ClipMind: A Framework for Auditing Short-Format Video Recommendations Using Multimodal AI Models

Aoyu Gong[1*], **Sepehr Mousavi**[2], **Yiting Xia**[3], **Savvas Zannettou**[4]
[1]EPFL, Lausanne, Switzerland
[2]Max Planck Institute for Software Systems, Saarbrücken, Germany
[3]Max Planck Institute for Informatics, Saarbrücken, Germany
[4]TU Delft, Delft, The Netherlands
aoyu.gong@epfl.ch, smousavi@mpi-sws.org, yxia@mpi-inf.mpg.de, s.zannettou@tudelft.nl

Abstract

We are witnessing a significant shift in social media platforms; we are transitioning from chronological social media feeds to feeds that are driven by AI recommendation systems. While the main goal of AI recommendation systems is to suggest engaging content to users, there are also some associated risks: AI recommendation systems can promote extreme content, causing negative consequences like online polarization and user radicalization. Overall, there is a pressing need to design powerful techniques that allow us to audit AI recommendation systems. Motivated by this, our work introduces *ClipMind*, a scalable and generalizable framework using advanced AI models to audit these recommendation algorithms on short-format video platforms like TikTok and YouTube Shorts. We demonstrate the merits of our framework by collecting social media feeds from TikTok. Our analysis shows that TikTok's recommendation algorithm increasingly recommends similar videos when a user expresses interest in mainstream topics like Food and Beauty Care. On the other hand, by investigating niche interests (War and Mental Health), we find no evidence of informational rabbit holes of extreme content on TikTok. Our work contributes to efforts that leverage AI for social good, as our framework can be used by several interested stakeholders, including users, social media platforms, regulators, and researchers, to understand and audit video-based algorithmic recommendations.

1 Introduction

Over the past few years, social media platforms have undergone a substantial makeover across two crucial dimensions: the format of content and the way of delivering content to users. Regarding the format of content, in the past, most of the content was either long (e.g., videos on YouTube) or non-multimedia-related (e.g., textual posts on Twitter, Facebook, etc.). Now, we are witnessing the rise of short-format videos as one of the main types of content disseminated on platforms like TikTok, YouTube, Facebook, and Instagram. Regarding the way of delivering content to users, in the past, social media platforms were creating chronologically ordered social media feeds based on users' social connections. However, nowadays, most platforms incorporate powerful AI recommendation systems that aim to recommend videos and create tailored and personalized social media feeds of short-format videos based on users' activities and interests (TikTok 2020). Overall, we live in an era where powerful AI recommendation systems create endless social media feeds of short-format videos.

AI recommendation systems on short-format video platforms have to generate many recommendations in a short period, mainly because of the format of content (i.e., videos that last below a minute). At the same time, users have less freedom to choose their recommendations, as the next video is determined exclusively by the recommendation system without users' explicit input (this is in contrast to YouTube, for example, as a user can select a recommended video out of 10-20 recommendations). Due to the large number of recommendations over a short period of time and the seemingly fewer explicit interactions by users, several concerns are raised on how these AI recommendation systems can drive users towards dangerous paths. Previous research work, journalistic investigations, and anecdotal evidence highlighted the dangers of such AI recommendation systems on short-format platforms like TikTok and YouTube Shorts. For instance, journalistic investigations from the Wall Street Journal (WSJ Staff 2021) demonstrate that TikTok's recommendation algorithm can drive users towards informational rabbit holes of extreme content (e.g., many similar, potentially extreme videos like depression videos within a short period). Another example is from SkyNews (Burgess 2023), which demonstrates that both YouTube Shorts and Instagram algorithms pushed misogynistic content on "teens" profiles created specifically for this investigation. Previous research work also highlighted the dangers of AI recommendation systems in radicalizing and polarizing online users (Ribeiro et al. 2020; Ledwich and Zaitsev 2019; Papadamou et al. 2022, 2020; Hussein, Juneja, and Mitra 2020). Overall, there is a pressing need to systematically audit these algorithms to assess their role and potential negative effects on our society.

While previous efforts are paramount, they are limited for various reasons. First, they focus on specific topics, including misinformation content (Hussein, Juneja, and Mitra 2020; Papadamou et al. 2022), political content (Ledwich and Zaitsev 2019; Ribeiro et al. 2020), or kids' videos (Papadamou et al. 2020). Second, previous work heavily relies on manual annotations to understand and analyze video content; hence, their

[*]Work done during an internship at the Max Planck Institute for Informatics.

approaches do not scale well. These limitations highlight the need for a generalizable and scalable framework to demystify algorithmic recommendations and effectively illuminate their potential negative effects.

Having this motivation in mind, this work presents *ClipMind*, a scalable and generalizable framework for understanding algorithmic recommendations on short-format video platforms like TikTok and YouTube Shorts. We use state-of-the-art (SOTA) AI models to audit AI recommendation systems on short-format video platforms. Specifically, we use AI models, including ImageBind (Girdhar et al. 2023), Video-LLaMA (Zhang, Li, and Bing 2023), GPT-4 (OpenAI 2023), and ADA-002 (Greene et al. 2022), to generate contextually and semantically rich embeddings for short-format videos, which allow us to assess the similarity of content. Then, by applying temporal and graph analysis, we demonstrate how to model sequences of videos that are recommended by AI algorithms and how this methodology helps us audit AI recommendation systems. We demonstrate the applicability and merits of ClipMind on a TikTok dataset. Specifically, we collect diverse traces of video recommendations on TikTok, including traces where users show interest in mainstream and niche topics like Food, Beauty Care, Mental Health, and War. By applying ClipMind to the collected traces, we find that: 1) When a user shows interests in mainstream topics like Food and Beauty Care, a significant percentage of recommendations within a sliding window are about the topic of interest (between 60% and 80%). 2) When a user shows interest in niche topics like War and Mental health, we do not find a significant percentage of recommendations that are related to these niche topics. In other words, we find no evidence of informational rabbit holes of extreme content.

Contributions & Implications. Our paper makes a notable contribution with important societal implications. We demonstrate and make publicly available ClipMind,[1] a powerful framework for auditing algorithmic recommendations on short-format video platforms. The developed framework has important implications for various stakeholders, including social media platforms, users, regulators, and researchers. Social media platforms can leverage ClipMind to understand how their recommendation algorithms behave on a large scale and improve their algorithms. Users can leverage ClipMind to analyze their own social media feeds, hence getting informed about potential signs of informational rabbit holes (i.e., many similar, potentially extreme videos within a short period). Finally, ClipMind can be used by regulators and the research community that aim to audit recommendation algorithms and assess the platforms' compliance with emerging and important regulations like the Digital Services Act (European Commission 2023). Overall, our work contributes to efforts that leverage AI for social good by proposing a framework that can help audit AI recommendation algorithms with the goal of raising more awareness about their issues and preventing the creation of fragmented societies because of user radicalization online.

[1]Available at https://github.com/aygong/ClipMind

2 Background

Here, we overview SOTA AI models used in our framework.

