What Makes For A Good Thumbnail? Video Content Summarization Into A Single Image*

Jasmine Yang[†] Oded Netzer[‡]

May 10, 2026

Abstract

Thumbnails serve as representations of video content when viewers are “in the dark" before watching. Yet it remains unclear whether thumbnails should summarize the video, like a synopsis, or tease the video. We study how thumbnails, relative to the video content they represent, affect video decisions. We propose a thumbnail-video mining procedure that uses multimodal LLMs and computer vision to transform unstructured thumbnails and video content into interpretable features, and then construct theory-based measures to characterize both the thumbnail itself and its relationship with the video it represents. We use secondary data from YouTube to document real-world relationships between thumbnails and video performance, and then experimental data from CTube, a video platform we built to randomize thumbnail exposure, to estimate a joint model of video choice and watchtime. Our results reveal a fundamental click–watchtime tradeoff. “Teaser-style" thumbnails that are visually appealing and emotionally engaging increase clicks, but these features do not sustain viewing. Instead, thumbnails affect watchtime through how the video unfolds relative to the thumbnail: viewers are more likely to quit when video content diverges from the thumbnail, but this penalty weakens after viewers observe the moment revealed by the thumbnail in the video. We simulate performance of counterfactual thumbnails and show that there is no one-size-fits-all thumbnail. Effective thumbnails require balancing visual impact with content alignment and timing, while tailoring each thumbnail to viewers and video content.

Keywords: thumbnails, video analysis, multimodal large language models, computer vision, unstructured data, content marketing

*We thank Olivier Toubia, Andrey Simonov, Hortense Fong, Asim Ansari, and seminar participants at Columbia University, Singapore Management University, HKUST, Yale University, Duke University, University of Houston, National University of Singapore, the Ohio State University, University of Wisconsin-Madison, the University of Texas at Dallas, the University of Utah, the Chinese University of Hong Kong, Marketing Science, Artificial Intelligence, Machine Learning, and Business Analytics Conference, China Marketing International Conference, AI in Management Conference, ASA Joint Statistical Meetings, and AMS Annual Conference for helpful comments and suggestions. We thank Daniel Joseph Merlau for excellent research assistance. We are grateful for the Deming Center for financial support.

[†]Chinese University of Hong Kong and Meta. Email: yumingyang@meta.com

[‡]Columbia University. Email: onetzer@gsb.columbia.edu

1 Introduction

Thumbnails, reduced-size preview images, GIFs, or clips of videos, are one of the first and most important cues viewers encounter when deciding whether to watch a video, beyond its title (YouTube Help, 2026). Their role is especially important in today's crowded video environment. On YouTube alone, around 500 hours of video are uploaded every minute (Wytlabs, 2026). Viewers must quickly decide which videos deserve their time and attention based on limited pre-consumption information. In this decision process, thumbnails provide a quick visual gateway that helps videos stand out and provide viewers with a first glimpse of the content.

Although sharing similarities with traditional visual marketing objects such as book covers, movie posters, and trailers (e.g. Luo, 2026; Kim et al., 2025; Yang, 2023; Liu et al., 2018), thumbnails are tied more directly to a specific, dynamic video experience that viewers may *immediately* consume after clicking. Whether selected from the video itself or designed separately, a thumbnail *previews* the video and may thus create expectations about what the viewer is about to see. This makes thumbnail selection not only a problem of visual appeal or click-through, but also a problem of representation: should a thumbnail *summarize* the video (much like a synopsis) to help viewers accurately infer its content, or should it *tease* the video by highlighting a visually appealing, emotional, surprising, or climactic moment? These objectives may conflict.

Consider a creator uploading a 60-second wildlife safari video and is interested in choosing a thumbnail from its video frame sequence (see Figure 1).[1] The video begins with a dawn safari scene, moves into a series of wildlife encounters, with many shots of a leopard in the trees, and ends with night-safari footage. The creator is considering among several stylized options (but not limited to those) for the thumbnail: (1) the first (non-black) frame, which represents early default choices used by some video systems, (2) the most aesthetic frame, which may be a shot of a visually striking sunset framed by trees, (3) the most emotional frame, which may be a scene that features giraffes silhouetted against the sun, which can elicit awe and excitement, or (4) the most content-representative frame, which may be a shot of the leopard, which best reflects the dominant wildlife content repeatedly shown in the video. Which frame should the creator choose as the thumbnail to represent *the entire video*? This question is the heart of our research.

[1] In this research, we only consider image (as opposed to short video) thumbnails and thumbnails that are a frame of the video itself to make it a constrained optimization. See Web Appendix Figure G.0.1 for an example. However, our framework can be extended to customized thumbnails and video clips (or hooks).

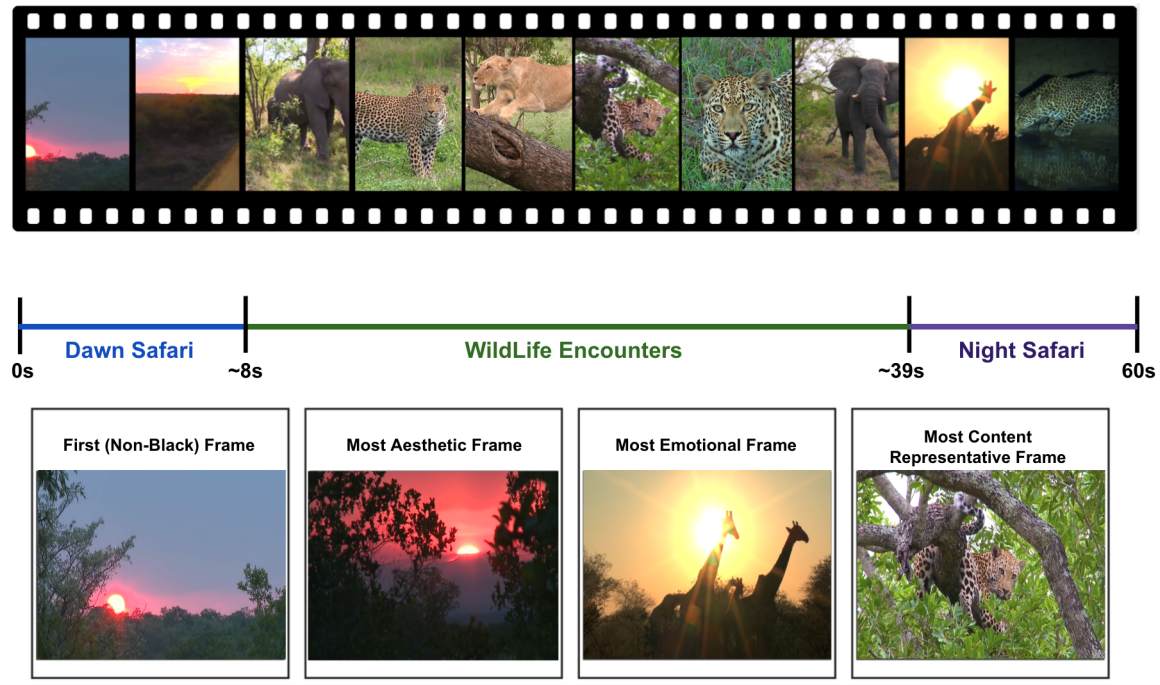


Figure 1: **Sample thumbnail candidates from a video sequence**.

In this paper, we study how thumbnails (relative to video content) affect viewers' video decisions (e.g., views, engagement like watchtime, preference match) and, subsequently, how to select and design thumbnails that balance views and engagement. To do so, we develop a thumbnail-video mining procedure that treats a thumbnail not as an isolated image, but as a visual representation of the video. Specifically, we characterize each video as a sequence of frames or scenes (a set of frames with limited visual change). We then transform high-dimensional, unstructured thumbnail and video content into interpretable and consumer-relevant features using multimodal large language models, computer vision, and deep learning methods. These features capture three core dimensions, semantic content, affective emotions, and visual aesthetics, through which thumbnails shape viewers' video experience. Using these features extracted at the thumbnail and video frame level, we then construct theory-based measures to characterize both the thumbnail itself and its relationship to the underlying video it represents.

First, we use secondary data from YouTube to document real-world relationships between thumbnails and video performance. We find evidence that there exists a tension in thumbnail selection. Thumbnails that evoke higher arousal or greater surprise are associated with more views. At the same time, thumbnails that are less representative of the underlying video content are associated with fewer views and low likeability. We hypothesize that by revealing an

emotionally charged or unexpected moment of a video, these thumbnails may function more as teasers that attract attention, while withholding the context to generate curiosity and motivate viewers to watch the video to resolve the gap about what is happening in the video. However, to improve video preference, thumbnails cannot simply maximize attention; they also need to set expectations that are supported by the subsequent video experience early on.

Secondary data are limited in the sense that we only observe aggregate video outcomes, the analysis can only be cross-sectional, and results are subject to endogeneity. To move beyond limitations of secondary data, we further build an experimental video platform called "CTube" to study the causal effect of thumbnails. Our experimental platform allows us to randomize thumbnails viewers see to start a video and observe individual-level behavior to capture heterogeneous response to thumbnails. Since each thumbnail is not a monolithic treatment, but rather a composite of attributes, we treat the randomization as providing exogenous variation in different thumbnail features and adopt a modeling approach to understand how thumbnails, decomposed into features, affect video decisions.

Using the experimental data, we estimate a joint copula model of video choice and watchtime using these experimental data. Our results reveal that when it comes to deciding which video to watch, thumbnails that are more aesthetic, higher in arousal, emotionally expressive, and more different from the video content are more likely to be clicked. When it comes to watchtime, static thumbnail attributes no longer directly sustain viewing. However, thumbnails can still affect watchtime through how the video unfolds relative to the thumbnail. Viewers are more likely to quit when the video content observed moment-by-moment diverges from the thumbnail, but this penalty on mismatch is reduced after viewers reach the moment depicted by the thumbnail. Thus, effective thumbnails need not be literal summaries of the video. Rather, thumbnails need to balance visual impact with content alignment and timing so that they generate enough interest to induce the click while being not so disconnected or delayed that viewers quit before the video delivers the moment the thumbnail led them to expect.

We use the estimated model to evaluate alternative thumbnail candidates to suggest thumbnail selection and design strategies that help improve video clicks, expected watchtime conditional on clicking, and joint performance. Our results show that simple thumbnail heuristics often trade off clicks and watchtime. High-arousal and aesthetically appealing frames substantially increase choice probability, but do not necessarily sustain watchtime; more representative or

summarizing frames perform better for watchtime but are less effective at generating clicks. Creator-selected thumbnails perform roughly on par with random frames, suggesting that current selection strategy leaves substantial room for improvement.

In contrast, our model-optimized thumbnails improve performance by combining the strengths of these different strategies. The model-suggested optimal thumbnails outperform the actual creator-selected frame, random frames, stylized rules, and even the best of the three platform-recommended frames across all dimensions. Their feature profiles suggest that the best thumbnails for joint performance balance visual impact, content fit, and timing in a way that converts initial attention into sustained viewing without losing viewers after the click. A held-out predictive analysis further shows that viewer heterogeneity is important for prediction, and that personalized thumbnail selection can generate additional performance gains for a meaningful share of viewer-video pairs. These improvements can compound at scale, with each additional second of expected watchtime per click, aggregated across millions of daily views, translating into thousands of additional hours of engagement. Together, these results show that there is no one-size-fits-all thumbnail for all viewers or performance metrics. Effective thumbnail selection requires balancing the features that attract initial attention with the features that sustain viewing, while accounting for heterogeneity in viewer preferences and video content.

The remainder of the paper is organized as follows. Section 2 discusses related literature. Section 3 introduces our thumbnail–video analysis framework, where we describe our video mining pipeline, interpretable feature extraction methods, and theory-driven measures that we construct to characterize both the thumbnail itself and its relationship to the video it represents. Section 4 presents evidence from YouTube secondary data on the relationship between thumbnail features and aggregate video outcomes, exploring hypotheses of thumbnails worthwhile to examine further. Section 5 discusses an experiment on a CTube, an experimental video platform we built, to study the causal effect of thumbnails. We develop a model in Section 6 and use the experimental data from CTube to estimate the effect of thumbnails on viewers' decisions in Section 7. We use the estimated model to evaluate alternative thumbnail candidates for different video outcomes, benchmark our model-selected thumbnails against current practice, and then re-calibrate the model to validate our model's predictive ability for held-out videos in Section 8. Section 9 concludes with managerial implications, limitations, and directions for future research.

2 Literature Review

Our paper contributes to several streams of literature. First, we contribute to the literature on thumbnail selection and performance. Prior work in computer science has primarily studied thumbnail selection as an automated frame-selection or generation problem, developing scalable methods that rank or generate candidate frames using proxy objectives such as visual quality and aesthetics, semantic relevance to the video or query, face presence and character prominence, user-interest or popular-topic similarity, and predicted clickability (e.g., Fantini, 2024; Pornpanvattana et al., 2024; Apostolidis et al., 2023; Yang et al., 2023; Ardeshir et al., 2022; Liu, 2022; Yu and Shi, 2021; Shimono et al., 2020; Song et al., 2016; Liu et al., 2015). However, this literature typically treats thumbnail selection as an algorithmic image-scoring or generation task rather than a consumer-response problem. A small but growing literature in marketing and information systems studies thumbnails, covering topics such as whether platform-selected thumbnails reflect biases (Turner et al., 2024), and how thumbnail attributes relate to video views (Cui et al., 2024; Koh and Cui, 2022), pre-roll advertising effectiveness (Li et al., 2022) and video-clip selection (Yoon and Kim, 2019). This stream of literature connects thumbnails to consumer response. However, most papers provide only correlational evidence of thumbnail effects, and treat thumbnails as a standalone image without considering its relationship with the video.

We differ from existing work on thumbnails in several important ways. First, we study whether and how thumbnails affect not only viewers' initial click decisions but also their post-click video experience, such as watchtime and preference match. This requires moving beyond treating thumbnails as standalone images and instead conceptualizing them as representations of the video content viewers are about to see. This perspective connects our paper to literature on creative content summarization (e.g., Toubia, 2021), teasers and spoilers (e.g., Ryoo et al., 2021; Menon and Soman, 2002), and behavioral theories of reference-dependent evaluation, expectation disconfirmation, and curiosity (e.g., Kahneman and Tversky, 1979; Oliver, 1977, 1980; Anderson and Sullivan, 1993; Loewenstein, 1994). A thumbnail can summarize the video by accurately previewing its content, thereby helping viewers form expectations that are more closely aligned with the subsequent viewing experience to avoid mismatch. At the same time, a thumbnail can also tease the video by highlighting an appealing, surprising, or later-occurring moment that becomes a reference point as the video unfolds. We draw on this literature to discipline our

theory-driven measure construction to capture thumbnails' relationship to the underlying video. Second, to address the confounding concerns that commonly arise in correlational studies, we build our own experimental video platform, randomize thumbnail exposure, and track individual viewers' clickstream behavior to back out video choice and watchtime decisions. This design builds on recent marketing research that creates realistic online environments to study content effects, such as Morozov and Tuchman (2024)'s simulated online bookstore experiment, and extends this approach to video consumption. To the best of our knowledge, we are the first to provide a comprehensive analysis of thumbnails that combines secondary data, a controlled realistic experiment, and behavioral modeling. This allows us not only to study how viewers respond to different thumbnails but also to use the model estimates to simulate counterfactual thumbnails to characterize "optimal" thumbnails for different video objectives.

Our paper is also related to a broad and growing literature on visual and video content marketing. Prior work has shown that visual content shapes consumer responses across many marketing contexts, such as social media engagement (e.g., Cao et al., 2025; Ceylan et al., 2024; Li and Xie, 2020), consumer and brand perceptions (e.g., Exner et al., 2025; Hartmann et al., 2021; Liu et al., 2020), product aesthetic design and logo design (e.g., Sisodia et al., 2024; Burnap et al., 2023; Dew et al., 2022), and product demand (e.g., Luo, 2026; Zhang and Luo, 2023; Zhang et al., 2022; Dzyabura et al., 2023). Related work on video analytics examines how video content shapes attention, persuasion, engagement, and demand in contexts such as video advertising, online videos, livestreaming, and influencer content (e.g., Yang et al., 2025; Overgoor et al., 2025; Jin et al., 2025; Chakraborty et al., 2024; Tian et al., 2024; Luo et al., 2024; Zhou et al., 2021; Rajaram and Manchanda, 2020). Our work combines images and videos in a relatively understudied context to understand how thumbnails represent the video content and shape consumer decisions for video watching.

Finally, we contribute to an emerging frontier of using multimodal large language models (MLLMs) for video measurement (e.g., Zhang et al., 2026; Yang et al., 2026; Overgoor et al., 2025; Jin et al., 2025). Feature extraction from complex video data has traditionally benefited from advances in deep learning models, particularly leveraging architectures such as convolutional neural networks (CNNs) and Transformers. These models have been especially successful for well-defined perceptual tasks with large-scale labeled datasets and clear prediction targets. Recent advances in MLLMs expand this measurement toolkit by enabling researchers to conduct

generalized video understanding tasks at virtually unlimited scale (see Jin et al. (2025) for reviews of different techniques). Our approach builds on this development by using a hybrid measurement approach. When strong task-specific models exist, such as NIMA for visual aesthetics, we use these specialized fine-tuned deep learning models because they are trained on high-quality large-scale human-annotated datasets and perform well for clearly defined perceptual tasks (Talebi and Milanfar, 2018). When the construct is more open-ended, semantic and contextual, such as scene interpretation, content representation, and emotions, we use MLLMs to transform thumbnails and video frames into structured, interpretable annotations. This choice builds on recent work in multimodal affective computing that has begun to use MLLMs for emotion annotation and affective reasoning, especially when emotion labels are costly to collect, or difficult to standardize across domains (e.g., Ruder et al., 2025; Shou et al., 2025; Niu et al., 2024). This approach allows us to retain the strengths of traditional fine-tuned deep learning models where supervised benchmarks are strong, while leveraging MLLMs to measure richer aspects of video content that are difficult to annotate manually or standardize across domains. In doing so, we show how MLLMs can serve as a flexible measurement layer for video marketing research, complementing existing feature extraction methods and enabling scalable analysis of videos.

3 Thumbnail-Video Analysis

Analyzing thumbnails and videos requires the decomposition of unstructured video data into meaningful features that are human-interpretable and consumer-relevant. In this section, we begin by describing our video mining pipeline, followed by the construction of interpretable features along three dimensions extracted utilizing state-of-the-art multimodal large language models, computer vision, text mining, and deep learning techniques to understand how thumbnails relative to videos affect consumer reactions.

3.1 Video Mining Pipeline

We conceptualize a video as a sequence of consecutive frames or scenes (a set of frames that exhibit limited visual change). Each frame can be perceived as a tensor of (width $\times$ height

$\times$ 3), where the last dimension represents the RGB channels. Since the raw pixel data is not directly interpretable, a critical step lies in converting videos into *human-interpretable* and *consumer-relevant* features.

Figure 2 illustrates our video mining pipeline. At a high conceptual level, the procedure involves three steps. In **Step 1**, each video is segmented into a sequence of frames sampled at fixed intervals of every second or at every scene (depending on the method used for feature detection). In **Step 2**, we extract interpretable features from both the video's frames and its thumbnail, covering semantic content, affective emotions, and aesthetic dimensions. When a feature is computed at every scene, all sampled frames from that scene will receive the same value. In **Step 3**, we construct theory-driven measures to study not just the thumbnail itself but how the thumbnail's features compare against the video by computing frame-level comparisons and then averaging across frames. We next detail our feature extraction techniques.

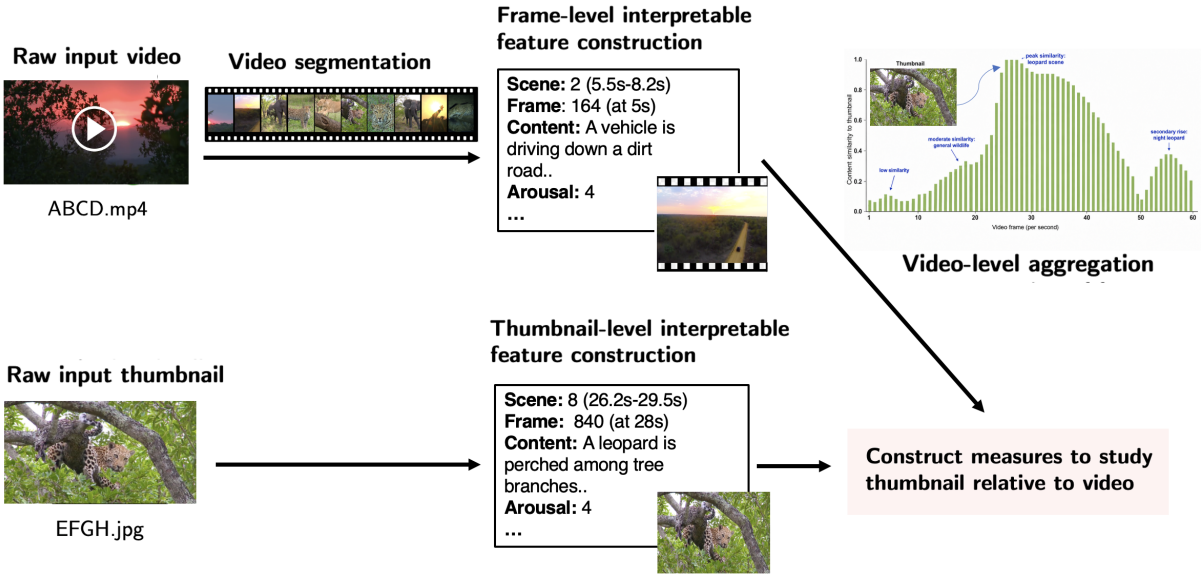


Figure 2: **Flowchart of the video mining procedure.** This figure shows the flowchart of our proposed video mining procedure. The features are extracted at both the sampled frame level (at either every second or every scene) and the thumbnail level. The extracted features are used to construct measures to study thumbnails relative to video content they represent.

3.2 Feature Extraction

We organize our features along three dimensions that reflect the primary channels through which a thumbnail can shape video decisions: (i) semantic content (what is depicted and how well

it represents the video), (ii) affective emotions (what it makes viewers feel), and (iii) visual aesthetics (how much it attracts viewers). We discuss the extraction of each dimension in turn.

Content. Given the strong performance of multimodal large language models (MLLM) for semantic content understanding that go beyond what traditional object detection or concept labeling models can provide, we generate descriptions of each scene and the thumbnail using LLaVA-Video-72B-Qwen2 (Zhang et al., 2024), an open-source vision-language MLLM that supports contextual reasoning across video and text modalities (see Web Appendix A.3 for model details). LLaVA-Video achieves state-of-the-art performance among open-source models on video understanding benchmarks for short videos (less than 2 minutes). Each scene (and the thumbnail) is input to the model with a structured prompt that instructs it to describe the main content in one sentence (see Web Appendix A.4 for prompt details). We use these descriptions as an input for measuring content difference between the thumbnail and the video (i.e., content disconfirmation), and as the basis for latent topic extraction in Section 3.3.

Affective Emotions. Analyzing video affective emotions is challenging, given that (1) a person may experience mixtures of different emotions rather than a single emotion, and (2) the same scene can evoke different emotional responses across individuals. Accurately measuring such affective responses therefore requires representations that preserve mixed emotions and contextual understanding rather than single-label classifications.

We again leverage LLaVA-Video-72B-Qwen2 (Zhang et al., 2024) for this task, as MLLMs can jointly interpret facial expressions, body language, contextual cues to infer the likely emotional response—capabilities that specialized single-label emotion classifiers, typically trained on small annotated datasets, lack. We directly input each scene to the model and prompt it to rate the intensity of Ekman (1999)'s six basic emotions (anger, disgust, fear, joy, sadness, and surprise) including a neutral class, as well as valence and arousal on a 1-7 scale (see Web Appendix A.4 for prompt details).

Visual Aesthetics. We evaluate the visual aesthetics of thumbnails and video frames using the Neural Image Assessment (NIMA) model (Talebi and Milanfar, 2018), a convolutional neural network trained on the large-scale Aesthetic Visual Analysis (AVA) dataset containing over 255,000 images evaluated by both professionals and amateurs (Marchesotti et al., 2015). NIMA is the current state-of-the-art and is preferable for this task since it is calibrated to human judgments of image aesthetic perception. It produces a continuous aesthetic score in the range of

1 to 10, where higher values indicate higher predicted aesthetics. We apply the pre-trained model to obtain aesthetic scores for both the thumbnail and each sampled video frame. These scores are later used to construct the aesthetic disconfirmation measures and the thumbnail aesthetics.

3.3 Feature Construction

3.3.1 Theory-Driven Features

We construct the following measures through which we rationalize the effect of thumbnails. See Web Appendix Panel A of Table G.0.1 for summary statistics.

Content Disconfirmation. For each video v, we construct *content disconfirmation* to reflect the degree of mismatch between the expected content (based on the thumbnail) and the actual content of the video. Let $\mathcal{F}_v$ denote the set of sampled frames for video v, with $|\mathcal{F}_v|$ frames indexed by $f = 1, \ldots, |\mathcal{F}_v|$. Content descriptions are extracted at the scene level, which is encoded into 768-dimensional dense vector representations using GIST-Embedding-v0 (Solatorio, 2024), a state-of-the-art sentence embedding model that supports scalable similarity computation beyond surface-level lexical overlap. Since the same scene description applies to all frames within that scene, each sampled frame f inherits the embedding $\mathbf{e}_{vf}$ of its parent scene. The thumbnail description is similarly encoded as $\mathbf{e}_v^{\text{thumb}}$. We define content disconfirmation as[2]:

$$\text{CD}_v = 1 - \frac{1}{|\mathcal{F}_v|} \sum_{f=1}^{|\mathcal{F}_v|} \cos\left(\mathbf{e}_v^{\text{thumb}}, \mathbf{e}_{vf}\right). \tag{1}$$

We note that, the cosine similarity between GIST embeddings is always non-negative in our case, and thus we subtract the cosine similarity from 1 to effectively invert the similarity measure.

Aesthetics Disconfirmation. We measure *aesthetic disconfirmation* as the discrepancy between the aesthetics conveyed by a video's thumbnail and the perceived aesthetics experienced during playback. Unlike content features, aesthetics has inherent valence, and thus we can decompose aesthetic disconfirmation into positive and negative components, capturing cases in which the thumbnail is aesthetically superior to, or inferior to, the subsequent video content.

Let A_v^{thumb} denote the thumbnail aesthetic score for video v, and let a_{vf} denote the aesthetic

[2]Because frames within the same scene share the same embedding, this is equivalent to a duration-proportional average across scenes.

score of sampled frame $f \in \mathcal{F}_v$. We define frame-level aesthetic disconfirmation as

$$\Delta a_{vf} = A_v^{\text{thumb}} - a_{vf}. \tag{2}$$

Let $\mathcal{F}_v^+ = \{f \in \mathcal{F}_v : \Delta a_{vf} \geq 0\}$ and $\mathcal{F}_v^- = \{f \in \mathcal{F}_v : \Delta a_{vf} < 0\}$, we decompose aesthetic disconfirmation into positive and negative components by averaging aggregate positive and negative discrepancies across sample frames:

$$\text{AD}_v^+ = \frac{1}{|\mathcal{F}_v^+|} \sum_{f \in \mathcal{F}_v^+} \Delta a_{vf}. \tag{3}$$

$$\text{AD}_v^- = \frac{1}{|\mathcal{F}_v^-|} \sum_{f \in \mathcal{F}_v^-} |\Delta a_{vf}|. \tag{4}$$

AD^+ captures the extent to which the thumbnail's aesthetics exceeds that of the video, i.e., the degree of aesthetic "overselling." AD^- captures the extent to which the video's aesthetics exceeds the thumbnail, i.e., the degree to which the video "outperforms" its thumbnail. When no frames satisfy a given condition, the corresponding measure is set to zero. Both measures are non-negative by construction, and larger values indicate a greater magnitude of aesthetic mismatch.

Thumbnail Location. We identify where in the video's timeline the thumbnail frame appears. For each video, we locate the frame that matches the thumbnail using a two-stage, coarse-to-fine matching procedure that first uses DINOv3 image embeddings (DINOv3 ViT-B/16; Siméoni et al. (2025)) to narrow down the right *region* of the video where the thumbnail is, and then conducts classical template matching (e.g., Lewis et al., 1995; Brunelli, 2009) to *precisely locate* the thumbnail's exact position (See Web Appendix A.1 for details). We record both the absolute timestamp t_{thumb} (in seconds) and the percentile position within the video t_{thumb}/T, where T is the video length.

