these studies illustrate that visual techniques function as deliberate meaning-making strategies. However, they rarely examine how these techniques influence audience engagement and community support—an area that remains underexplored given that social media self-presentation is often motivated by seeking connection and peer support (Andalibi, 2017). Understanding which visual storytelling strategies effectively foster supportive community responses becomes essential for creators navigating vulnerable disclosure and platforms designing supportive environments. This study therefore uses computational methods to quantify visual storytelling techniques and examine their relationships with audience engagement metrics and expressions of social support.

## 3 METHODS

To answer the research questions, this study started with qualitative analysis of videos that generate categories of visual storytelling; based on the identified categories, we used a novel AI-assisted video annotation to quantify the visual features to prepare regression and statistical testing of visual techniques on engagement and support.

### 3.1 Empirical Setting

We focused on schizophrenia narratives because the condition is prevalent among serious mental health disorders (National Institute of Mental Health, 2023b). Schizophrenia profoundly interferes with daily functioning through symptoms like hallucinations and delusions (National Institute of Mental Health, 2023a), often resulting in social withdrawal (Lee et al., 2022) and urgent needs for safe spaces to express experiences. Online platforms offer unique benefits for people with Schizophrenia, who actively engage in digital communities to exchange coping strategies (Mojtabai & Olfson, 2006) and personal experiences (Reavley & Jorm, 2011), serving as powerful conduits for social support (Cohen & Wills, 1985; Sangeorzan et al., 2019). We collected vlog videos rather than general educational content because vlogs offer rich, first-person accounts that reveal authentic self-presentation strategies and emotional expression patterns essential for our research objectives (Huh et al., 2014).

### 3.2 Data collection

To understand the effects of multimodal storytelling on video engagement, we collected a comprehensive dataset representing the full spectrum of visibility rather than focusing solely on popular videos, which might overlook less visible content. Using the YouTube Data API v3, we employed search queries combining "schizophrenia" with related terms such as "psychosis" and "schizophrenic," paired with content descriptors like "vlog," "vlogging," and "story." We restricted our search to English-language videos of medium length (4-20 minutes) uploaded between 2022 and 2023. Data collection was conducted on June 20, 2024, yielding 555 videos. According to the video ID, we downloaded the videos using the YouTubeDownloader [1] and generated video transcripts using the YouTube Transcript API [2]. We filtered out incomplete data entries (e.g., videos with comments disabled, missing metadata) and institution-created videos. Also, we excluded videos with 0 view count in metadata, resulting in 401 videos for further analysis. The metadata of the video includes the video ID, title, description, channel name, and publish date. The average video duration is 9m35s.

### 3.3 Qualitative Analysis

As described in Section 2.2, we integrated self-presentation theory (Goffman, 1959) and semiotics (Barthes, 1968) to examine how meaning is constructed through the interplay of visual, auditory, and textual elements in YouTube videos. The qualitative analysis unfolded through multiple iterative phases (Corbin & Strauss, 2014), with codes being organized in Excel. The first author began by watching 50% of the video corpus, documenting detailed observations of both spoken and visual elements and viewers' comment interactions. This initial phase revealed emerging patterns while maintaining attention to the multimodal nature of the data. Code development progressed through constant comparative analysis

[1] https://github.com/Tyrrrz/YoutubeDownloader
[2] https://github.com/jdepoix/youtube-transcript-api

(Strauss & Corbin, 1997). To maintain analytical rigor, five members of the team conducted weekly 60-minute collaborative sessions where emerging codes were discussed while viewing selected video segments together.

*3.3.1 Verbal Narrative Analysis.* We analyzed creators' verbal narration by applying thematic analysis to video transcripts alongside the original recordings (Corbin & Strauss, 2014). This examination focused on identifying topics and narrative strategies that characterized these vlogs. For example, initial codes such as "hallucination experiences" and "therapy" were developed and later refined into higher-level themes such as "symptoms," which encompassed various dimensions of disease experiences living with schizophrenia. We also identified the potential associations between verbal disclosure and visual presentation choices. This insight informed the subsequent computational analysis of narrative content. Table 4 lists the codebook of verbal storytelling topics.

*3.3.2 Visual Analysis.* Inspired by previous studies on video-sharing platforms (Niu et al., 2021; Wu et al., 2025), we integrated self-presentation theory for staging analysis and semiotics for frame-level visual analysis. Table 5 lists the codebook of visual storytelling techniques.

Following Goffman's dramaturgical framework (Goffman, 1959), we analyzed how vloggers orchestrate their "performance stages" to understand identity construction strategies in mental health disclosure. We coded for staging approaches to examine how creators employ different presentation modes, as Goffman's framework suggests that performers strategically manage their "front stage" impression through setting, appearance, and manner. Accordingly, we analyzed the staging patterns by examining personal spaces, activities, and social interactions. We defined "visual staging" as the deliberate arrangement and disclosure of environmental contexts within which vloggers position themselves during mental health disclosures. For example, we coded contextual disclosure ranging from blank backgrounds, such as walls and ceilings, to detailed environments, including decorated bookshelves and art creation spaces, to examine how spatial choices function as impression management strategies and connect to audience engagement patterns.

Drawing on Barthes' semiotic framework (Barthes, 1968), we analyzed video frames to decode "visual grammar," which we defined as the use of visual elements—such as color, composition, lighting, and spatial organization—to construct meaning and convey emotions in videos. We sampled one frame every 30 seconds from each video, creating a dataset of 3,800 frame images that captured the visual evolution of each vlog across narrative segments. For instance, the chromatic choices and lighting patterns stood out in the analysis: we noticed that color serves as a powerful semiotic resource for conveying psychological states and contributes to emotional communication beyond verbal narratives. We examined how these visual choices shape meaning-making and audience interpretation, understanding that visual rhetoric operates alongside spoken discourse in video-mediated communication, with creators consciously or unconsciously employing these elements as communication tools.

*3.3.3 Comment and Audience Support Analysis.* While watching the videos, the first author developed open codes for the top 30 comments, noting how commenters responded to the videos. For example, we coded how commenters acknowledged visual settings, appreciated vloggers showcasing daily routine activities, and engaged in reciprocal disclosure. Using thematic analysis to generate open codes, we later organized these codes using the framework for annotating depression-tagged posts on Instagram (Andalibi, 2017). We grouped comments into three overarching

categories closely related to our data: network support, informational support, and affirming behavior. Table 6 lists the categories of supportive behaviors in comments.

### 3.4 AI-Assisted Content Annotation

We operationalized the identified multimodal factors as quantitative variables to examine their relationships with viewership and comments. All variables are listed in Table 1. The following describes details of the AI-assisted video content annotation and measurement.

Table 1. Variables in the Regression Model

| Category | Variable Name | Description | Measurement |
|---|---|---|---|
| Verbal Narration | Emotions, Psychiatry, Relationships, Substance Use | The four variables are based on the emerging themes in qualitative analysis of verbal narration as described in Section 3.3.1. Table 4 includes the definitions and examples of each topic. | The value of each variable is derived by the similarity score using guided BERTopic modeling of video transcripts. |
| Visual Narration | Aesthetics, Colorfulness, Clarity, Brightness | The four variables are based on the emerging themes in qualitative analysis of visual narration as described in Section 3.3.2. Table 5 contains the definitions and examples of each feature. | The value of each visual variable represents the percentage of frames that are annotated as having high corresponding features. |
| Engagement Matrix | #Views, #Likes, #Comments | The numbers of subscribers, views, likes, and comments. | These numbers are from YouTube API 3 |
| Comment Support | Network support, Informational support, Affirming behavior | The categories of comment supports are based on a codebook developed (Andalibi, 2017) and validated (Pater et al., 2016) by previous studies on online self-disclosure of mental health. Table 6 includes the definitions and examples of each type of comment support. | The value of these variables represents the number of comments that include such supports. |

*3.4.1 Verbal Features Annotation.* Based on the developed codebook of verbal narration from the qualitative analysis, for example, relationships, substance use, and psychiatry, we applied topic modeling based on embeddings[3], which yielded the probability of each narration category related to a video. The results helped us group videos of each narration topic. We then leveraged the group's videos to accelerate the evidence searching and synthesizing process, which results are shown in Section 4.2.

To understand the emotions and sentiments in the video transcript, we employed a pre-trained emotion detection model to analyze emotional content in video transcripts. The model, based on DistilRoBERTA and fine-tuned for emotion classification, can identify Ekman's six basic emotions: anger, disgust, fear, joy, sadness, and surprise. For each video transcript, the model computed probability scores across all emotion categories, allowing us to determine the dominant emotional tone of each video. Given the context of schizophrenia vlogs, three emotions emerged as most prevalent and relevant: fear, joy, and sadness. We excluded surprise, disgust, and anger as less contextually relevant to mental health narratives and focused our subsequent analyses on the three primary emotions. To validate model performance,

[3]https://platform.openai.com/docs/guides/embeddings/

we manually annotated emotions for 50 randomly sampled videos. The automated detection achieved 92% accuracy compared to human annotation.