ImageBind. ImageBind (Girdhar et al. 2023) is a novel approach able to bind data from six modalities: images/videos, audio, texts, depth, thermal, and inertial measurement units. It learns one joint embedding space by leveraging the binding property of images/videos and aligning the embeddings of other modalities to image/video embeddings. It attains SOTA performance on various tasks across modalities. We use this model to generate video and audio embeddings of dimensions $d_{\text{video}} = d_{\text{audio}} = 1024$.

Video-LLaMA. Video-LLaMA (Zhang, Li, and Bing 2023) is a multi-modal framework able to understand both visual and auditory content in videos. It utilizes pre-trained visual and auditory encoders to generate video and audio embeddings. To align the embeddings with the embedding space of pre-trained large language models (LLMs), Video-LLaMA was trained on massive image/video-caption pairs and fine-tuned on visual-instruction datasets. Given its strong performance in video understanding, we use Video-LLaMA, which takes videos as inputs and generates responses based on prompts.

GPT-4. GPT-4 (OpenAI 2023) is a multimodal model capable of accepting text and image inputs and producing text outputs. It was trained on publicly available data and data licensed from third-party providers, and then fine-tuned using reinforcement learning from human feedback. GPT-4 outperforms prior LLMs and most SOTA systems on a suite of NLP benchmarks. We use this model to process user-defined metadata and generate keywords.

ADA-002. ADA-002 (Greene et al. 2022) is OpenAI's new embedding model capable of producing numerical representations of texts. It outperforms all OpenAI's old embedding models on tasks such as text similarity. We use this model that generates text embeddings of dimensions $d_{\text{text}} = 1536$.

3 Related Work

Here, we overview work on auditing algorithmic recommendation algorithms and analyzing social media feeds.

Auditing Recommendation Algorithms. Prominent social media platforms have started using AI-powered recommendation systems in their feeds to serve content to their end-users (TikTok 2020; Meta 2019). However, this shift in delivering content brought about potentially problematic side effects for the end-users across multiple social media platforms. For example, these AI-powered recommendation systems may trap the users in a rabbit hole of sad content (WSJ Staff 2021) or suggest extreme content to them (Ribeiro et al. 2020). Therefore, concerns about these AI-based recommendation systems have increased over time, and legislators have called for periodic audits on these AI-powered recommendation systems to address the rising concerns. (Bandy 2021; European Commission 2023). Motivated by this, previous work has audited these recommendation systems to understand better how they work and what factors contribute to the content selection of user feeds. For example, a journalistic investigation by the Wall Street Journal showed that

the video-watching time is a strong signal that immensely affects TikTok video recommendations (WSJ Staff 2021). Also, researchers empirically studied different factors that affect user feed personalization on TikTok (Boeker and Urman 2022). Klug et al. performed a mixed-method study on user assumptions about the TikTok algorithm (Klug et al. 2021). They found out that the time a TikTok video gets posted also influences the recommendation system algorithm.

Analyzing Social Media Feeds. Researchers have been employing different methodologies to analyze social media content. One major technique is to analyze the content description, especially hashtags, to evaluate content similarity, how user's feed is shaped, or to detect if content is about a specific topic of interest (Boeker and Urman 2022; Klug et al. 2021; Pilař et al. 2021; Ling, Gummadi, and Zannettou 2023; Weimann and Masri 2023; Shen et al. 2022). Another popular methodology for analyzing social media content is to annotate the posts by researchers or external annotators (Ali et al. 2023; Song et al. 2021; Jasser et al. 2023; Paudel et al. 2023; Papadamou et al. 2022). Also, research recently proposed methodologies for analyzing social media content by incorporating multi-modal features, such as visual, audio, and textual content. For instance, a recent work introduces TikTec, which is a multi-modal framework for detecting misinformation in videos (Shang et al. 2021).

4 Data Collection

To validate and demonstrate the applicability of our framework, we collect social media feeds (i.e., traces) from TikTok. Each trace contains a sequence of videos recommended by TikTok's recommendation algorithm on the "For You" feed. To collect the data, we use automated accounts (i.e., bots) that scroll through the "For You" feed and collect video content and associated metadata. Each automated account has a pre-defined "bootstrapping" and "watching" configuration. The former aims to bootstrap an account with some pre-defined interests, while the latter determines which videos are watched until the end by the bot (this is an important feature that provides input to the recommendation algorithm (TikTok 2020)). We implemented our automated accounts using the Playwright library (Microsoft 2024), and we ran our data collection between November 2023 and January 2024 from a US-based IP address. Overall, we collected five different traces from TikTok: one *random trace* and four *topic-specific traces*, using five separate automated accounts. We explain the traces below.

Random Trace. We collect a *Random* trace that acts as a baseline dataset. We use one account without any bootstrapping that watches videos with a random viewing duration for each video. In total, we collected the first 500 videos that were recommended by TikTok's algorithm.

Topic-Specific Traces. We aim to collect TikTok traces for accounts with pre-defined topics of interest. We created four separate accounts, each bootstrapped with one of the following topics: 1) Food; 2) Beauty Care; 3) Mental Health; and 4) War/Military. We chose these four topics to cover a wide span of TikTok content, including mainstream and potentially problematic topics, (see Appendix A for the topic definitions). Mainstream topics refer to subjects that are widely popular, socially accepted, and commonly consumed by the general public. The Food and Beauty Care traces were chosen as they align with cultural trends and are officially featured on TikTok's Explore page. In contrast, potentially problematic topics refer to subjects that may give rise to controversy, ethical concerns, or societal issues. The Mental Health and War/Military traces were selected based on recent research concerns (Braghieri, Levy, and Makarin 2022; Pretorius, McCashin, and Coyle 2022; Liaropoulos 2023; Badola 2023). Such content can have profound ethical implications for end-users. For instance, inappropriate recommendations related to mental health might expose vulnerable users to harmful or misleading content, while recommendations on war/military could reinforce political and societal divisions. These implications highlight potential risks associated with such topics, ultimately influencing user perceptions and behaviors in ways that could undermine trust, safety, and well-being.

For each topic, we first collect a set of 40 related hashtags (see Appendix A for details about the hashtag collection). Then, we bootstrapped each account with a different topic by performing hashtag searches and watching videos. Specifically, we selected five hashtags from the set of 40 related hashtags and performed searches on TikTok, finding popular videos including these hashtags. Then, each account watches between ten to fifteen videos, including these hashtags, for twice the duration of each video. After completing the bootstrapping phase, each account visits the "For You" feed and watches videos. In particular, if a video has a hashtag that appears in our set of 40 hashtags per topic, the account watches the video for twice its duration; otherwise, it goes to the next video immediately. This setup is based on the evidence that social media algorithms prioritize longer engagement to personalize recommendations (Shahbaznezhad, Dolan, and Rashidirad 2021), ensuring that each account efficiently develops a strong interest in the selected topic. We collect the video content and metadata for each video that the automated account encounters. Overall, we collected a total of 2000 videos, 500 videos for each automated account and topic.