Thumbnail Aesthetics and Emotions. Besides measures that capture the thumbnail relative to video, we also include features that characterize the thumbnail itself. We capture the absolute aesthetic appeal of the thumbnail using its NIMA aesthetic score. In addition, we capture the affective content of the thumbnail as perceived by a viewer using the emotion detection pipeline described in Section 3.2 to obtain scores on six discrete emotion dimensions (anger, disgust, fear,

joy, sadness, and surprise) plus neutral, as well as valence and arousal.

3.3.2 Control Variables

We control for several categories of confounds when analyzing the impact of thumbnails (see Web Appendix Panel B of Table G.0.1 for summary statistics and Web Appendix Table G.0.2 for the full list of controls with definitions).

Video Caption. Video captions may work together with thumbnails in shaping viewers' expectations of a video. We account for video captions along three sub-dimensions: (i) *Caption-Thumbnail (Video) Congruence*, which captures how well the caption aligns with thumbnail or video content (averaged similarities across frames) using Contrastive Language-Image Pre-Training model (Radford et al., 2021), a neural network that learns a multimodal embedding space by jointly training an image encoder and a text encoder to effectively predict image-text similarity, (ii) *Caption Sentiment*, which classifies caption sentiment into negative, positive or neutral using a RoBERTa-based classifier fine-tuned on the TweetEval benchmark (Barbieri et al., 2020; Liu et al., 2019),[3] which is well-suited to the short, informal style of user-generated video captions, (iii) *Caption Basic Features*, such as word count, the share of affect (e.g., happy, crying, awesome) and netspeak (e.g., lol, haha), capitalized words, and the share of numbers, punctuation (e.g., question marks, exclamation points), that are referred by creators to affect video decisions.[4]

Video Topics. To control for video content, we train a Latent Dirichlet Allocation (LDA; Blei et al., 2003) topic model on the concatenated scene-level descriptions generated by LLaVA-Video-72B-Qwen2 (Zhang et al., 2024) for each video. We train the model using online variational inference algorithm with 5-fold cross-validation and obtain an optimal of eight topics (human & vlogging, entertainment, water, event, nature, hospitality, transportation, food) based on average predictive per-word perplexity across folds and topic interpretability (see Web Appendix A.5 for

[3]We used a Transformer-based model pre-trained on 160GB of English texts (e.g., the Wikipedia corpus) in a self-supervised fashion with the TweetEval benchmark provided by Barbieri et al. (2020). The version we used is https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment.

[4]For example, to attract views, it is recommended by creators to use numbers in captions, such as "How I Paid Off 85,000 in Debt This Year". Incorporating affect or positively toned words, like "awesome", "thrilling", and "eye-opening," are suggested to elicit an emotional response and increase the appeal of the caption and thus content. Additionally, shorter captions and the use of punctuation (e.g., exclamation marks) are also believed to draw more attention to viewers. Sources: https://mktoolboxsuite.com/youtube-titles-that-get-views/; https://rankmath.com/blog/power-words/.

technical details, the topics and representative words).

Video Mechanics. We control for video length. We also detect video scenes using the content-aware detector from PySceneDetect with an adaptive threshold (see Web Appendix A.2 for details), and from the scene cuts calculate the number of scenes and the average scene length. Together these features proxy for the degree of visual variation and thus captures the likelihood that viewers experience boredom.

Channel and Timing. We also control for channel popularity using log(subscriber count + 1) and the number of days since the video was uploaded at the time of scraping.

4 Thumbnails and Consumers' Video Viewing Behavior: YouTube Data

In this paper, we first conduct observational data analysis to explore interesting hypotheses on how thumbnails relate to video reactions. We further build an experimental video website and randomize thumbnails shown to viewers to examine the causal effects in Section 5 and Section 6.

4.1 Data Collection

We retrieved a random sample of 5000 videos in the Travel & Events category published between March and September 2025 in the United States using the YouTube API v3.[5] We restrict the sample to English-language videos between 30 and 120 seconds in duration. We limit our analysis to relatively short videos because our analysis involves segmenting each video into a sequence of frames sampled at a rate every second (or every scene). Extracting rich and interpretable features across the video is costly and requires high computational power, which restricts the use of videos with manageable lengths.

We downloaded each video's MP4 files given video URLs using youtube-dl package. In addition, we crawled each video's associated thumbnail and video caption at the time of scraping. We also retrieved each video's metadata, including channel information (e.g., channel subscriber count), video information (e.g., video length and time since posting), and consumer reactions to content (number of views, likes, and comments).[6]

[5] https://developers.google.com/youtube/v3, accessed on October 12, 2025.

[6] YouTube made dislikes private on December 13, 2021. We tried to supplement the data using the Return

To make thumbnail selection a constrained optimization problem, we dropped videos for which thumbnails are not drawn from its video frame sequence by applying a two-stage coarse-to-fine template matching procedure (see Web Appendix A.1 for technical details). 11.2% of the videos were dropped because the thumbnail is not in the video sequence. We also dropped videos with downloading errors and videos for which at least one of the statistics was made unavailable. The resulting dataset contains 4,121 videos. Table 1 reports the summary statistics.

Variable	Mean	Std. Dev.	Min	Median	Max	% Zero	N
Views	603,786	5,330,538	151	1,819	181,549,701	0.0	4,121
Likes	14,224	129,036	0	31	4,274,670	1.9	4,121
Comments	129	1,130	0	2	46,347	38.0	4,121
Video length (sec)	57.4	20.9	30	55	120	—	4,121

Table 1: **Summary statistics for YouTube data.**

We note that the sample shows substantial variation in the number of views, likes, dislikes, and comments across videos, suggesting that our dataset covers videos with different levels of popularity and levels of engagement, with a few videos of extremely high popularity. Given the right-skewed distribution of these variables, we log transform and construct the following main outcome variables for analysis: (1) *Views.* We use log(views) to capture whether the thumbnail attracts people to watch the video, (2) *Likeability.* While a thumbnail that is not representative of a video may attract clicks, viewers may not like the video when the content does not fulfill their initial expectation set by the thumbnail. We capture this using $= \log\left(\frac{\text{likes}+1}{\text{views}}\right)$, which is normalized by views to allow for comparison across videos with different levels of popularity. 98.1% of videos in our sample have at least one like, so this measure is not conflated with low engagement, and (3) *Commenting*. We use $\log(\frac{\text{comments}+1}{\text{views}})$ to capture engagement in the comment section.

4.2 Thumbnail Features and Video Reactions

We first use regression analysis to examine the relationship between thumbnails and video reactions. Since each video v in our cross-sectional data has exactly one thumbnail, we estimate

YouTube Dislike API (https://returnyoutubedislike.com), which estimates dislikes from historical data and browser extension users who opt in to share their dislike activity. However, since all videos in our sample were published after this change, the dislike data is limited: 73% of videos report zero dislikes, likely reflecting measurement limitations of the API rather than true viewer sentiment.

the following main OLS regression specification:

$$y_v = \boldsymbol{\alpha}\,\text{Controls}_v + \boldsymbol{\beta}\,\text{Thumbnail Features}_v + \varepsilon_v, \tag{5}$$

where y_v denotes the outcome variable for video v. $Controls_v$ include caption basic features (e.g., word count, share of network/affect words), caption-thumbnail and caption-video congruence, caption sentiment, video mechanical features (e.g., video length, number of scenes and average scene length), video topics, channel features, days since posting (see Section 3.3.2 and Web Appendix Table G.0.2). Thumbnail Features$_v$ includes the theory-driven features described in Section 3.3.1: content disconfirmation, thumbnail aesthetics, positive and negative aesthetic disconfirmation, thumbnail location, arousal, and six discrete emotions (anger, disgust, fear, joy, sadness, surprise).[7]

We focus our discussion below on the estimates for video views and likeability, reported in Table 2. Web Appendix Table G.0.5 reports the corresponding estimates for commenting; commenting appears to be more closely related to channel and video content shown rather than the thumbnail that is used.

Video Views. Table 2 (Columns 1-3) reports the results for video views. First, we find that thumbnails that are less content representative of the video (reflected by higher content disconfirmation) is negatively associated with views. Recall that on YouTube, a "view" is counted only when a viewer watches at least 30 seconds of a video. This effect thus suggests that if the mismatch is salient within the first 30 seconds, viewers may leave before a view is counted. Supporting this interpretation, the effect is robust when content disconfirmation is measured using only the first 30 seconds of a video (see Web Appendix Table G.0.4). Following existing theories, one hypothesis is that viewers may use the thumbnail as a frame of reference for comparative video judgment (Kahneman and Tversky, 1979). Their reactions to the video could be a result of a (mis)match between their realized video experience and a-priori expectation set by the thumbnail (Oliver, 1977, 1980), which we will investigate more cleanly in the experiment section.

In addition, we find that thumbnails evoking higher energy (arousal) or greater surprise are associated with more views. A higher arousal thumbnail can be one that captures an

[7]We exclude thumbnail valence from the specification since it is highly correlated with several of the discrete emotion variables (e.g., $r = 0.62$ with joy), resulting in elevated variance inflation factors (VIF > 5).

	log views[a]			log((likes + 1)/views)		
	(1) S1	(2) S2	(3) S3	(4) S1	(5) S2	(6) S3
Content Disconfirmation (CD; mean-centered)		−1.169**	−1.254**		−0.673**	−0.616*
		(0.446)	(0.448)		(0.240)	(0.242)
Thumbnail Aesthetics		0.011	0.010		−0.060	−0.059
		(0.147)	(0.147)		(0.079)	(0.079)
Positive Aes. Disconf. (AD^+)		0.327	0.320		0.228	0.233
		(0.295)	(0.295)		(0.159)	(0.159)
Negative Aes. Disconf. (AD^-)		0.314	0.305		0.119	0.125
		(0.272)	(0.272)		(0.147)	(0.147)
Thumbnail Location (percentile)		−0.107	−0.106		−0.127*	−0.128*
		(0.115)	(0.115)		(0.062)	(0.062)
Thumbnail Arousal		0.159**	0.161**		0.011	0.010
		(0.060)	(0.060)		(0.032)	(0.032)
Thumbnail Anger		0.120	0.116		0.025	0.028
		(0.123)	(0.123)		(0.067)	(0.067)
Thumbnail Disgust		−0.071	−0.069		−0.010	−0.011
		(0.096)	(0.096)		(0.052)	(0.052)
Thumbnail Fear		−0.041	−0.043		0.042	0.044
		(0.069)	(0.069)		(0.037)	(0.037)
Thumbnail Joy		0.032	0.033		0.030	0.029
		(0.037)	(0.037)		(0.020)	(0.020)
Thumbnail Sadness		0.074	0.076		0.073	0.071
		(0.083)	(0.083)		(0.045)	(0.045)
Thumbnail Surprise		0.193***	0.195***		0.013	0.012
		(0.036)	(0.036)		(0.020)	(0.020)
CD (mean-centered)			2.205			−1.445*
× Thumbnail Location (percentile)			(1.235)			(0.666)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.4751	0.4832	0.4836	0.1408	0.1456	0.1466
Adjusted R^2	0.4724	0.4790	0.4793	0.1364	0.1387	0.1395
N	4,121	4,121	4,121	4,121	4,121	4,121

Table 2: **Relationship between thumbnail features and video reactions.** OLS estimates. S1: controls only. S2: controls + theory-driven features. S3: S2 + CD×Location interaction. All specifications include controls for caption features, video topics, video mechanical features, video channel features, and days since upload; see Table G.0.3 for OLS estimates on the control variables. Standard errors in parentheses. Significance levels: $^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. [a]On YouTube, a "view" is counted only when a viewer watches at least 30 seconds of the video.

exciting or emotionally charged moment of a video. Surprise, in contrast, usually captures unexpected, unusual or not immediately self-explanatory content, e.g., an exaggerated expression or a dramatic scene, that can make viewers want to understand what is happening in the video and why the scene occurred. Rather than fully revealing the underlying video, such thumbnails may function more as teasers that create uncertainty, which can generate curiosity and motivate viewers to view the content to resolve the uncertainty (Loewenstein, 1994; Wiggin et al., 2019).

Finally, we note that thumbnail aesthetics is not significantly associated with views in the secondary data analysis. One possibility is that aesthetics captures surface-level appeal that can attract clicks but is not necessary enough to sustain viewing past 30 seconds. Our experiment in Section 5 is designed to separate clicking from watchtime, allowing us to examine how aesthetics affects video decisions.

Video Likeability. Table 2 (Columns 4-6) reports the results for video likeability. First, consistent with our previous discussion, thumbnails that are content-wise more different from videos tend to have fewer likes, indicating that thumbnail-content mismatch hurts viewer approval, possibly because the content does not deliver the "promise" set by the thumbnail. Second, we find that videos whose thumbnails appear later in the video receive fewer likes. One interpretation is that viewers are "in the dark" before watching a video and they may use the thumbnail as an anchor in the viewing experience. Before seeing the moment depicted by the thumbnail, viewers worry about when the video will deliver what we expect to see from the thumbnail. Choosing a later thumbnail means that there will be a longer period of time when this expectation is unresolved, thereby hurting viewer approval. We provide supporting evidence of this hypothesis by examining the interaction between content disconfirmation and thumbnail location. The negative and significant interaction indicates that content disconfirmation is *more* damaging to video likeability when the thumbnail is taken from a later part of the video. In other words, when a thumbnail previews later-stage content and shows information that has more mismatch with the video content, viewers appear to respond less favorably.

The secondary data analysis is useful because it examines thumbnail characteristics and video outcomes in natural viewing contexts, allowing us to identify patterns of thumbnails that may operate in real world. However, they are subject to common limitations. First, since outcome variables are at the aggregate level, we lack insights such as whether individuals quit before or after seeing the thumbnail to examine the relationship between content disconfirmation and thumbnail location. Second, the absence of randomization introduces potential endogeneity. The estimates may be confounded by the platform's recommendation algorithm, the presentation of content on the website, or other supply-side factors such as the creator's choice of which thumbnail to use. Thus, while these results provide broad and real-world descriptive evidence for the relationship between thumbnail characteristics and viewing behavior, we are limited in our ability to make causal claims for the optimal choice of thumbnails. To rule out confounding

factors, we move on to building a video website and collecting experimental data, which we discuss next.

5 Creation of a Video Platform for Experimental Design: "CTube" Data

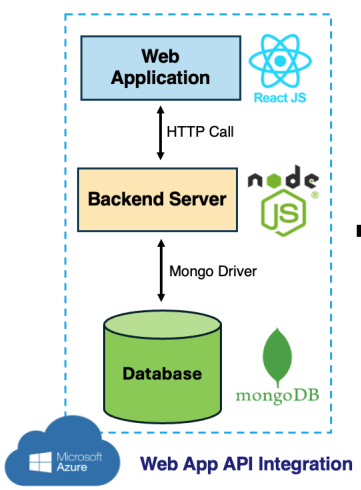
We built an experimental video platform called CTube, a simplified video watching website inspired by YouTube, to collect individuals' click-stream data.[8] Our experimental design provides three key advantages relative to the YouTube secondary data: First, CTube enables us to obtain high-frequency clickstream data at the individual level with respect to video viewing and watching decisions. Second, CTube is not subject to the influence of recommendation algorithms and advertising, therefore providing a clean effect of thumbnails on individuals' video choice and watchtime decisions. Finally, to avoid the creator's choice of the thumbnail, which may be endogenous, we select four different thumbnails (and thus there is variation in underlying thumbnail features) for each video and exogenously randomize the thumbnails each viewer sees. Thus, the experiment provides us a clean randomization to assess the causal effect of thumbnails on viewing behavior. We describe our website design and experiment next.

5.1 Video Platform Design

We hosted 40 short travel videos, totaling 60 minutes of watchtime, for participants to explore on our website. These videos cover a variety of content within the travel category. Participants can freely navigate our website by clicking on the arrow on the right side of each page to browse more videos and clicking on specific videos to watch the content. We utilized several plugins and tools to track any action participants take on our website, including which videos they click on, how long they watch each video, how many video pages they visit, and their page dwell time. Figure 3 illustrates the website architecture and the high-frequency clickstream data we track for each viewer browsing our website. See Web Appendix B for all additional details.

Video Choice Decision. The typical video list page displays four videos at a time, arranged in a

[8] All experimental data were collected with the approval of the Columbia University Institutional Review Board (Protocol AAAT8037). This experiment was preregistered on AsPredicted (#175073, https://aspredicted.org/D3N_T19).



High-Frequency Clickstream Data

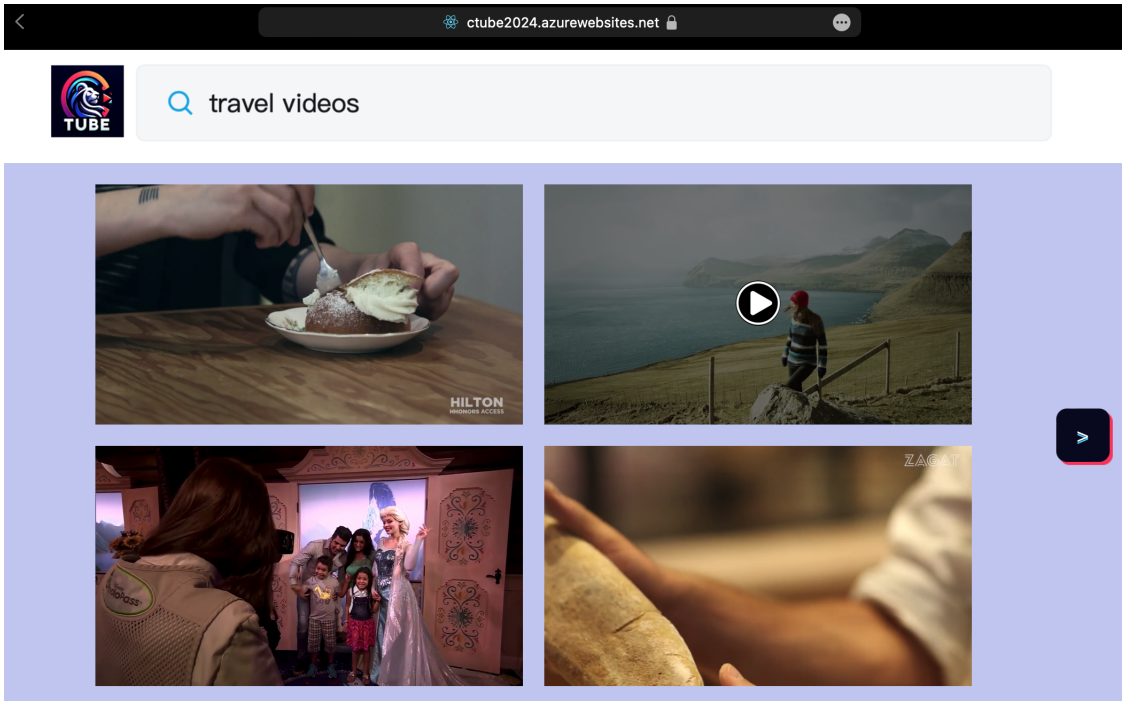
ID	ActionType	VideoId	Timestamp	VideosOnPage
1	Clicked Continue (to Start Page)		2024-05-18 16:52:55	
1	Clicked Right		2024-05-18 16:53:06	S2/V8, S2/V6, S2/V7, S2/V5
1	Clicked Video	V29	2024-05-18 16:53:49	S8/V31,S8/V29, S8/V32, S8/V30
1	Video Played	V29	2024-05-18 16:54:53	
1	Clicked Back	V29	2024-05-18 16:55:26	
1	Clicked Popup Submit		2024-05-18 16:55:28	S8/V31,S8/V29, S8/V32, S8/V30

Figure 3: **Graphical Illustration of The Platform Design and Napkin High-Frequency Data.** This figure shows the platform development and the high-frequency clickstream data of any viewer that is tracked when coming to the "CTube" platform. We hosted CTube on Microsoft Azure (the backend server) within a personal Azure tenant, with the domain managed through DNS bindings. CTube's front end is built with React, a popular JavaScript library for creating user interfaces with dynamic and interactive elements. The data is managed through MongoDB (a flexible NoSQL database that handles data storage and retrieval for real-time access). We integrated an API endpoint to facilitate data download after all users' browsing sessions. The backend event logs are accessible as CSV files.

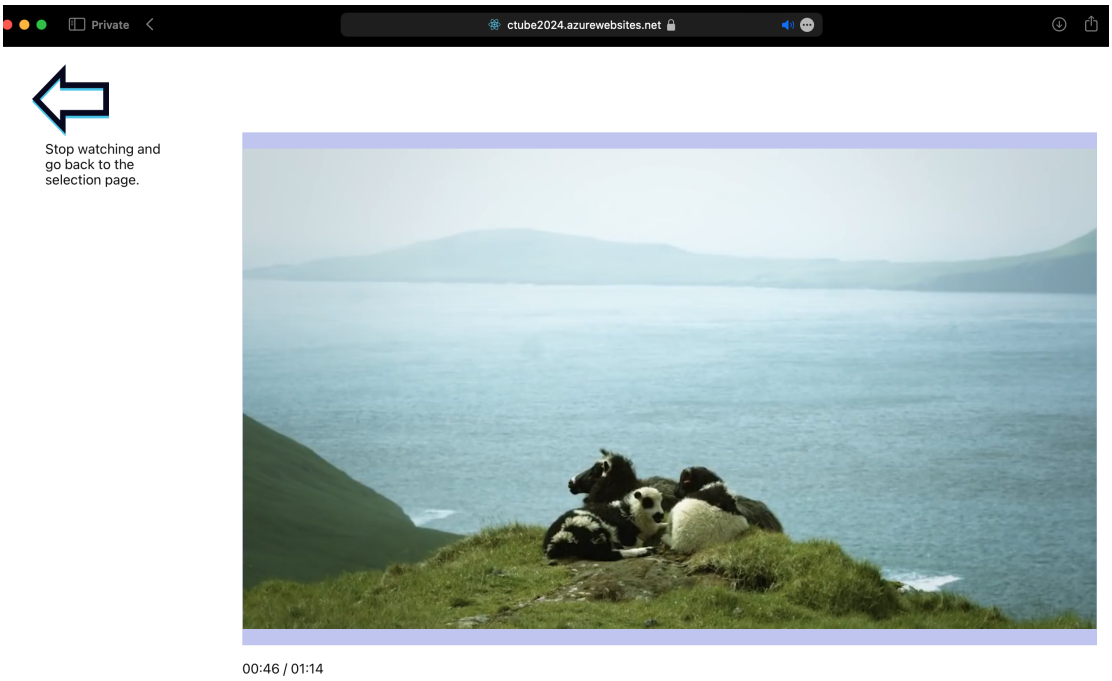
2×2 carousel. We design the video list page so that participants only see the thumbnails of the videos.[9] Participants can choose to watch any or none of the videos on a page. They can also navigate to the next screen by clicking on the arrow at the right side of the screen to see another set of four videos. See Figure 4a for an example of the video choice page.

CTube has several noteworthy features. First, it restricts participants from returning to previously viewed videos once they advance to the next screen. Second, upon finishing watching all four videos from the current page, participants are automatically redirected to the next video choice page. Third, the carousel remains as the 2×2 layout for all video display screens, unlike typical online video platforms where the number of displayed videos can vary when users resize their window or screen size. These designs are implemented to avoid repetitive watching and to simplify the definition of a choice set. Participants are informed about this design and are reminded of this feature before and during the experiment (which we implement as mouseover tooltips where a text message appears when participants hover over the right-click button). See Web Appendix B.3 for all website design details as well as additional website features (e.g.,

[9]This helps simplify our website design, but in typical video websites, viewers might also see information such as the video's creator, title, and number of views or likes.



(a) Video Choice Page



(b) Actual Video Page

Figure 4: **CTube Interface**. Panel (a) shows the interface for the video choice page where participants select travel-related videos to watch. Panel (b) shows the video page after participants select a video to watch.

attention check questions, gaming and inactivity detection).

Video Watchtime Decision. Upon clicking on a video to watch, the participant is directed to the video page, where the video plays automatically upon landing. A widget below the videos displays the watchtime (starting at 00:00/XX:XX and syncing with the video load), helping participants track how long they have watched the video and the remaining time until the end of the video. See Figure 4b for an example of the video page. While watching the video, participants cannot interact with the video navigation bar, i.e., no skipping forward or rewinding is allowed. However, participants can leave the video at any point if they do not wish to continue watching and want to explore other content. Once participants exit the video, they land back at the same choice task the video came from, with one exception: if they have watched all four videos from the current set, they are automatically redirected to the next set of four videos.

Experimental Conditions, and Randomization. We created four versions of thumbnails for each video: (1) the first (non-blank) frame; (2) the most content representative frame; (3) the most aesthetic frame; and (4) the most emotional frame (defined using our extracted features in Section 3.2).[10] We group the 40 videos into 10 pre-defined sets of 4 videos (there are thus $4 \times 40 = 160$ different thumbnails). We randomly assign participants to one of four conditions, which determined which thumbnail they saw for each video. The assignment was structured so that within any set of four videos, each video displays a different thumbnail type, and across conditions, every video is paired with every thumbnail type exactly once. Let S_i be the ith set of four videos on each page and V_{ij} be the j-th video in set S_i, where $j = 1, 2, 3, 4$. For example, in set S_1, a participant in Condition 1 sees the first (non-black) frame for V_{11}, the content-representative frame for V_{12}, the aesthetic frame for V_{13}, and the emotional frame for V_{14}; a participant in Condition 2 sees a different rotation of thumbnails for the same four videos, and so on (see Figure 5 for an example and Web Appendix Table G.0.6 for complete assignment). Each participant remained in the same condition across all 10 video sets. We also randomized the order of video sets and the order of videos within each set across participants, while holding the composition of each set fixed.

[10]The first (non-blank) frame is detected by evaluating the brightness of the frames. We convert the starting frames to grayscale and compute the proportion of pixels exceeding a defined brightness threshold. In our case, we set the threshold to 0.5. The most content representative and aesthetic frame correspond to the frame with lowest content disconfirmation and highest aesthetics from the video sequence. The most emotional frame is the frame with the highest emotion score on any of the six emotions: anger, joy, surprise, sadness, fear, disgust. When ties occur, we randomly select one of the tied candidates.

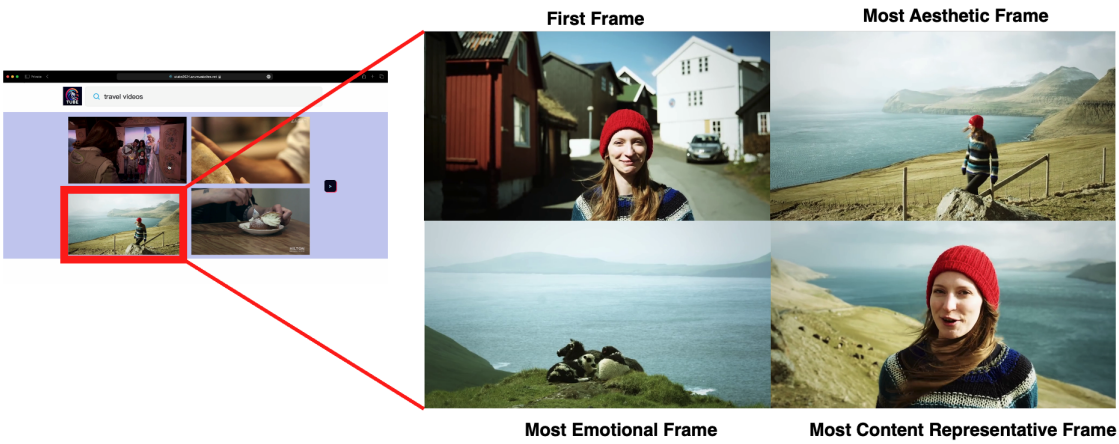


Figure 5: **An Example of Experimental Conditions (Four Versions of Video Thumbnails)**. This figure shows four versions of thumbnails participants could see to start a video. Before participants started browsing any videos on our website, they were randomly assigned to an experimental condition which determined the thumbnail version they would see.

5.2 Experimental Design and Data Collection

The experiment consists of a single session divided into two parts. In the *first* part, participants browse CTube at their own pace, just as they would in a video streaming platform (e.g., YouTube). In the *second* part, after finishing their website browsing and video watching, participants are re-directed to complete a survey about their demographics, video-watching habits, preferences for thumbnails and videos.