*3.4.2 Visual Narration Annotation.* We examined the visual narration of the videos through two dimensions: framing and visual appeal. Based on the qualitative codebook, the first author went over all the videos and annotated framing manually since our current workflow can not directly capture this video-level feature. For the features related to visual appeal, we employed a human-in-the-loop LLM-assisted keyframe annotation workflow to annotate the visual information (J. ( Liu et al., 2024). We sampled one keyframe every 30 seconds from each video and extracted visual narration features from these keyframes. We employed a multimodal LLM, LLaVa-1.6[4] (an open-sourced model with demonstrated potential for automated video annotation (J. ( Liu et al., 2024; Y. Liu et al., 2024)) for the annotation. Based on the qualitative codebook, we experimented with LLM annotations of 11 features that reflect the visual appeal (e.g., innovative and craftsmanship). We conducted multiple rounds of prompt testing on our dataset (J. ( Liu et al., 2024). Specifically, we first randomly selected 200 sample keyframes and manually annotated them as ground-truth. Then we tested different prompt versions and compared the performance based on the ground-truth. Our results indicated that concise prompts without any term explanation and examples led the model to generate the best results. Using 80% as the accuracy threshold and qualitative categories, we kept four visual appeal features (i.e., aesthetics, cleanliness, colorfulness, and brightness) that are relevant to our research questions for our following quantitative analysis. An example prompt used to annotate the aesthetics of a keyframe was: "Does this picture have an overall aesthetic appeal? Answer 'Yes' or 'No'." The full list of prompts is included in the appendix. Subsequently, we applied the optimized prompts to annotate all the keyframes in our dataset. Based on LLaVa's binary assessment of the keyframes, we could finally quantify the aesthetic features for each video using the mean value of the binary outputs of all the keyframes for each feature. For example, a video getting a score of 0.5 for its brightness means that 50% of the keyframes were assessed as bright by LLaVa.

*3.4.3 Audience Comments Annotation.* Similar to the method in visual narration annotation, we leveraged the human-in-the-loop workflow with GPT-4o mini to find out the optimized prompts to annotate video comments. Using a subset of 200 comments with our manual annotation as ground truth, all the LLM annotation accuracy for the four features was higher than 0.8 (the threshold). An example of our final prompts reads: "Esteem support includes boosting someone's confidence or self-worth, such as praising their efforts or abilities. Does this comment offer esteem support? Respond with Yes or No." This structure combines a clear definition with a binary question, facilitating precise and consistent annotations. The full list of prompts used in our comment annotation process is included in the appendix. We used the number of specific support comments to represent the quantified strength of that feature. For example, a video with 20 network support comments has a score of 20 for the feature of network support. The annotation results were used for the following quantitative analysis.

### 3.5 Quantitative Data Analysis

To understand how visual and verbal storytelling interact (RQ2), we conducted statistical analyses to examine the relationships between visual elements (continuous), narrative topics (categorical), and six DVs (ordinal) in schizophrenia

[4]https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

vlogs. We employed Kruskal-Wallis tests to examine how continuous visual features (aesthetics, colorfulness, brightness, clarity) related to ordinal engagement and support outcomes, chosen because this non-parametric test handles non-normally distributed ordinal data without assuming linear relationships. Chi-square tests of independence examined associations between the four categorical topic clusters from BERTopic and the ordinal dependent variables, revealing whether certain narrative themes co-occur with specific audience response patterns.

To answer RQ3, we fit Multiple Linear Regression (MLR) models on the six target-dependent variables (DVs) related to audience engagement and support. Many of our data are highly skewed and sparse, e.g., the #Views, #Likes, and #Comments as shown in Table 8. To minimize the effect of data skewness on our data analysis, we applied binary binning for independent variables (e.g., Aesthetics) and 5-level binning for dependent variables (e.g., #Views) (Y. Wang et al., 2022; Zafar et al., 2016).

Table 1 lists the measurements of the variables included in the regression. Given a dependent variable (eg, #Views), we first included all the selected factors as main effects and two-way interactions in a regression model. Then we applied the stepwise approach to select the best model by minimizing the Bayesian Information Criterion (BIC) (Burnham & Anderson, 2004). Finally, we conducted Tukey's honestly significant difference test (Tukey's HSD) on the terms of the final regression model. The statistical analysis was assisted with JMP Pro 18 and R.

## 4 RESULTS

We present the qualitative analysis of visual storytelling and audience reactions in Section 4.1. Section 4.2 examines the quantitative correlations between visual and verbal storytelling elements, specifically focusing on topics and emotions in verbal narratives. Finally, Section 4.3 constructs regression models to understand the relationships between these elements and audience engagement and comment supportive behaviors.

### 4.1 RQ1: Visual Storytelling Techniques used by Vloggers

Our qualitative analysis reveals that visual storytelling techniques potentially influence viewer responses to videos, which in turn shape vloggers' perceptions and motivations for continued vlogging.

*4.1.1 Two Dimensions of Visual Storytelling.* Two primary categories of visual storytelling techniques emerged from the qualitative analysis: visual staging and visual grammar. Visual staging encompasses the performance space that vloggers construct; visual grammar refers to aesthetic and technical choices such as color palettes and compositional elements. Together, these dimensions reveal how vloggers strategically orchestrate both what they show and how they show it to construct mental health narratives. The following subsections examine each dimension in detail.

*Visual Staging as a Tool for Authenticity Construction.* Staging emerged as a key mechanism for constructing perceived authenticity. Schizophrenia vloggers employed approaches ranging from minimal, static contexts to rich settings that integrate environment and self-presentation (Figure 1). We identified two main formats: *talk-to-camera* staging, where vloggers address the audience directly in a static environment, and *in-the-moment* staging, which situates disclosure within real-time surroundings and daily activities.

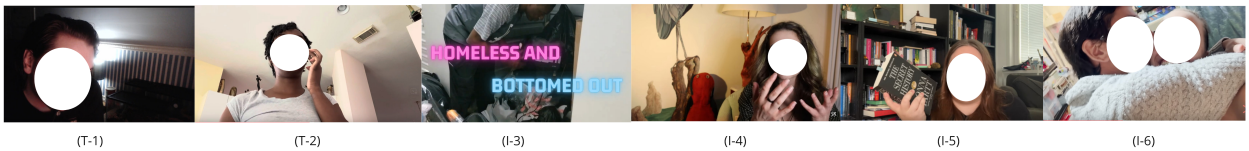



Fig. 1. Examples of different visual staging approaches in mental health vlogs.

In talk-to-camera staging, vloggers often position themselves against plain backgrounds such as walls (T-1) or ceilings (T-2), with their faces dominating the frame and surrounding contexts largely absent. This style produces a confessional atmosphere that emphasizes direct verbal delivery while minimizing environmental cues. By contrast, vloggers using in-the-moment staging (I-3 to I-6) adopt richer approaches, actively showcasing personal spaces, outdoor settings, and background details, and integrating these environments into their verbal narratives. For instance, the vlogger in I-5 discussed her recent reading as part of a coping strategy, situating the disclosure within her personal space, while the vlogger in I-6 emphasized family support by framing her story in a shared domestic environment.

It seems that videos that employ in-the-moment staging create enhanced opportunities for viewer interaction and community building, as evidenced by the comments their videos receive. Viewers expressed appreciation for contextual details, with many commenters responding to vlogger (I-5)'s book-filled background by asking about "*the books on the shelf and collection of CDs*." In-the-moment staging elicits many affirmative and supportive comments related to how vloggers visually "show" specific activities, creating deeper disclosure opportunities that extend beyond emotional expression to include relatable experiences that viewers can connect with. For example, one vlogger consistently presented his artwork across multiple videos, which generated affirmative comments celebrating artistic expression as a coping mechanism for schizophrenia. This visual disclosure prompted one commenter to share their own creative journey: "*It's definitely what I do when I hope that people who don't have schizophrenia take away from them... it's nice to draw pictures and do artwork and express myself. In 2019 when I started my TikTok, I was expressing myself because nobody would listen and people treated me weird.*" This example illustrates how strategic visual staging can foster community formation around shared experiences and coping strategies.

*Visual Grammar as Affective and Stylistic Choices.* Visual grammar refers to the structured use of visual elements to shape meaning and emotional tone in videos. Guided by color theory and semiotics (Metallinos, 2013; Seckler et al., 2015), our analysis focuses on four key attributes summarized in Table 5: chromatic elements (brightness, colorfulness) and compositional features (clarity, aesthetics). As shown in Figure 2, vlogs vary widely in their visual grammar. These choices shape both the emotional tone of videos and their alignment with verbal storytelling.


Fig. 2. Examples of different visual grammars.