5 ClipMind Framework

This section presents our framework for analyzing video sequences on short-format video platforms. Given a sequence of videos, we use SOTA multimodal AI models to: 1) **Generating Embeddings:** Perform feature extraction and map videos into a high-dimensional vector space (i.e., convert videos into embeddings capturing their semantics); 2) **Assessing Similarity:** Assess the similarity of videos based on their embeddings; and 3) **Video Sequence Analysis:** Analyze sequences of videos using temporal and graph analysis techniques. Table 1 provides an overview of the used features and SOTA AI models. Below, we explain each component.

5.1 Generating Embeddings

Videos on short-format video platforms consist of various multimodal features, including video visual content, audio content, and user-defined metadata (e.g., video hashtags). To fully capture the semantics of short-format videos, we take

into account all available multimodal features and create five high-level features corresponding to: 1) Video visual content; 2) Video audio content; 3) User-defined metadata; 4) LLM-generated description; 5) LLM-generated keywords. The first three features are readily available on short-format video platforms, while the last two features leverage SOTA LLMs to generate additional context about each video. The final embedding for each video is the concatenation of individual feature embeddings. Below, we describe how we generate embeddings for each feature.

Video Visual and Audio Content Visual and auditory signals enter a user's brain when viewing short-format videos. The former incorporates information conveyed through continuous frames, and the latter sounds like music and dialogue. For a given video x, we use ImageBind to represent its visual and auditory signals numerically, i.e., obtain its video and audio embeddings, denoted by $\mathbf{e}_{\text{v},x}$ and $\mathbf{e}_{\text{a},x}$ respectively:

$$\mathbf{e}_{\text{v},x} = \text{ImageBind}(x) \in \mathbb{R}^{d_{\text{video}} \times 1},$$
$$\mathbf{e}_{\text{a},x} = \text{ImageBind}\big(\text{AudioConverter}(x)\big) \in \mathbb{R}^{d_{\text{audio}} \times 1},$$

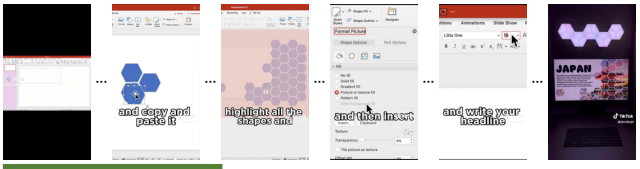
where AudioConverter is an MP4-to-MP3 converter.

LLM-Generated Description ClipMind considers both video visual and audio content using ImageBind by generating separate embeddings for each modality. While these embeddings provide important semantic information, they lack the general knowledge of audio-visual LLMs that combine visual and audio content and can interpret various video events. Motivated by this, our framework uses Video-LLaMA to generate rich and contextual descriptions based on the video's visual and audio content. We use the prompt "Describe this video" and generate descriptions for short-format videos. For a given video x, we pass its description to ADA-002 and obtain its description embedding, denoted by $\mathbf{e}_{\text{d},x}$:

$$\mathbf{e}_{\text{d},x} = \text{ADA-002}\big(\text{Video-LLaMA}(x)\big) \in \mathbb{R}^{d_{\text{text}} \times 1}.$$

User-Defined Metadata An important feature of short-format videos on social media platforms is user-defined metadata like titles and/or hashtags associated with videos. This information is usually defined by users (i.e., the uploader/creator of the video) and aims to summarize video content to attract views from other users. Therefore, user-defined metadata provides important information that allows us to assess the semantics of the video and assess the similarity between videos (see Fig. 1 for examples). Our framework concatenates user-defined metadata, particularly the video's title and a set of hashtags, into a single document and then uses the ADA-002 model to generate embeddings.

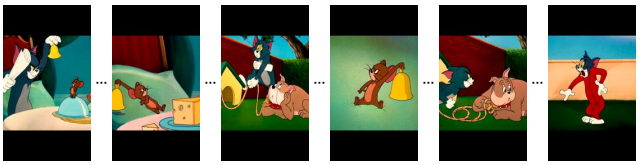
Although user-defined metadata provides crucial context, some limitations must be addressed via preprocessing steps. First, videos may contain meaningless hashtags unrelated to their content. These hashtags are commonly used in many videos to influence the platform's algorithm; hence, we can easily identify and remove them. Second, videos may lack user-defined metadata. To overcome this limitation, we use SOTA LLMs to generate metadata for videos that lack it. Third, user-defined metadata are diverse in terms of language,



User-Defined Metadata

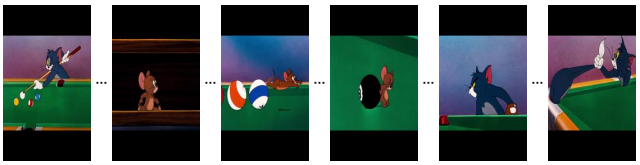
10x your PowerPoint presentations this semester with this trick ✨‼ A super simple PowerPoint tutorial to level up your design 💻 ...

(a) This short-format video is about PowerPoint tutorials and the user-defined metadata clearly describes its content.



User-Defined Metadata

#tom #jerry #tomandjerry #jerryandtom #animation #childhoodaanimation #carttoon #carttoon #carrtoonnetwork #cartoonclips #cartoonclips2023 ...



User-Defined Metadata

#tom #jerry #tomandjerry #jerryandtom #animation #childhoodaanimation #carttoon #carrtoonnetwork #cartoonclips #cartoonclips2023 ...

(b) The text embeddings of the user-defined metadata of the two short-format videos have a cosine similarity of 0.9971.

Figure 1: Examples of how user-defined metadata provides important semantic information.

which might affect the performance of our approach. Although ADA-002 and other SOTA embedding models are multilingual, the text embeddings of metadata in the same language tend to be closer. To overcome this, we translate non-English metadata. We describe each step below.

Filtering Hashtags. To filter meaningless hashtags, we use a semi-automatic approach. First, given a sequence of videos, we find the occurrence frequency for each hashtag and manually examine the top $\theta\%$ in terms of their occurrences in the sequence. Then, we create a set of meaningless hashtags (not likely to contribute to capturing video semantics) and preprocess the user-defined metadata, removing all hashtag occurrences from this set.

Missing Metadata. For videos lacking user-defined metadata, we use GPT-4 to generate metadata based on LLM-generated descriptions. We use the prompt:

> *I will provide you with the description of a {platform} video. I want you to give a title to this video.*
> *The description is {LLM-generated description}.*
> *Your answer is:*

Translation. We use Google Translate to translate the metadata into English for videos with non-English metadata.