Study Procedure. Upon landing on our website, participants were presented with a cover story stating that we were collaborating with a company to build a video streaming website and were interested in testing the content's appeal to viewers like them. Participants were asked to freely allocate 8 minutes of their time to our website to watch videos. They were informed that the website hosted numerous short, travel-related videos totaling 60 minutes of watchtime, from which they could freely choose. Participants were told that they could select any videos they wanted to watch, skip videos, or stop watching a video and choose another video to watch at any time, just like they would do on any video website.

We randomly assigned participants to one of four conditions, from which they saw a different combination of thumbnails for the same set of videos (the videos were always the same across conditions). Once participants completed 8 minutes of video watching on the website, they had the option to continue browsing any content or leave at their discretion. A logout feature was also made available on every video selection page and was made visible after participants watched at

least one video for a few seconds, after which they can leave and complete a final survey.

Subjects. We recruited $N = 523$ participants, out of which $N = 361$ participants had complete and valid browsing and survey data[11] from Prolific, with the restriction that participants need to complete the study on a standard desktop device. The study took on average 18 minutes to complete (12.24 minutes for actual time on the website; remaining on instructions and survey). We also conducted randomization checks on viewers' demographic data. In Web Appendix Table G.0.7, we show that the four experimental conditions are demographically balanced according to gender, age, race/ethnicity, validating randomization in our experiment.

5.3 Descriptive Statistics

Table 3 shows the summary statistics on the video engagement data collected from the CTube experiment, including total and per-page videos watched, total and average video watchtime, and the number of pages visited. First of all, we note that there exists heterogeneity in platform usage across viewers. On average, viewers click through 7 pages and watch 9 videos. Among pages where viewers watched at least one video, they typically watch on average 2 videos per page. Viewers spend an average of 538.4 seconds (approximately 8.9 minutes) watching videos. This matches closely our experimental instructions where viewers are asked to freely allocate 8 minutes of time on watching videos.[12]

To assess whether our experimental platform elicits realistic watchtime behavior, we compare watchtime on CTube with an external benchmark from the YouTube Engagement '16 Dataset (Wu et al., 2018), that reports average watchtime percentage of videos over the first 30 days after upload.[13] After filtering the dataset to the same selection criteria (e.g., English-language travel

[11]Our website does not have a feature to automatically filter out participants who fail attention checks during the study. Therefore, we removed participants by accessing the data only after the data collection. From the 523 responses, we removed 87 (16.6%) participants who failed attention checks (judged either by answers to attention check questions, which are implemented as pop-up windows on our website (see Web Appendix B.3 for details) or based on their data, such as staying on a video choice page for more than 29.86 seconds (95th percentile) without any interactions after the video finishes playing). Additionally, we removed 33 participants (6.3%) who reported experiencing technical issues (e.g., videos not loading) in our final survey. We also removed 42 (8.0%) participants who did not report technical issues but were observed from our high-frequency clickstream data to have multiple videos not loading. We note that we have conducted many tests and pilots on the website before launching it. This is not our server-side issue, but more due to browser compatibility or device limitations on the participants side.

[12]We note that some viewers watch below or beyond 8 minutes which is expected, as an option is given to viewers to continue browsing beyond the 8-minute mark or log out at any time after watching at least one video for a few seconds. This flexibility is deliberately built into the experimental design to ensure realistic watchtime decisions.

[13]YouTube discontinued its web interface for daily watchtime tracking in 2018, so we can no longer obtain videos' watchtime metadata.

videos, 30-120 seconds in length, and current thumbnails, as of October 2025, drawn from video frames), the 3,286 videos left in the sample have an average watchtime proportion of 71%.[14] This closely aligns with the 74% average watchtime proportion we observe from CTube, providing internal validity that our platform can elicit realistic watchtime behavior.

Variable	Mean	Std. Dev.	Min	Median	Max	No. Obs.
No. videos watched	9	4	2	8	32	361
No. videos watched on a page	2	0.7	1	2	4	361
Total video watchtime (sec.)	538.4	197.8	20.0	511.0	1,311.0	361
Average video watchtime (sec.)	62.6	19.1	10.0	62.3	113.3	361
Average video watchtime proportion	0.74	0.21	0.10	0.74	1.0	361
No. pages visited	7	2.7	1	9	10	361
No. pages with videos watched	5	1.9	1	5	9	361
Video length (sec.)	87.4	23.9	34.0	93.0	119.0	40

Table 3: **Summary statistics for CTube data**. Average video watchtime and average video watchtime proportion are computed based on the average values across all watched videos across participants. Watchtime proportion for each video is computed as the actual watchtime of a video divided by its video length.

Table 4 reports the choice probability and average watchtime proportion for each thumbnail type. The most aesthetic frame has the highest click probability, while the first non-black frame performs worst. For watchtime, the most emotional frame receives the lowest watchtime proportion, whereas the other three types are comparable. We also assess statistical significance using paired t-tests at the video level. For each video, we compute the within-video difference in average outcomes between two thumbnail types and test whether the mean difference across the 40 videos is zero. Table 5 reports all six pairwise comparisons with Bonferroni-adjusted p-values. For choice probability, the most aesthetic frame significantly outperforms the first non-black frame ($\Delta = 0.075$, $p_{\text{adj}} < 0.001$) and the most emotional frame ($\Delta = 0.061$, $p_{\text{adj}} < 0.001$). Other choice differences are not significant after adjustment, although the most representative frame is marginally higher than the first non-black frame ($\Delta = 0.041$, $p_{\text{adj}} = 0.082$). For watchtime, no pairwise difference is significant at the 5% level.

We note that simple statistics is informative but limited: First, each thumbnail is a not monolithic treatment, but rather a composite of attributes. These attributes not only vary across thumbnails *within* a video, such as content disconfirmation, but also *across* videos. Thus, using

[14] Mean = 0.71, Median = 0.72, SD = 0.14, Min=0.15, Max=1.0

Frame Type	Choice Probability	Avg. Watchtime Proportion
First non-black frame	0.197	0.706
Most aesthetic frame	0.276	0.706
Most representative frame	0.236	0.699
Most emotional frame	0.211	0.666

Table 4: **Performance of thumbnail types.** Choice probability is the fraction of times a video is clicked when displayed with a given thumbnail type. Watchtime proportion is the fraction of the video watched, conditional on clicking.

	Choice Probability		Watchtime Proportion	
Comparison	Difference	p_{adj}	Difference	p_{adj}
Most aesthetic − First non-black	+0.075	$< 0.001^{***}$	−0.009	1.000
Most aesthetic − Most emotional	+0.061	$< 0.001^{***}$	+0.030	1.000
Most aesthetic − Most representative	+0.035	0.126	−0.002	1.000
Most representative − First non-black	+0.041	0.082^{*}	−0.007	1.000
Most representative − Most emotional	+0.026	0.455	+0.031	0.073^{*}
Most emotional − First non-black	+0.015	1.000	−0.038	0.487

Table 5: **Pairwise tests across thumbnail types.** Each cell reports the mean within-video difference in average outcomes between two thumbnail types, tested using paired t-tests across the 40 videos ($df = 39$). p_{adj} is the Bonferroni-adjusted p-value for six simultaneous comparisons ($p_{\text{adj}} = \min(p \times 6, 1)$). $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

thumbnail types alone cannot capture the complexity of a thumbnail. To explain why certain thumbnails perform better and recommend better thumbnails, we next adopt a modeling approach that decomposes thumbnails into interpretable theory-driven features and estimates how these features affect viewers' video decisions. In this framework, the four experimental thumbnails provide exogenous variation in thumbnail features, allowing us to identify optimal thumbnails for specific videos and viewers. We discuss the model next.

6 Model

Each viewer in the experiment makes two types of decisions: decision of which video to watch at every video selection page, and the decision of when to quit a video she has decided to watch. We link these two decisions using a joint model that uses a multinomial choice model for video selection with a discrete-time survival model for video watchtime through a Frank copula. We

describe each component in turn next.

6.1 Video Choice

Viewer $i \in \{1, \ldots, N\}$ arrives at the first video selection page at occasion $\tau = 1$. On each page, the viewer is presented with 4 videos and can choose to watch any video v. If the viewer does not choose any video to watch, we use $v = 0$ to denote outside option. Let $c_{i,\tau}$ denote viewer i's choice at occasion τ and $C_{i,\tau}$ denote her choice set at τ, $\tau \in \{1, ..., \mathcal{T}_i\}$.[15] Each viewer is randomly assigned to one experimental condition. The condition determines which thumbnail version the viewer sees for each video. Let $\kappa_i(v) \in \{1, 2, 3, 4\}$ denote the thumbnail version of video v shown to viewer i. Thus, although the choice alternative is the video v, the observed thumbnail features vary across viewers through $\kappa_i(v)$. For notational simplicity, we index alternatives by v and write the thumbnail feature vector as $\boldsymbol{x}^c_{iv} \equiv \boldsymbol{x}^c_{v,\kappa_i(v)}$.

We model video choice using a mixed multinomial logit. For each video alternative $v \in C_{i,\tau}$, viewer i receives utility

$$U^c_{iv} = V^c_{iv} + \epsilon^c_{iv}, \qquad V^c_{iv} = \boldsymbol{x}^{c\prime}_{iv}\boldsymbol{\beta}^c_i + \gamma^c_i + \xi^c_v, \tag{6}$$

where $\boldsymbol{x}^c_{iv}$ is the vector of thumbnail features corresponding to the thumbnail version of video v shown to viewer i, including aesthetics, arousal, six specific emotions (anger, disgust, fear, joy, sadness, and surprise), and content disconfirmation, as defined in Section 3.3.1. $\boldsymbol{\beta}^c_i$ captures unobserved heterogeneity in viewers' preferences. γ^c_i is an individual-level random intercept that captures unobserved viewer heterogeneity, ξ^c_v are video fixed effects, and ϵ^c_{iv} is an idiosyncratic error that follows the Type I extreme value distribution. The choice probability of viewer i chooses video $j \in C_{i,\tau}$ is

$$\Pr(c_{i,\tau} = j \mid \gamma^c_i, \boldsymbol{\beta}^c_i, \boldsymbol{\xi}^c) = \frac{\exp(V^c_{ij})}{1 + \sum_{v \in C_{i,\tau}} \exp(V^c_{iv})}. \tag{7}$$

where the utility of the outside option $j = 0$ is normalized to zero.

[15] If the viewer watches a video from the current page, the choice set at the next occasion $\tau + 1$ consists of the videos she hasn't watched from the page: $C_{i,\tau+1} = C_{i,\tau} \setminus \{c_{i,\tau}\}$. If the viewer moves to a new page, $C_{i,\tau+1}$ becomes a new set of 4 videos. By our experimental set-up, a viewer can arrive at a new page either when she chooses the outside option in the last choice occasion or when she finishes watching all 4 videos on the last video selection page.

6.2 Video Watchtime

We model watchtime using a discrete-time hazard framework, which naturally accommodates *time-varying* covariates and *censored* observations. In our setting, right-censoring occurs when viewers finish watching the entire video, so that their latent quit time beyond the video length is unobserved. Because videos differ in length, we normalize watchtime by each video's total duration and estimate the hazard on a standardized relative-time scale. Specifically, we divide the unit interval into $M = 10$ contiguous periods, where period t corresponds to the interval $((t-1)/M, t/M]$ of the video's duration, i.e., every 10% of video progress. This normalization reflects viewing progression where viewers' continuation decisions usually depend on how far they have progressed relative to the video than on the exact number of seconds elapsed. It also allows the baseline hazard to capture viewing dynamics comparable across videos.

Let T_{iv} denote the latent discrete survival time for viewer i watching video v. The discrete hazard, i.e., the probability that viewer i quits video v at the end of period t, conditional on not having quit before period t, is:

$$\lambda_{iv}(t \mid \boldsymbol{x}_{ivt}) = \Pr(T_{iv} = t \mid T_{iv} \geq t, \boldsymbol{x}_{ivt}). \tag{8}$$

We use the complementary log-log link function[16] to parametrize the link between the discrete-time hazard with the covariates and time intervals, with hazard specified as:

$$\log(-\log(1 - \lambda_{iv}(t \mid \boldsymbol{x}_{ivt}))) = \alpha_t^s + \xi_v^s + \gamma_i^s + \boldsymbol{x}'_{ivt}\boldsymbol{\beta}_i^s, \tag{9}$$

where α_t^s is a nonparametric specification for baseline hazard (i.e., time intercepts), ξ_v^s are video fixed effects, γ_i^s captures individual-level unobserved heterogeneity in quitting propensity, and $\boldsymbol{\beta}_i^s$ are viewer-specific random coefficients which capture heterogeneity in preferences for thumbnail features. $\boldsymbol{x}_{ivt}$ contains both thumbnail theory-driven attributes and dynamic measures of the video content observed up until period t, which we discuss next. For notational simplicity, we use the common subscript (i, v, t) for all covariates entering the hazard.

For $\boldsymbol{x}_{ivt}$, we first include all thumbnail theory-driven features as used in observational study:

[16] Another common link function is the logit link function. For short intervals, the two specifications yield nearly identical estimates (Thompson Jr, 1977). Complementary log-log link provides more intuitive interpretation, since the exponential term of any of its estimated coefficients can be directly described as the changed value in hazard.

content disconfirmation, positive and negative aesthetic disconfirmation, thumbnail location, thumbnail aesthetics, arousal, six specific emotions (anger, disgust, fear, joy, sadness, and surprise),[17] and the interaction between content disconfirmation and thumbnail location. These features preserve finer-grained temporal information and are constructed using frames sampled at every second or at every scene (as described in Section 3.3.1) and are then mapped to the corresponding period t, and when viewers quit in the middle of a period, features are computed only using frames actually observed. Several changes apply to adapt the features to hazard model: First, since we can now observe the exact second at which a viewer quits, we construct content disconfirmation as rolling averages of content embeddings over the most recent five seconds, up to and including the current second, before computing the disconfirmation against the thumbnail.[18] Aesthetic disconfirmation is constructed in a similar way. In the hazard setting, the outcome is the risk of quitting at a particular point in the viewing process, and thus we construct disconfirmation to capture whether the recently observed content confirms or violates the expectation induced by the thumbnail. Second, for thumbnail location, we replace the raw percentile location from the observational study with a post-thumbnail indicator PostThumb_{ivt}, which equals zero before the viewer reaches the thumbnail moment in the video and one thereafter.

We further include two time-varying covariates that capture content engagement dynamics. Both measures are constructed from the same GIST embeddings $\mathbf{e}_{vf}$ introduced in Section 3.3.1. Let $\mathcal{F}_{vt} \subseteq \mathcal{F}_v$ denote the set of frames in video v that fall within period t, and let $d(\cdot,\cdot) = 1-\cos(\cdot,\cdot)$ denote the reverse of cosine similarity.[19] The first measure we construct is Boredom_{vt}, which captures accumulative content repetitiveness from the beginning of the video through the end of period t:

$$\text{Boredom}_{vt} = -\sum_{f=2}^{\max \mathcal{F}_{vt}} d\left(\mathbf{e}_{vf}, \mathbf{e}_{v,f-1}\right). \tag{10}$$

A video with frequent content changes accumulates a more negative value, indicating lower boredom. The second is Novelty_{vt}, which captures how far the current content departs from prior content. Let $\bar{\mathbf{e}}_v^{(<f)} = \frac{1}{f-1}\sum_{g=1}^{f-1}\mathbf{e}_{vg}$ be the cumulative prior embedding that summarizes the

[17] For thumbnail-level features that are not time varying within videos, they vary across thumbnails for the same video since viewers are randomly assigned to different experimental thumbnail conditions.

[18] For the first four seconds of a video, the average is computed over all content observed up and including that second. Results are robust to alternative rolling window sizes of ten and twenty seconds; see Web Appendix Tables E.0.2 and E.0.3.

[19] In our case, the cosine similarity between GIST embeddings is always non-negative, so we effectively subtract it from 1 to invert the measure.

content shown before frame f. Content novelty at period t is

$$\text{Novelty}_{vt} = 1 - \frac{1}{|\mathcal{F}_{vt}^{+}|} \sum_{f \in \mathcal{F}_{vt}^{+}} \cos\left(\mathbf{e}_{vf},\ \bar{\mathbf{e}}_{v}^{(<f)}\right), \quad t > 1, \tag{11}$$

where $\mathcal{F}_{vt}^{+} = \{f \in \mathcal{F}_{vt} : f > 1\}$ and $\text{Novelty}_{v1} = 0$. We set $\text{Novelty}_{v1} = 0$ for the first period. As noted, when a viewer quits in the middle of a period, all time-varying covariates for that period are computed using only the frames the viewer actually observed. Since the quitting decision in the discrete-time hazard model is modeled at the end of each period, this ensures that the covariates reflect the content that have actually informed the viewer's decision.

Let $(t_{iv}, \{\boldsymbol{x}_{ivs}\}_{s \le t_{iv}}, \delta_{iv})$ denote the observed survival data for viewer i watching video v, where $t_{iv} \in \{1, \ldots, M\}$ is the last observed period, $\boldsymbol{x}_{ivs}$ is the vector of covariates we discussed before, and δ_{iv} is the event indicator. Specifically, $\delta_{iv} = 1$ if the viewer quits before the video ends, and $\delta_{iv} = 0$ if the viewer watches the video to completion without quitting and is therefore right-censored. Given the discrete hazard in Equation (9), the corresponding survival function is

$$S_{iv}(t) = \Pr(T_{iv} > t) = \prod_{s=1}^{t} \left[1 - \lambda_{iv}(s)\right], \tag{12}$$

whereas for a right-censored observation, the contribution is the probability of surviving beyond the observed final period $S_{iv}(M) = \Pr(T_{iv} > t_{iv} = M)$.

6.3 Joint Model via Copula

Clicking and watching can be two linked stages of the same viewing process. If there are unobserved factors that jointly affect a viewer's click and watchtime decisions even after accounting for observed covariates, viewer heterogeneity and video fixed effects, separate estimation may yield biased survival coefficients. Whether such dependence exists is an empirical question to study.

To allow for residual dependence, our main specification links choice and watchtime decisions through a copula. By Sklar's theorem, the copula separates the dependence structure from the marginal distributions, allowing us to preserve the functional forms of the two marginal models while introducing a dependence parameter θ (e.g. Nelsen, 2006; Trivedi and Zimmer, 2007). In our setting, both outcomes are discrete and thus the joint probabilities are obtained from finite

differences of the copula CDF evaluated at the relevant marginal CDF intervals (e.g. Cameron et al., 2004; Genest and Nešlehová, 2007).

Each viewer-occasion (i, τ) produces a discrete choice $c_{i,\tau} \in C_{i,\tau} \cup \{0\}$, where 0 denotes the outside option. Conditional on clicking video v, the viewer also produces a discrete watchtime outcome t_{iv}. To construct the joint probability, we map each realized outcome into an interval on the unit interval. For the choice margin, let the alternatives in $C_{i,\tau} \cup \{0\}$ follow a fixed ordering.[20] Define the cumulative choice probability as $F^c_{i\tau}(j) = \sum_{m=0}^{j} \Pr(c_{i,\tau} = m), j = 0, 1, \ldots, |C_{i,\tau}|$. Thus, the realized choice $c_{i,\tau} = j^*$ corresponds to the choice-side interval denoted by $(a^-, a^+] = \left(F^c_{i\tau}(j^* - 1), F^c_{i\tau}(j^*)\right]$. For the survival margin, let $F^s_{iv}(t) = 1 - S_{iv}(t)$ denote the CDF of the discrete exit time. An uncensored exit at period t_{iv} corresponds to the survival-side interval denoted by $(b^-, b^+] = \left(F^s_{iv}(t_{iv} - 1), F^s_{iv}(t_{iv})\right]$. For right-censored observations, in which the viewer finishes watching the video and does not quit, the relevant survival event is $T_{iv} > M$, with $\left(F^s_{iv}(t_{iv}), 1\right]$.

We use the Frank copula which allows both positive and negative dependence and treats the two tails symmetrically. Its CDF is

$$C_\theta(a, b) = -\frac{1}{\theta} \log \left[1 + \frac{(e^{-\theta a} - 1)(e^{-\theta b} - 1)}{e^{-\theta} - 1}\right] \tag{13}$$

When $\theta = 0$, there is no dependence and the joint model reduces to separate marginal estimation. Because both margins are discrete, the joint probability mass is obtained from differences of the copula CDF over the relevant CDF rectangles (Cameron et al., 2004). Each viewer-occasion contributes to the likelihood according to one of three cases:

1. *Click and exit (uncensored, $\delta_{iv} = 1$).* The viewer clicks video j^* and quits at period t_{iv}:

$$\Pr(c_{i,\tau} = j^*, T_{iv} = t_{iv}) = C_\theta(a^+, b^+) - C_\theta(a^-, b^+) - C_\theta(a^+, b^-) + C_\theta(a^-, b^-). \tag{14}$$

2. *Click and watch to end (right-censored, $\delta_{iv} = 0$).* The viewer clicks j^* and watches the

[20] Because the multinomial logit is a nominal rather than ordered choice model, this CDF construction requires a fixed ordering of alternatives. We assess sensitivity to alternative orderings in Web Appendix Table E.0.1 and find that the copula parameter θ and the marginal coefficients are virtually unchanged.

full video:

$$\Pr(c_{i,\tau} = j^*,\, T_{iv} > M) = (a^+ - a^-) - \left[C_\theta(a^+, b^+) - C_\theta(a^-, b^+)\right], \tag{15}$$

where $v^+ = F^s_{iv}(M)$.

3. *No click ($c_{i,\tau} = 0$).* No watchtime is observed, and the contribution reduces to the marginal probability of choosing the outside option $\Pr(c_{i,\tau} = 0)$.

The full joint log-likelihood sums over all viewer-occasions:

$$\begin{aligned}\log L = &\sum_{(i,\tau):\, \delta_{iv}=1} \log \Pr(c_{i,\tau} = j^*, T_{iv} = t_{iv}) + \sum_{(i,\tau):\, \delta_{iv}=0} \log \Pr(c_{i,\tau} = j^*, T_{iv} > M) \\ &+ \sum_{(i,\tau):\, c_{i,\tau}=0} \log \Pr(c_{i,\tau} = 0).\end{aligned} \tag{16}$$

6.4 Estimation Procedure

We denote the full parameter set of the joint copula model as $(\boldsymbol{\Theta}, \boldsymbol{\Psi}, \theta)$, where $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_i : i = 1, \ldots, N\}$, where $\boldsymbol{\Theta}_i = (\gamma_i^c, \boldsymbol{\beta}_i^{c\prime}, \gamma_i^s, \boldsymbol{\beta}_i^{s\prime})'$ represent random effect parameters that vary across individuals and $\boldsymbol{\Psi} = \{\boldsymbol{\xi}^c, \boldsymbol{\xi}^s, \boldsymbol{\alpha}^s\}$ represent parameters that are common across individuals, and θ is the copula dependence parameter. We show that all parameters including θ are recoverable via simulation studies in Web Appendix C.

We estimate the joint model using a hierarchical Bayes approach (Rossi and Allenby, 2003). Heterogeneity is introduced into the model for the random effect parameters

$$\boldsymbol{\Theta}_i \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \boldsymbol{\mu}^c \\ 0 \\ \boldsymbol{\mu}^s \end{pmatrix},\ \operatorname{diag}\left(\sigma_\gamma^{2,c},\ \boldsymbol{\sigma}^{2,c},\ \sigma_\gamma^{2,s},\ \boldsymbol{\sigma}^{2,s}\right)\right), \tag{17}$$

where $(\boldsymbol{\mu}^c, \boldsymbol{\sigma}^{2,c}, \sigma_\gamma^{2,c}, \boldsymbol{\mu}^s, \boldsymbol{\sigma}^{2,s}, \sigma_\gamma^{2,s})$ are hyperparameters. We place weakly informative priors on the hyperparameters (normal on population means, half-normal on population standard deviations), diffuse priors on the fixed parameters $\boldsymbol{\Psi}$, and a weakly informative prior on the copula parameter θ. The joint posterior does not have a closed form, so we sample from it using

Hamiltonian Monte Carlo (HMC) with the No-U-Turn Sampler (NUTS). We run four parallel chains, each with 2,000 warm-up and 2,000 sampling iterations. Convergence diagnostics indicate the chains have converged and the posterior estimates are reliable: all monitored parameters have the $\hat{R}$ statistic (Gelman and Rubin, 1992) less than 1.01, and no divergent transitions are detected. See Web Appendix D for additional estimation details.

7 Results

Table 6 reports the posterior means and posterior standard deviations of the joint copula model. The results reveal a central asymmetry: thumbnail features that attract clicks are *fundamentally different* from those that sustains watchtime.

Effect on video choice. Panel A reports the choice estimates. Thumbnail aesthetics is the strongest predictor of clicking, with more visual appealing thumbnails significantly attract more clicks. This is consistent with marketing literature showing that visual design and aesthetics shape consumer attention and product decisions (Bloch, 1995; Hagtvedt and Patrick, 2008; Pieters et al., 2010). Arousal also matters, with high-energy thumbnails being more effective at driving clicks. Notably, all six discrete emotions are positive and significant, regardless of valence. This suggests that emotionally salient thumbnails attract clicks relative to showing no emotion (neutral as baseline). Together these suggest that affective stimuli are preferentially processed and can influence consumer response (Pham et al., 2013).

In terms of heterogeneity, we find that viewers show remarkably little variation in *what* features makes them click. In other words, what makes thumbnails attractive is universal; viewers differ mainly in how actively they engage with the platform (as reflected by viewer random intercepts).

Finally, although viewers have not yet seen the video content when making the click decision, we include content disconfirmation and find that it also significantly increases clicking. Imagine that creators are strategic in their decisions and use a thumbnail that differs from the rest of the video (where the information gap could stand out as more novel or weird and induce curiosity (Loewenstein, 1994)), this is suggesting that such thumbnails could increase viewers' incentive to click. However, next for the survival-side estimates, we show that the same feature that attract clicks may reduce watchtime: content disconfirmation makes the thumbnail more clickable ex

	Mean		Heterogeneity (SD)	
	Estimate	95% CI	Estimate	95% CI
Panel A: Video Choice				
Thumbnail Aesthetics	0.500	[0.388, 0.612]	0.038	[0.002, 0.102]
Thumbnail Arousal	0.196	[0.070, 0.323]	0.033	[0.001, 0.093]
Thumbnail Anger	0.194	[0.105, 0.283]	0.029	[0.001, 0.082]
Thumbnail Joy	0.170	[0.086, 0.257]	0.038	[0.002, 0.102]
Thumbnail Sadness	0.142	[0.087, 0.196]	0.039	[0.002, 0.101]
Thumbnail Disgust	0.310	[0.218, 0.400]	0.034	[0.001, 0.095]
Thumbnail Fear	0.140	[0.088, 0.193]	0.041	[0.002, 0.103]
Thumbnail Surprise	0.147	[0.090, 0.205]	0.029	[0.001, 0.082]
Content Disconfirmation	0.130	[0.087, 0.173]	0.026	[0.001, 0.069]
Individual Random Intercepts (γ_i^c)	—		1.506	[1.274, 1.766]
Video Fixed Effects (ξ_v^c)		✓		
Viewers / Videos / Choice Occasions		361 / 40 / 4,137		
Panel B: Video Watchtime				
Content Disconfirmation (CD, mean-centered)	0.347	[0.252, 0.445]	0.190	[0.045, 0.295]
Post-Thumbnail Indicator	−0.123	[−0.191, −0.055]	0.126	[0.012, 0.232]
CD (mean-centered) × Post-Thumbnail Indicator	−0.141	[−0.233, −0.048]	0.131	[0.009, 0.244]
Pos. Aesthetic Disconfirmation	−0.143	[−0.248, −0.038]	0.152	[0.019, 0.264]
Neg. Aesthetic Disconfirmation	0.007	[−0.058, 0.071]	0.048	[0.002, 0.129]
Boredom	0.363	[0.256, 0.472]	0.308	[0.225, 0.390]
Novelty	−0.087	[−0.154, −0.019]	0.322	[0.254, 0.389]
Thumbnail Aesthetics	0.117	[−0.019, 0.254]	0.087	[0.005, 0.204]
Thumbnail Arousal	−0.006	[−0.095, 0.082]	0.246	[0.147, 0.338]
Thumbnail Joy	−0.061	[−0.162, 0.041]	0.081	[0.004, 0.189]
Thumbnail Anger	0.015	[−0.055, 0.085]	0.072	[0.005, 0.155]
Thumbnail Disgust	0.066	[−0.012, 0.143]	0.180	[0.075, 0.283]
Thumbnail Fear	−0.003	[−0.092, 0.087]	0.108	[0.007, 0.233]
Thumbnail Sadness	−0.075	[−0.170, 0.019]	0.106	[0.006, 0.221]
Thumbnail Surprise	−0.038	[−0.114, 0.034]	0.067	[0.003, 0.173]
Individual Random Intercepts (γ_i^s)	—		1.058	[0.950, 1.175]
Video Fixed Effects (ξ_v^s)		✓		
Time Fixed Effects (Baseline Hazards) (α_t^s)		✓		
Viewers / Videos / Viewer-Video Pairs / Obs. / Periods		361 / 40 / 3,312 / 23,225 / 10		
Copula Parameter				
θ (Frank)	−0.160	[−0.785, 0.474]		

Table 6: **Joint copula model estimation results.** The estimated time intercepts are baseline hazard in the discrete-time survival model; see Web Appendix Figure G.0.2. In our data, 28.3% of viewer-video pairs watched the videos to completion (right-censored).

ante, but once viewing begins and viewers see the information gap as the viewer unfolds, they may quit from the video.