The choices of visual grammar appear to be closely tied to the emotional tone of the videos and to the verbal storytelling. Low-brightness, muted palettes—such as example (a)—create shadowy, introspective atmospheres that often accompany

vulnerable disclosures, for instance the vlogger quietly recounting a depressive episode. Similar subdued settings appear in (c) and (e), where strained family relationships are discussed in neutral light that conveys heaviness and reflection. By contrast, high-brightness, high-colorfulness presentations—such as (b) and (d)—reinforce more optimistic narratives. In (b), the vlogger discusses his growing ability to maintain social connections while walking through a brightly lit outdoor space and saying, "*I feel that I am doing better compared to last week. I recently discovered this park... Getting close to nature is really good for me.*" The vibrant colors and sunny skies in the videos reinforce a sense of progress and sociability. In (d), warm saturated tones transform a cooking scene into a hopeful narrative of everyday resilience. Likewise, vloggers in (f)–(h) situate recovery stories in lively, colorful spaces that signal vitality and openness.

At the same time, our analysis shows that visual grammar reflects not only production quality but also each vlogger's stylistic choices. One vlogger, for instance, consistently filmed unedited close-ups while lying on a sofa, presenting his videos as personal journals and signaling a sense of intimacy through the raw style. In contrast, the vlogger in (h) maintained high brightness and carefully composed aesthetics even while crying during a discussion of severe hallucinations, illustrating how polished visuals can themselves become a deliberate mode of self-presentation. While individual stylistic choices may be made either deliberately or unconsciously, these attributes nonetheless shape the audience's viewing experience and interpretation of narratives. For example, in general, vlogs such as (f) and (h), which display clear visuals, clean backgrounds, steady lighting, and a sense of visual harmony, seem to reinforce narrative coherence, compared to vlogs like (e) and (g), with cluttered settings and shaky framing. The quantitative correlations between these visual grammar features and audience engagement are presented in Section 4.2.

#### *4.1.2 Visual Storytelling and Audience Engagement Dynamics.*

*Positive engagement creates supportive feedback Loops.* We observed a positive feedback loop between vloggers' storytelling and commenters' affirmation, reciprocal self-disclosure, and support that fosters the vlogger-viewer bond. The vloggers post videos, as described "*I just want to thank everyone who watched this. I hope this is enlightening or helpful to somebody.*" In turn, the supportive comments can enhance a vlogger's sense of self-worth and happiness. For example, one vlog appreciated the "*All the amazing comments I'm getting are boosting my self-esteem, and that part is making me very happy. I have been feeling much happier than I ever have since the onset of my schizophrenia.*"

*Negative impact of low audience engagement.* Vloggers seek interactions with their audiences, and we observe potential harms of low engagement (i.e., low #Views, #Likes, and #Comments). Some vloggers expressed feelings of self-doubt and inadequacy when their content failed to attract viewers or bring comments. For instance, one vlogger posted 15 videos, but the average view number is below 10. Despite maintaining a consistent routine of content creation, he felt unfulfilled by the process due to consistently low viewership. He explained, "*making videos stopped making me feel very fulfilled, as when I started doing these in the last 16 days.*" He describes feeling "*pathetic*" about making videos and stopped posting videos. These experiences highlight the emotional toll that low audience engagement can have on schizophrenia vloggers.

### 4.2 RQ2: Relationships between Verbal and Visual Storytelling Elements

We investigated associations between verbal content and visual storytelling techniques, specifically examining how visual elements relate to topics and sentiments in verbal narratives.

*4.2.1 Visual Staging Associates with Verbal Sentiments.* We examined the relationship between visual staging techniques and emotional expression in vlog transcripts using pair-wise chi-square tests. Results revealed that in-the-moment staging was significantly more prevalent in videos expressing joy compared to those expressing sadness (p=.0027) and fear (p=.0003). No significant difference was found between sadness and fear groups (p=.7702).

*4.2.2 Visual Grammar Associates with Verbal Sentiments.* We conducted ANOVA and Tukey's HSD tests to examine relationships between visual grammar and sentiments. Statistically significant comparisons are illustrated in Figure 3. Videos expressing joy demonstrated significantly higher levels of aesthetics, colorfulness, and brightness compared to videos expressing sadness and fear. However, no differences were observed in video clarity across emotional categories. Between the two negative emotions (fear and sadness), no significant differences were found across any visual grammar features. These findings demonstrate that high-key lighting and vibrant colors typically align with joyful content, while low-key, subdued visuals characterize videos expressing fear and sadness.

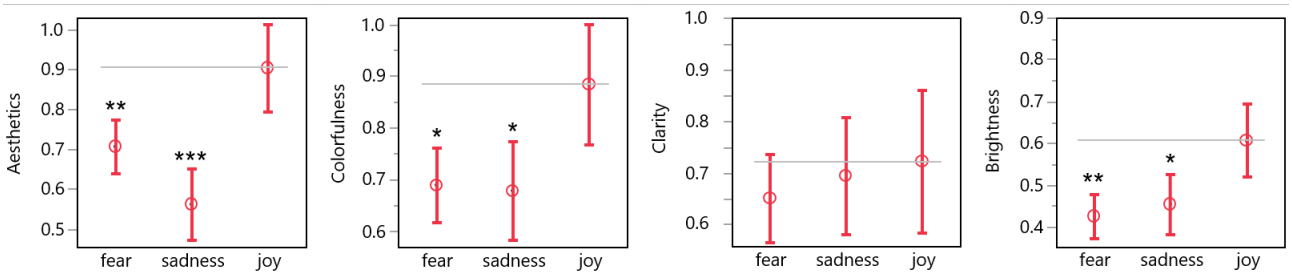



Fig. 3. The comparisons of visual appeal features in videos expressing different emotions. The significance levels shown in sub-figures are between joy and other emotion groups. * indicates p<.05, ** indicates p<.01, *** indicates p<.001.

### 4.3 RQ3: Regression of Verbal and Visual Storytelling on Audience Engagement and Comment Support

We examined the quantitative relationships between verbal narratives, visual elements, and audience responses to identify patterns in multimodal communication. Using regression analysis, we investigated how visual staging and visual grammar correlate with engagement metrics (views, likes, comments) and comment-based social support (network, informational, and affirming support). The analysis reveals three key findings: First, visual elements demonstrate independent effects on engagement beyond verbal content, suggesting that how mental health narratives are visually presented matters as much as what is verbally disclosed. Second, specific visual techniques—particularly aesthetics and colorfulness—show stronger associations with engagement than others. Third, the relationships between visual elements and different types of social support vary, suggesting that visual choices may selectively attract specific forms of audience response. These findings illuminate how visual storytelling shapes both the visibility and reception of mental health narratives on video platforms.

Specifically, for each dependent variable, two models were constructed: Model 1 included only verbal features, while Model 2 incorporated both verbal and visual features. All models controlled for the number of subscribers. The regression results are summarized in Table 2 for the three audience engagement metrics (Views, Likes, Comments) and in Table 3 for the three types of supportive comments (network, informational, and affirming support). Overall, Model 2 consistently outperformed Model 1 across all six outcome variables. The improvement in explanatory power was particularly evident in predicting engagement metrics, compared to comment support features.