Overall, for a given video x, we pass its preprocessed

Feature	Generating or Processing Model	Embedding Model	Embedding Notation	Embedding Dimensions
Video visual content	ImageBind	ImageBind	$\mathbf{e}_{\mathrm{v},x}$	1024
Video audio content	ImageBind	ImageBind	$\mathbf{e}_{\mathrm{a},x}$	1024
LLM-generated description	Video-LLaMA	ADA-002	$\mathbf{e}_{\mathrm{d},x}$	1536
User-defined metadata	GPT-4	ADA-002	$\mathbf{e}_{\mathrm{m},x}$	1536
LLM-generated keywords	GPT-4	ADA-002	$\mathbf{e}_{\mathrm{k},x}$	1536

Table 1: Overview of our features. We report the models for generating/processing features/embeddings, and the embeddings' notation/dimensions. We use ImageBind, Video-LLaMA, GPT-4, and ADA-002 to represent a video with an embedding.

user-defined metadata to ADA-002 and obtain its metadata embedding, denoted by $\mathbf{e}_{\mathrm{m},x}$:

$$\mathbf{e}_{\mathrm{m},x} = \text{ADA-002}\big(\text{MetadataPreprocessor}(x)\big) \in \mathbb{R}^{d_{\text{text}} \times 1}.$$

LLM-Generated Keywords Despite the potential of LLM-generated descriptions and user-defined metadata, they have inherent limitations due to how they are produced. First, the descriptions are generated by Video-LLaMA, the performance of which is strong but still inferior to that of humans. In consequence, the generated descriptions may contain inaccurate information. Moreover, since the generation is only performed by Video-LLaMA, the sentence structure of such descriptions may be similar, thus leading to close text embeddings for dissimilar videos. On the other hand, the metadata is defined by users whose creativity is very diverse. Even for the same video, different users may define various metadata, thus leading to distant text embeddings for similar videos.

Inspired by latent semantic indexing, which uses concepts for retrieval (Barde and Bainwad 2017), we instruct GPT-4 to generate keywords for videos from a given pool, denoted by {keyword pool}. This constraint avoids inconsistent representations of the same content. By standardizing keywords, we ensure that identical content is represented consistently, addressing inaccuracies and limitations in LLM-generated descriptions while mitigating diversity in user-defined metadata. Based on descriptions generated by Video-LLaMA and metadata defined by users, we adopt a prompt as follows:

> *I will give you the description, metadata, and auxiliaries of a {platform} video. I will give you a pool of keywords: {keyword pool}. I want you to select keywords related to the video from this pool based on the description, metadata, and auxiliaries. I want you to only reply with the keywords and nothing else. Note that the metadata and auxiliaries may be empty or may provide no additional information.*
> *The description is {LLM-generated description}.*
> *The metadata is {user-defined metdata}.*
> *The auxiliaries are {auxiliaries}.*
> *Your answer is:*

where {*auxiliaries*} denote other (optional) types of metadata that may also help generate keywords. The responses of GPT-4 are considered as generated keywords. For a given video x, we pass its keywords to ADA-002 and obtain its keyword embedding, denoted by $\mathbf{e}_{\mathrm{k},x}$:

$$\mathbf{e}_{\mathrm{k},x} = \text{ADA-002}\big(\text{KeywordGenerator}(x)\big) \in \mathbb{R}^{d_{\text{text}} \times 1}.$$

Remark. For general analysis, the pool can be created using official topics from social media platforms or trending ideas identified by third-party analysis websites. Furthermore, this pool can be further customized to suit specific use cases. For example, it can be designed to intentionally include detailed categories or topics within a specific domain, which enables the detection and measurement of biases or the evaluation of diversity in recommended content. By refining the keyword pool, the finer keywords enable more precise similarity assessment and deeper recommendation analysis. In addition, the pool can be tailored to reflect users' preferences, ensuring that the generated keywords are constrained to their interests. Correspondingly, our framework will place greater emphasis on videos that align more closely with users' interests. Such a pool can potentially adapt our framework to various usages.

5.2 Assessing Video Similarity

In the previous section, we described the various features we consider and how to extract an embedding for each feature. At the end, a video can be represented as a single embedding by concatenating all or a subset of the feature embeddings. This embedding captures the semantics of the video and allows us to assess similarity across videos. Here, having a single embedding for each video, we aim to identify the best combination of features and similarity threshold that will help us label videos accurately and confidently as similar. We perform sampling of pairs of videos, manual annotations to construct a ground truth dataset of similar/dissimilar video pairs, and then an evaluation to identify the best feature combination and similarity threshold based on the ground truth dataset. Recall that we introduced five features and their embeddings; let $\mathcal{F}$ denote the set of these features, i.e.,

$$\begin{aligned}\mathcal{F} = \{&\text{video visual content, video audio content,}\\ &\text{LLM-generated description, user-defined metadata,}\\ &\text{LLM-generated keywords}\}.\end{aligned}$$

Let $\mathcal{P}(\mathcal{F})$ represent the set of all subsets of $\mathcal{F}$. We consider each set $\mathcal{S} \in \mathcal{P}(\mathcal{F}) \setminus \emptyset$ as a *combination* of features. Given an arbitrary combination, we fuse its features by normalizing their corresponding embeddings with the L^2 norm and concatenating the normalized embeddings. For instance, for a given video x, if using the combination $\mathcal{F}$, we have

$$\begin{aligned}\mathbf{e}_{\mathcal{F},x} = &\frac{\mathbf{e}_{\mathrm{v},x}}{\|\mathbf{e}_{\mathrm{v},x}\|} \oplus \frac{\mathbf{e}_{\mathrm{a},x}}{\|\mathbf{e}_{\mathrm{a},x}\|} \oplus \frac{\mathbf{e}_{\mathrm{d},x}}{\|\mathbf{e}_{\mathrm{d},x}\|} \oplus \frac{\mathbf{e}_{\mathrm{c},x}}{\|\mathbf{e}_{\mathrm{c},x}\|} \oplus \frac{\mathbf{e}_{\mathrm{k},x}}{\|\mathbf{e}_{\mathrm{k},x}\|}\\ &\in \mathbb{R}^{d_{\text{video}}+d_{\text{audio}}+3d_{\text{text}}} \qquad (1)\end{aligned}$$

where $\oplus$ represents the concatenation operator. Since there are 31 combinations, we aim to find the best combination for assessing video similarity. Note that each individual feature is generated and/or processed using SOTA AI models, which serve as baselines in our comparison.

Sampling. Given a sequence of videos, we first generate the concatenated embedding based on the combination $\mathcal{F}$ (i.e., all five features) and Equation (1). Then, we calculate the cosine similarity for all video pairs in the sequence and assume that such values are distributed in the interval $[l, u]$. To make sampled video pairs as balanced as possible, we divide the interval $[l, u]$ into A intervals of equal length and uniformly sample B video pairs from each interval. After this, we have a dataset that consists of AB video pairs.

Annotation. Two annotators independently annotate the dataset. For each video pair, they are provided with two videos along with their user-defined metadata. Using such information, they are required to annotate whether the two videos are similar or not. The information provided here is similar to what users encounter when browsing videos on social media platforms. After completing the annotations, a quantitative assessment is conducted by calculating Cohen's Kappa coefficient. To resolve disagreements, the annotators discuss and determine the final annotations.