Effect on video watchtime. Panel B reports the survival estimates, where positive coefficients increase the quit hazard (i.e., reduce watchtime). The results reveal a striking asymmetry: thumbnails can continue to affect watchtime, though not through the same way that affects clicks. None of the static thumbnail features—aesthetics, arousal, or the six discrete emotions—

significantly affects the quit hazard. Once a viewer begins watching, the thumbnail's visual appeal and emotional intensity become irrelevant for continued engagement. However, thumbnails may still matter for watchtime by shaping viewers' expectations and how they interpret the video content as it unfolds. We find several empirical patterns that support the claims.

First, we find that viewers are more likely to quit when the unfolding video content diverges from the thumbnail. This effect reveals a moment-by-moment expectation violation from what the thumbnail promised: viewers may form an expectation of the video content based on the thumbnail; when the video delivers content not well aligned with their expectation, viewers feel the mismatch and are more likely to quit the video.

Second, the thumbnail may serve as temporal anchor within the watching experience. Once the viewer reaches the point in the video where the thumbnail's depicted scene appears, the quit hazard drops. This is consistent with the resolution of the information gap: before the thumbnail moment, viewers face uncertainty about whether the video will deliver on its promise, and this uncertainty may contribute to quitting. After the thumbnail moment, the promise has been fulfilled and the uncertainty resolved, reducing the viewer's propensity to leave.

Third, the interaction between content disconfirmation and thumbnail location indicator reveals how these two forces resolve over time. Up to the location in the video of the frame that was used as the thumbnail, viewers are more likely to quit when the recently observed content differs from the thumbnail. After the thumbnail moment appears, viewers are less concerned with subsequent content differences. This pattern is consistent with the idea that once the video delivers the expectations promised by the thumbnail, later deviations are less likely to be perceived as mismatch.

In addition, we find that while thumbnail aesthetics itself does not directly affect watchtime, a more visually appealing thumbnail over its video (positive aesthetic disconfirmation) may create a positive halo that carries over into the watching experience, thereby reducing quit. This finding is consistent with the halo effect literature, in which an initial favoriate impression can shape subsequent evaluations and experiences (Thorndike, 1920; Nisbett and Wilson, 1977). We do not find evidence that the halo effect can also operate in the opposite direction, suggesting that viewers may not penalize a plain thumbnail that leads to an aesthetically pleasing video, but they do reward a beautiful thumbnail that sets a pleasing experience from the outset.

Finally, the two time-varying covariates that capture content dynamics, independent of

the thumbnail, affects quitting. Boredom reflected by repetitive content drives viewers away. Conversely, content novelty, which measures how different the current content is from everything the viewer has seen so far, significantly reduces quitting. Together, these effects indicate that content variety is also another primary driver of sustained watchtime.

We note that in contrast to video choices, the survival-side estimates reveal substantial viewer heterogeneity in *how* they respond to video content. The largest heterogeneity are in boredom and novelty, where some viewers tolerate repetitive content while others quit rapidly. Content disconfirmation, positive aesthetic disconfirmation, and location indicator and the interaction term also show meaningful heterogeneity, implying that the effect operates with different intensities across viewers.

Dependence between clicking and watching. In this experiment, we find limited negative residual dependence ($\hat{\theta} = -0.163$, 95% CI $[-0.805, +0.454]$) that is not statistically different from zero between the click and watchtime decisions after accounting for all other factors such as observed covariates and viewer heterogeneity.[21] This suggests that the thumbnail effects on choice and watchtime reported above are unlikely to be driven by unobserved selection into watching that remains after these controls.[22]

8 Thumbnail Selection

8.1 Counterfactual Thumbnail Evaluation

Through the experimental design, we have shown that thumbnails affect both video choice and watchtime. We next use the estimated joint copula model described in Section 6.3 to evaluate how to select thumbnails to optimize different outcomes. For each focal video v, we replace a viewer's observed thumbnail at each choice occasion with an alternative frame k from thumbnail candidate pool, which we describe shortly, and compute the model-implied outcomes by averaging over posterior draws. Because a viewer's choice depends on the thumbnails of all videos on the page, we average over all possible experimental-thumbnail configurations of the

[21] For the Frank copula, $\theta \in (-\infty, \infty)$, with $\theta = 0$ corresponding to independence.

[22] We note that if we apply an alternative click or view definition requiring at least 30 seconds of watching, the copula parameter becomes significantly negative ($\hat{\theta} = -1.398$, 95% CI $[-2.214, -0.595]$). This suggests that the model is capable of detecting dependence when it exists and that the null result under the main specification reflects a genuine absence of residual selection, not a lack of statistical power.

other three videos on the same page when computing for clicking, based on our experimental design.[23] For each candidate frame k, we report three performance measures: choice probability ($\widehat{P}_{vk}$), expected watchtime conditional on clicking ($\widehat{T}_{vk}$), and joint performance ($\widehat{J}_{vk}$), defined as the expected watchtime weighted by the probability of clicking,[24] where

$$\widehat{P}_{vk} = \Pr(c = v \mid k), \tag{18}$$

$$\widehat{T}_{vk} = \mathbb{E}[T_v \mid c = v, k]\,, \tag{19}$$

$$\widehat{J}_{vk} = \mathbb{E}[\mathbf{1}\{c = v\}T_v \mid k] \tag{20}$$

For each focal video, because scenes within a video are relatively homogeneous, we construct a candidate thumbnail pool by selecting a frame from each scene. We then compute two types of model-optimized frames: (i) the *global-best* frame, which is selected per video by searching over that video's candidate pool and choosing the frame that maximizes the predicted outcome of interest, averaged across viewers; (ii) the *personalized* frame, which allows the maximizing candidate to vary across viewer–video pairs. We benchmark these model-predicted optima against several alternatives that reflect current practice. These include: (i) the *actual frame* used by the video's creator on YouTube; (ii) three *platform-recommended frames* generated by YouTube, including the best-performing frame among the three per metric (see Web Appendix F); (iii) a set of *stylized frames*—first non-black, most aesthetic, most content-representative, most emotional, most surprising, highest-arousal, and most summarizing frame,[25] and, finally, (iv) a *random frame* selected from the video sequence.

Table 7 reports the mean estimated performance of each candidate frame per metric (see Web Appendix Figures G.0.3 and G.0.4 for the distributions). First, there is a click-watchtime tradeoff when it comes to thumbnail selection. The highest-arousal and most aesthetic frames rank the highest on choice probability, but perform near the bottom on watchtime. Conversely, the

[23] In the experiment, each video can appear with one of four experimental thumbnails. Thus, for each focal-video candidate, we average over $4^3 = 64$ experimental-thumbnail configurations, yielding the focal video's expected performance under all experimental condition combinations.

[24] $\widehat{J}_{vk}$ is computed under copula rectangle probabilities. It coincides with $\widehat{J}_{vk} = \widehat{P}_{vk} \times \widehat{T}_{vk}$ under independence $\theta = 0$. Since our estimated $\theta = -0.16$ is near zero, they'll be numerically very close.

[25] See Footnote 10 for frame definitions. The most surprising and highest-arousal frames are intuitive. The most summarizing frame is constructed using Gemini-2.5-Pro (Comanici et al., 2025) by prompting the model: "Please identify the frame that best summarizes the video's story." The output is a frame denoted by whole second. We manually inspect the 40 videos and their summaries to ensure that the selection reflects human intuition.

most content-representative and summarizing frame, which performs well on watchtime among stylized candidates, does not work well for clicks. Indeed, when we turn to the average feature profiles of these frames in Table 8), we notice that frames that attract clicks are characterized by high aesthetics and high arousal, aligning with our model estimates from Table 6. Frames that sustain watchtime are on average characterized by low content disconfirmation and early-to-mid video location, which are features that align well with the thumbnail's promise with the video's content and resolve viewer expectations quickly. The most emotional frame illustrates the cost of ignoring this distinction: despite depicting the most emotionally intense moment, it has relatively low aesthetics and surprise, high content disconfirmation, and is typically drawn from late in the video (at 86% of duration), creating a prolonged expectation gap that increases quitting before the viewer reaches the depicted scene. This suggests that the most effective strategy is to seek frames that combine high visual and emotion impact with reasonable content alignment and an early-to-mid location in the video to improve both clicks and watchtime.

Second, a randomly selected frame performs nearly identically to the actual creator-chosen frame across metrics, suggesting that creators' current thumbnail choices are roughly on par with chance. To understand what drives creators' intuition, we examine how often creator's actual choice overlaps with each candidate (see Web Appendix Table G.0.8). The creator's actual frame most frequently overlaps with the platform's recommended middle frame (which has low content disconfirmation and is also the platform default; see Web Appendix F) and the most content-representative frame. This suggests that creators gravitate more toward showing "what the video is about" to sustain watchtime. However, those who choose based on "what catches the eye" can do substantially better for joint performance, because the click gains outweigh the modest watchtime costs. Creators could afford to tilt slightly toward visual impact over content representativeness without sacrificing watchtime when selecting thumbnails. YouTube's platform-recommended frames, which appear to prioritize temporal coverage (left frame capturing content close to the start ~14%, middle frame ~50%, and right frame ~80%) over visual or emotional features (See Table 8), also leave potential click gains on the table than feature-based selection.[26] For creators who prefer to select from the platform's default recommendations rather than manually searching for frames, the results suggest that they could

[26]The overlap of creator's frame and platform's frames is about 70%, suggesting that the platform-recommended frames were recorded may not perfectly reflect the same recommendations available to creators when they originally uploaded their videos.

choose to adopt the left frame more, which on average achieves the highest watchtime and joint performance without hurting too much on clicks.

Finally, while heuristic strategies often face tradeoff, our model-optimized thumbnails can substantially improve *both* simultanesously, outperforming even the best of the three platform-recommended frames on each metric. While the global-best thumbnail is already optimal for the majority of viewers, personalization yields additional watchtime for 39% of viewer-video pairs, reaching up to 39 seconds gains for specific viewers. This suggests that personalization can help unlock watchtime and thus joint performance improvements without sacrificing clicks. Even modest per-exposure improvements compound at scale, where each additional second of expected watchtime per click, aggregated across millions of daily views, can translate to thousands of additional hours of daily engagement.

Thumbnail Candidate (Per Video)	Choice Prob. ($\widehat{P}$)	Watchtime ($\widehat{T}$)	Joint Perf. ($\widehat{J}$)
Optimal frames			
Global best (per metric)	0.355	8.12	2.536
Personalized (per viewer and metric)	0.355	8.23	2.575
Baselines			
Random frame	0.212	7.35	1.566
Actual (creator-chosen)	*0.205*	*7.49*	*1.550*
Platform-recommended frames			
Best of 3 Platform (per metric)	0.231	7.75	1.693
Left Frame (∼14%)	0.191	7.70	1.501
Middle Frame (∼50%)	0.200	7.22	1.465
Right Frame (∼80%)	0.206	7.05	1.461
Stylized frames			
Highest Arousal	0.263	7.25	1.918
Most Aesthetic	0.254	7.43	1.878
Most Surprising	0.232	7.17	1.669
Most Emotional	0.204	7.00	1.453
Most Representative	0.207	7.51	1.571
Most Summarizing	0.193	7.43	1.443
First Non-Black	0.203	7.42	1.529

Table 7: **Average Predicted performance of thumbnail candidates.**

Candidate	Aesthetics	Arousal	CD	Anger	Disgust	Fear	Joy	Sadness	Surprise	Max Emo	Location
Optimal frames											
Global (for click)	4.41	4.67	0.37	1.35	1.20	1.45	3.85	1.50	2.15	4.25	46s (52%)
Global (for watchtime)	4.43	4.15	0.34	1.02	1.00	1.20	4.83	1.48	1.88	4.97	12s (14%)
Global (for joint)	4.48	4.67	0.37	1.27	1.18	1.45	3.85	1.55	2.23	4.33	33s (38%)
Personalized (for click)	4.42	4.69	0.37	1.34	1.20	1.45	3.82	1.50	2.19	4.25	46s (52%)
Personalized (for watchtime)	4.40	4.18	0.33	1.02	1.01	1.20	4.83	1.43	1.91	4.83	14s (17%)
Personalized (for joint)	4.49	4.63	0.37	1.25	1.14	1.44	3.99	1.54	2.21	4.35	33s (38%)
Baselines											
Random Frame	4.14	4.10	0.34	1.05	1.00	1.23	4.83	1.25	2.00	4.92	42s (49%)
Actual (creator)	4.24	4.30	0.26	1.02	1.02	1.20	4.95	1.30	1.82	4.95	43s (50%)
Platform-recommended frames											
Best of 3 (for click)	4.25	4.50	0.29	1.05	1.05	1.30	5.17	1.20	1.90	5.20	51s (59%)
Best of 3 (for watchtime)	4.21	4.17	0.27	1.00	1.00	1.23	4.95	1.23	1.95	4.95	23s (25%)
Best of 3 (for joint)	4.28	4.35	0.28	1.05	1.05	1.32	5.03	1.23	1.93	5.05	44s (49%)
Left Frame	4.14	4.12	0.28	1.02	1.00	1.18	4.88	1.23	1.90	4.88	13s (14%)
Middle Frame	4.09	4.47	0.28	1.02	1.02	1.25	5.12	1.12	1.88	5.15	43s (50%)
Right Frame	4.16	4.28	0.28	1.02	1.02	1.25	5.15	1.20	1.88	5.15	71s (80%)
Stylized frames											
Highest Arousal	4.48	**5.10**	0.28	1.07	1.00	1.38	4.62	1.27	2.00	4.72	55s (62%)
Most Aesthetic	**4.79**	4.17	0.29	1.02	1.05	1.23	4.85	1.30	1.90	4.85	49s (54%)
Most Surprising	4.17	4.28	0.31	1.02	1.00	1.20	4.50	1.15	**2.83**	4.95	54s (62%)
Most Emotional	4.17	4.35	0.31	1.02	1.00	1.02	5.92	1.15	1.80	**5.92**	76s (86%)
Most Representative	4.09	4.40	**0.26**	1.05	1.00	1.30	4.75	1.32	2.02	4.90	38s (47%)
Most Summarizing	3.99	4.33	0.28	1.05	1.00	1.23	5.05	1.20	1.90	5.05	41s (49%)
First Non-Black	4.17	4.05	0.31	1.02	1.05	1.15	4.70	1.15	2.02	4.70	1s (1%)

Table 8: **Feature profiles of thumbnail candidates.**

8.2 Predictive Performance and Personalization

Finally, we explore the value of customized thumbnails for individuals using the model by leaving one video out for each individual for prediction and treating the other watched videos as user histories when re-calibrating the joint copula model.

Predictive Ability. We first evaluate the model's ability to predict video choice and watchtime given the thumbnails viewers saw in the experiment. Two findings emerge from model comparison in Table 9. First, the joint model predicts held-out choices and watchtime significantly better than the homogeneous specification, suggesting that viewer heterogeneity is essential for predictive performance. Second, the copula has little effect on marginal predictive accuracy, consistent with the estimated dependence parameter being close to zero (see Table 4). This does not undermine the role of the copula in the joint model: the copula is included to allow residual dependence between clicking and watching, and we empirically find that there is unlikely unobserved selection into watching that remains after accounting for thumbnail effects and viewer heterogeneity, and more.

We also note that without heterogeneity, the hit rate of a homogeneous specification is same to that of a random choice (hit rate 22.0%) that assigns equal probability to the alternatives available in each occasion (which vary across occasions depending on remaining videos on the page); for watchtime, it's also not significantly better than a random-period rule calibrated to the empirical distribution of quit and completion outcomes in the calibration sample (35.6%). A priori, the accuracy of the prediction will depend on the number of choice alternatives and watchtime periods. For occasions where there are only 2 alternatives, the hit rate can go up to 88.9%.

	Full model	No Copula	No Heterogeneity
Calibration:−2log likelihood	19,370	19,377	34,032
Validation:−2log likelihood	2,388	2,390	5,748
Choice hit rate	31.0%	30.7%	22.0%
Watchtime hit rate (Within the same period)	67.4%	67.8%	40.1%

Table 9: **Fit and Predictive Ability Measures.** Both choice and watchtime (periods) are discrete. For watchtime, the hit rate metric respects the discrete-time hazard structure, where quitting decisions are modeled at period boundaries rather than continuously within a period.

Value of Personalization. We note that our calibration and full-sample estimates are nearly identical (mean absolute difference of 0.019 across all parameters), suggesting that removing the last choice occasion does not materially distort the model. In Web Appendix Table G.0.9, we report the predicted personalized best thumbnail against key alternative thumbnails and confirm that one-size-fits-all thumbnail selection leaves substantial value on the table.

9 General Discussion

In this paper, we provide a comprehensive analysis to understand how thumbnails (relative to the video they represent) impact video decisions. We conceptualize a thumbnail not merely as a static image, but as a visual representation of the underlying video viewers will experience. To address this research question, we use both broad secondary data from YouTube across thousands of videos and thumbnails, as well as an experiment using a video platform we created to fully control the randomization of thumbnails across videos and track users' viewing behaviors. Using secondary and experimental data and combining video analysis, multimodal LLM, computer

vision, and a model estimated using experimental data, we provide several insights on how one should select and design thumbnails.

First, thumbnails create a tension between attracting viewers and sustaining their engagement. To increase views, one should consider choosing visually appealing, high-arousal, and emotionally engaged thumbnails. These “teaser-style" thumbnails are effective at generating initial interest and can help videos stand out in crowded choice environments. However, these same features do not necessarily translate into longer watchtime after the click.

Second, thumbnails can still affect watchtime through how well they represent, summarize, or preview the video content and, in doing so, set viewers' expectations about the video experience. To increase watchtime, one should therefore consider not only whether a thumbnail is visually appealing enough compared to the video, but also whether it accurately communicates what viewers will encounter after clicking and whether the promised content appears early enough to sustain their interest. Viewers are more likely to stop watching when the content they observe diverges from the thumbnail, but this penalty is reduced after seeing the moment depicted in the thumbnail.

Third, to balance views and watchtime, one should avoid simple thumbnail-selection heuristics. Choosing the most aesthetically pleasing, emotional, high-arousal, or content-representative frame can improve one performance dimension, but often creates tradeoffs along another. Instead, the effective thumbnails for joint performance is to balance visual impact, content alignment, and the timing of the moment chosen to depict in the thumbnail. Our model-optimized thumbnails are aimed to balance these design rules, which outperform stylized rules, as well as creator-selected and even the best of the three platform-recommended thumbnails

Fourth, we emphasize that there is no one-size-fits-all thumbnail that works well for all viewers, videos, and performance metrics. The best thumbnail depends on whether the goal is to maximize clicks, watchtime, or a joint measure of both. It also depends on the content trajectory of the video and on viewer-level heterogeneity in how consumers respond to visual appeal, emotion, content mismatch, and delayed fulfillment of the thumbnail promise. Consistent with this, we show that allowing thumbnails customized to viewers can yield additional watchtime gains for a meaningful share of observations. Thus, thumbnail design should be customized not only to the video and the objective, but also, when possible, to the viewer. Rather than asking which thumbnail is best in general, creators should ask which thumbnail best communicates this

video to this particular target audience for this performance goal.

Finally, leaving thumbnail selection to simple platform defaults and creator intuition can be costly. Creator-selected thumbnails on average perform roughly on par with random frames. We suggest that thumbnail selection should be treated as an integral part of the video creative and promotion workflow, rather than as a default or afterthought.

Our research provides several promising avenues for future research. First, to make it a constrained optimization problem, we only focus on static thumbnails which are a part of a video's frames. One exciting future direction would be to extend our analysis to animated or customized thumbnails, video previews or video hooks. One could even combine insights from our research with generative AI to help streamline the production of thumbnails or video hooks. Second, we focus on within-video thumbnail selection because we think it is a first-order question to understand what makes an effective thumbnail for each video. Future research can study thumbnail selection in competitive environments and how the optimal thumbnail depends on the platform ranking algorithms and competitors' thumbnail strategies. Finally, there are opportunities to extend our work to other contexts such as book covers, movie posters or trailers to provide suggestions on how to design or choose an optimal image for such products. We hope that our work will spark more research on these important topics.

References

Eugene W Anderson and Mary W Sullivan. The antecedents and consequences of customer satisfaction for firms. *Marketing science*, 12(2):125–143, 1993.

Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. Selecting a diverse set of aesthetically-pleasing and representative video thumbnails using reinforcement learning. In *Proceedings of the 2023 IEEE International Conference on Image Processing*, 2023. doi: 10.1109/ICIP49359.2023.10222743.

Shervin Ardeshir, Nagendra Kamath, and Hossein Taghavi. Character-focused video thumbnail retrieval, 2022.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*, 2020.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Peter H. Bloch. Seeking the ideal form: Product design and consumer response. *Journal of Marketing*, 59(3):16–29, 1995.

Roberto Brunelli. *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons, 2009.

Alex Burnap, John R. Hauser, and Artem Timoshenko. Product aesthetic design: A machine learning augmentation. *Marketing Science*, 42(6):1029–1056, 2023. doi: 10.1287/mksc.2022.1429.

A. Colin Cameron, Tong Li, Pravin K. Trivedi, and David M. Zimmer. Modeling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *The Econometrics Journal*, 7(2):566–584, 2004.

Jingcun Cao, Xiaolin Li, and Lingling Zhang. Is relevancy everything? a deep-learning approach to understand the effect of image-text congruence. *Management Science*, 71(12):10579–10602, 2025. doi: 10.1287/mnsc.2022.01896.

Gizem Ceylan, Kristin Diehl, and Davide Proserpio. Words meet photos: When and why photos increase review helpfulness. *Journal of Marketing Research*, 61(1):5–26, 2024. doi: 10.1177/00222437231169711.

Ishita Chakraborty, Khai Chiong, Howard Dover, and K Sudhir. Can ai and ai-hybrids detect persuasion skills? salesforce hiring with conversational video interviews. *Marketing Science*, 2024.

George Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Geng Cui, Sebastian Yu-Ho Chung, Ling Peng, and Qiaofei Wang. Clicks for money: Predicting video views through a sentiment analysis of titles and thumbnails. *Journal of Business Research*, 183:114849, 2024.

Ryan Dew, Asim Ansari, and Olivier Toubia. Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science*, 41(2): 401–425, 2022. doi: 10.1287/mksc.2021.1326.

Daria Dzyabura, Siham El Kihal, John R. Hauser, and Marat Ibragimov. Leveraging the power of images in managing product return rates. *Marketing Science*, 42(6):1125–1142, 2023. doi: 10.1287/mksc.2023.1451.

Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

Yannick Exner, Jochen Hartmann, Oded Netzer, and Shunyuan Zhang. Ai in disguise-how ai-generated ads' visual cues shape consumer perception and performance. *Available at SSRN*, 5096969, 2025.

Elia Fantini. Automating video thumbnails selection and generation with multimodal and multistage analysis, 2024.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

Christian Genest and Johanna Nešlehová. A primer on copulas for count data. *ASTIN Bulletin*, 37(2):475–515, 2007.

Henrik Hagtvedt and Vanessa M. Patrick. Art infusion: The influence of visual art on the perception and evaluation of consumer products. *Journal of Marketing Research*, 45(3): 379–389, 2008.

Jochen Hartmann, Mark Heitmann, Christina Schamp, and Oded Netzer. The power of brand selfies. *Journal of Marketing Research*, 58(6):1159–1177, 2021. doi: 10.1177/00222437211037258.

Yepeng Jin, Ishita Chakraborty, and Kevin Chung. Advancing marketing video analysis with multimodal llms and context-aware graphrag. *Available at SSRN 5517439*, 2025.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

Y. Kim et al. From visuals to value: Leveraging generative ai to explore visual elements of movie posters and commercial success. *Journal of Business Research*, 2025.

Byungwan Koh and Fuquan Cui. An exploration of the relation between the visual attributes of thumbnails and the view-through of videos: The case of branded video content. *Decision Support Systems*, 160:113820, 2022.

John P Lewis et al. Fast template matching. In *Vision interface*, volume 95, pages 15–19. Quebec City, QC, Canada, 1995.

Yiling Li, Hye-jin Kim, Boram Do, and Jeonghye Choi. The effect of emotion in thumbnails and titles of video clips on pre-roll advertising effectiveness. *Journal of Business Research*, 151: 232–243, 2022.

Yiyi Li and Ying Xie. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1):1–19, 2020.

Bo Liu. Towards micro-video thumbnail selection via a multi-label visual-semantic embedding model, 2022.

Liu Liu, Daria Dzyabura, and Natalie Mizik. Visual listening in: Extracting brand image portrayed on social media. *Marketing Science*, 39(4):669–686, 2020.

Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3707–3715, 2015.

Xuan Liu, Savannah Wei Shi, Thales Teixeira, and Michel Wedel. Video content marketing: The making of clips. *Journal of Marketing*, 82(4):86–101, 2018.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

George Loewenstein. The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116(1):75–98, 1994.

Lan E. Luo. How visual designs drive success: Interpretable generative ai for data-driven design. Working paper, 2026.

Xueming Luo, Nan Jia, Erya Ouyang, and Zheng Fang. Introducing machine-learning-based data fusion methods for analyzing multimodal data: An application of measuring trustworthiness of microenterprises. *Strategic Management Journal*, 45(8):1597–1629, 2024.

Luca Marchesotti, Naila Murray, and Florent Perronnin. Discovering beautiful attributes for aesthetic image analysis. *International journal of computer vision*, 113(3):246–266, 2015.

Satya Menon and Dilip Soman. Managing the power of curiosity for effective web advertising strategies. *Journal of Advertising*, 31(3):1–14, 2002. doi: 10.1080/00913367.2002.10673672.

Ilya Morozov and Anna Tuchman. Where does advertising content lead you? we created a bookstore to find out. *Marketing Science*, 43(5):986–1001, 2024.

Roger B Nelsen. *An introduction to copulas*. Springer, 2006.

Richard E. Nisbett and Timothy D. Wilson. The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4):250–256, 1977.

Minxue Niu, Yara El-Tawil, Amrit Romana, and Emily Mower Provost. Rethinking emotion annotations in the era of large language models. *arXiv preprint arXiv:2412.07906*, 2024.

Richard L Oliver. Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of applied psychology*, 62(4):480, 1977.

Richard L Oliver. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of marketing research*, 17(4):460–469, 1980.

Gijs Overgoor, Samsun Knight, and Yakov Bart. (mis) measuring the drivers of ad performance. *SMU Cox School of Business Research Paper*, 2025.

Michel Tuan Pham, Maggie Geuens, and Patrick De Pelsmacker. The influence of ad-evoked feelings on brand evaluations: Empirical generalizations from consumer responses to more than 1000 tv commercials. *International Journal of Research in Marketing*, 30(4):383–394, 2013.

Rik Pieters, Michel Wedel, and Rajeev Batra. The stopping power of advertising: Measures and effects of visual complexity. *Journal of Marketing*, 74(5):48–60, 2010.