Table 2. Multiple Linear Regression Models - Engagement Metrics

| | View Count | | Like Count | | Comment Count | |
|---|---|---|---|---|---|---|
| | Model 1 (1) | Model 2 (2) | Model 1 (3) | Model 2 (4) | Model 1 (5) | Model 2 (6) |
| Subscriber Count | **0.576**$^{***}$ (0.039) | **0.484**$^{***}$ (0.036) | **0.671**$^{***}$ (0.036) | **0.575**$^{***}$ (0.033) | **0.738**$^{***}$ (0.038) | **0.649**$^{***}$ (0.035) |
| Emotions | **-0.689**$^{***}$ (0.117) | **-0.539**$^{***}$ (0.105) | **-0.453**$^{***}$ (0.108) | **-0.318**$^{**}$ (0.096) | **-0.585**$^{***}$ (0.113) | **-0.440**$^{***}$ (0.102) |
| Relationships | 0.042 (0.112) | 0.071 (0.099) | **-0.239**$^{*}$ (0.103) | -0.165 (0.090) | -0.190 (0.107) | -0.140 (0.096) |
| Substance Use | **0.362**$^{**}$ (0.110) | **0.292**$^{**}$ (0.097) | **0.302**$^{**}$ (0.101) | **0.199**$^{*}$ (0.089) | **0.407**$^{***}$ (0.106) | **0.325**$^{***}$ (0.094) |
| Psychiatry | **-0.263**$^{*}$ (0.116) | -0.145 (0.103) | -0.156 (0.107) | -0.043 (0.094) | -0.153 (0.112) | -0.044 (0.100) |
| Aesthetics | | **1.231**$^{***}$ (0.166) | | **1.135**$^{***}$ (0.151) | | **1.192**$^{***}$ (0.161) |
| Colorfulness | | 0.172 (0.150) | | **0.532**$^{***}$ (0.137) | | **0.314**$^{*}$ (0.146) |
| Brightness | | **0.562**$^{***}$ (0.140) | | **0.259**$^{*}$ (0.128) | | **0.371**$^{**}$ (0.135) |
| Clarity | | -0.044 (0.158) | | 0.039 (0.145) | | -0.007 (0.154) |
| Constant | **1.505**$^{***}$ (0.168) | 0.363 (0.203) | **1.224**$^{***}$ (0.156) | -0.025 (0.186) | **0.434**$^{**}$ (0.162) | **-0.736**$^{***}$ (0.197) |
| $R^2$ | 0.416 | 0.560 | 0.509 | 0.638 | 0.539 | 0.647 |
| Adjusted $R^2$ | 0.409 | 0.550 | 0.503 | 0.630 | 0.533 | 0.639 |
| F Statistic | **56.294**$^{***}$ (df = 5; 395) | **55.324**$^{***}$ (df = 9; 391) | **81.810**$^{***}$ (df = 5; 395) | **76.526**$^{***}$ (df = 9; 391) | **92.224**$^{***}$ (df = 5; 395) | **79.558**$^{***}$ (df = 9; 391) |
| *Model Comparison Statistics* | | | | | | |
| $\Delta R^2$ | | **0.144**$^{***}$ | | **0.129**$^{***}$ | | **0.108**$^{***}$ |
| F-test (nested models) | | **25.34**$^{***}$ (df = 4; 391) | | **22.87**$^{***}$ (df = 4; 391) | | **19.76**$^{***}$ (df = 4; 391) |

*Note:* $^{***}$ p<0.001, $^{**}$ p<0.01, $^{*}$ p<0.05. Standard errors in parentheses.
Model 1: Verbal only; Model 2: Verbal + Visual variables.

*Certain Visual Grammars Elicit More Views, Likes, and Comments.* Aesthetics consistently emerged as the strongest visual predictor across all engagement metrics. Videos with higher aesthetic quality attracted significantly more views ($\beta$ = 1.231, p<0.001), likes ($\beta$ = 1.135, p<0.001), and comments ($\beta$ = 1.192, p<0.001). This finding suggests that visual appeal serves as a fundamental gateway for audience attention and interaction. Brightness also demonstrated positive effects across engagement metrics, with particularly strong effects on view count ($\beta$ = 0.562, p<0.001) and moderate effects on likes ($\beta$ = 0.259, p<0.05) and comments ($\beta$ = 0.371, p<0.01). Colorfulness showed more selective effects, significantly influencing likes ($\beta$ = 0.532, p<0.001) and comments ($\beta$ = 0.314, p<0.05) but not views, suggesting different visual elements may serve distinct functions in the engagement process.

Table 3. Multiple Linear Regression Models - Support and Behavior Metrics

| | Network Support | | Informational Support | | Affirming Behavior | |
|---|---|---|---|---|---|---|
| | Model 1 (7) | Model 2 (8) | Model 1 (9) | Model 2 (10) | Model 1 (11) | Model 2 (12) |
| Subscriber Count | **0.244**$^{***}$ | **0.219**$^{***}$ | **0.327**$^{***}$ | **0.290**$^{***}$ | **0.374**$^{***}$ | **0.325**$^{***}$ |
| | (0.022) | (0.022) | (0.022) | (0.022) | (0.022) | (0.020) |
| Emotions | -0.037 | 0.011 | **-0.284**$^{***}$ | **-0.228**$^{***}$ | **-0.257**$^{***}$ | **-0.181**$^{**}$ |
| | (0.065) | (0.065) | (0.065) | (0.063) | (0.064) | (0.059) |
| Relationships | -0.071 | -0.048 | **-0.136**$^{*}$ | -0.114 | -0.113 | -0.081 |
| | (0.062) | (0.062) | (0.062) | (0.060) | (0.061) | (0.056) |
| Substance Use | **0.313**$^{***}$ | **0.285**$^{***}$ | **0.268**$^{***}$ | **0.233**$^{***}$ | **0.186**$^{**}$ | **0.138**$^{*}$ |
| | (0.061) | (0.060) | (0.061) | (0.058) | (0.060) | (0.054) |
| Psychiatry | -0.061 | -0.035 | 0.058 | 0.094 | -0.067 | -0.013 |
| | (0.065) | (0.064) | (0.064) | (0.062) | (0.064) | (0.058) |
| Aesthetics | | **0.381**$^{***}$ | | **0.457**$^{***}$ | | **0.628**$^{***}$ |
| | | (0.103) | | (0.100) | | (0.093) |
| Colorfulness | | 0.120 | | 0.131 | | **0.204**$^{*}$ |
| | | (0.094) | | (0.090) | | (0.085) |
| Brightness | | 0.018 | | 0.113 | | 0.145 |
| | | (0.087) | | (0.084) | | (0.078) |
| Clarity | | 0.034 | | 0.095 | | 0.069 |
| | | (0.099) | | (0.095) | | (0.089) |
| Constant | **0.556**$^{***}$ | 0.173 | **0.585**$^{***}$ | 0.081 | **0.696**$^{***}$ | 0.027 |
| | (0.094) | (0.126) | (0.093) | (0.122) | (0.092) | (0.114) |
| $R^2$ | 0.281 | 0.323 | 0.422 | 0.488 | 0.476 | 0.585 |
| Adjusted $R^2$ | 0.272 | 0.308 | 0.415 | 0.476 | 0.469 | 0.575 |
| F Statistic | **30.844**$^{***}$ | **20.754**$^{***}$ | **57.769**$^{***}$ | **41.395**$^{***}$ | **71.624**$^{***}$ | **61.153**$^{***}$ |
| | (df = 5; 395) | (df = 9; 391) | (df = 5; 395) | (df = 9; 391) | (df = 5; 395) | (df = 9; 391) |
| *Model Comparison Statistics* | | | | | | |
| $\Delta R^2$ | | 0.042$^{**}$ | | 0.066$^{***}$ | | 0.109$^{***}$ |
| F-test (nested models) | | 5.82$^{**}$ | | 9.48$^{***}$ | | 20.84$^{***}$ |
| | | (df = 4; 391) | | (df = 4; 391) | | (df = 4; 391) |

*Note:* $^{***}$p<0.001; $^{**}$p<0.01; $^{*}$p<0.05
Model 1: Topics + Subscriber Count; Model 2: Topics + Subscriber Count + Visual variables

Visual elements influenced social support in distinct ways. Aesthetics consistently predicted all three types of support—network ($\beta$ = 0.381, p < .001), informational ($\beta$ = 0.457, p < .001), and affirming behavior ($\beta$ = 0.628, p < .001)—with the strongest effect observed for affirming responses. Colorfulness was uniquely associated with affirming behavior ($\beta$ = 0.204, p < .05), suggesting that visually vibrant content may foster positive emotional engagement. Other visual features, such as brightness and clarity, showed no significant effects.

*Verbal Topics Elicit Distinct Engagement Patterns and Support Responses.* Verbal narratives demonstrated nuanced relationships with engagement metrics. Direct emotion expression consistently reduced engagement across all metrics (views: $\beta$ = -0.539, p<0.001; likes: $\beta$ = -0.318, p<0.01; comments: $\beta$ = -0.440, p<0.001), suggesting that explicit emotional disclosure may create barriers to audience engagement. In contrast, substance use narratives positively influenced all engagement metrics (views: $\beta$ = 0.292, p<0.01; likes: $\beta$ = 0.199, p<0.05; comments: $\beta$ = 0.325, p<0.001), indicating audience interest in these specific mental health experiences. Relationship-focused content showed mixed effects, negatively affecting likes in the verbal-only model ($\beta$ = -0.239, p<0.05) but becoming non-significant when visual elements were included.

Different verbal narratives elicited distinct forms of social support, particularly around mental health topics. Substance use disclosures prompted the most comprehensive response—positively predicting network support ($\beta$ = 0.285, p < .001), informational support ($\beta$ = 0.233, p < .001), and affirming behavior ($\beta$ = 0.138, p < .05). This suggests that such content

mobilizes community engagement, aligning with the role of peer support in recovery contexts. Informational responses were especially strong for videos that included detailed questions about treatment, prompting viewers to share practical advice. In contrast, direct emotional expressions (e.g., "I have been sad lately") were associated with reduced informational support ($\beta$ = −0.228, p < .001) and affirming behavior ( $\beta$ = −0.181, p < .01), with no significant effect on network support.