Best Feature Combination/Similarity Threshold. Given the annotated (groundtruth) dataset, we aim to identify the best feature combination and similarity threshold for identifying whether a pair of videos is similar or not. To do this, For each combination $\mathcal{S} \in \mathcal{P}(\mathcal{F}) \setminus \emptyset$, we generate concatenated embeddings and calculate the cosine similarity for the AB video pairs. For each combination, we further introduce a threshold, denoted by $\epsilon_{\mathcal{S}}$, so that a video pair is considered to be similar if the cosine similarity is greater than $\epsilon_{\mathcal{S}}$ and dissimilar otherwise. Let $\mathcal{E} = \{0, \Delta, 2\Delta, \ldots, 1\}$ represent the set of all possible values of $\epsilon_{\mathcal{S}}$, where $\Delta \in (0, 1]$. For each $\epsilon_{\mathcal{S}} \in \mathcal{E}$, we consider the annotations to be ground truth and compute four metrics including accuracy, precision, recall, and F_1 score, denoted by $A_{\mathcal{S}}(\epsilon_{\mathcal{S}})$, $P_{\mathcal{S}}(\epsilon_{\mathcal{S}})$, $R_{\mathcal{S}}(\epsilon_{\mathcal{S}})$, $F_{\mathcal{S}}(\epsilon_{\mathcal{S}})$, respectively. The optimal threshold $\epsilon^*_{\mathcal{S}}$ is further defined by $\epsilon^*_{\mathcal{S}} = \arg\max_{\epsilon_{\mathcal{S}} \in \mathcal{E}} F_{\mathcal{S}}(\epsilon_{\mathcal{S}})$.

5.3 Video Sequence Analysis

Thus far, we described how ClipMind generates embeddings for short-format videos and assesses the semantic similarity between videos. Here, we will describe how we analyze entire video sequences on short-format video platforms to demystify video recommendations and whether recommendation algorithms on short-format video platforms drive users toward information rabbit holes.

Let $\mathcal{X}$ represent a sequence of videos an AI algorithm recommends on a short-format video platform. The sequence of videos in this set can be written as $(x_n)_{n \in \{1,2,\ldots,|\mathcal{X}|\}}$, where x_n is the n-th video. For each $x_n \in \mathcal{X}$, we prepare its features in the best combination $\mathcal{S}^*$ and the corresponding embeddings as introduced in Section 5.1. Then, we fuse its features by generating its concatenated embedding $\mathbf{e}_{\mathcal{S}^*, x_n}$. Given the optimal threshold $\epsilon^*_{\mathcal{S}^*}$, we are ready to conduct similarity analysis with the aim of identifying

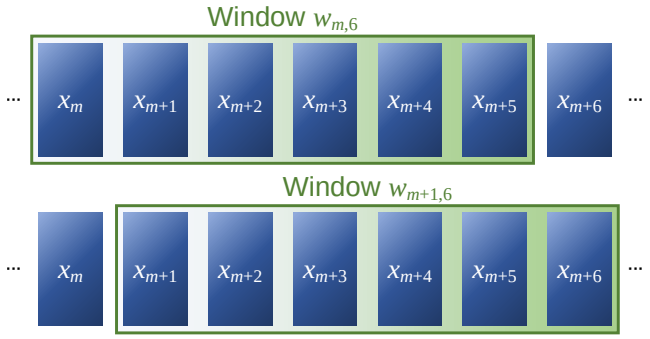


Figure 2: An example of a sliding window with $L = 6$.

similar video pairs and understanding how the system's recommendations change over time. To analyze sequences of videos, we perform a sliding window analysis. We consider that a window with a length of $L \geq 2$ contains L consecutive videos. Let the sequence $w_{m,L} = (x_m, x_{m+1}, \ldots, x_{m+L-1})$ be the window in which the first video is the m-th video in $(x_n)_{n \in \{1,2,\ldots,|\mathcal{X}|\}}$, where $m \in \{1, 2, \ldots, |\mathcal{X}| - L + 1\}$. An example for $L = 6$ is presented in Fig. 2. Note that a window with $L = 2$ contains two consecutive videos, while a window with $L = |\mathcal{X}|$ contains all videos in a sequence.

We represent all videos as an undirected graph for each window $w_{m,L}$, where nodes are videos that are connected together if they are similar. Specifically, for each $i, j \in \{m, m+1, \ldots, m+L-1\}$ and $i \neq j$, we assume there is an edge linking videos x_i and x_j if their cosine similarity is greater than $\epsilon_{\mathcal{S}}$. Then, we extract all connected components, for each sliding window. A connected component is defined as a subgraph where a path exists between every pair of videos, and no edges connect to videos outside the subgraph. Based on this definition, we assume that videos within a connected component are similar. The main idea is that each sliding window's graph will consist of one or more connected components, each representing videos with a different topic. By analyzing these connected components, we can understand the kind of videos recommended by the short-format video platform. Let $C = (\mathcal{V}, \mathcal{E})$ represent a connected component, where $\mathcal{V}$ is a set of similar videos and $\mathcal{E}$ is a set of edges. The order of a connected component is defined as $|\mathcal{V}|$ (i.e., the number of nodes in the connected component). Assuming multiple connected components exist within a sliding window, we denote them as $C_1, C_2, C_3, \ldots$ to represent distinct components.

To understand and analyze the different connected components within each sliding window, we use the LLM-generated keywords for each video (see Section 5.1). For a connected component C, we consider the keywords of its videos as a collection and count the frequencies of all keywords. We define *component keywords* as those whose frequencies are equal to $|\mathcal{V}|$, i.e., those possessed by each video in the component. An example for $L = 6$ is shown in Fig. 3. This approach is enabled by the standardized LLM-generated keywords, which offer consistent representations of identical content.

Taken together, using AI models to generate video embeddings, cosine similarity to assess the similarity of videos, sliding windows, and graph analysis, ClipMind allows us

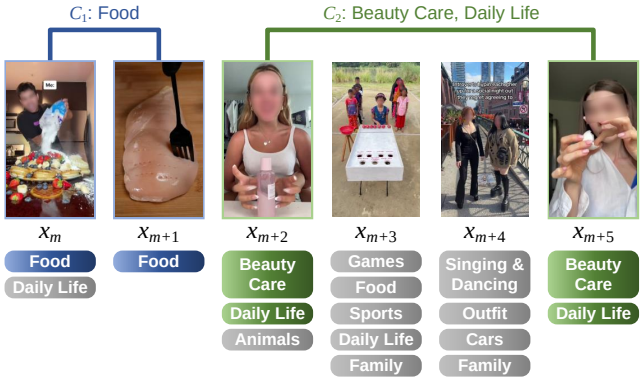


Figure 3: Example of connected components and their keywords ($L = 6$). There are two connected components C_1 and C_2. The component keyword of C_1 is "Food", while those of C_2 are "Beauty Care" and "Daily Life".

to analyze algorithmic video recommendations over time. We aim to analyze whether specific sliding windows have very large connected components, which would indicate that recommendation systems in short-format video platforms recommend many videos with similar topics. This can assist us in demystifying algorithmic recommendations and shed light on the existence of the phenomenon of informational rabbit holes in short-format video platforms.

6 Experiments

Here we present our analysis and results after running ClipMind on the collected TikTok video traces. All experiments were run on an Intel Xeon CPU and an NVIDIA A100 GPU.

6.1 Experimental Setup for TikTok

Generating Embeddings For videos in our traces, we prepare their features and the embeddings as reported in Section 5.1. We use the following preprocessing and prompts.

Filtering user-defined metadata. For every collected trace, we report the frequencies of the top 1% hashtags and the pool of meaningless ones in Appendix B.