Anyamanee Pornpanvattana, Metpiya Lertakkakorn, Peerat Pookpanich, Khodchapan Vitheethum, and Thitirat Siriborvornratanakul. Youtube thumbnail design recommendation systems using

image-tabular multimodal data for thai's youtube thumbnail. *Social Network Analysis and Mining*, 14(1):181, 2024. doi: 10.1007/s13278-024-01317-7.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Prashant Rajaram and Puneet Manchanda. Unboxing engagement in youtube influencer videos: An attention-based approach. *arXiv preprint arXiv:2012.12311*, 2020.

Peter E Rossi and Greg M Allenby. Bayesian statistics and marketing. *Marketing Science*, 22(3): 304–328, 2003.

Deniss Ruder, Andero Uusberg, and Kairit Sirts. Assessing the reliability and validity of gpt-4 in annotating emotion appraisal ratings. *arXiv preprint arXiv:2503.16883*, 2025.

Jun Hyun Ryoo, Xin Wang, and Shijie Lu. Do spoilers really spoil? using topic modeling to measure the effect of spoiler reviews on box office revenue. *Journal of Marketing*, 85(2): 70–88, 2021.

Akari Shimono, Yuki Kakui, and Toshihiko Yamasaki. Automatic youtube-thumbnail generation and its evaluation. In *Proceedings of the 2020 Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia*, pages 25–30, 2020. doi: 10.1145/3379173.3393711.

Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. Multimodal large language models meet multimodal emotion recognition and reasoning: A survey. *arXiv preprint arXiv:2509.24322*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Ankit Sisodia, Alex Burnap, and Vineet Kumar. Generative interpretable visual design: Using disentanglement for visual conjoint analysis. *Journal of Marketing Research*, 2024.

Aivin V Solatorio. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*, 2024.

Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the 25th ACM*

International on Conference on Information and Knowledge Management, pages 659–668, 2016.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

WA Thompson Jr. On the treatment of grouped observations in life studies. *Biometrics*, pages 463–470, 1977.

Edward L. Thorndike. A constant error in psychological ratings. *Journal of Applied Psychology*, 4(1):25–29, 1920.

Zijun Tian, Ryan Dew, and Raghuram Iyengar. Mega or micro? influencer selection using follower elasticity. *Journal of Marketing Research*, 61(3):472–495, 2024.

Olivier Toubia. A poisson factorization topic model for the study of creative documents (and their summaries). *Journal of Marketing Research*, page 0022243720943209, 2021.

Pravin K. Trivedi and David M. Zimmer. Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, 1(1):1–111, 2007.

Broderick Turner, Yandu Lu, and Huessein Eslam. Algorithmic emotional expression selection. *working paper*, 2024.

Kyra L. Wiggin, Martin Reimann, and Shailendra P. Jain. Curiosity tempts indulgence. *Journal of Consumer Research*, 45(6):1194–1212, 2019. doi: 10.1093/jcr/ucy055.

Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Beyond views: Measuring and predicting engagement in online videos. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

Wytlabs. Youtube statistics. https://wytlabs.com/blog/youtube-statistics/, 2026. Accessed: 2026-05-02.

Jasmine Yang, Poppy Zhang, and Shawndra Hill. Mllm-vadstory: Domain knowledge-driven multimodal llms for video ad storyline insights. *arXiv preprint arXiv:2601.07850*, 2026.

Jeremy Yang, Juanjuan Zhang, and Yuhan Zhang. Engagement that sells: Influencer video advertising on tiktok. *Marketing Science*, 44(2):247–267, 2025.

S. Yang. Analysis of top box office film poster marketing scheme based on data mining and deep learning in the context of film marketing. *PLOS ONE*, 2023.

Tao Yang, Fan Wang, Junfan Lin, Zhongang Qi, Yang Wu, Jing Xu, Ying Shan, and Changwen Chen. Toward human perception-centric video thumbnail generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6653–6664, 2023. doi: 10.1145/3581783.3612434.

Sang-Hyeak Yoon and Hee-Woong Kim. What content and context factors lead to selection of a video clip? the heuristic route perspective. *Electronic Commerce Research*, 19:603–627, 2019.

YouTube Help. Thumbnail & title tips. https://support.google.com/youtube/answer/12340300, 2026. Accessed: 2026-05-02.

Zhifeng Yu and Nanchun Shi. A multi-modal deep learning model for video thumbnail selection. *arXiv preprint arXiv:2101.00073*, 2021.

Kunpeng Zhang, Poppy Zhang, Shawndra Hill, and Amel Awadelkarim. Decoding the hook: A multimodal llm framework for analyzing the hooking period of video ads. *arXiv preprint arXiv:2602.22299*, 2026.

Mengxia Zhang and Lan Luo. Can consumer-posted photos serve as a leading indicator of restaurant survival? evidence from yelp. *Management Science*, 69(1):25–50, 2023.

Shunyuan Zhang, Dokyun Lee, Param Vir Singh, and Kannan Srinivasan. What makes a good image? airbnb demand analytics leveraging interpretable image features. *Management Science*, 68(8):5644–5666, 2022. doi: 10.1287/mnsc.2021.4175.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Mi Zhou, George H Chen, Pedro Ferreira, and Michael D Smith. Consumer behavior in the online classroom: Using video analytics and machine learning to understand the consumption of video courseware. *Journal of Marketing Research*, 58(6):1079–1100, 2021.

Web Appendix

Table of Contents

Web Appendix A Additional Video Analysis Details

A.1 Identifying Thumbnail Location from Video Frame Sequence

We propose a two-stage, coarse-to-fine visual matching procedure to (i) identify whether each thumbnail is a frame from its own video frame sequence and, if so, (ii) pin down the thumbnail's exact location within the video (frame index, time in seconds, and location percentile relative to video length). The procedure first uses the state-of-the-art DINOv3 image embeddings to narrow down the right *region* of the video where the thumbnail is, and then conducts classical template matching to *precisely locate* the thumbnail's exact position.

Stage 1: Coarse semantic matching in DINOv3 visual embedding space. For each video, we uniformly sample frames at a fixed time interval (0.1s in our implementation)[27] with frames downsized to a maximum width of 720 pixels to reduce computation.[28] We extract frame embeddings using the ViT-based DINOv3 encoder (DINOv3 ViT-B/16; Siméoni et al. (2025)) to obtain image embeddings for both the thumbnail and each sampled video frame. DINOv3 is an open-source, self-supervised vision model developed by Meta trained on diverse image corpora, producing high-quality visual features without relying on human labeled data (e.g., Siméoni et al., 2025; Meta AI, 2025). It provides state-of-the-art, high-quality visual features for a wide range of downstream tasks, making it a strong choice for our image–image matching task. We compute cosine similarity between the thumbnail embedding and each frame embedding. Rather than simply taking the highest-similarity frame as the thumbnail, we select the top-K frames (K=12 in our implementation) within the highest cosine similarities as our *candidate matches*. This alleviates the issue of picking a near miss rather than the exact source frame when the video revisits similar scenes or when several frames are nearly indistinguishable in embedding space.

Stage 2: Localized template-matching refinement. While the first stage acts a high-recall filter that narrows down the thumbnail search to a small set of visually plausible locations, in the second stage, we apply pixel-level, multi-scale template matching around each candidate frame to enforce geometric consistency and determine the thumbnail's exact location in the video. For each candidate frame identified in Stage 1, we refine the match in a small temporal neighborhood (± 4 frames) around its frame index. We precompute a grayscale version of each frame in the neighborhood and its Canny edge map (Canny, 2009),[29] and we similarly compute

[27]This sampling rate corresponds to every 3 frames on a 30 frames/second video. It is dense enough that, even for fast-changed videos (cuts, zooms, etc), we still get reasonable visually similar candidates in the embedding space later though we might not hit the exact source frame).

[28]Videos in our dataset are High Definition (HD) with typical resolutions around 1920 (width) × 1080 (height) in pixels. We downscale the frames to widely used video resolutions (preserving video aspect ratio) to make large-scale processing tractable, while preserving details clear enough to reliably match the thumbnail.

[29]These are two common pre-processing steps for image analysis, which helps reduce distractions to make matching more stable and robust. Grayscale removes color variations to preserve detailed intensity and texture patterns, while edge maps emphasize structural alignment under brightness/contrast changes.

those for the thumbnail. To tolerate black borders (if any) around the (downloaded) thumbnail and slight crops, we first pad the thumbnail at several padding ratios (up to 30% of its size). For each padded thumbnail, we then perform multi-scale normalized cross-correlation (NCC) template matching (e.g., Lewis et al., 1995; Brunelli, 2009) over a set of scales between 0.6× and 1.4× the original size, a common technique to account for small geometric discrepancies (e.g., zoom) between matched images. For every frame in the neighborhood, we obtain (i) the best NCC score on grayscale images across all scales and paddings and (ii) the best NCC score on edge images. These two scores are averaged to obtain a single refined match score, which is high only when the thumbnail and frame align well in both intensity and structure. Across all frames in the candidate's neighborhood, we keep the frame with the highest refined score; across all candidates, we then select the overall best frame.

For every video–thumbnail pair, we record down (i) the matched frame index (if any), (ii) the corresponding timestamp in seconds, (iii) the thumbnail's relative position within the video, computed as the matched frame index divided by the total number of frames. These outputs allow us both to detect whether a thumbnail is drawn from its own video and to precisely locate where in the video that frame occurs for downstream analysis. In particular, we classify a thumbnail as from its own video frame sequence if the refined match score exceeds a threshold of 0.6. We manually inspected 100 randomly sampled video–thumbnail pairs to verify that this threshold yields accurate detection results.

A.2 Video Scene Detection Using an Adaptive Threshold

To detect scene changes, we use PySceneDetect's content-aware detector,[30] which triggers a scene (sometimes shot) cut when adjacent frames experience a color change that exceeds a certain threshold λ. However, a fixed threshold of $\lambda = 27$, which tends to work well for movies, may sometimes miss cuts or over-trigger scene changes for short and user-generated videos. To address this, we propose an *adaptive threshold* procedure: First, we read in the video frames, convert them into grayscale (to simplify computation), and calculate the absolute pixel difference between successive grayscale frames. Then, we set the threshold as the minimum between 27 and the mean pixel difference multiplied by a factor (1.5 in our implementation; determined experimentally). This yields scene timestamps with high accuracy for videos in our dataset.

A.3 LLaVA-Video-72B-Qwen2

We use LLaVA-Video-72B-Qwen2 (Zhang et al., 2024b), an open-source large multimodal video instruction-following model for video content and emotion understanding. The model is based on the Qwen2 language-model backbone and is designed to process both visual and textual

[30]Source: https://www.scenedetect.com/api/

inputs. At inference, each video is represented by a sequence of uniformly sampled frames, with the model supporting up to 64 frames per video. These visual inputs allow the model to reason over video content in combination with the text prompt.

Architecturally, LLaVA-Video-72B-Qwen2 follows the standard vision-language design used in the LLaVA family. A pretrained visual encoder first extracts representations from the sampled video frames. These visual representations are then mapped into the language model's token embedding space through a trainable projection module and concatenated with the text tokens. The Qwen2 transformer subsequently performs autoregressive decoding conditioned on both the visual and textual information, enabling unified reasoning over video frames and natural-language instructions.

LLaVA-Video-72B-Qwen2 is trained using large-scale visual instruction-tuning data, including LLaVA-Video-178K,[31] a synthetic video instruction dataset designed for detailed video captioning and multimodal question answering, and visual instruction data from LLaVA-OneVision.[32] This training setup improves the model's ability to follow natural-language prompts, describe visual scenes, answer video-based questions, and infer high-level semantic and emotional content from short videos. LLaVA-Video-72B-Qwen2 achieves strong performance among open-source video-language models on video understanding benchmarks such as Video-MME[33] and LongVideoBench.[34]

A.4 MLLM Prompts

Prompt: Describing video scene content

You are given a scene of a video. Your task is to describe the main content, actions or events happening in the scene using one sentence (maximum 25 words). Do not mention camera movements, editing styles, transitions, animation techniques, visual aesthetics, and evoked emotions.

Examples:
"Scene": "A herd of elephants is walking through a forest, with the largest elephant in the front and several smaller ones following closely behind."
"Scene": "A group of people are gathered around a campfire, sitting on the ground and roasting marshmallows."
"Scene": "A woman is walking down a red carpet, being photographed by a crowd and posing for pictures."

Respond using JSON format only:

[31]Source: https://huggingface.co/datasets/lmms-lab/LLaVA-Video-178K
[32]Source: https://huggingface.co/datasets/lmms-lab/LLaVA-OneVision-Data
[33]Source: https://video-mme.github.io/home_page.html
[34]Source: https://longvideobench.github.io

```
```json
{{
 "Scene": "<scene description>"
 }}
```
```

Don't start with any introduction and provide only json output. Start with "'json.

Prompt: Detecting video scene emotions

You are an average viewer and you are given a scene of a video. Your task is to determine the emotions you feel after seeing the video scene. You can pick from the following six basic emotions (anger, disgust, fear, joy, sadness and surprise) plus neutral (if you feel none of the six emotions). Please rate your evoked emotions (anger, disgust, fear, joy, sadness, surprise, and neutral) after seeing the video scene using a scale of 1 to 7, where 1 means not at all and 7 means a great deal.
Please also indicate your perceived valence and arousal levels after seeing the video scene on a scale of 1 to 7.

Note that:
- You may feel more than one emotion after seeing the video scene, or none at all (neutral).
- You are not rating on the emotions of the people in the video scene (if there are any) unless those emotions are clearly used to make you feel something about the scene.

Definitions:
Anger: a feeling of antagonism, frustration, or hostility, often triggered by perceived offense or mistreatment.
Disgust: a feeling of intense displeasure or revulsion, often stemming from something offensive or repulsive.
Fear: a primal emotion that can trigger the fight-or-flight response, arising from perceived threats or dangers.
Joy: a feeling characterized by happiness, contentment, or pleasure.
Sadness: a feeling of sorrow, grief, or unhappiness.
Surprise: a brief, often positive, emotion, usually triggered by unexpected events.
Neutral: feels none of the basic six emotions (anger, disgust,fear,joy, sadness, surprise)
Valence: measures how positive or negative a viewer feels; higher valence maps to a more positive feeling.
Arousal: measures how energetic a viewer feels; higher arousal maps to greater excitement and energy.

Respond using JSON format only:

```
```json
{{
"anger": "[score from 1 to 7]",
"disgust": "[score from 1 to 7]",
"fear": "[score from 1 to 7]",
"joy": "[score from 1 to 7]",
"sadness": "[score from 1 to 7]",
"surprise": "[score from 1 to 7]",
"neutral": "[score from 1 to 7]",
"valence": "[score from 1 to 7]",
"arousal": "[score from 1 to 7]",
 }}
```
```

Don't start with any introduction and provide only json output. Start with "'json.

A.5 Video Latent Topics

We use content descriptions generated by LLaVA-Video-72B-Qwen2 (Zhang et al., 2024a) across scenes to train a Latent Dirichlet Allocation (LDA; Blei et al., 2003) topic model. We define each video as a *document* and construct its bag-of-words representation from the concatenated scene descriptions after tokenizing, removing English stopwords, and applying part-of-speech-aware lemmatization to collapse morphological variants (e.g., *watching*, *watched* → *watch*) while preserving valid English forms. We further prune the vocabulary by removing words that occur in fewer than 20 videos or more than 80% of all videos from the observational study.

We train the model using online variational inference algorithm (Hoffman et al., 2010) with 5-fold cross-validation, using priors from Griffiths and Steyvers (2004). We varied the number of topics between 2 to 30 and selected an optimal model of eight topics based on the average predictive perplexity per word across the 5 folds (lower perplexity means better model fit) and topic interpretability. The video-level topic proportion is computed as the mean of the LDA-inferred topic proportions across all scenes within the video. This yields an 8-dimensional topic proportion vector for each video, with proportions summing to 1. Table A.5.1 shows the eight topics and the most representative words associated with each topic based on a relevance score of $\lambda = 0.5$ (Sievert and Shirley, 2014).

LDA Topics (Topic Frequency)	Most Representative Concepts with $\lambda = 0.5$
Human & Vlogging (22.3%)	woman, man, camera, hair, wearing, speaking, shirt, standing, sitting, black, holding, talking, front, looking, white, young, gesturing, long, smiling
Entertainment and Design (3.5%)	roller, coaster, train, game, video, track, airplane, text, screen, displayed, character, logo
Water Activities (7.9%)	water, beach, body, boat, swimming, ocean, shore, waves, floating, pool, sand
Event and Performance (3.8%)	people, group, crowd, gathered, stage, others, dancing, watching, performing, standing, around, playing
Natural Landscape and Scenery (18.0%)	trees, rocky, mountains, landscape, dirt, path, forest, field, sky, grassy, background
Hospitality and Architecture (21.5%)	building, large, shown, featuring, room, stone, chairs, view, surrounded, cityscape, trees, roof, white, house
Transportation (6.3%)	road, driving, car, street, parked, buildings, busy, passing, city, side, riding, motorcycle, vehicles
Food (16.8%)	food, plate, eating, table, bowl, meat, person, preparing, piece, sauce, metal, vegetables, cooking

Table A.5.1: **LDA topics and most representative concepts with $\lambda = 0.5$.** Topic names are assigned based on the most representative words with a relevance score of $\lambda = 0.5$, which balances uniqueness and frequency of concepts under each topic. Topic frequency is the percentage of videos for which the topic has the highest probability.

Web Appendix B Additional Website and Experiment Details

B.1 Website Construction and Viewer Tracking

We build the video website using open-source website technologies, with the frontend constructed using the standard combination of HTML for organization, CSS for styling and JavaScript for interactive user behavior. Our main framework is built using React, which provides a tidy and compact architecture in the form of a single-page application (SPA). It allows the website to dynamically render content with the entire website acting as “one" page, conceptually akin to a whiteboard on which the site draws, erases, and redraws different content, significantly reducing video and image load times and enhancing user experience. On the backend, we leverage a runtime environment, Node.js, which empowers the website to handle server-side operations such as recording and storing user data and creating API endpoints to facilitate real-time data download after users' browsing sessions. We manage all data using MongoDB, which is set up and integrated with the website to handle data storage and retrieval for real-time access in a non-relational (JSON-like) form. Finally, the entire ensemble is hosted on Microsoft Azure cloud infrastructure within a personal Azure tenant, including storage for the database JSON blob and for images, videos and other assets.

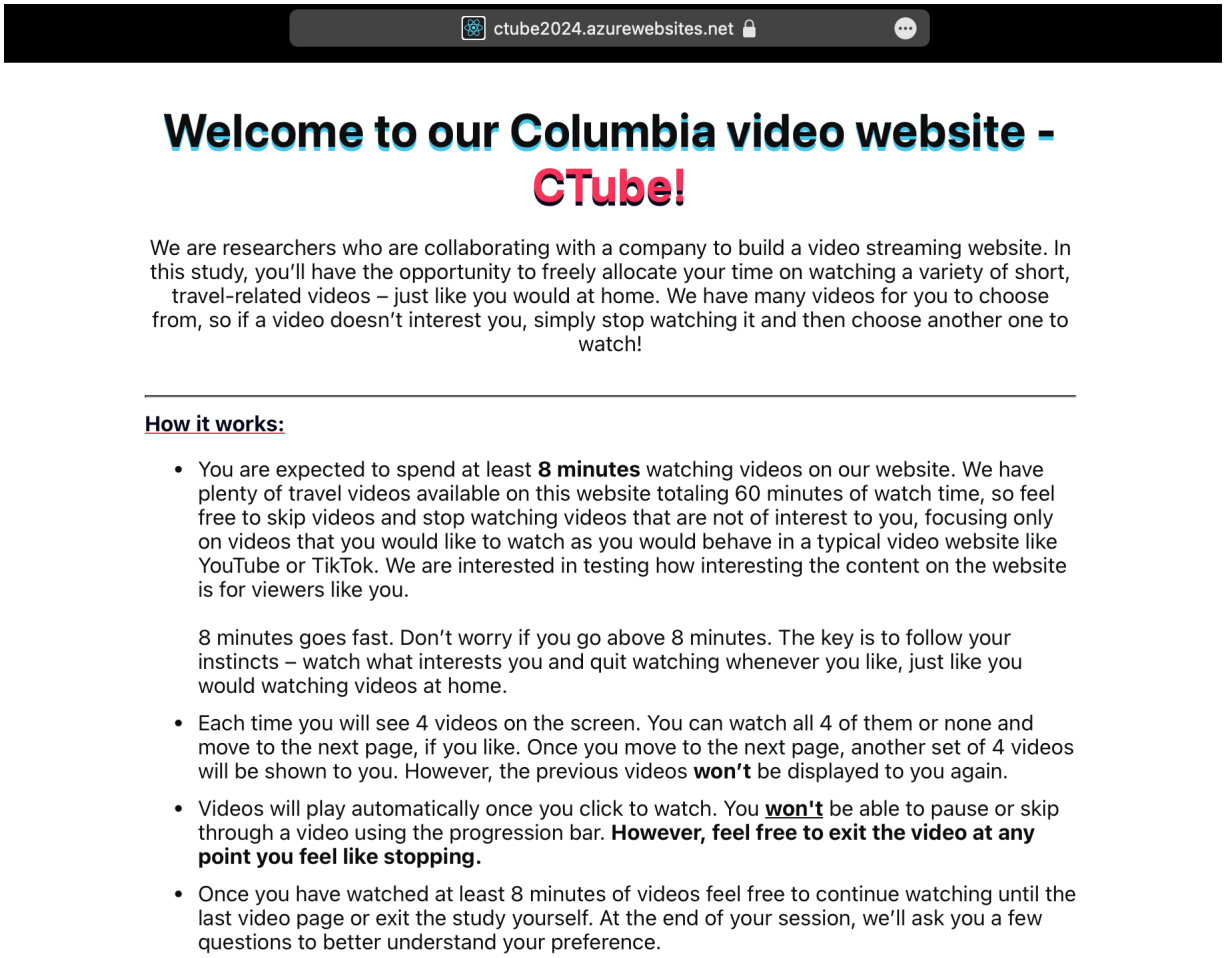
B.2 Cover Story and Instructions

Participants see the cover story and instructions before continuing to explore video content on our website, as shown in Figure B.2.1.

B.3 Website and Experimental Features

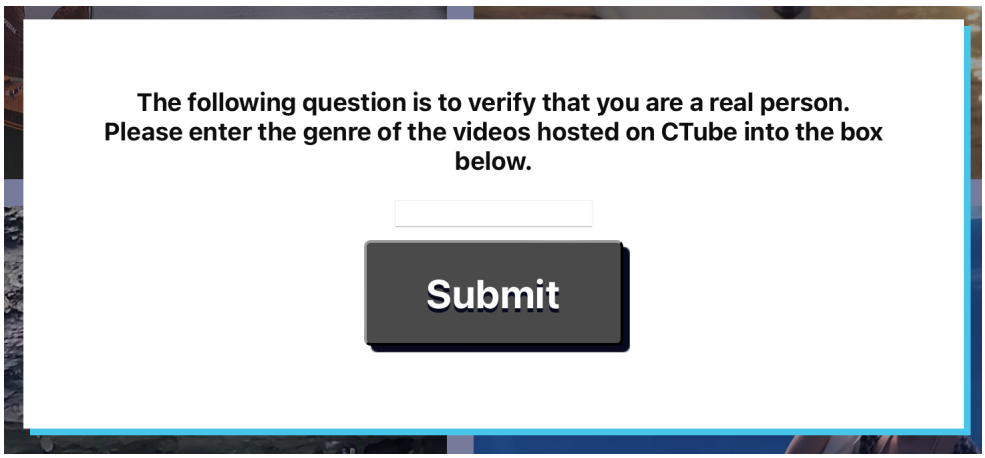
Attention Check Question. To ensure participant engagement, we included an attention check question. Participants were asked to confirm they were real users by entering the genre of the video hosted on CTube into a text box (See Figure B.3.1). This was triggered under two conditions: if a viewer clicked the navigation arrow five times consecutively without watching any videos, or if a viewer watched any three videos consecutively for more than 30 seconds each. Participants had to submit a response before the carousel was unlocked for further navigation.
Gaming and Inactivity Detection. We included two popup checks to monitor gaming behavior and inactivity. The gaming detection popup appeared if a viewer clicked the navigation arrow five times consecutively without watching any videos or watched any three videos consecutively for more than 30 seconds each. The inactivity alert appeared if a user clicked through the first five pages without watching any videos, prompting: “Alert: You may want to slow down and watch some videos to not exhaust the videos in your queue before your 8

Figure B.2.1: Cover Story and Instructions



Notes. This figure shows the cover story and instructions that participants see upon landing on our website.

Figure B.3.1: Attention Check Question

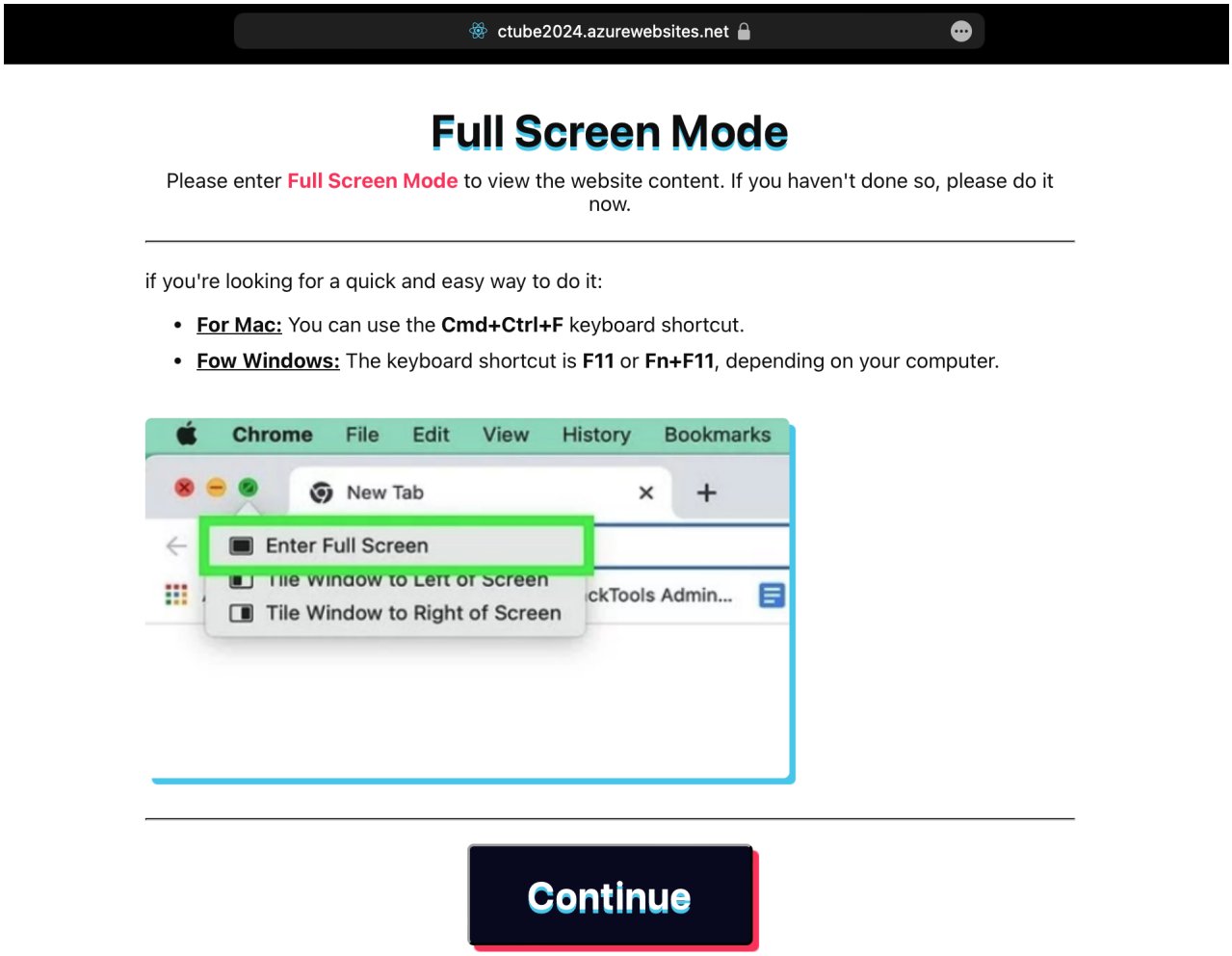


minutes are up." If both conditions triggered simultaneously, the inactivity alert was given priority.