*Aesthetics Are Important for Audience Engagement for Vloggers with Low #Subscribers.* Besides the main effects, we also observed interesting interactive effects between #subscriber and aesthetics on audience engagement. As shown in

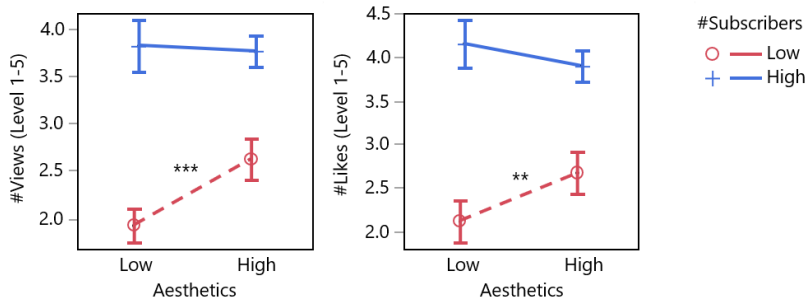



Fig. 4. Interaction Effects of #Subscriber and Aesthetics on #Views and #Likes. Dotted lines indicate statistically significant differences (** indicates p<.01, *** indicates p<.001); solid lines indicate no significance at p>.05. Error bars indicate a 90% confidence interval.

Figure 4, the aesthetics of videos play a more important role in getting audience engagement for vloggers with fewer #Subscribers, for example, vloggers who are new to YouTube. After vloggers have built a stable audience base and maintained a relationship with their followers, the visual narration tends to be less important.

## 5 DISCUSSION

### 5.1 Visual storytelling techniques on video sharing platforms

Our analysis identifies two interrelated dimensions of visual storytelling in mental health disclosure videos: *visual staging* and *visual grammar*. Drawing on theories of self-presentation and semiotics, these dimensions formalize how creators communicate authenticity and emotions through visual means. This contributes to growing scholarship that decodes the visual language of online video, which has been shown to shape discourses of credibility (Peng et al., 2023), conspiracy (Z. Wang et al., 2025), and stigma (Wu et al., 2025).

Visual staging functions as environmental rhetoric that transforms intimate spaces into communicative resources. Our findings resonate with broader studies of visual self-presentation in stigmatized domains such as sexuality, aggression, and personal failure (Farber, 2003; Park & Lee, 2021), showing how staging strategies can signal vulnerability while managing audience trust. Schizophrenia vloggers foreground bedrooms, personal objects, and daily routines not as incidental backdrops but as semiotic evidence of authenticity. This extends prior work on parasocial connection (Niu et al., 2021), demonstrating how environmental cues invite viewers into forms of mediated intimacy otherwise unavailable in text-based disclosure.

Visual grammar operates as an affective coding system, shaping the emotional resonance of verbal narratives. Through choices in color palettes, lighting, and framing, creators construct atmospheres that amplify, contradict, or nuance spoken content. Prior studies similarly demonstrate how visual patterns communicate affect: Manikonda and De Choudhury (2017)

identified that emotional distress is often conveyed through muted tones and close-up framing, while raw vulnerability is marked by stark lighting (Andalibi, 2017). We propose a taxonomy with two primary axes: (1) *chromatic elements* such as brightness and saturation, and (2) *compositional features* such as clarity and aesthetics. These visual codes extend semiotic readings of multimodal discourse, where meaning is not merely transmitted but enacted through the interplay of visual and verbal signs.

One potential reason of the divide of engagement of different visual storytelling is that visual function as heuristics (Sundar, 2008) for attention and credibility. Prior research on health video consumption (J. Liu & Zhang, 2024b) demonstrates that judgments of authenticity and production quality are central to engagement decisions. Thus, staging, and editing choices may invite rapid assessments before verbal content is processed. Thus, visual storytelling becomes a primary site where authenticity is negotiated and where disclosure oscillates between candid expression and strategic performance (Pinch, Fiers, et al., 2024).

By formalizing visual staging and grammar as dimensions of mental health disclosure, this study offers analytical tools for examining multimodal self-presentation online. These findings extend self-presentation (Goffman, 1959) and semiotic (Barthes, 1968) approaches by showing how visual cues operate alongside verbal narratives in sensitive disclosure. Future work can apply this taxonomy to explore audience engagement in health communication and other domains.

### 5.2 Visual Disparity, Creation Labor, and Video-Sharing Platforms

Prior studies of self-presentation in gaming (Meriläinen & Ruotsalainen, 2024), beauty (Foster & Baker, 2022), and travel (Arthur, 2022) highlight how visual strategies function as tools of monetization and revenues (Bishop, 2018), directly shaping engagement. However, our study shows that even in less commodified communities of sensitive disclosure and care-seeking, engagement is strongly conditioned by visual language. This raises concerns about the extent to which video-sharing platforms can serve as equitable spaces for online support.

Our quantitative analysis uncovers a concerning "visual disparity,"- videos that use certain visual languages tend to get more audience engagement and support. Regression models in Section 4.3 reveal that videos filmed in visually dynamic, “in-the-moment” settings—such as walking outdoors—received significantly higher engagement than static close-ups against plain backdrops. Similarly, brighter palettes and emotionally evocative framing consistently amplified audience response. These findings indicate that, when verbal content is similar in topics, audience judgments are shaped by visible signals. In other words, visual storytelling choices—color, setting, and composition—do not merely accompany verbal content but strongly influence its reception.

This study contributes to research on platform affordances by highlighting how visual disparities operate on YouTube, a content-generation platform in which visibility is structured through discovery rather than pre-existing ties (Kaplan & Haenlein, 2010). In contrast to support-seeking on social networking sites such as Facebook or Instagram, where disclosure circulates within established social relationships (Kumar, 2014), discovery-based platforms like YouTube rely heavily on visual presentation as a gatekeeping mechanism (Kaplan & Haenlein, 2010). This dynamic disadvantages vloggers with smaller subscriber bases, for whom polished visuals can partly compensate for limited reach. Without such production, creators risk invisibility. As Section 4.3 shows, repeated disappointments with low engagement led some vloggers to stop posting altogether, exacerbating emotional distress (Johnson, 2019).

In this context, our findings underscore the uneven creation labor demanded of vulnerable communities on general-purpose video-sharing platforms. Producing visually appealing content requires not only technical skill and editing expertise (Huh et al., 2014) but also significant emotional investment (Crilley & Chatterje-Doody, 2020). For individuals managing chronic conditions, this burden can be prohibitive. Schizophrenia vloggers, for example, may lack the energy to design elaborate staging or sustain intensive editing, leaving them structurally disadvantaged in gaining visibility (National Institute of Mental Health, 2023a). Moreover, certain forms of disclosure—such as videos addressing sadness or personal failure—are systematically marginalized because they deviate from the polished, affectively bright visual scripts rewarded by the platform.

Taken together, these dynamics suggest that algorithmic and audience preferences for specific visual languages create a cycle in which those most in need of recognition and support often receive the least engagement, compounding both emotional and social marginalization.

### 5.3 Computational Approach to Studies of Online Videos

This study makes a methodological contribution by applying AI-assisted video content analysis to operationalize a theory-informed codebook of visual storytelling techniques at scale. We extracted 3,800 frames from 401 schizophrenia vlogs and employed multimodal LLMs to annotate visual features, including staging (e.g., talk-to-camera vs. in-the-moment filming) and visual grammar choices (e.g., color schemes, aesthetics). This automated annotation enables systematic, replicable coding of visual languages, moving beyond small-scale qualitative analyses that have dominated prior work. Our workflow integrates frame-level sampling, multimodal annotation, and regression-based analysis to capture links between visual languages and audience response.

The multimodal framework allows us to identify how verbal and visual elements serve distinct communicative functions in videos. As Section 4.3 shows, visual features broadly shape engagement metrics such as views, likes, and comments, but exhibit weaker associations with supportive behaviors—including informational, network, and affirming support. By contrast, verbal narratives are closely tied to peer support outcomes. For example, the regression results in Table 3 show that vloggers who shared coping strategies or detailed personal experiences elicited more informational and network support in viewer comments. This may suggest that audience interactions unfold progressively: while visual features capture and sustain attention, verbal disclosure is more closely related to specific types of supportive behaviors.

Recent studies have begun to explore the potential of large language models for analyzing visual information at scale. For example, researchers have employed LLMs to annotate credibility cues in online images (Peng et al., 2023, 2025), support journalists in detecting visual misinformation (S. J. Kim et al., 2025), and identify audiovisual and thematic markers of unsafe or sensational content in short videos (Xue, 2024; Xue et al., 2025). Our workflow echos with these emerging efforts by formalizing a theory-informed codebook of visual storytelling features and demonstrating how multimodal LLMs can systematically annotate and analyze videos at scale. This approach not only improves the scalability and consistency of annotation (J. ( Liu et al., 2024; Ziems et al., 2024) but also strengthens the interpretability of computational analyses by grounding feature discovery in communication theory.

Beyond our study context, future research can adapt the workflow to investigate other forms of online disclosure, compare visual strategies across long- and short-form video platforms, such as TikTok and Instagram, or track changes in

visual storytelling practices over time. Such applications can deepen understanding of how multimodal communication shapes expression, support, and community formation across diverse digital contexts.