Prompts for generating user-defined metadata and generating keywords. We set the parameters as follows:

> $\{platform\} = TikTok$
> $\{auxiliaries\} = channel\ tags$
> $\{keyword\ pool\} = Singing\ and\ Dancing, Comedy, Sports, Anime\ and\ Comics, Relationship, Shows, Lipsync, Daily\ Life, Beauty\ Care, Games, Society, Outfit, Cars, Food, Animals, Family, Drama, Fitness\ and\ Health, Education, Technology$

and instantiate the prompts. Here, without loss of generality, we set the value of $\{keyword\ pool\}$ with 20 official topics on the TikTok Explore page (TikTok 2024). This setup ensures a representative selection of widely recognized topics, providing a robust foundation for video similarity assessment and video sequence analysis. Moreover, *channel tags* used here are another type of metadata specific to short-format videos on TikTok. As about 50% of collected videos lack such information, we only use it as an auxiliary for generating keywords instead of as a feature. Examples of generating user-defined metadata and keywords are shown in Appendixes C and D.

Assessing Video Similarity We calculate the cosine similarity for all video pairs in the Random trace and present the cumulative distribution function (CDF) of these values in Appendix E, where $[l, u] = [0.4, 0.9]$. Setting $A = 5$ and $B = 100$, we obtain a dataset that consists of 500 video pairs. After completing the annotations, a quantitative assessment reveals that the two annotators agreed on 94% of the annotations with Cohen's Kappa coefficient of 0.80, indicating a substantial agreement. After a discussion between the annotators to resolve disagreements, the final dataset has 101 similar video pairs and 399 dissimilar ones. Further, we set $\Delta = 10^{-4}$ and report $\epsilon^*_{\mathcal{S}}$, $A_{\mathcal{S}}(\epsilon^*_{\mathcal{S}})$, $P_{\mathcal{S}}(\epsilon^*_{\mathcal{S}})$, $R_{\mathcal{S}}(\epsilon^*_{\mathcal{S}})$, $F_{\mathcal{S}}(\epsilon^*_{\mathcal{S}})$ for all $\mathcal{S} \in \mathcal{P}(\mathcal{F}) \setminus \emptyset$ in Table 2. We find that the best feature combination with the smallest dimensionality is the one that uses visual video content, user-defined metadata, and LLM-generated keywords. The optimal threshold is 0.7921, and our framework has a performance of 0.9580, 0.9255, 0.8614, and 0.8923 in terms of accuracy, precision, recall, and F_1 score, respectively.

Video Sequence Analysis To analyze sequences of videos, we use sliding windows of length 10 and 20 (i.e., set $L = 10$ and 20).

6.2 Results

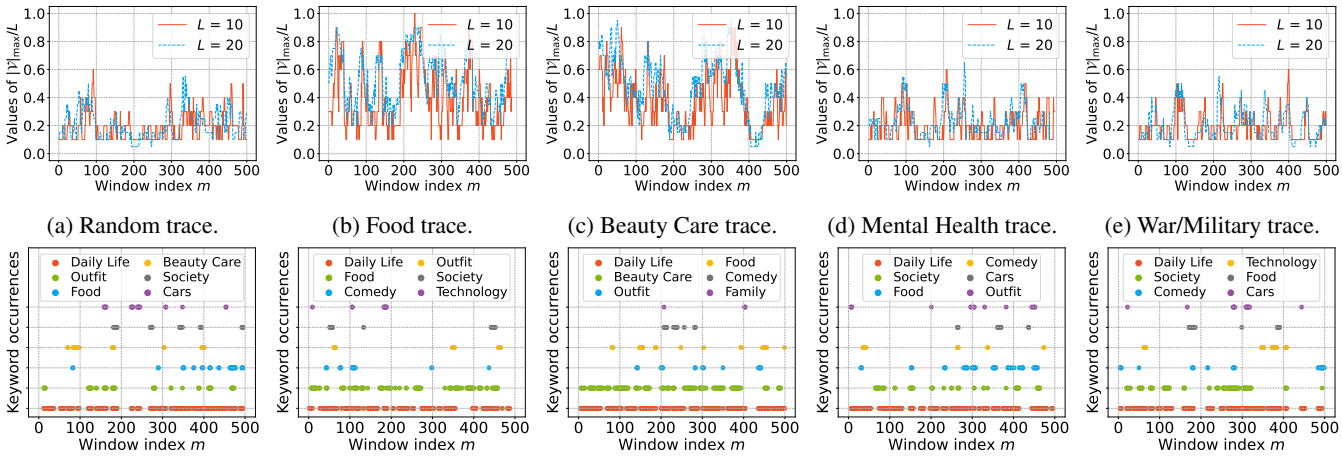
Here, we present and analyze results on the TikTok traces (see Section 4). Fig. 4a–4e show the values of $|\mathcal{V}|_{\max}/L$ for each window index m, where $|\mathcal{V}|_{\max}$ represents the order of the largest connected component (LCC) in a given window and L the size of the sliding window. The CDF of these values is shown in Fig. 5. To visualize the content of the LLCs in a trace, we define *window keywords* as component keywords belonging to the LLC in a given window. For each trace, we count the frequencies of all window keywords and focus on the top-6 frequent ones. Then, we show their occurrences in each window m when $L = 10$ in Fig. 4g–4j. Below, we analyze the results on the random trace, which serves as our baseline. Then, we delve into the results of the four topic-specific traces. For these traces, we further provide detailed information about the connected components in the randomly sampled windows in Appendix F.

Random trace. As shown in Fig. 4a and 5, for $L = 10$ and 20, $|\mathcal{V}|_{\max}/L \geq 0.5$ holds for 2% and 1% of windows. The averages of the values of $|\mathcal{V}|_{\max}/L$ are 0.19 and 0.20, which is expected given that the automated account used for data collection followed a random watching configuration and there is likely little user personalization. Such results will serve as our baseline for further comparison.

Food trace. As shown in Fig. 4b and 5, for $L = 10$ and 20, $|\mathcal{V}|_{\max}/L \geq 0.5$ holds for 36% and 56% of windows. The averages of the values of $|\mathcal{V}|_{\max}/L$ are 0.41 and 0.53. This means that, on average, between 41% and 53% of the videos within a window are part of a large connected component (i.e., they are similar). Compared with the Random trace, much more similar videos are recommended in the Food trace. Moreover, it is shown from Fig. 4g that the keyword