Video Timer Widget. We added an unobtrusive widget below the video displaying the time in progression. Starting at 00:00/XX:XX (XX:XX is a video's length), it counted seconds from when the video is loaded. This provided participants with information on how long they had watched a video without interacting with the actual video controls. The timer synchronized correctly with the video, waiting at 00:00 until the video started playing.

Video Carousel and Navigation. We implemented a 2×2 video carousel on desktops of any resolution and instructed participants to enter the website in full-screen mode. This is because our website was designed to be only compatible on desktops or laptops browsing. We restrict entering the website using other devices such as iPads, phones, widescreen projectors to ensure proper website display. In addition, the full-screen mode was configured to ensure hiding of back/forward/reload buttons, forcing users to use the built-in navigation, rather than using browser-level actions to bypass parts of the workflow (see Figure B.3.2).

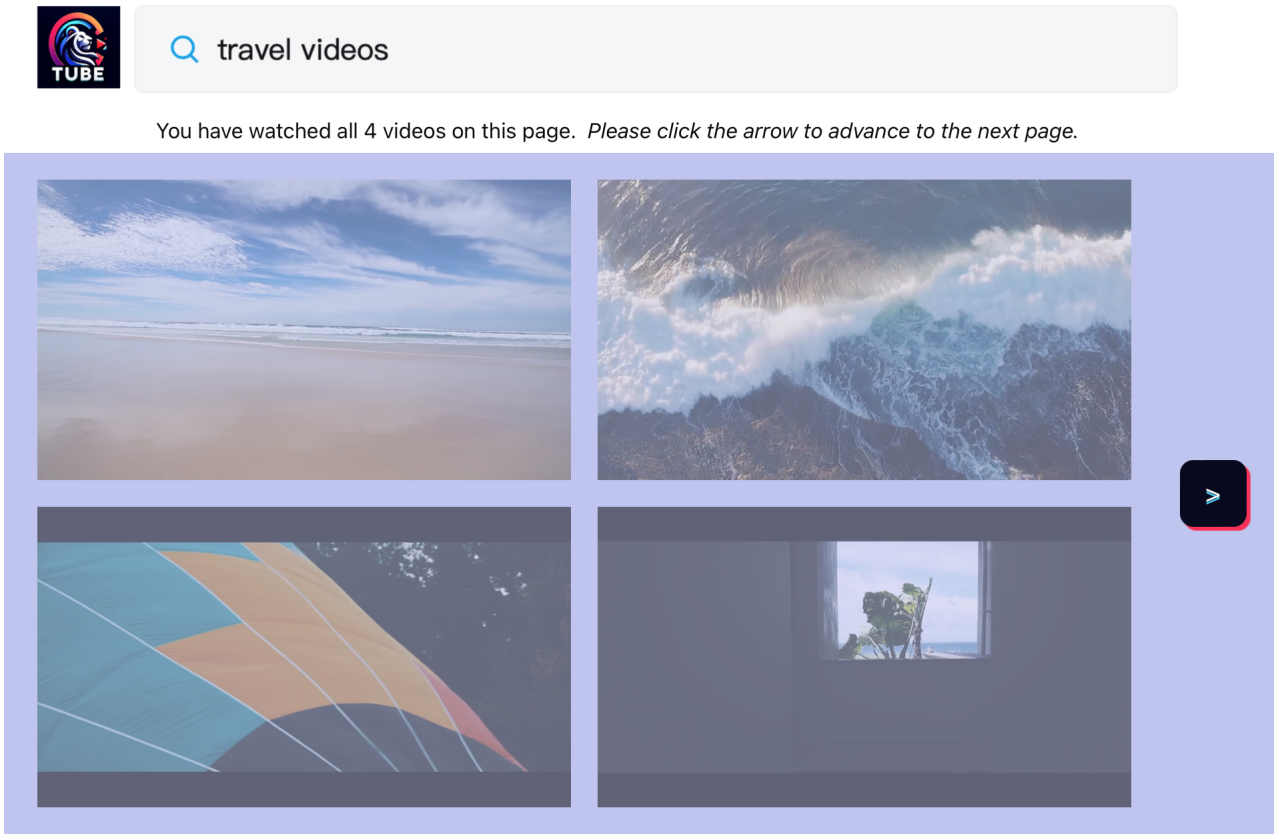
Figure B.3.2: Full Screen Mode



After watching all four videos in a particular set, we implemented a design such that users returning to the carousel would see the interface disabled (videos grayed out) and be prompted

to click the right arrow to proceed to the next set of videos (see Figure B.3.3).

Figure B.3.3: Lock Screen Mode



B.4 Exit the Website and Survey

Participants are free to exit the website at any time. A log out button will be displayed at the lower right corner on the carousel on the video choice page, after a viewer has interacted with at least one video. A popup will appear to confirm if viewers want to exit. In addition, if users explore the website extensively and exhaust all available videos hosted on the website, they will also be exited from the website and be redirected to the final survey. If any of those two cases, users will see an ending message explaining the situation, before we redirect them to the survey: "Thank you for participating in our study! You've just explored the range of videos currently in your queue. We have a few questions at the end for you to answer. When you are ready, please click Continue to proceed to the survey".

The final survey contains questions about consumers' video watching habits, thumbnail and video preferences, in addition to basic demographics (age, gender, racial/ethnic background).

- Video watching habit:

- How often do you watch videos online (via platforms such as YouTube, TikTok and Netflix)? *1=Never; 2=Rarely (e.g., once a month); 3=Sometimes (e.g., 1-2 times per week); 4=Often (e.g., every day, but less than 15 minutes per day); 5=Very often (e.g., every day, between 15 and 1 hr per day); 6=Frequently (e.g., more than 1 hour per day)*
- Do you have a YouTube account? *1=Yes, I do; 0=No, I don't.*
- Approximately how many minutes per day (on average) do you watch on YouTube? Please enter a number only. *Open entry*

- Preference over thumbnails:
  - To what extent each of the following video characteristics affects your decision of which video to watch on YouTube? *1=Not at all; 7=Very much*
    - Videos' face cover images (thumbnails)
    - Video's title
    - Video's content creator
  - How important are the following factors of video face covers (thumbnails) in affecting your decision to watch a video? *1=Not important at all; 7=Extremely important*
    - The face cover is visually appealing (beautiful).
    - The face cover makes me feel emotional (e.g., happy/sad/angry/scared).
    - The face cover summarizes the story of the video well.
    - The face cover makes me curious about the video.
    - I believe that the face cover aligns well with the video I'm about to see.
    - The face cover has people in it.
    - The face cover surprises me.
    - The face cover contains the greatest number of key concepts/elements of the video.

- Preference over videos:
  - We want to learn about your preferences in watching each of the following content. All content are subcategories within the travel video category. *1 = I rarely watch videos related to this content; 7= I frequently watch video related to this content*
    - Travel vlogs: personal experiences and stories from travels, focusing on the vlogger's journey and interactions
    - Culinary travel: videos focusing on local cuisine, street food, unique eateries, and culinary experiences in different locations

* Cultural insights: videos that explore local traditions, festivals, customs, cultural aspects of different places, as well as local projects and community initiatives
* Adventure travel: videos focusing on adventure activities like hiking, diving, safari, extreme sports, etc., in various travel destinations
* City guide and walkthroughs: short guides or walking tours of cities, highlighting key attractions, landmarks, and hidden gems
* Nature and scenery: videos showing breathtaking natural landscapes, scenic views, national parks, beaches, etc.
* Historical sites: videos exploring historical landmarks, ancient ruins, historical towns, and heritage sites.
* Luxury travel: videos featuring high-end travel experiences, luxury accomodations, exclusive destinations, etc.
* Theme parks and entertainment: videos exploring popular theme parks, amusement parks, and other entertainment attractions
* Wildlife and conservation: videos related to wildlife sanctuaries, wild animals in their natural habitat, or conservation efforts and eco-friendly practices

Web Appendix C Simulation Studies

We conduct simulation studies to verify that the joint copula model correctly recovers all model parameters $(\boldsymbol{\Theta}, \boldsymbol{\Psi}, \theta)$. The simulated data consist of $N = 200$ viewers, each facing $\mathcal{T} = 10$ choice occasions. At each occasion, the viewer is presented with $V = 4$ video alternatives and an outside option, $j = 0$. Each video alternative is characterized by $K_c = 3$ choice covariates drawn independently from standard normal distributions. For each clicked video, the viewer's watchtime is observed over up to $T_{\max} = 20$ survival periods, with $K_s = 3$ time-varying survival covariates.

Parameter values. The true parameters are set as follows. The individual-level random effect parameters $\boldsymbol{\Theta}_i = (\gamma_i^c, \boldsymbol{\beta}_i^{c\prime}, \gamma_i^s, \boldsymbol{\beta}_i^{s\prime})'$ are drawn from:

$$\boldsymbol{\Theta}_i \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \boldsymbol{\mu}^c \\ 0 \\ \boldsymbol{\mu}^s \end{pmatrix},\ \operatorname{diag}\left(\sigma_\gamma^{2,c},\ \boldsymbol{\sigma}^{2,c},\ \sigma_\gamma^{2,s},\ \boldsymbol{\sigma}^{2,s}\right)\right), \tag{21}$$

with hyperparameters:

$$\begin{aligned} &\boldsymbol{\mu}^c = (0.8,\ -0.5,\ 0.3)', \quad \boldsymbol{\sigma}^c = (0.5,\ 0.4,\ 0.3)', \quad \sigma_\gamma^c = 1.0, \\ &\boldsymbol{\mu}^s = (0.4,\ -0.3,\ 0.2)', \quad \boldsymbol{\sigma}^s = (0.4,\ 0.3,\ 0.3)', \quad \sigma_\gamma^s = 0.8. \end{aligned}$$

The fixed parameters $\boldsymbol{\Psi} = \{\boldsymbol{\xi}^c, \boldsymbol{\xi}^s, \boldsymbol{\alpha}^s\}$ are set to[35]:

$\boldsymbol{\xi}^c = (0.6,\ -0.4,\ 0.3,\ -0.2)'$,

$\boldsymbol{\xi}^s = (0.0,\ 0.3,\ -0.5,\ 0.2)'$ (first normalized to zero),

$\boldsymbol{\alpha}^s = (-3.5,\ -3.39,\ \ldots,\ -1.5,\ 0.5)'$ (linearly spaced for $t = 1, \ldots, 19$, with $\alpha_{20}^s = 0.5$).

The last entry of $\boldsymbol{\alpha}^s$ is set substantially higher to generate a concentration of exits near the end of the video, consistent with the empirical pattern in our data. We simulate data under three dependence configurations: $\theta \in \{0, 3, -3\}$, corresponding to independence, positive dependence, and negative dependence.

Simulation procedure. For each viewer i, we draw $\boldsymbol{\Theta}_i$ from the multivariate normal in Equation (21). The copula-based data generating process then proceeds as follows:

[35]For the choice-side video fixed effects ξ_v^c,it measures the baseline attractiveness of video v relative to the outside option, which is already normalized to 0. For the survival-side video fixed effects $\boldsymbol{\xi}^s$, we impose the normalization $\xi_1^s = 0$, because the absolute level of $\boldsymbol{\xi}^s$ is not separately identified from the period intercepts $\boldsymbol{\alpha}^s$.

1. For each viewer-occasion (i, τ), compute the choice utilities $V_{iv}^c = \boldsymbol{x}_{iv}^{c\prime}\boldsymbol{\beta}_i^c + \gamma_i^c + \xi_v^c$ for all alternatives v (with $V_{i0}^c = 0$ for the outside option) and derive the cumulative choice probabilities $F_{i\tau}^c(j) = \sum_{m=0}^{j} \Pr(c_{i,\tau} = m)$. Draw a latent uniform $W_i \sim \text{Uniform}(0, 1)$.

2. Draw a second latent uniform V_i from the conditional Frank copula distribution:

$$V_i = C_{V|W}^{-1}(r \mid W_i; \theta), \quad r \sim \text{Uniform}(0, 1), \tag{22}$$

   where $C_{V|W}^{-1}$ is the inverse of the conditional copula CDF. For the Frank copula, the conditional CDF has a closed-form inverse:

$$V = -\frac{1}{\theta} \log\left(1 + \frac{e^{-\theta} - 1}{(r^{-1} - 1)\, e^{-\theta W_i} + 1}\right). \tag{23}$$

   When $\theta = 0$, W_i and V_i are independent, and the joint model reduces to separate marginal estimation.

3. Map W_i to a discrete choice outcome: the observed choice is j^* such that $F_{i\tau}^c(j^* - 1) < W_i \leq F_{i\tau}^c(j^*)$.

4. Map V_i to a discrete watchtime outcome: for the chosen video $v = j^*$, compute the survival CDF $F_{iv}^s(t) = 1 - \prod_{s=1}^{t}[1 - \lambda_{iv}(s)]$ at each period using the hazard from Equation (9), and assign the exit time as the smallest t such that $F_{iv}^s(t-1) < V_i \leq F_{iv}^s(t)$. If $V_i > F_{iv}^s(T_{\max})$, the observation is right-censored. If $j^* = 0$ (outside option), no watchtime is observed.

This procedure ensures that the marginal distributions of choice and watchtime exactly match the models in Sections 6.1 and 6.2, while the copula parameter θ controls the residual dependence between the two decisions.

Estimation and results. For each simulated dataset, we estimate the joint copula model using the Hamiltonian Monte Carlo (HMC) algorithm described in Web Appendix D. Specifically, we place weakly informative priors on the hyperparameters:

$$\begin{aligned} \mu_k^c &\sim \mathcal{N}(0, 2^2), \quad \sigma_k^c \sim \text{HalfNormal}(1), \quad \sigma_\gamma^c \sim \text{HalfNormal}(1), \\ \mu_k^s &\sim \mathcal{N}(0, 2^2), \quad \sigma_k^s \sim \text{HalfNormal}(1), \quad \sigma_\gamma^s \sim \text{HalfNormal}(1), \end{aligned}$$

and diffuse priors on the fixed parameters: $\xi_v^c \sim \mathcal{N}(0, 2^2)$, $\xi_v^s \sim \mathcal{N}(0, 2^2)$ (with $\xi_1^s = 0$), $\alpha_t^s \sim \mathcal{N}(0, 5^2)$. The copula parameter receives $\theta \sim \mathcal{N}(0, 10^2)$. We run 4 chains with 1,000 warmup and 1,000 sampling iterations using the No U-Turn Sampling (NUTS) sampler.

Parameter recovery. Table C.0.1 reports the posterior estimates for all parameters under the positive dependence configuration ($\theta = 3$). All true values fall within the 95% credible intervals, and the posterior means are close to the true values. Both the choice-side parameters ($\boldsymbol{\mu}^c$, $\boldsymbol{\sigma}^c$, σ_γ^c, $\boldsymbol{\xi}^c$) and the survival-side parameters ($\boldsymbol{\mu}^s$, $\boldsymbol{\sigma}^s$, σ_γ^s, $\boldsymbol{\xi}^s$) are recovered accurately, confirming that the joint estimation does not distort the marginal parameters. The copula parameter θ is also recovered. Convergence diagnostics indicate $\hat{R} = 1.00$ for all parameters, and no divergent transitions.

Figure C.0.1 presents the trace plots for eight key parameters across all four chains. The chains are well-mixed and stationary, with no visible trends or divergences. The true parameter values (dashed red lines) fall well within the range of the posterior draws.

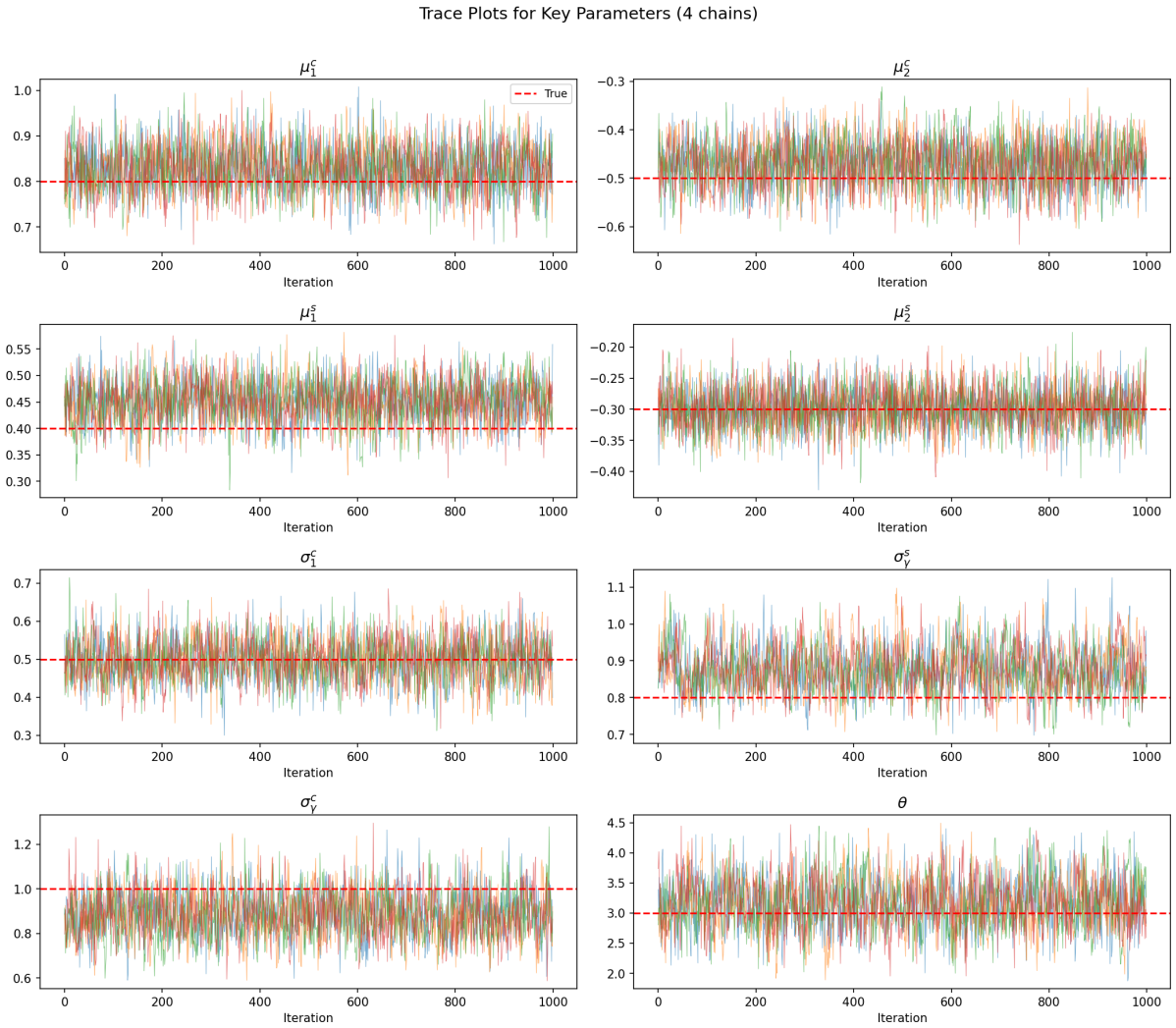


Figure C.0.1: **Trace plots for key parameters in the copula simulation** ($\theta = 3$)**.** Each panel shows 1,000 posterior draws from four parallel chains. Red dashed lines indicate the true parameter values. All chains exhibit good mixing and stationarity.

Figure C.0.2 further compares the true population distributions of the random coefficients

Parameter	True	Estimate	95% Credible Interval
Choice parameters			
μ_1^c	0.80	0.829	(0.731, 0.930)
μ_2^c	−0.50	−0.471	(−0.566, −0.379)
μ_3^c	0.30	0.316	(0.236, 0.399)
σ_1^c	0.50	0.498	(0.392, 0.605)
σ_2^c	0.40	0.461	(0.361, 0.563)
σ_3^c	0.30	0.360	(0.258, 0.463)
σ_γ^c	1.00	0.884	(0.689, 1.096)
ξ_1^c	0.60	0.442	(0.247, 0.645)
ξ_2^c	−0.40	−0.569	(−0.791, −0.345)
ξ_3^c	0.30	0.101	(−0.101, 0.318)
ξ_4^c	−0.20	−0.306	(−0.521, −0.083)
Survival parameters			
μ_1^s	0.40	0.452	(0.370, 0.529)
μ_2^s	−0.30	−0.298	(−0.363, −0.233)
μ_3^s	0.20	0.212	(0.147, 0.282)
σ_1^s	0.40	0.416	(0.336, 0.501)
σ_2^s	0.30	0.265	(0.183, 0.345)
σ_3^s	0.30	0.282	(0.193, 0.369)
σ_γ^s	0.80	0.877	(0.764, 1.008)
ξ_2^s	0.30	0.176	(0.000, 0.350)
ξ_3^s	−0.50	−0.554	(−0.756, −0.360)
ξ_4^s	0.20	0.153	(−0.072, 0.364)
Copula parameter			
θ (Frank)	3.00	3.161	(2.343, 4.002)

Table C.0.1: **Posterior estimates for the joint copula simulation study** ($\theta_{\text{true}} = 3$)**.** The table reports true values, posterior means, and 95% credible intervals for the choice-side and survival-side parameters, and the Frank copula dependence parameter. The first survival video fixed effect is normalized to $\xi_1^s = 0$. Period intercepts α_t^s, $t = 1, \ldots, 20$, are recovered but omitted for brevity (See Figure C.0.3 for details). All parameters satisfy $\hat{R} = 1.00$, and no divergent transitions are detected.

with the estimated distributions. For each covariate k, the true density $\mathcal{N}(\mu_k, \sigma_k^2)$ (solid blue) is overlaid with the posterior predictive density $\mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k^2)$ (dashed red), along with a histogram of the 200 simulated individual-level draws (gray). The close overlap between the true and estimated densities confirms that the estimation procedure correctly recovers both the location and scale of the population distributions.

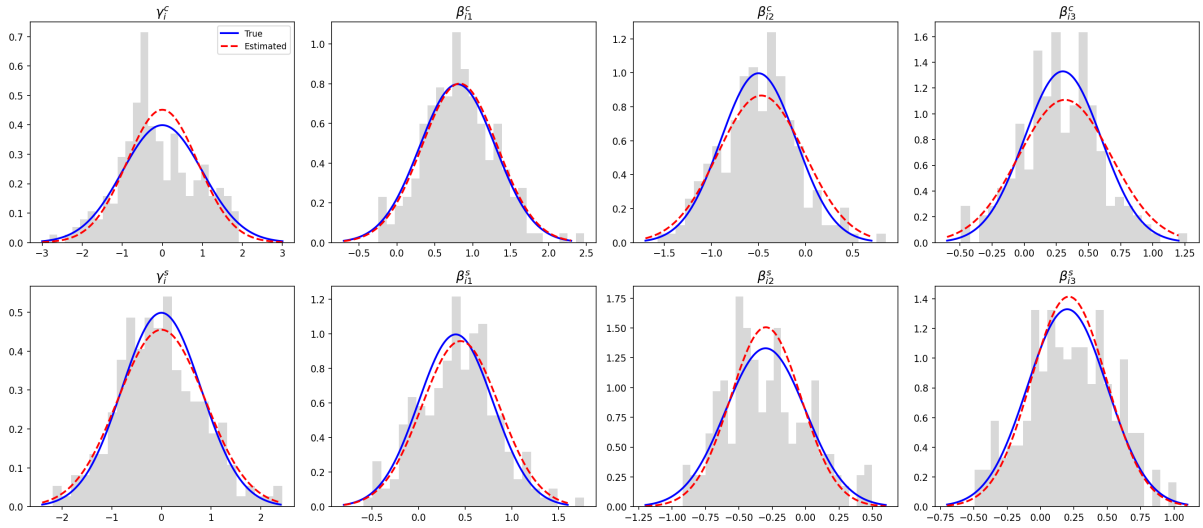


Figure C.0.2: **True vs estimated population distributions of random effects.** Top row: choice random intercept γ_i^c and random slopes $(\beta_{i1}^c, \beta_{i2}^c, \beta_{i3}^c)$. Bottom row: survival random intercept γ_i^s and random slopes $(\beta_{i1}^s, \beta_{i2}^s, \beta_{i3}^s)$. Solid blue: true distribution. Dashed red: estimated distribution using posterior mean hyperparameters. Gray histogram: simulated individual-level draws.

Figure C.0.3 compares the baseline hazard rates across periods in the simulated and predicted datasets (which we construct using the posterior mean estimates) for two sample videos. The baseline hazard is computed by setting $\gamma_i^s = 0$ and $\boldsymbol{x}_{ivt} = \mathbf{0}$, so that only the period intercepts α_t^s and video fixed effects ξ_v^s contribute. The trajectories are very similar across the two datasets, indicating that the estimated parameters predict quitting decisions well. In particular, the model correctly recovers the low and gradually increasing hazard through periods 1–19 and the sharp spike at period 20.

Copula parameter recovery across configurations. We also tested different dependence configurations to check whether the estimation procedure can detect both positive and negative dependence, as well as correctly identify independence when $\theta = 0$. Table C.0.2 reports the recovery of θ across all configurations. In all cases, the true θ falls within the 95% credible interval.

Configuration	True θ	Estimate $\hat{\theta}$	95% Credible Interval
Independence	0	+0.159	(−0.497, 0.836)
Positive dependence	3	3.161	(2.343, 4.002)
Negative dependence	−3	−2.535	(−3.283, −1.822)

Table C.0.2: **Copula parameter recovery across simulation configurations.**

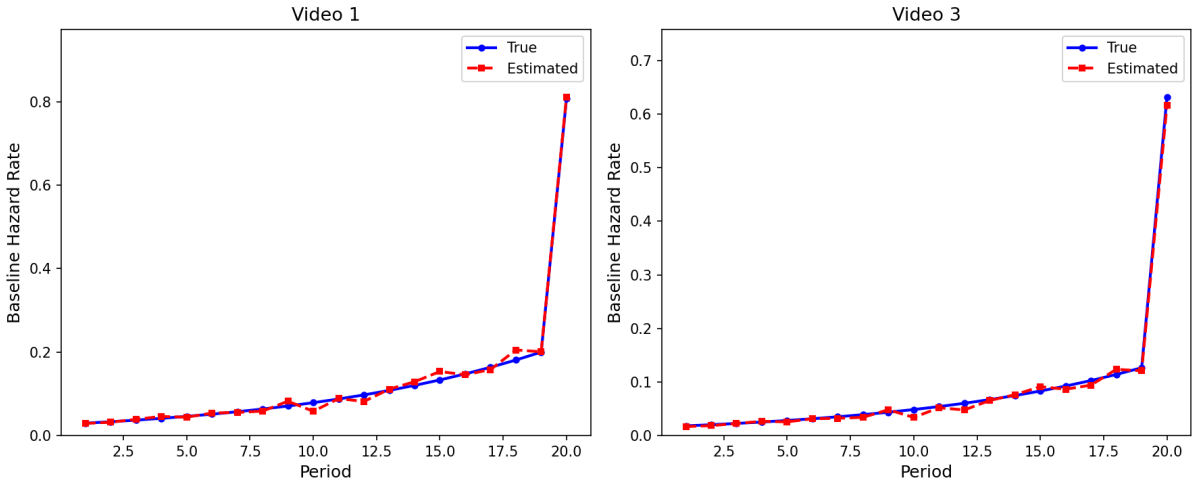


Figure C.0.3: **Comparison of simulated vs predicted baseline hazard rates.** True hazard rates (solid blue) computed from the true $(\alpha_t^s + \xi_v^s)$ vs estimated hazard rates (dashed red) computed from the posterior means $(\hat{\alpha}_t^s + \hat{\xi}_v^s)$, with all other parameters set to zero.