### 5.4 Limitations, Implications and Future Work

This study has several limitations. The categorical classification of emotions through computational methods relies on video transcripts, which may not capture the full emotional nuance of the content. We attempted to mitigate these limitations by contextualizing vloggers' disclosed emotions through qualitative analysis of the videos. To provide a concise presentation of results, we did not include findings about audio features in this submission. Additionally, BertTopic uses a pre-trained language model, so clustering may reflect prior associations embedded in the training data rather than patterns derived specifically from our study context.

Despite these limitations, our findings highlight the impediments of current engagement-centered algorithmic systems in supporting vulnerable communities. This study builds upon growing research examining marginalization and injustice inherent in platform-mediated illness construction (Feuston & Piper, 2019; Pendse et al., 2023) and serves as a cautionary note to platform researchers and designers creating more inclusive environments for technology-mediated care-seeking (Chordia et al., 2024; Pendse et al., 2022). While maximizing views and watch time serves platform business models, it may systematically disadvantage creators who lack high-production resources but offer valuable peer support and authentic experiences. One promising direction involves curating specialized spaces where algorithms prioritize support-seeking intentions (J. Liu & Zhang, 2024a; Pyle et al., 2021) over entertainment value, boosting visibility based on community support potential rather than conventional engagement metrics. Human-computer interaction research could also develop Generative AI tools (Y. Wang et al., 2023) that simplify visual editing for individuals with mental illness, reducing the substantial labor barriers to creating disclosure videos.

## 6 CONCLUSION

In conclusion, this study sheds light on how individuals with severe mental illnesses, particularly schizophrenia, leverage video blogging as a medium for emotional disclosure and support-seeking. We observed various types of relationships between verbal and visual narration features with audience engagement, among which an imperative finding is the 'visual appeal disparity' in audience engagement. Future research should explore ways to mitigate the visual appeal disparity and investigate how these findings might apply to other mental health conditions and online platforms.

## REFERENCES

Aldao, A., Nolen-Hoeksema, S., & Schweizer, S. (2010). Emotion-regulation strategies across psychopathology: A meta-analytic review. *Clinical Psychology Review*, *30*(2), 217–237. https://doi.org/10.1016/j.cpr.2009.11.004

Andalibi, N. (2016). Social Media for Sensitive Disclosures and Social Support: The Case of Miscarriage. *Proceedings of the 19th International Conference on Supporting Group Work*, 461–465. https://doi.org/10.1145/2957276.2997019

Andalibi, N. (2017). Self-disclosure and Response Behaviors in Socially Stigmatized Contexts on Social Media: The Case of Miscarriage. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 248–253. https://doi.org/10.1145/3027063.3027137

Andalibi, N., Haimson, O. L., De Choudhury, M., & Forte, A. (2016). Understanding Social Media Disclosures of Sexual Abuse Through the Lenses of Support Seeking and Anonymity. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3906–3918. https://doi.org/10.1145/2858036.2858096

Andalibi, N., Ozturk, P., & Forte, A. (2017). Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of #Depression. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1485–1500. https://doi.org/10.1145/2998181.2998243

Anjani, L., Mok, T., Tang, A., Oehlberg, L., & Goh, W. B. (2020). Why do people watch others eat food? An Empirical Study on the Motivations and Practices of Mukbang Viewers. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376567

Arthur, T. O. (2022). "we bring home the roots": Black women travel influencers, digital culture bearing, and african internationalism in instagram. *Social Media+ Society*, *8*(2), 20563051221103843.

Barthes, R. (1968). *Elements of Semiology* [Google-Books-ID: OVJhOA6iWxEC]. Macmillan.

Bishop, S. (2018). Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm [Publisher: SAGE Publications Ltd]. *Convergence*, *24*(1), 69–84. https://doi.org/10.1177/1354856517736978

Boyd, D. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In *A networked self* (pp. 47–66). Routledge.

Burgess, J., & Green, J. (2018, August). *YouTube: Online Video and Participatory Culture* [Google-Books-ID: mg1rDwAAQBAJ]. John Wiley & Sons.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological methods & research*, *33*(2), 261–304.

Chancellor, S., Kalantidis, Y., Pater, J. A., De Choudhury, M., & Shamma, D. A. (2017). Multimodal Classification of Moderated Online Pro-Eating Disorder Content. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3213–3226. https://doi.org/10.1145/3025453.3025985

Chaudoir, S. R., & Fisher, J. D. (2010). The disclosure processes model: Understanding disclosure decision-making and post-disclosure outcomes among people living with a concealable stigmatized identity. *Psychological bulletin*, *136*(2), 236–256. https://doi.org/10.1037/a0018193

Chordia, I., Baltaxe-Admony, L. B., Boone, A., Sheehan, A., Dombrowski, L., Le Dantec, C. A., Ringland, K. E., & Smith, A. D. R. (2024). Social Justice in HCI: A Systematic Literature Review. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–33. https://doi.org/10.1145/3613904.3642704

Cohen, S., & Wills, T. A. (1985). Stress, social support, and the buffering hypothesis [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *98*(2), 310–357. https://doi.org/10.1037/0033-2909.98.2.310

Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.

Crilley, R., & Chatterje-Doody, P. N. (2020). Emotions and war on YouTube: Affective investments in RT's visual narratives of the conflict in Syria [Publisher: Routledge _eprint: https://doi.org/10.1080/09557571.2020.1719038]. *Cambridge Review of International Affairs*, *33*(5), 713–733. https://doi.org/10.1080/09557571.2020.1719038

De Choudhury, M., Morris, M. R., & White, R. W. (2014). Seeking and sharing health information online: Comparing search engines and social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1365–1376. https://doi.org/10.1145/2556288.2557214

Ditchfield, H., & Vicari, S. (2025). Identity roles and sociality on tiktok: Performance in hereditary cancer content (# brca and# lynchsyndrome). *Social Media+ Society*, *11*(2), 20563051251340862.

Doyle, P. C., & Campbell, W. K. (2020). *Linguistic markers of self-disclosure: Using YouTube coming out videos to study disclosure language* [Publisher: PsyArXiv]. Retrieved January 17, 2024, from https://psyarxiv.com/tvgs9/download?format=pdf

Duffy, B. E., Poell, T., & Nieborg, D. B. (2019). Platform practices in the cultural industries: Creativity, labor, and citizenship. *Social media+ society*, *5*(4), 2056305119879672.

Elliot, A. J., & Maier, M. A. (2012). Color-in-context theory. In *Advances in experimental social psychology* (pp. 61–125, Vol. 45). Elsevier.

Farber, B. A. (2003). Patient self-disclosure: A review of the research [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jclp.10161]. *Journal of Clinical Psychology*, *59*(5), 589–600. https://doi.org/10.1002/jclp.10161

Ferwerda, B., Schedl, M., & Tkalcic, M. (2016). Using Instagram Picture Features to Predict Users' Personality. In Q. Tian, N. Sebe, G.-J. Qi, B. Huet, R. Hong, & X. Liu (Eds.), *MultiMedia Modeling* (pp. 850–861). Springer International Publishing. https://doi.org/10.1007/978-3-319-27671-7_71

Feuston, J. L., & Piper, A. M. (2018). Beyond the Coded Gaze: Analyzing Expression of Mental Health and Illness on Instagram [Number: CSCW]. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–21. https://doi.org/10.1145/3274320

Feuston, J. L., & Piper, A. M. (2019). Everyday Experiences: Small Stories and Mental Illness on Instagram. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300495

Foster, J., & Baker, J. (2022). Muscles, makeup, and femboys: Analyzing tiktok's "radical" masculinities. *Social Media+ Society*, *8*(3), 20563051221126040.

Goffman, E. (1959). *The presentation of self in everyday life*. Anchor Books.

Green, M., Bobrowicz, A., & Ang, C. S. (2015). The lesbian, gay, bisexual and transgender community online: Discussions of bullying and self-disclosure in YouTube videos [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0144929X.2015.1012649]. *Behaviour & Information Technology*, *34*(7), 704–712. https://doi.org/10.1080/0144929X.2015.1012649

Hogan, B. (2010). The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*, *30*(6), 377–386.