$\mathcal{S}$	$\epsilon^*_{\mathcal{S}}$	$A_{\mathcal{S}}(\epsilon^*_{\mathcal{S}})$	$P_{\mathcal{S}}(\epsilon^*_{\mathcal{S}})$	$R_{\mathcal{S}}(\epsilon^*_{\mathcal{S}})$	$F_{\mathcal{S}}(\epsilon^*_{\mathcal{S}})$
{visual content, user-defined metadata, keywords}	0.7921	0.9580	0.9255	0.8614	0.8923
{audio content, visual content, description, user-defined metadata, keywords}	0.8022	0.9580	0.9255	0.8614	0.8923
{audio content, visual content, description, user-defined metadata}	0.7465	0.9500	0.8519	0.9109	0.8804
{audio content, visual content, user-defined metadata, keywords}	0.7573	0.9500	0.8519	0.9109	0.8804
{visual content, description, user-defined metadata}	0.7739	0.9520	0.9231	0.8317	0.8750
{visual content, description, user-defined metadata, keywords}	0.8133	0.9520	0.9231	0.8317	0.8750
{visual content, description}	0.7725	0.9460	0.8558	0.8812	0.8683
{visual content, description, keywords}	0.8145	0.9460	0.8558	0.8812	0.8683
{audio content, visual content, user-defined metadata}	0.7129	0.9420	0.8462	0.8713	0.8585
{visual content, user-defined metadata}	0.7291	0.9460	0.9205	0.8020	0.8571
{visual content, keywords}	0.7893	0.9400	0.8257	0.8911	0.8571
{audio content, visual content, description, keywords}	0.7836	0.9380	0.8365	0.8614	0.8488
{visual content}	0.6950	0.9380	0.8571	0.8317	0.8442
{description, user-defined metadata, keywords}	0.8408	0.9380	0.8571	0.8317	0.8442
{description, user-defined metadata}	0.8092	0.9360	0.8632	0.8119	0.8367
{audio content, visual content, keywords}	0.7652	0.9320	0.8384	0.8218	0.8300
{audio content, visual content, description}	0.7412	0.9260	0.7909	0.8614	0.8246
{user-defined metadata, keywords}	0.8454	0.9280	0.9012	0.7228	0.8022
{user-defined metadata}	0.7596	0.9280	0.9114	0.7129	0.8000
{audio content, visual content}	0.6521	0.9020	0.7000	0.9010	0.7879
{audio content, description, user-defined metadata, keywords}	0.8081	0.8940	0.7034	0.8218	0.7580
{description, keywords}	0.8825	0.8840	0.6807	0.8020	0.7364
{audio content, user-defined metadata, keywords}	0.7741	0.8640	0.6138	0.8812	0.7236
{audio content, description, user-defined metadata}	0.7501	0.8580	0.5987	0.9010	0.7194
{description}	0.8626	0.8660	0.6574	0.7030	0.6794
{audio content, user-defined metadata}	0.7001	0.8380	0.5676	0.8317	0.6747
{audio content, description, keywords}	0.7810	0.8280	0.5460	0.8812	0.6742
{audio content, keywords}	0.7364	0.8040	0.5085	0.8911	0.6475
{keywords}	0.9222	0.8620	0.6702	0.6238	0.6462
{audio content, description}	0.7176	0.8020	0.5057	0.8713	0.6400
{audio content}	0.5500	0.7740	0.4691	0.9010	0.6169

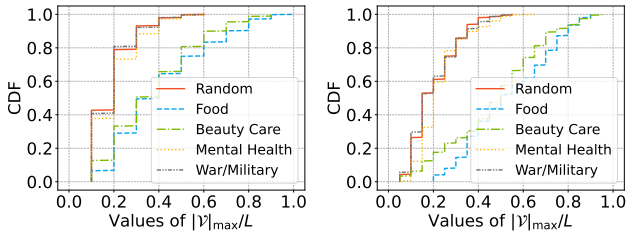
Table 2: The optimal thresholds and the corresponding four metrics of all the 31 combinations. For ease of readability, we abbreviate "video visual content" as "visual content", "video audio content" as "audio content", "LLM-generated description" as "description", and "LLM-generated keywords" as "keywords" in this table.



(a) Random trace. (b) Food trace. (c) Beauty Care trace. (d) Mental Health trace. (e) War/Military trace.

(f) Random trace. (g) Food trace. (h) Beauty Care trace. (i) Mental Health trace. (j) War/Military trace.

Figure 4: The values of $|\mathcal{V}|_{\max}/L$ for each window m when $L = 10, 20$ and the keyword occurrences in each m when $L = 10$.



(a) $L = 10$ (b) $L = 20$

Figure 5: CDF of the values of $|\mathcal{V}|_{\max}/L$ for the five traces.

"Food" occurs in 29% of windows. This indicates that, under the effect of the bootstrapping and recommendation system on TikTok, many food-related videos were continuously recommended. Then, among such windows, we randomly sample one and visualize it in Fig. 6. As shown in the figure, the window $w_{17,10}$ has one connected component C, and its component keyword is "Food". Here, our framework accurately detects similar videos, such as preparing food and beverages.

Beauty Care trace. As shown in Fig. 4c and 5, for $L = 10$ and 20, $|\mathcal{V}|_{\max}/L \geq 0.5$ holds for 34% and 53% of windows. The averages of the values of $|\mathcal{V}|_{\max}/L$ are 0.37 and 0.47. Moreover, it is shown from Fig. 4h that the keyword "Beauty Care" occurs in 37% of windows. We further conduct a case study. As shown in Fig. 7, the window $w_{163,10}$ has one connected component C, and its component keywords are "Beauty Care" and "Daily Life". Our framework accurately detects similar videos like shaving and makeup.

Mental Health trace. As shown in Fig. 4d and 5, for $L = 10$ and 20, $|\mathcal{V}|_{\max}/L \geq 0.5$ holds for 3% and 4% of windows. The averages of the values of $|\mathcal{V}|_{\max}/L$ are 0.20 and 0.23. Such results are comparable with those in the Random trace. This phenomenon can be attributed to two possible reasons. First, mental health is a niche topic, i.e., the number of videos related to it is much fewer than that of food and beauty care. Second, the system is sensitive to extreme content, and only a few videos are recommended, even if users show interest. It is shown from Fig. 4i that the keyword "Society" occurs in 14% of windows. Similar to before, we conduct a case study. As shown in Fig. 8, the window $w_{70,10}$ has two connected components C_1 and C_2, and their component keywords are "Drama", "Society" and "Society", "Education". Here, our framework accurately detects similar videos related to both negative social events and mental health self-assessments.

War/Military trace. As shown in Fig. 4e and 5, for $L = 10$ and 20, $|\mathcal{V}|_{\max}/L \geq 0.5$ holds for 2% and 1% of windows. The averages of the values of $|\mathcal{V}|_{\max}/L$ are 0.19 and 0.20. Such results confirm the characteristic of extreme content and how the system deals with it again. It is shown from Fig. 4j that the keyword "Society" occurs in 24% of windows. Similarly, we conduct a case study. As shown in Fig. 9, the window $w_{54,10}$ has one connected component C, and its component keywords are "Society" and "Daily Life". Our framework accurately detects similar videos about soldiers standing on duty in this case.

Also, we introduce a metric to evaluate how effectively recommendation systems balance the promotion of mainstream topics against the suppression of potentially problematic topics. Specifically, we define the *balance measurement* β as the ratio of the average values of $|\mathcal{V}_{\max}|/L$ between two arbitrary traces. When comparing a topic-specific trace to the Random trace, a higher β value signifies stronger promotion of the topic, whereas a lower β value indicates greater suppression. The pairwise balance measurements for all considered traces are summarized in Table 3. Notably, we observe that, under identical bootstrapping configurations, the recommendation algorithm promotes topics such as Food and Beauty Care with $\beta \approx 2$, while suppressing topics like Mental Health and War/Military with $\beta \approx 1$. Moreover, both Food and Beauty Care have similar β values, indicating a comparable degree of promotion across these topics, whereas Mental Health and War/Military exhibit consistently lower β values, reflecting similar levels of suppression. This metric can help evaluate the implicit biases in recommendation algorithms, assisting in developing more balanced recommendation systems.