Web Appendix D Estimation Procedure

Let $\mathcal{Y} = \left\{c_{i,\tau}, t_{iv}, \delta_{iv}\boldsymbol{x}_{iv}^{c}, \{\boldsymbol{x}_{ivs}\}_{s\le t_{iv}}\right\}$ denote the observed data, and the full set of model parameters be the joint vector $\boldsymbol{\Phi} = (\boldsymbol{\Theta}, \boldsymbol{\Psi}, \theta)$. In particular, $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_i : i = 1, \ldots, N\}$, where $\boldsymbol{\Theta}_i = (\gamma_i^c, \boldsymbol{\beta}_i^{c\prime}, \gamma_i^s, \boldsymbol{\beta}_i^{s\prime})'$ denote random effect parameters that vary across individuals, $\boldsymbol{\Psi} = \{\boldsymbol{\xi}^c, \boldsymbol{\xi}^s, \boldsymbol{\alpha}^s\}$ denote parameters that are common across individuals, and θ denote the copula dependence parameter. The posterior distribution is proportional to the product of the joint likelihood and the prior distributions:

$$p(\boldsymbol{\Phi} \mid \mathcal{Y}) \propto L(\mathcal{Y} \mid \boldsymbol{\Theta}, \boldsymbol{\Psi}, \theta) \prod_{i=1}^{N} p(\boldsymbol{\Theta}_i \mid \boldsymbol{\Omega}) p(\boldsymbol{\Omega}) p(\boldsymbol{\Psi}) p(\theta). \tag{24}$$

The likelihood $L(\mathcal{Y} \mid \boldsymbol{\Theta}, \boldsymbol{\Psi}, \theta)$ is the joint copula likelihood defined in Equation (16), and the priors are same as the ones used in simulation study in Web Appendix C.

As the posterior is not analytically tractable, we implement Hamiltonian Monte Carlo (HMC) with the No-U-Turn Sampler (NUTS) algorithm and utilizes a tuned, diagonal mass matrix to draw from the posterior distribution. The procedure goes as follows:

Step 1. Initialization and Warm-up. The sampler runs an initial adaptive phase (Warm-up) to tune the algorithm's control parameters before collecting final samples. The two primary control parameters are the integration step size ϵ and the Mass Matrix $\mathbf{M}$. The step size ϵ is tuned to achieve a target average acceptance rate (e.g., 0.8 to 0.9). The Mass Matrix $\mathbf{M}$ is a diagonal (or optionally dense) matrix estimated from the covariance structure of the posterior distribution during the Warmup phase. It is used to precondition the joint parameter space, effectively scaling and orienting the HMC trajectory to improve the efficiency of the Hamiltonian simulation by mitigating the effects of parameter correlation.

Step 2. Hamiltonian Dynamics Simulation. For each subsequent MCMC iteration t, the sampler generates a new state $\boldsymbol{\Phi}^{(t)}$ from the current state $\boldsymbol{\Phi}^{(t-1)}$ by simulating Hamiltonian dynamics. This process simultaneously updates all continuous parameters in a single, high-dimensional step: first, a set of auxiliary momentum variables $\mathbf{p}^{(t-1)}$ is drawn from the distribution $\mathcal{N}(\mathbf{0}, \mathbf{M})$, where $\mathbf{M}$ is the adapted mass matrix.

The Hamiltonian energy $H(\boldsymbol{\Phi}, \mathbf{p})$ is defined by the sum of the potential energy $U(\boldsymbol{\Phi})$ (negative log-posterior) and the kinetic energy $K(\mathbf{p})$:

$$H(\boldsymbol{\Phi}, \mathbf{p}) = U(\boldsymbol{\Phi}) + K(\mathbf{p}) = -\log \pi(\boldsymbol{\Phi} \mid \mathcal{Y}) + \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p},$$

where $U(\boldsymbol{\Phi})$ is computed using the likelihood function in Equation (16) and the hierarchical structure and prior distributions described in Section 6.4 and Web Appendix C.

A trajectory is then simulated using the Leapfrog integrator by solving Hamilton's equations

of motion, where the gradient of the potential energy is equal to the negative derivative of the log-posterior:

$$\frac{d\mathbf{\Phi}}{dt} = \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p}$$
$$\frac{d\mathbf{p}}{dt} = -\frac{\partial H}{\partial \mathbf{\Phi}} = \nabla_{\mathbf{\Phi}} \log \pi(\mathbf{\Phi} \mid \mathcal{Y})$$

The trajectory of the sampler is simulated using a series of tiny, discrete jumps, where the length of these jumps is determined by the adapted step size ϵ.

Step 3. No-U-Turn Criterion and Acceptance. The NUTS algorithm dynamically determines the length of the Leapfrog trajectory L by stopping when the trajectory begins to turn back on itself, ensuring efficient exploration. The simulation yields a single proposed state $\mathbf{\Phi}^*$ (and momentum $\mathbf{p}^*$) that simultaneously updates all continuous parameters. This proposed state is accepted as the new state $\mathbf{\Phi}^{(t)}$ with the following Metropolis-style probability:

$$\min\left(1, \exp\left(H(\mathbf{\Phi}^{(t-1)}, \mathbf{p}^{(t-1)}) - H(\mathbf{\Phi}^*, \mathbf{p}^*)\right)\right).$$

Web Appendix E Robustness Checks

Sensitivity to alternative ordering. Because the choice margin is a nominal multinomial logit (not an ordered model), converting the multinomial outcome into a univariate CDF interval requires imposing a fixed ordering of alternatives. On each video selection page, the viewer chooses among an outside option and up to four videos. The CDF is constructed by cumulating choice probabilities in a fixed order within each page. For example, if page 1 presents videos $\{1, 2, 3, 4\}$ and page 2 presents $\{5, 6, 7, 8\}$, the three orderings we test are:

Ordering	Page 1	Page 2	...
Ascending (video ID)	0, 1, 2, 3, 4	0, 5, 6, 7, 8	
Reversed	0, 4, 3, 2, 1	0, 8, 7, 6, 5	
Fixed random permutation	0, 3, 1, 4, 2	0, 7, 5, 8, 6	

where 0 denotes the outside option. Under independence ($\theta = 0$), the ordering does not affect the joint probability. When $\theta \neq 0$, the rectangle probability can in principle depend on where the chosen alternative's interval falls in the unit interval.

We re-estimate the full copula model under all three orderings (4 chains × 2,000 draws each). Table E.0.1 reports the copula parameter and all marginal models' posterior means. The copula parameter θ is virtually identical across all three orderings (-0.148 to -0.153), with nearly indistinguishable credible intervals. All choice and survival parameters are similarly unaffected. We conclude that the alternative ordering has negligible impact on our estimates.

	Ascending	Reversed	Random Permutation
Choice Parameters			
Thumbnail Aesthetics	0.500 [0.389, 0.610]	0.499 [0.389, 0.608]	0.499 [0.387, 0.611]
Thumbnail Arousal	0.196 [0.070, 0.321]	0.196 [0.066, 0.322]	0.196 [0.068, 0.322]
Thumbnail Anger	0.194 [0.102, 0.282]	0.194 [0.105, 0.285]	0.194 [0.104, 0.285]
Thumbnail Disgust	0.310 [0.221, 0.399]	0.310 [0.220, 0.400]	0.310 [0.220, 0.400]
Thumbnail Fear	0.140 [0.088, 0.193]	0.140 [0.087, 0.193]	0.140 [0.087, 0.193]
Thumbnail Joy	0.170 [0.083, 0.259]	0.170 [0.083, 0.259]	0.170 [0.084, 0.259]
Thumbnail Sadness	0.142 [0.088, 0.199]	0.142 [0.088, 0.198]	0.142 [0.089, 0.198]
Thumbnail Surprise	0.147 [0.089, 0.205]	0.147 [0.089, 0.207]	0.147 [0.089, 0.206]
Content Disconfirmation	0.130 [0.085, 0.174]	0.130 [0.085, 0.174]	0.130 [0.085, 0.174]
Survival Parameters			
Content Disconf. (CD, mc)	0.346 [0.250, 0.444]	0.346 [0.249, 0.443]	0.346 [0.251, 0.443]
Post-Thumbnail Indicator	−0.123 [−0.191, −0.052]	−0.123 [−0.193, −0.053]	−0.123 [−0.194, −0.053]
CD × Post-Thumbnail	−0.140 [−0.234, −0.048]	−0.141 [−0.233, −0.047]	−0.141 [−0.233, −0.050]
Thumbnail Aesthetics	0.117 [−0.020, 0.253]	0.117 [−0.028, 0.257]	0.117 [−0.023, 0.255]
Thumbnail Arousal	−0.005 [−0.094, 0.084]	−0.005 [−0.094, 0.083]	−0.005 [−0.092, 0.082]
Boredom	0.364 [0.258, 0.475]	0.364 [0.258, 0.471]	0.363 [0.257, 0.470]
Novelty	−0.087 [−0.154, −0.019]	−0.086 [−0.155, −0.019]	−0.086 [−0.155, −0.019]
Pos. Aesthetic Disconf.	−0.143 [−0.246, −0.040]	−0.143 [−0.243, −0.039]	−0.143 [−0.244, −0.042]
Neg. Aesthetic Disconf.	0.007 [−0.056, 0.070]	0.007 [−0.055, 0.070]	0.007 [−0.056, 0.070]
Copula Parameter			
θ (Frank)	−0.148 [−0.770, 0.457]	−0.149 [−0.778, 0.473]	−0.153 [−0.781, 0.451]

Table E.0.1: **Sensitivity to alternative ordering.** Posterior means and 95% credible intervals for all parameters in the joint copula model under three orderings of the video alternatives within each choice occasion.

Alternative rolling window size. We assess the robustness of the main results to alternative rolling window size used to construct content and aesthetic disconfirmation in the survival model. We re-estimate the full joint copula model under 10 and 20 seconds, and show that the results are robust to alternative rolling window size.

	Mean		Heterogeneity (SD)	
	Estimate	95% CI	Estimate	95% CI
Panel A: Video Choice				
Thumbnail Aesthetics	0.500*	[0.387, 0.612]	0.037	[0.001, 0.105]
Thumbnail Arousal	0.196*	[0.068, 0.326]	0.033	[0.001, 0.091]
Content Disconfirmation	0.130*	[0.086, 0.174]	0.027	[0.001, 0.072]
Thumbnail Anger	0.195*	[0.105, 0.283]	0.029	[0.001, 0.082]
Thumbnail Disgust	0.310*	[0.219, 0.401]	0.034	[0.001, 0.093]
Thumbnail Fear	0.141*	[0.089, 0.194]	0.040	[0.001, 0.103]
Thumbnail Joy	0.170*	[0.081, 0.257]	0.038	[0.002, 0.102]
Thumbnail Sadness	0.142*	[0.088, 0.198]	0.037	[0.001, 0.100]
Thumbnail Surprise	0.148*	[0.089, 0.207]	0.031	[0.001, 0.081]
Video Fixed Effects (ξ_v^c)		✓		
Individual Random Intercept (γ_i^c)		✓		
Viewers / Videos / Choice Occasions		361 / 40 / 4,137		
Panel B: Video Watchtime				
Content Disconfirmation (CD, mean-centered)	0.236*	[0.140, 0.333]	0.177	[0.052, 0.272]
Post-Thumbnail Indicator	−0.160*	[−0.229, −0.092]	0.088	[0.005, 0.192]
CD (mean-centered) × Post-Thumbnail Indicator	−0.130*	[−0.224, −0.036]	0.090	[0.004, 0.203]
Thumbnail Aesthetics	0.367*	[0.100, 0.622]	0.089	[0.004, 0.207]
Thumbnail Arousal	−0.037	[−0.127, 0.053]	0.244	[0.146, 0.336]
Boredom	0.346*	[0.239, 0.455]	0.315	[0.234, 0.397]
Novelty	−0.068	[−0.138, 0.000]	0.331	[0.262, 0.399]
Thumbnail Anger	0.029	[−0.040, 0.097]	0.075	[0.006, 0.158]
Thumbnail Disgust	0.070	[−0.007, 0.146]	0.181	[0.079, 0.281]
Thumbnail Fear	−0.012	[−0.101, 0.077]	0.116	[0.007, 0.239]
Thumbnail Joy	−0.050	[−0.155, 0.051]	0.080	[0.004, 0.190]
Thumbnail Sadness	−0.079	[−0.174, 0.017]	0.102	[0.006, 0.215]
Thumbnail Surprise	−0.042	[−0.121, 0.034]	0.070	[0.003, 0.182]
Pos. Aesthetic Disconfirmation	−0.323*	[−0.499, −0.140]	0.157	[0.020, 0.268]
Neg. Aesthetic Disconfirmation	0.071	[−0.028, 0.169]	0.050	[0.002, 0.136]
Individual Random Intercepts (γ_i^s)		✓		
Video Fixed Effects (ξ_v^s)		✓		
Time Fixed Effects (α_t^s)		✓		
Viewers / Videos / Pairs / Obs. / Periods		361 / 40 / 3,312 / 23,225 / 10		
Copula Parameter				
θ (Frank)	−0.177	[−0.811, 0.446]		

Table E.0.2: **10-second rolling window.** Max $\hat{R}$ = 1.01, min ESS = 444, 0 divergences. *95% CI excludes zero.

	Mean		Heterogeneity (SD)	
	Estimate	95% CI	Estimate	95% CI
Panel A: Video Choice				
Thumbnail Aesthetics	0.500*	[0.388, 0.613]	0.038	[0.002, 0.108]
Thumbnail Arousal	0.196*	[0.068, 0.327]	0.032	[0.001, 0.091]
Content Disconfirmation	0.130*	[0.086, 0.175]	0.027	[0.001, 0.074]
Thumbnail Anger	0.195*	[0.105, 0.283]	0.029	[0.001, 0.081]
Thumbnail Disgust	0.310*	[0.220, 0.402]	0.034	[0.001, 0.093]
Thumbnail Fear	0.140*	[0.088, 0.193]	0.041	[0.001, 0.104]
Thumbnail Joy	0.170*	[0.081, 0.256]	0.038	[0.001, 0.104]
Thumbnail Sadness	0.142*	[0.088, 0.198]	0.038	[0.002, 0.103]
Thumbnail Surprise	0.147*	[0.088, 0.206]	0.031	[0.001, 0.084]
Video Fixed Effects (ξ_v^c)		✓		
Individual Random Intercept (γ_i^c)		✓		
Viewers / Videos / Choice Occasions		361 / 40 / 4,137		
Panel B: Video Watchtime				
Content Disconfirmation (CD, mean-centered)	0.207*	[0.114, 0.302]	0.155	[0.027, 0.257]
Post-Thumbnail Indicator	−0.154*	[−0.224, −0.084]	0.084	[0.004, 0.189]
CD (mean-centered) × Post-Thumbnail Indicator	−0.128*	[−0.219, −0.039]	0.087	[0.004, 0.201]
Thumbnail Aesthetics	0.368*	[0.101, 0.627]	0.089	[0.005, 0.207]
Thumbnail Arousal	−0.044	[−0.135, 0.045]	0.243	[0.143, 0.334]
Boredom	0.341*	[0.234, 0.447]	0.313	[0.232, 0.394]
Novelty	−0.066	[−0.135, 0.002]	0.332	[0.263, 0.399]
Thumbnail Anger	0.031	[−0.038, 0.098]	0.074	[0.006, 0.159]
Thumbnail Disgust	0.070	[−0.007, 0.146]	0.177	[0.071, 0.281]
Thumbnail Fear	−0.018	[−0.108, 0.070]	0.121	[0.008, 0.244]
Thumbnail Joy	−0.053	[−0.158, 0.047]	0.078	[0.003, 0.182]
Thumbnail Sadness	−0.082	[−0.177, 0.013]	0.101	[0.006, 0.211]
Thumbnail Surprise	−0.045	[−0.124, 0.031]	0.073	[0.003, 0.186]
Pos. Aesthetic Disconfirmation	−0.325*	[−0.504, −0.144]	0.159	[0.025, 0.273]
Neg. Aesthetic Disconfirmation	0.074	[−0.026, 0.172]	0.052	[0.002, 0.138]
Individual Random Intercepts (γ_i^s)		✓		
Video Fixed Effects (ξ_v^s)		✓		
Time Fixed Effects (α_t^s)		✓		
Viewers / Videos / Pairs / Obs. / Periods		361 / 40 / 3,312 / 23,225 / 10		
Copula Parameter				
θ (Frank)	−0.167	[−0.796, 0.457]		

Table E.0.3: **20-second rolling window.** Max $\hat{R}$ = 1.01, min ESS = 470, 0 divergences. *95% CI excludes zero.

Web Appendix F Platform-Recommended Thumbnails

To explore the supply-side dimension of thumbnail selection—specifically, how platforms recommend thumbnails to creators, we uploaded the 40 videos hosted on our experimental platform to YouTube at the same hour daily for 2 consecutive days. For each video, we documented the three thumbnails suggested by YouTube's automatic thumbnail generation algorithm (see Figure F.0.1).

To avoid the influence of prior viewing history or previous uploads on the algorithm's recommendations, we created a new YouTube account and uploaded the videos in private mode, restarting the process for each new upload after recording the suggested thumbnails and their positions (in seconds) within the video. For the actual thumbnails used by the creators, we use the original thumbnails of the videos.

YouTube's algorithm generates three thumbnails per video, strategically selected to ensure broad coverage across different sections (see Figure F.0.2): the left thumbnail is typically selected near the beginning of the video (averaging around 14% of a video's length), the middle thumbnail is chosen around the midpoint (around 50%), and the right thumbnail is selected from the later portion (around 80%). The default thumbnail, the one displayed when the cursor is not hovering over alternatives, is the middle thumbnail.

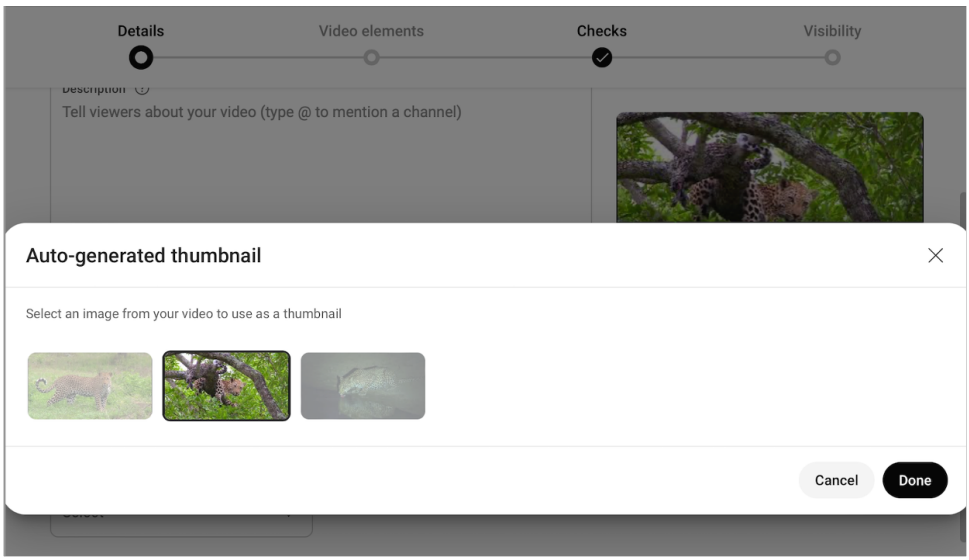


Figure F.0.1: **Auto-generated thumbnails from YouTube**. This figure shows an example of the three platform-recommended thumbnails for a video. The default thumbnail (where the cursor is left on) is the middle one.

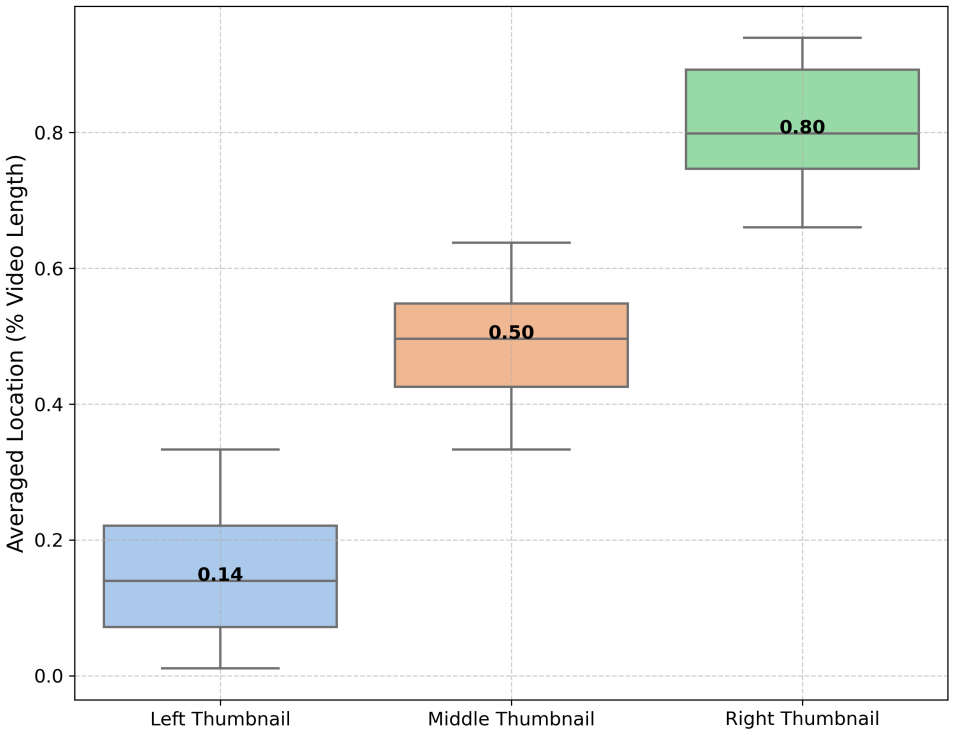


Figure F.0.2: **Location of Three Platform-Recommended Thumbnails (Box Plot)**. This figure illustrates the distribution of thumbnail locations within a video, based on YouTube's three auto-generated recommendations. The median value for each thumbnail is indicated on the plot. The left thumbnail is generally located closer to the beginning of the video, the middle thumbnail is around the halfway mark, and the right thumbnail is towards the end of the video.

Web Appendix G Additional Figures and Tables

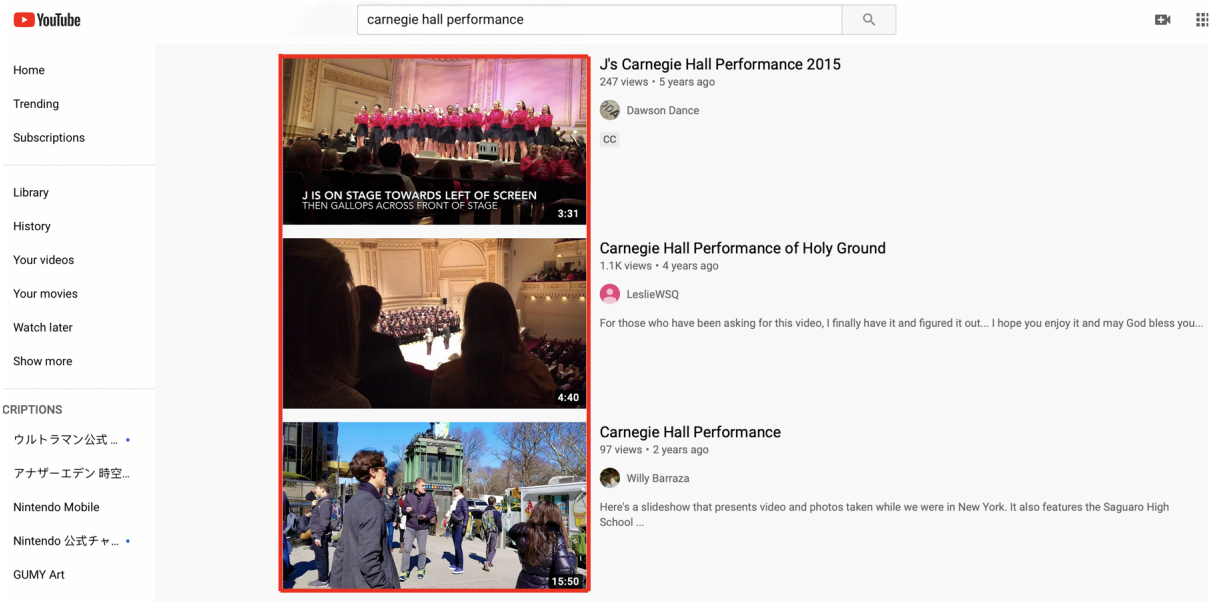


Figure G.0.1: **Thumbnail Examples on YouTube**.

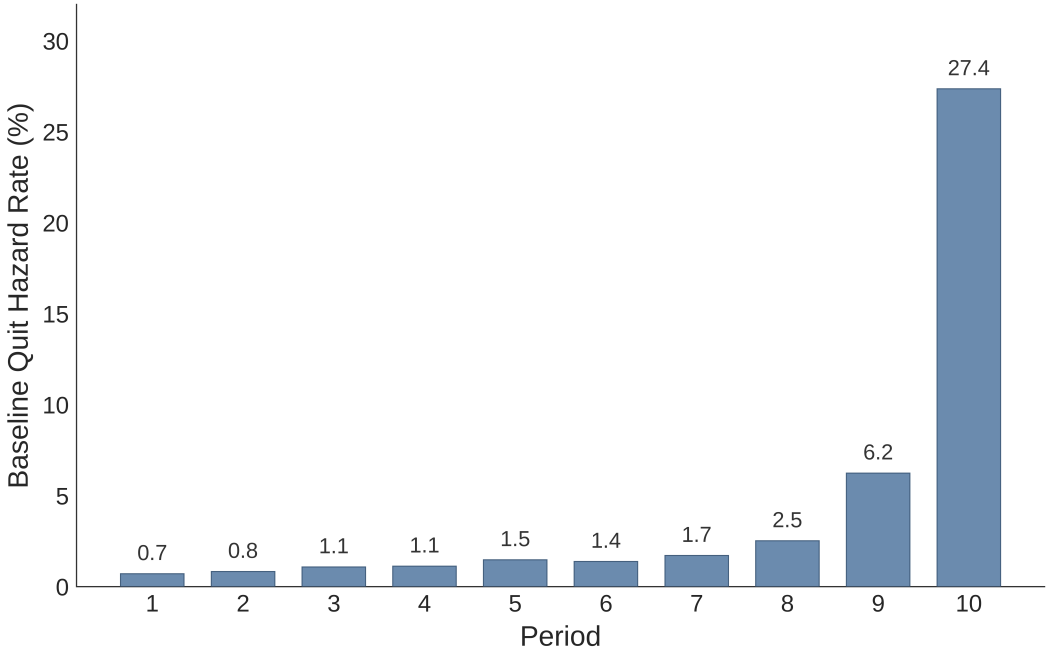


Figure G.0.2: **Estimated baseline quit hazard rate across video duration.** Each bar represents the posterior mean hazard rate $\lambda_t = 1 - \exp(-\exp(\hat{\alpha}_t))$ at the corresponding period (10% of video progression).

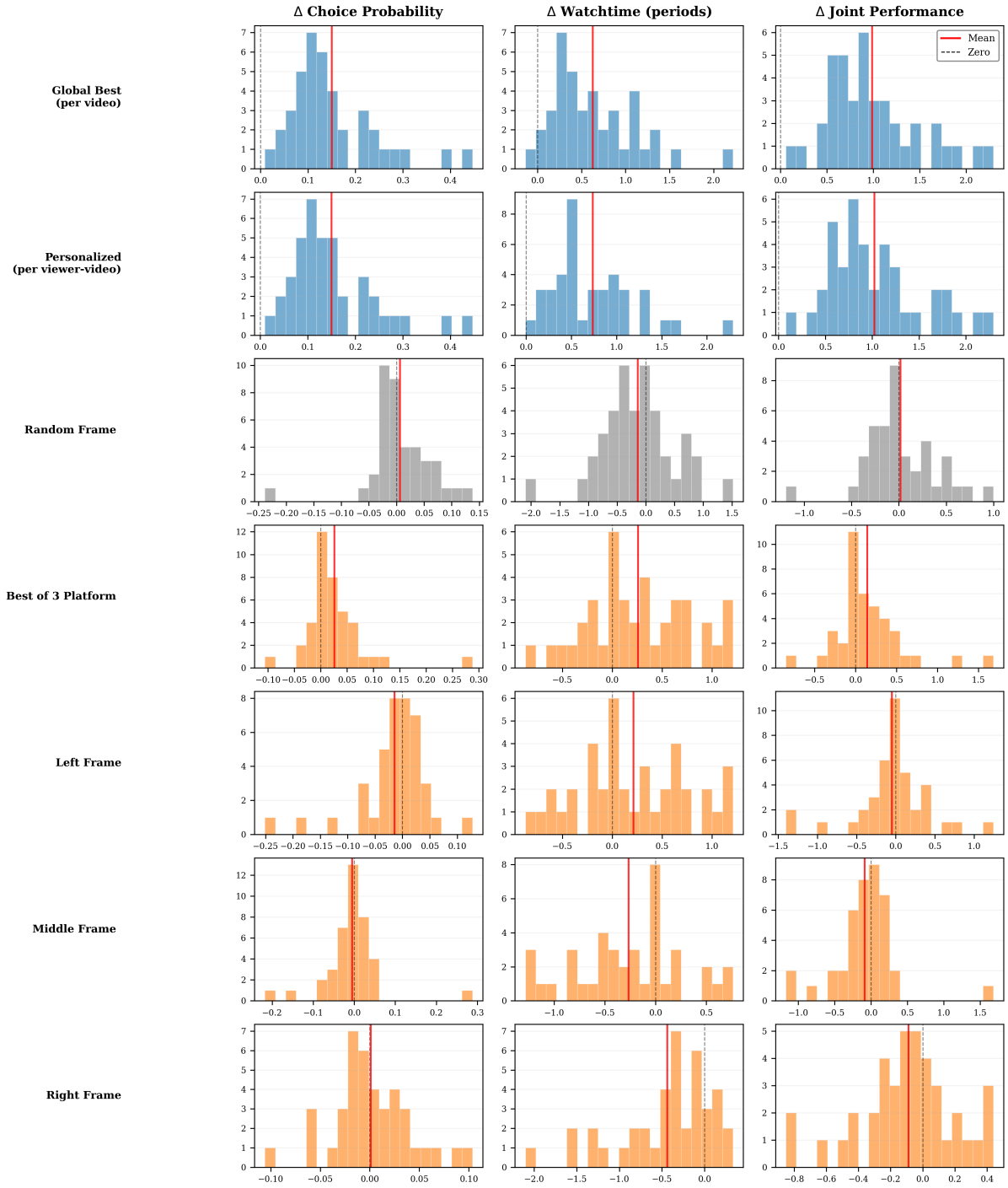


Figure G.0.3: **Distribution of gains: optimal, baseline and platform frames.** Each histogram shows the distribution of choice probability, watchtime and joint performance gains across 40 videos relative to the actual (creator-chosen) frame.