Hong, S., Jahng, M. R., Lee, N., & Wise, K. R. (2020). Do you filter who you are?: Excessive self-presentation, social cues, and user evaluations of Instagram selfies. *Computers in Human Behavior*, *104*, 106159. https://doi.org/10.1016/j.chb.2019.106159

Horton, D., & Richard Wohl, R. (1956). Mass Communication and Para-Social Interaction: Observations on Intimacy at a Distance [Number: 3]. *Psychiatry*, *19*(3), 215–229. https://doi.org/10.1080/00332747.1956.11023049

Hu, P., Lin, C., Li, J., Tan, F., Han, X., Zhou, X., & Hu, L. (2023). Making the Implicit Explicit: Depression Detection in Web across Posted Texts and Images [ISSN: 2156-1133]. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 4807–4811. https://doi.org/10.1109/BIBM58861.2023.10385590

Huh, J., Liu, L. S., Neogi, T., Inkpen, K., & Pratt, W. (2014). Health Vlogs as Social Support for Chronic Illness Management. *ACM Trans. Comput.-Hum. Interact.*, *21*(4), 23:1–23:31. https://doi.org/10.1145/2630067

Johnson, M. R. (2019). Inclusion and exclusion in the digital economy: Disability and mental health as a live streamer on Twitch.tv [Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2018.1476575]. *Information, Communication & Society*, *22*(4), 506–520. https://doi.org/10.1080/1369118X.2018.1476575

Joinson, A. N. (2001). Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.36]. *European Journal of Social Psychology*, *31*(2), 177–192. https://doi.org/10.1002/ejsp.36

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media [Number: 1]. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Kim, J., & Lee, J.-E. R. (2011). The Facebook Paths to Happiness: Effects of the Number of Facebook Friends and Self-Presentation on Subjective Well-Being [Publisher: Mary Ann Liebert, Inc., publishers]. *Cyberpsychology, Behavior, and Social Networking*, *14*(6), 359–364. https://doi.org/10.1089/cyber.2010.0374

Kim, S. J., Lu, Y., & Peng, Y. (2025). Unmasking Deception: How computer vision could empower journalists in unveiling visual misinformation. In *The Routledge Companion to Visual Journalism* (pp. 471–482). Routledge. Retrieved September 21, 2025, from https://www.taylorfrancis.com/chapters/edit/10.4324/9781003391340-45/unmasking-deception-sang-jung-kim-yingdan-lu-yilang-peng

Kumar, N. (2014). Facebook for self-empowerment? A study of Facebook adoption in urban India [Publisher: SAGE Publications]. *New Media & Society*, *16*(7), 1122–1137. https://doi.org/10.1177/1461444814543999

Lee, H., Jiang, R., Yoo, Y., Henry, M., & Cooperstock, J. R. (2022). The Sound of Hallucinations: Toward a more convincing emulation of internalized voices. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3491102.3501871

Li, S., Wu, S., Liu, T., Zhang, H., Guo, Q., & Peng, Z. (2024). Understanding the Features of Text-Image Posts and Their Received Social Support in Online Grief Support Communities. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*, 917–929. https://doi.org/10.1609/icwsm.v18i1.31362

Liu, J., & Zhang, Y. (2024a, January). Understanding and Facilitating Mental Health Help-Seeking of Young Adults: A Socio-technical Ecosystem Framework [arXiv:2401.08994 [cs]]. https://doi.org/10.48550/arXiv.2401.08994

Liu, J., & Zhang, Y. (2024b). Modeling Health Video Consumption Behaviors on Social Media: Activities, Challenges, and Characteristics. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 208:1–208:28. https://doi.org/10.1145/3653699

Liu, J. (, Wang, Y., Lyu, Y., Su, Y., Niu, S., Xu, X. ", & Zhang, Y. (2024). Harnessing LLMs for Automated Video Content Analysis: An Exploratory Workflow of Short Videos on Depression. *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, 190–196. https://doi.org/10.1145/3678884.3681850

Liu, L. S., Huh, J., Neogi, T., Inkpen, K., & Pratt, W. (2013). Health vlogger-viewer interaction in chronic illness management. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 49–58. https://doi.org/10.1145/2470654.2470663

Liu, Y., Li, S., Liu, Y., Wang, Y., Ren, S., Li, L., Chen, S., Sun, X., & Hou, L. (2024, March). TempCompass: Do Video LLMs Really Understand Videos? [Issue: arXiv:2403.00476 arXiv:2403.00476 [cs]]. https://doi.org/10.48550/arXiv.2403.00476

Lu, Z. (2019). Improving Viewer Engagement and Communication Efficiency within Non-Entertainment Live Streaming. *The Adjunct Publication of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 162–165. https://doi.org/10.1145/3332167.3356879

Lyu, Y., Cai, J., Dosono, B., Yadav, D., & Carroll, J. M. (2024). " i upload... all types of different things to say, the world of blindness is more than what they think it is": A study of blind tiktokers' identity work from a flourishing perspective. *arXiv preprint arXiv:2404.14305*.

Manikonda, L., & De Choudhury, M. (2017). Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 170–181. https://doi.org/10.1145/3025453.3025932

Meriläinen, M., & Ruotsalainen, M. (2024). Online disinhibition, normative hostility, and banal toxicity: Young people's negative online gaming conduct. *Social Media+ Society*, *10*(3), 20563051241274669.

Metallinos, N. (2013). *Television aesthetics: Perceptual, cognitive and compositional bases*. Routledge.

Milton, A., Ajmani, L., DeVito, M. A., & Chancellor, S. (2023). "I See Me Here": Mental Health Content, Community, and Algorithmic Curation on TikTok. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3544548.3581489

Misoch, S. (2014). Card Stories on YouTube: A New Frame for Online Self-Disclosure. *Media and Communication*, *2*(1), 2–12. https://doi.org/10.17645/mac.v2i1.16

Mojtabai, R., & Olfson, M. (2006). Treatment Seeking for Depression in Canada and the United States [Publisher: American Psychiatric Publishing]. *Psychiatric Services*, *57*(5), 631–639. https://doi.org/10.1176/ps.2006.57.5.631

Naslund, J. A., Grande, S. W., Aschbrenner, K. A., & Elwyn, G. (2014). Naturally Occurring Peer Support through Social Media: The Experiences of Individuals with Severe Mental Illness Using YouTube. *PLOS ONE*, *9*(10), 9.

National Institute of Mental Health. (2023a). Schizophrenia - National Institute of Mental Health (NIMH). Retrieved August 22, 2024, from https://www.nimh.nih.gov/health/topics/schizophrenia

National Institute of Mental Health. (2023b, March). Mental Illness. Retrieved April 16, 2023, from https://www.nimh.nih.gov/health/statistics/mental-illness

Niu, S., Bartolome, A., Mai, C., & Ha, N. B. (2021). #StayHome #WithMe: How Do YouTubers Help with COVID-19 Loneliness? *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3411764.3445397

Park, H., & Lee, J. (2021). Designing a Conversational Agent for Sexual Assault Survivors: Defining Burden of Self-Disclosure and Envisioning Survivor-Centered Solutions. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3411764.3445133

Pater, J. A., Haimson, O. L., Andalibi, N., & Mynatt, E. D. (2016). "Hunger Hurts but Starving Works": Characterizing the Presentation of Eating Disorders Online. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1185–1200. https://doi.org/10.1145/2818048.2820030

Pendse, S. R., Kumar, N., & De Choudhury, M. (2023). Marginalization and the Construction of Mental Illness Narratives Online: Foregrounding Institutions in Technology-Mediated Care. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 346:1–346:30. https://doi.org/10.1145/3610195

Pendse, S. R., Nkemelu, D., Bidwell, N. J., Jadhav, S., Pathare, S., De Choudhury, M., & Kumar, N. (2022). From Treatment to Healing:Envisioning a Decolonial Digital Mental Health. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–23. https://doi.org/10.1145/3491102.3501982

Peng, Y., Lu, Y., & Shen, C. (2023). An Agenda for Studying Credibility Perceptions of Visual Misinformation [Number: 2 Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2023.2175398]. *Political Communication*, *40*(2), 225–237. https://doi.org/10.1080/10584609.2023.2175398

Peng, Y., Qian, S., Lu, Y., & Shen, C. (2025, April). Large Language Model-Informed Feature Discovery Improves Prediction and Interpretation of Credibility Perceptions of Visual Content [arXiv:2504.10878 [cs]]. https://doi.org/10.48550/arXiv.2504.10878

Pinch, A., Birnholtz, J., Macapagal, K., Kraus, A., & Moskowitz, D. (2024). The Subtleties of Self-Presentation: A study of sensitive disclosure among sexual minority adolescents. *Proc. ACM Hum.-Comput. Interact.*, *8*(CSCW1), 131:1–131:27. https://doi.org/10.1145/3637408

Pinch, A., Fiers, F., Birnholtz, J., Fisher, J., & Reilly, B. (2024). "Is it time for me to be authentic?": Understanding, performing, and evaluating authenticity on BeReal [Publisher: SAGE Publications]. *New Media & Society*, 14614448241267731. https://doi.org/10.1177/14614448241267731

Poquet, O., Lim, L., Mirriahi, N., & Dawson, S. (2018). Video and learning: A systematic review (2007–2017). *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 151–160. https://doi.org/10.1145/3170358.3170376

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

Postigo, H. (2016). The socio-technical architecture of digital labor: Converting play into YouTube money [Publisher: SAGE Publications]. *New Media & Society*, *18*(2), 332–349. https://doi.org/10.1177/1461444814541527