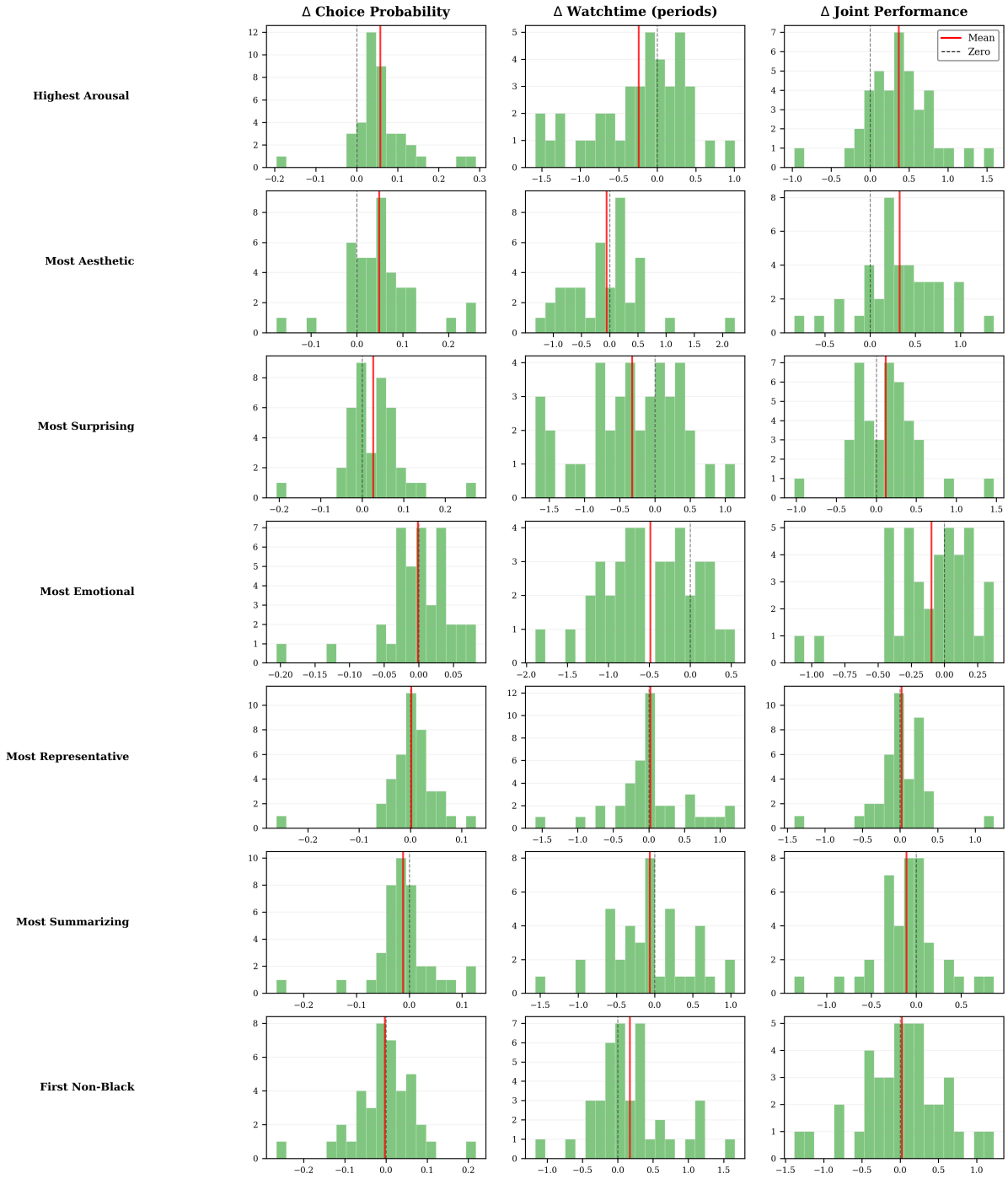


Figure G.0.4: **Distribution of gains: stylized frames.** Each histogram shows the distribution of choice probability, watchtime and joint performance gains across 40 videos relative to the actual (creator-chosen) frame.

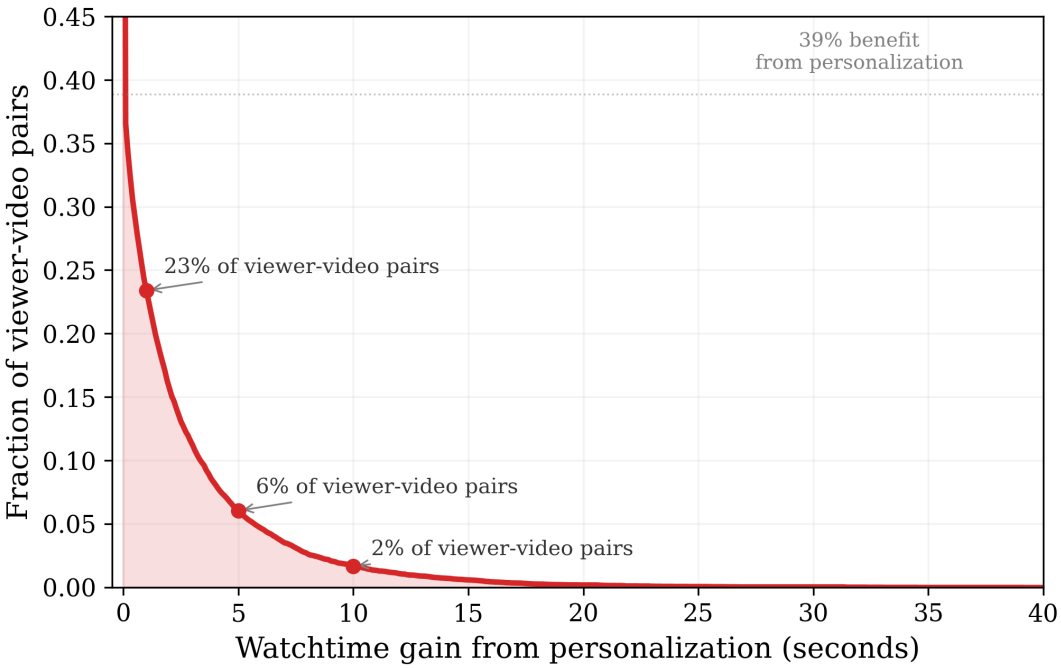


Figure G.0.5: **Distribution of watchtime gains from personalized thumbnail selection.** The curve shows the fraction of viewer-video pairs (out of $40 \times 361 = 14{,}440$) for which personalized thumbnail selection improves watchtime, relative to the global-best thumbnail for the same video. The global-best frame is already optimal for 61% of viewer-video pairs. Among the 39% who benefit, gains reach over 5 seconds for 5% of pairs and over 10 seconds for 1%, reflecting substantial viewer heterogeneity in responses to personalized thumbnails.

Variable	Mean	Std. Dev.	Median	Min	Max
Panel A: Thumbnail Theory-Driven Features					
Content Disconfirmation	0.257	0.092	0.272	0.000	0.485
Positive Aes. Disconf.	0.268	0.186	0.226	0.000	2.63
Negative Aes. Disconf.	0.244	0.130	0.226	0.000	0.977
Thumbnail Aesthetics	4.08	0.371	4.04	3.13	6.40
Thumbnail Location (Percentile)	0.377	0.280	0.398	0.000	1.000
Thumbnail Valence	5.52	0.779	6.00	2.00	7.00
Thumbnail Arousal	4.18	0.578	4.00	1.00	7.00
Thumbnail Anger	1.05	0.306	1.00	1.00	6.00
Thumbnail Disgust	1.05	0.378	1.00	1.00	7.00
Thumbnail Fear	1.13	0.536	1.00	1.00	6.00
Thumbnail Joy	5.14	1.06	5.00	1.00	7.00
Thumbnail Sadness	1.10	0.475	1.00	1.00	6.00
Thumbnail Surprise	1.89	0.921	2.00	1.00	7.00
Panel B: Control Variables					
Caption Features					
Caption Word Count	9.82	3.11	10.0	1.00	22.0
Caption Netspeak Share	0.002	0.013	0.000	0.000	0.200
Caption Affect Share	0.040	0.067	0.000	0.000	0.600
Caption Punctuation Share[a]	0.328	0.229	0.300	0.000	1.67
Caption Capitalized Share	0.132	0.119	0.119	0.000	1.00
Caption-Thumbnail Congruence	0.629	0.022	0.628	0.559	0.723
Caption-Video Congruence	0.622	0.014	0.622	0.571	0.692
Caption Negative Sentiment	0.022	0.107	0.001	0.000	0.947
Caption Neutral Sentiment	0.619	0.359	0.778	0.001	0.991
Caption Positive Sentiment	0.360	0.363	0.175	0.002	0.997
Video Mechanics					
Video Length	57.4	20.9	55.0	30.0	120
Number of Scenes	18.9	13.2	16.0	2.00	87.0
Average Scene Length	5.47	6.32	3.17	0.967	59.4
Video Topics					
Video Topic: Human & Vlogging	0.210	0.263	0.095	0.000	0.999
Video Topic: Entertainment	0.054	0.126	0.000	0.000	0.987
Video Topic: Water	0.082	0.173	0.000	0.000	0.998
Video Topic: Event	0.057	0.134	0.000	0.000	0.996
Video Topic: Nature	0.159	0.254	0.000	0.000	0.999
Video Topic: Hospitality	0.187	0.257	0.051	0.000	0.999
Video Topic: Transportation	0.084	0.168	0.000	0.000	0.997
Video Topic: Food	0.163	0.288	0.000	0.000	0.999
Channel and Timing					
Log Subscribers	8.61	3.17	8.42	0.000	17.0
Days Since Upload	106	62.3	107	0.000	212
N			4,121		

Table G.0.1: **Summary statistics for all variables.** [a]Note that in LIWC, punctuation output is expressed as a percentage of total words.

Variable	Type	Description
Caption Basic Features		
Caption Word Count	Numeric	Word count of video caption
Caption Netspeak Share	Numeric	Proportion of internet-specific language in caption (LIWC; rescaled to [0,1])
Caption Affect Share	Numeric	Proportion of affective/emotional language in caption (LIWC; rescaled to [0,1])
Caption Punctuation Share	Numeric	Proportion of punctuation marks in caption (LIWC; rescaled to [0,1]). It is expressed as a percentage of total words in LIWC.
Caption Capitalized Share	Numeric	Proportion of capitalized words in caption
Caption-Thumbnail/Video Features		
Caption-Thumbnail Congruence	Numeric	Cosine similarity between caption text and thumbnail
Caption-Video Congruence	Numeric	Cosine similarity between caption text and video content
Caption Sentiment (reference group = Neutral)		
Caption Negative Sentiment	Probability	Probability that caption sentiment is negative; positive, negative, and neutral probabilities sum to 1, with neutral as the reference category
Caption Positive Sentiment	Probability	Probability that caption sentiment is positive
Video Topics (reference group = Human & Vlogging)		
Video: Entertainment	Probability	Topic probability from topic model; probabilities across all topics (including human & vlogging) sum to 1 for each video
Video: Water	Probability	Topic probability for water activities
Video: Event	Probability	Topic probability for events and performance
Video: Nature	Probability	Topic probability for nature landscape and scenery
Video: Hospitality	Probability	Topic probability for hospitality and architecture
Video: Transportation	Probability	Topic probability for transportation
Video: Food	Probability	Topic probability for food
Video Mechanics		
Video Length	Numeric	Total video duration in seconds
Number of Scenes	Numeric	Number of detected scene cuts in the video
Average Scene Length	Numeric	Average scene duration in seconds
Channel and Timing Features		
Log(Subscribers+1)	Numeric	Log of channel subscriber count, log(subscribers + 1)
Days Since Upload	Numeric	Number of days since video was published

Table G.0.2: **Control variables included in all regression specifications.**

	log views			log((likes + 1)/views)		
	(1) S1	(2) S2	(3) S3	(4) S1	(5) S2	(6) S3
Thumbnail Features		✓	✓		✓	✓
Interaction Terms			✓			✓
Video Topic: Entertainment	0.112	−0.094	−0.084	−0.268	−0.328*	−0.334*
	(0.265)	(0.270)	(0.270)	(0.143)	(0.145)	(0.145)
Video Topic: Water	−1.332***	−1.448***	−1.457***	−0.154	−0.192	−0.186
	(0.202)	(0.207)	(0.207)	(0.108)	(0.112)	(0.112)
Video Topic: Event	−0.662***	−0.902***	−0.903***	−0.055	−0.097	−0.096
	(0.246)	(0.256)	(0.255)	(0.132)	(0.138)	(0.138)
Video Topic: Nature	−1.501***	−1.616***	−1.617***	0.454***	0.410***	0.410***
	(0.158)	(0.165)	(0.164)	(0.085)	(0.089)	(0.089)
Video Topic: Hospitality	−1.844***	−1.938***	−1.936***	0.298***	0.263**	0.262**
	(0.155)	(0.160)	(0.160)	(0.083)	(0.086)	(0.086)
Video Topic: Transportation	−1.764***	−1.790***	−1.772***	0.314**	0.306**	0.294**
	(0.208)	(0.210)	(0.210)	(0.112)	(0.113)	(0.113)
Video Topic: Food	−0.045	0.052	0.056	0.266***	0.283***	0.280***
	(0.150)	(0.153)	(0.153)	(0.081)	(0.083)	(0.083)
Caption Word Count	−0.027**	−0.025*	−0.026*	0.030***	0.030***	0.030***
	(0.010)	(0.010)	(0.010)	(0.005)	(0.005)	(0.005)
Caption Netspeak Share	−1.660	−1.880	−1.840	−1.214	−1.182	−1.208
	(2.410)	(2.390)	(2.390)	(1.292)	(1.292)	(1.291)
Caption Affect Share	−0.450	−0.400	−0.380	−0.877***	−0.953***	−0.966***
	(0.490)	(0.500)	(0.500)	(0.263)	(0.268)	(0.268)
Caption Punctuation Share	−0.130	−0.140	−0.150	0.441***	0.430***	0.434***
	(0.140)	(0.140)	(0.140)	(0.075)	(0.076)	(0.076)
Caption Capitalized Share	−0.029	−0.052	−0.064	−0.111	−0.101	−0.093
	(0.267)	(0.266)	(0.266)	(0.143)	(0.144)	(0.144)
Caption-Thumbnail Congruence	8.336***	7.108***	7.584***	−0.753	−1.439	−1.751
	(1.818)	(1.936)	(1.953)	(0.976)	(1.044)	(1.054)
Caption-Video Congruence	0.720	0.289	−0.415	−2.090	−2.131	−1.670
	(3.037)	(3.162)	(3.186)	(1.630)	(1.706)	(1.718)
Caption Negative Sentiment	1.647***	1.536***	1.536***	0.478**	0.463**	0.463**
	(0.294)	(0.296)	(0.296)	(0.158)	(0.160)	(0.160)
Caption Positive Sentiment	−0.137	−0.146	−0.147	0.121*	0.127*	0.127*
	(0.093)	(0.094)	(0.094)	(0.050)	(0.051)	(0.051)
Video Length	−0.011***	−0.011***	−0.011***	0.007***	0.007***	0.008***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Number of Scenes	0.018***	0.019***	0.019***	0.002	0.003	0.003
	(0.004)	(0.004)	(0.004)	(0.002)	(0.002)	(0.002)
Average Scene Length	−0.009	−0.015*	−0.014*	0.003	−0.000	−0.001
	(0.007)	(0.007)	(0.007)	(0.004)	(0.004)	(0.004)
Log(Subscribers+1)	0.448***	0.444***	0.443***	0.101***	0.099***	0.099***
	(0.011)	(0.011)	(0.011)	(0.006)	(0.006)	(0.006)
Days Since Upload	0.003***	0.003***	0.003***	0.001***	0.001***	0.001***
	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)
R^2	0.4751	0.4832	0.4836	0.1408	0.1456	0.1466
Adjusted R^2	0.4724	0.4790	0.4793	0.1364	0.1387	0.1395
N	4,121	4,121	4,121	4,121	4,121	4,121

Table G.0.3: **Extension of Table 2: Control variable estimates.** OLS estimates for control variables corresponding to the specifications in Table 2. Video topics reference category is human & vlogging. Caption sentiment baseline is neutral. Standard errors in parentheses. Significance levels: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

	log views		
	(1) S1	(2) S2	(3) S3
Content Disconfirmation (CD; first 30s; mean-centered)		−0.944*	−0.944*
		(0.414)	(0.414)
Thumbnail Aesthetics		−0.074	−0.075
		(0.139)	(0.139)
Pos. Aesthetic Disconf. (first 30s)		0.448	0.449
		(0.275)	(0.275)
Neg. Aesthetic Disconf. (first 30s)		0.104	0.106
		(0.251)	(0.251)
Thumbnail Location (first 30s Indicator)		−0.013	−0.017
		(0.075)	(0.075)
Thumbnail Arousal		0.163**	0.164**
		(0.060)	(0.060)
Thumbnail Anger		0.120	0.121
		(0.123)	(0.123)
Thumbnail Disgust		−0.071	−0.073
		(0.096)	(0.096)
Thumbnail Fear		−0.043	−0.044
		(0.069)	(0.069)
Thumbnail Joy		0.032	0.032
		(0.037)	(0.037)
Thumbnail Sadness		0.076	0.077
		(0.083)	(0.083)
Thumbnail Surprise		0.196***	0.196***
		(0.036)	(0.036)
CD (first 30s; mean-centered) × Thumbnail Location (first 30s indicator)			−0.855
			(0.658)
Controls	Yes	Yes	Yes
R^2	0.4750	0.4829	0.4832
Adjusted R^2	0.4723	0.4788	0.4789
N	4,121	4,121	4,121

Table G.0.4: **First 30 seconds robustness check.** Content disconfirmation and aesthetic disconfirmation are computed using only the first 30 seconds of each video. Thumbnail location is replaced with a binary indicator which equals to one if the thumbnail falls within the first 30 seconds. All other variables and controls are identical to Table 2. Standard errors in parentheses. Significance levels: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

	log((comments + 1)/views)		
	(1) S1	(2) S2	(3) S3
Content Disconfirmation (CD; mean-centered)		−0.124	−0.074
		(0.314)	(0.315)
Thumbnail Aesthetics		−0.195	−0.195
		(0.103)	(0.103)
Pos. Aesthetic Disconf. (AD^+)		0.126	0.131
		(0.208)	(0.208)
Neg. Aesthetic Disconf. (AD^-)		−0.552**	−0.547**
		(0.191)	(0.191)
Thumbnail Location (percentile)		0.107	0.107
		(0.081)	(0.081)
Thumbnail Arousal		−0.045	−0.046
		(0.042)	(0.042)
Thumbnail Anger		0.107	0.110
		(0.087)	(0.087)
Thumbnail Disgust		0.092	0.091
		(0.068)	(0.068)
Thumbnail Fear		−0.026	−0.024
		(0.049)	(0.049)
Thumbnail Joy		−0.027	−0.027
		(0.026)	(0.026)
Thumbnail Sadness		0.011	0.010
		(0.059)	(0.059)
Thumbnail Surprise		−0.042	−0.044
		(0.026)	(0.026)
CD (mean-centered) × Thumbnail Location (percentile)			−1.298
			(0.869)
Video Topic: Entertainment	−0.627***	−0.543**	−0.549**
	(0.186)	(0.190)	(0.190)
Video Topic: Water	0.292*	0.372*	0.378**
	(0.141)	(0.146)	(0.146)
Video Topic: Event	0.067	0.114	0.115
	(0.173)	(0.180)	(0.180)
Video Topic: Nature	0.708***	0.756***	0.756***
	(0.111)	(0.116)	(0.116)
Video Topic: Hospitality	0.684***	0.726***	0.725***
	(0.109)	(0.113)	(0.113)
Video Topic: Transportation	0.799***	0.759***	0.748***
	(0.146)	(0.148)	(0.148)
Video Topic: Food	−0.095	−0.140	−0.142
	(0.105)	(0.108)	(0.108)

Caption Word Count	0.015*	0.016*	0.016*
	(0.007)	(0.007)	(0.007)
Caption Netspeak Share	−0.118	0.024	0.000
	(1.687)	(1.685)	(1.685)
Caption Affect Share	0.065	−0.118	−0.129
	(0.344)	(0.350)	(0.350)
Caption Punctuation Share	0.307**	0.277**	0.281**
	(0.098)	(0.099)	(0.099)
Caption Capitalized Share	−0.198	−0.178	−0.171
	(0.187)	(0.187)	(0.187)
Caption-Thumbnail Congruence	−2.413	−1.584	−1.864
	(1.274)	(1.362)	(1.375)
Caption-Video Congruence	−4.120	−4.472*	−4.058
	(2.128)	(2.225)	(2.242)
Caption Negative Sentiment	0.126	0.052	0.052
	(0.206)	(0.208)	(0.208)
Caption Positive Sentiment	−0.016	0.008	0.009
	(0.065)	(0.066)	(0.066)
Video Length	0.013***	0.012***	0.012***
	(0.001)	(0.001)	(0.001)
Number of Scenes	−0.012***	−0.010***	−0.011***
	(0.003)	(0.003)	(0.003)
Average Scene Length	0.010*	0.008	0.008
	(0.005)	(0.005)	(0.005)
Log(Subscribers+1)	−0.138***	−0.140***	−0.140***
	(0.008)	(0.008)	(0.008)
Days Since Upload	−0.001*	−0.001*	−0.001*
	(0.000)	(0.000)	(0.000)
R^2	0.2051	0.2109	0.2113
Adjusted R^2	0.2010	0.2045	0.2047
N	4,121	4,121	4,121

Table G.0.5: **Commenting full estimates.** OLS estimates with the outcome commenting defined as $\log((\text{comments} + 1)/\text{views})$. All specifications are identical to Table 2. Video topics reference category is human & vlogging. Caption sentiment baseline is neutral. Standard errors in parentheses. Significance levels: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Panel A: General Case (∀ **Set** S_i)				
	First Video	**Second Video**	**Third Video**	**Fourth Video**
Condition 1	1st Thumbnail	2nd Thumbnail	3rd Thumbnail	4th Thumbnail
Condition 2	2nd Thumbnail	1st Thumbnail	4th Thumbnail	3rd Thumbnail
Condition 3	3rd Thumbnail	4th Thumbnail	1st Thumbnail	2nd Thumbnail
Condition 4	4th Thumbnail	3rd Thumbnail	2nd Thumbnail	1st Thumbnail
Panel B: Specific Example (Set S_1, Videos V_{11} to V_{14})				
	V_{11}	V_{12}	V_{13}	V_{14}
Condition 1	T_{111}	T_{122}	T_{133}	T_{144}
Condition 2	T_{112}	T_{121}	T_{134}	T_{143}
Condition 3	T_{113}	T_{124}	T_{131}	T_{142}
Condition 4	T_{114}	T_{123}	T_{132}	T_{141}

Table G.0.6: **Assignment of Thumbnails to Videos Under Four Experimental Conditions**. Panel A shows the assignment of thumbnails to any set of four videos under the four experimental conditions. Panel B provides a specific example to illustrate the thumbnail assignment, using the first set of four videos. We denote S_i as ith set of four videos; V_{ij} as the j-th video in set S_i, where j ranges from 1 to 4; T_{ijk} as the k-th thumbnail for video V_{ij}, where k ranges from 1 to 4 (experimental conditions). In this study, we have 40 videos on the website grouped into 10 sets of videos (S_1 to S_{10}), and each video has four versions of thumbnails. Thus, there are 160 different thumbnails.

Variable	All Experimental Conditions	Condition 1	Condition 2	Condition 3	Condition 4
Age below 25	0.07	0.07	0.03	0.06	0.12
Age 25–34	0.30	0.33	0.25	0.31	0.34
Age 35–44	0.29	0.26	0.35	0.29	0.29
Age 45–54	0.20	0.18	0.26	0.23	0.09
Age 55–64	0.12	0.15	0.08	0.10	0.17
Age above 64	0.01	0.01	0.02	0.01	0.00
Female	0.56	0.51	0.68	0.65	0.36
Male	0.43	0.48	0.31	0.34	0.64
Other Gender	0.01	0.01	0.01	0.01	0.00
White	0.72	0.72	0.68	0.73	0.73
Black	0.10	0.12	0.15	0.07	0.06
Asian	0.08	0.08	0.06	0.10	0.09
Hispanic	0.06	0.03	0.06	0.06	0.08
Other Ethnicity	0.04	0.04	0.06	0.03	0.04
No. Obs.	361	98	88	98	77

Table G.0.7: **Distribution of respondents by demographics**. This table shows the distribution of respondents based on demographic characteristics for all subjects and separately by condition. The column "Other Ethnicity" includes "Indigenous, American Indian or Alaska Native", "I prefer to self-describe", and "I don't wish to answer". Note that our experimental condition is a combination of thumbnails assigned to participants to choose from among the set of four videos.

Thumbnail Candidate	Overlap
Optimal frames	
Global Best	12%
Platform-recommended frames	
Best of 3 Platform	70%
Left Frame	20%
Middle Frame	32%
Right Frame	25%
Stylized frames	
Highest Arousal	25%
Most Aesthetic	22%
Most Surprising	8%
Most Emotional	15%
Most Representative	30%
Most Summarizing	18%
First Non-Black	10%

Table G.0.8: **Overlap between candidate thumbnails and creator-chosen frame.** % videos (out of 40) in which the candidate frame matches with the creator-chosen thumbnail. The platform-recommended frames were recorded at the time of our data collection and may not reflect the options creators saw when they originally uploaded their videos and thus the 70% agreement. The relative rank of overlap matters more since it informs us creators' strategies.

				Platform-recommended (YouTube)				
	Personalized	Assigned	Creator-chosen	Left	Middle	Right	Best of 3	Random
Click Prob.	0.401	*0.279*	0.268	0.257	0.272	0.269	0.308	0.268
Watchtime Periods	8.19	*7.57*	7.60	7.75	7.30	7.15	7.85	7.44
Joint	2.956	*2.095*	2.034	2.011	1.989	1.915	2.277	1.982

Table G.0.9: **Holdout thumbnail performance**

References (Web Appendix)

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Roberto Brunelli. *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons, 2009.

John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 2009.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235, 2004.

Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23, 2010.

John P Lewis et al. Fast template matching. In *Vision interface*, volume 95, pages 15–19. Quebec City, QC, Canada, 1995.

Meta AI. Dinov3: Self-supervised learning for vision at unprecedented scale. https://ai.meta.com/blog/dinov3-self-supervised-vision-model/, 2025. Accessed: 2025-11-10.

Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024a. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.