Pyle, C., Roosevelt, L., Lacombe-Duncan, A., & Andalibi, N. (2021). LGBTQ Persons' Pregnancy Loss Disclosures to Known Ties on Social Media: Disclosure Decisions and Ideal Disclosure Environments. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3411764.3445331

Reavley, N. J., & Jorm, A. F. (2011). Stigmatizing Attitudes towards People with Mental Disorders: Findings from an Australian National Survey of Mental Health Literacy and Stigma. *Australian & New Zealand Journal of Psychiatry*, *45*(12), 1086–1093. https://doi.org/10.3109/00048674.2011.621061

Sangeorzan, I., Andriopoulou, P., & Livanou, M. (2019). Exploring the experiences of people vlogging about severe mental illness on YouTube: An interpretative phenomenological analysis. *Journal of Affective Disorders*, *246*, 422–428. https://doi.org/10.1016/j.jad.2018.12.119

Schaadhardt, A., Fu, Y., Pratt, C. G., & Pratt, W. (2023). "Laughing so I don't cry": How TikTok users employ humor and compassion to connect around psychiatric hospitalization. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3544548.3581559

Seckler, M., Opwis, K., & Tuch, A. N. (2015). Linking objective design factors with subjective aesthetics: An experimental study on how structure and color of websites affect the facets of users' visual aesthetic perception. *Computers in Human Behavior*, *49*, 375–389.

Song, S., Zhang, Y., & Yu, B. (2021). Interventions to support consumer evaluation of online health information credibility: A scoping review. *International Journal of Medical Informatics*, *145*, 104321. https://doi.org/10.1016/j.ijmedinf.2020.104321

Strauss, A., & Corbin, J. M. (1997, March). *Grounded Theory in Practice* [Google-Books-ID: TtRMolAapBYC]. SAGE.

Sundar, S. S. (2008). The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. *Digital Media, Youth, and Credibility*, 73–100. https://doi.org/10.1162/dmal.9780262562324.073

Wan, Q., & Lu, Z. (2024). Investigating VTubing as a Reconstruction of Streamer Self-Presentation: Identity, Performance, and Gender. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–22. https://doi.org/10.1145/3637357

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., & Cranor, L. F. (2011). "I regretted the minute I pressed share": A qualitative study of regrets on Facebook. *Proceedings of the Seventh Symposium on Usable Privacy and Security*, 1–16. https://doi.org/10.1145/2078827.2078841

Wang, Y., Shen, S., & Lim, B. Y. (2023). Reprompt: Automatic prompt editing to refine ai-generative art towards precise expressions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581402

Wang, Y., Venkatesh, P., & Lim, B. Y. (2022). Interpretable directed diversity: Leveraging model explanations for iterative crowd ideation. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–28.

Wang, Z., Zhu, J., Zuo, W., Jiang, Z., Lei, J., & Wang, Z. (2025). Living with covid-19 conspiracies on chinese tiktok (douyin): Unpacking multimodal features, national identity, and user engagement. *Journal of Information Technology & Politics*, 1–13.

Woloshyn, V., & Savage, M. J. (2020). Features of YouTube™ videos produced by individuals who self-identify with borderline personality disorder [Publisher: SAGE Publications Ltd]. *DIGITAL HEALTH*, *6*, 2055207620932336. https://doi.org/10.1177/2055207620932336

Wu, P., Zou, S., Chen, C., & Song, Y. (2025). Hotbed of stigmatization or source of support: A multimodal analysis of mental health-related videos on Douyin. *Computers in Human Behavior*, *172*, 108716. https://doi.org/10.1016/j.chb.2025.108716

Xiao, S., Metaxa, D., Park, J. S., Karahalios, K., & Salehi, N. (2020). Random, messy, funny, raw: Finstas as intimate reconfigurations of social media. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13.

Xue, H. (2024). *Leveraging Multimodal Features in Short Videos to Communicate Risks in a Sensational Media Environment* [PhD Thesis]. UC Davis. Retrieved September 21, 2025, from https://escholarship.org/uc/item/7422m1w6

Xue, H., Nishimine, B., Hilbert, M., Cingel, D., Vigil, S., Shawcroft, J., Thakur, A., Shafiq, Z., & Zhang, J. (2025, August). Catching Dark Signals in Algorithms: Unveiling Audiovisual and Thematic Markers of Unsafe Content Recommended for Children and Teenagers [arXiv:2507.12571 [cs]]. https://doi.org/10.48550/arXiv.2507.12571

Yi, H., & Xian, L. (2024, September). The Informal Labor of Content Creators: Situating Xiaohongshu's Key Opinion Consumers in Relationships to Marketers, Consumer Brands, and the Platform [arXiv:2409.08360]. https://doi.org/10.48550/arXiv.2409.08360

Zafar, M. B., Gummadi, K., & Danescu-Niculescu-Mizil, C. (2016). Message impartiality in social media discussions. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 466–475.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237–291.

## A CODEBOOKS OF QUALITATIVE ANALYSIS

Table 4. The codebook for verbal storytelling.

| Code | Definition | Example Quotes |
|---|---|---|
| Emotions | Explicit verbal articulation of emotional states, feelings, and affective experiences related to mental health conditions. | "I'm overeating out of anxiety and out of fear. I'm paranoid out of my delusions and my auditory hallucinations." |
| Psychiatry | References to formal mental health treatment, medical interventions, psychiatric medications, therapy sessions, and professional therapeutic services. | "My psychiatrist suggested therapy to help manage my anxiety." |
| Relationships | Discussion of interpersonal connections, social interactions, family relationships, and relational dynamics in the context of mental health experiences. | "I feel like I'm a failure as a daughter and she [vlogger's mother] is like why do you feel like that?" |
| Substance Use | References to the use of psychoactive substances, including recreational drugs, alcohol, and their relationship to mental health experiences. | "let's uh one advice I would give everybody, do not even touch any of those substances that are called medicine psychedelics." |

## B PROMPTS FOR LLAVA

## C DESCRIPTIVE ANALYSIS

Table 5. The codebook of visual characteristics and categories.

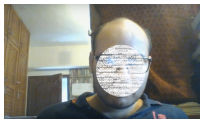
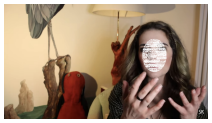
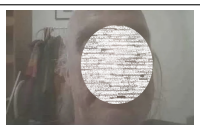
| Category | Definition | Description | Low Example | High Example |
|---|---|---|---|---|
| Chromatic Elements | Brightness | This describes the perceived intensity of light, affecting how light or dark the image appears [46]. |  |  |
| | Colorfulness | This refers to the diversity and saturation of colors used in the images [46]. |  |  |
| Compositional Structure | Clarity | This refers to how some layouts were easy to grasp and well-constructed, while others appeared cluttered or less organized. |  |  |
| | Aesthetics | This refers to the overall beauty and visual appeal of the images [78]. |  |  |

Table 6. The codebook of categories of supportive behaviors in comments.

| Category | Definition | Example Comments |
|---|---|---|
| Network Support | Expressions of emotional connection, empathy, and solidarity that foster a sense of community and shared understanding. | "I relate to this so much. You're not alone in this journey." |
| Informational Support | Provision of advice, resources, coping strategies, or factual information related to mental health experiences and treatment. | "Have you tried mindfulness meditation? It really helped me manage my anxiety symptoms." |
| Affirming Behavior | Positive reinforcement, validation of experiences, and encouragement that acknowledges the vlogger's courage and authenticity. | "Thank you for being so brave and sharing your story. Your honesty is inspiring." |

Table 7. Prompts for Keyframe and Comment Annotation

| Model | Code Name | Prompt |
|---|---|---|
| LLaVA | **Aesthetics** | Does this picture have an overall aesthetic appeal? |
| | **Colorfulness** | Is this picture colorful? |
| | **Clarity** | Is this picture clear in layout? |
| | **Brightness** | Is this picture bright in appearance? |
| GPT-min | **Network support** | Network support is providing personal connections such as willingness to chat or contact information. Does this comment offer network support (Respond with Yes or No)?<br>Comment: {} |
| | **Informational support** | Informational support is giving relevant information, advice, suggestions to help someone. Does this comment offer informational support (Respond with Yes or No)?<br>Comment: {} |
| | **Affirm behavior** | Does the comment affirm the feelings or the behaviors of the vlogger (Respond with Yes or No)?<br>Comment: {} |

Table 8. The descriptive statistics of #View, #Likes, and #Comments.

| | Mean | Std. Error | Skewness | Kurtosis |
|---|---|---|---|---|
| **#Views** | 16650.62 | 6760.97 | 9.99 | 103.52 |
| **#Likes** | 467.13 | 178.27 | 11.15 | 131.52 |
| **#Comments** | 69.82 | 23.58 | 11.15 | 130.80 |