Discussions. Our results show that when an account is bootstrapped with mainstream topics like Food and Beauty Care, the recommendation algorithm recommends many similar videos. On the other hand, when an account shows interest in niche and potentially harmful topics like Mental Health and War/Military content, the algorithm only recommends to a few windows many videos that are similar. Our TikTok evaluation shows no evidence of the algorithm driving people towards informational rabbit holes of extreme content. On the other hand, for mainstream topics, the algorithm recommends many similar videos with substantially less diversity of content compared to the other traces. Such results show that ClipMind can empower researchers to conduct unbiased audits and produce insights into recommendation algorithms that are employed by popular social media platforms like TikTok, YouTube, and Instagram. Moreover, regulators can leverage our framework to assess the compliance of social media platforms with current regulations like the Digital Service Act (European Commission 2023), which emphasizes the need to conduct audits to assess potential harm that may arise from the use of recommendation algorithms. In addition, social media platforms can employ our framework to improve their recommendation systems. First, they can address inter-topic biases by analyzing video sequences across various topics to uncover discrepancies in algorithmic behavior. Second, they can mitigate intra-topic biases by evaluating component keywords across sliding windows within a video sequence. Third, they can adjust platform-user interactions by quantifying the influence of engagement patterns on generated recommendations. For example, this can involve examining the relationship between hashtag pool configurations and the proposed metrics. Finally, end-users can obtain a copy of their personal data from platforms like TikTok (e.g., following the method from Zannettou et al. (2024)) and apply our framework to understand how their content is curated and recommended. Specifically, end-users can control their engagement or adapt to their preferences by modifying the hashtag pools or the keyword pools: the former enables them to showcase different interests during data collection, while the latter adjusts topic granularity during similarity

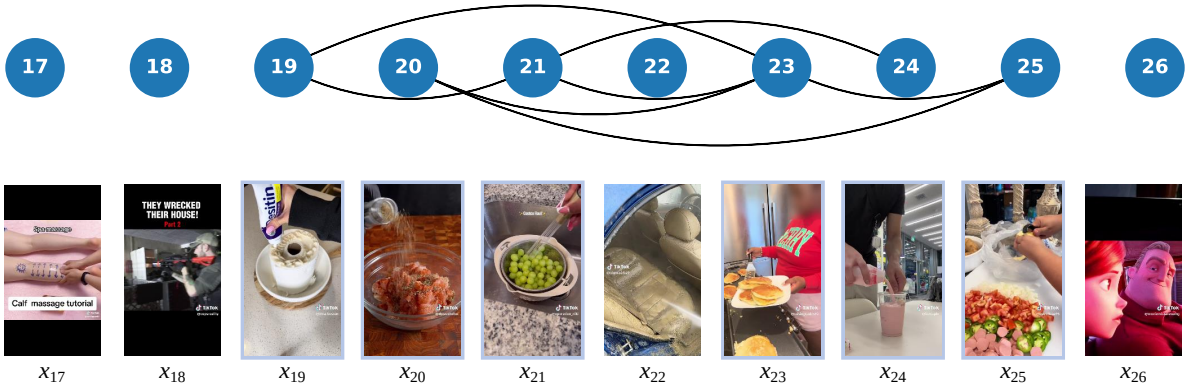


Figure 6: The visualization for the window $w_{17,10}$ in the Food trace.

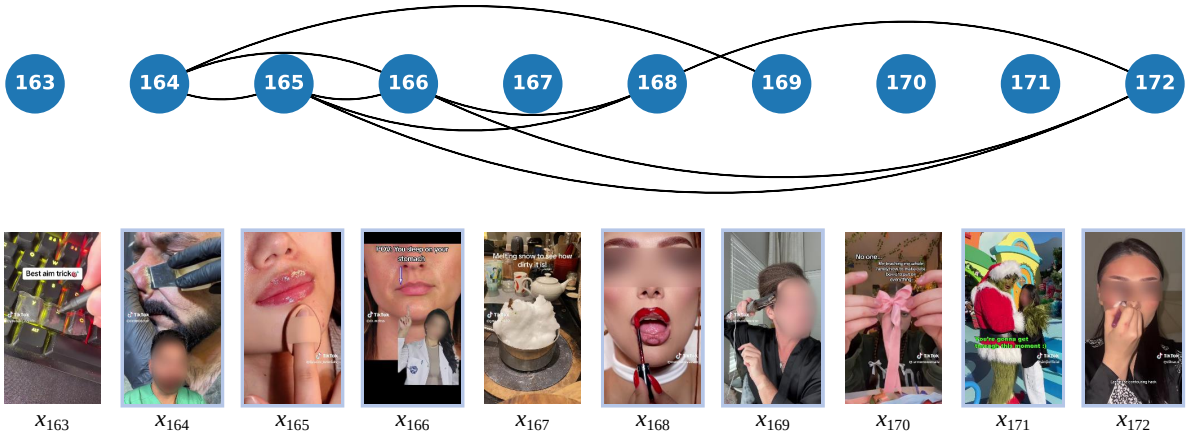


Figure 7: The visualization for the window $w_{163,10}$ in the Beauty Care trace.

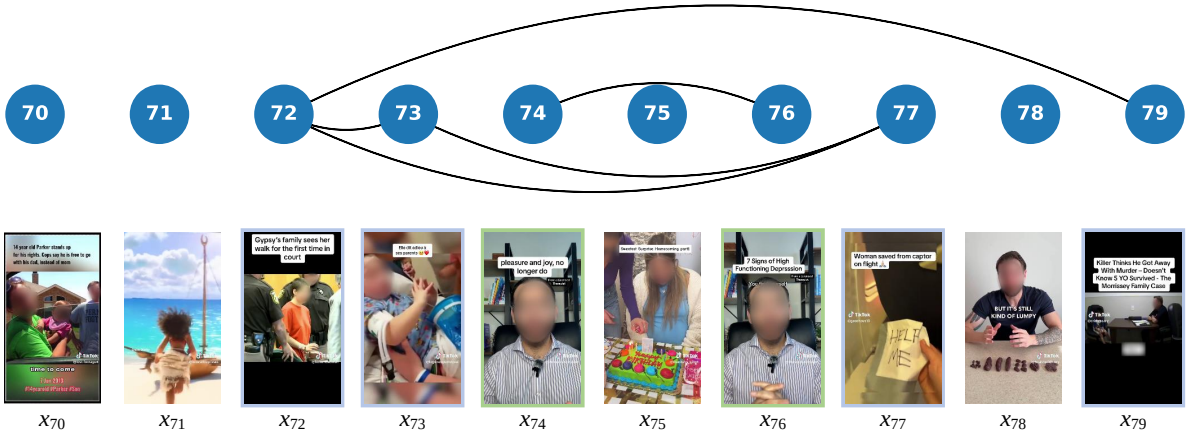
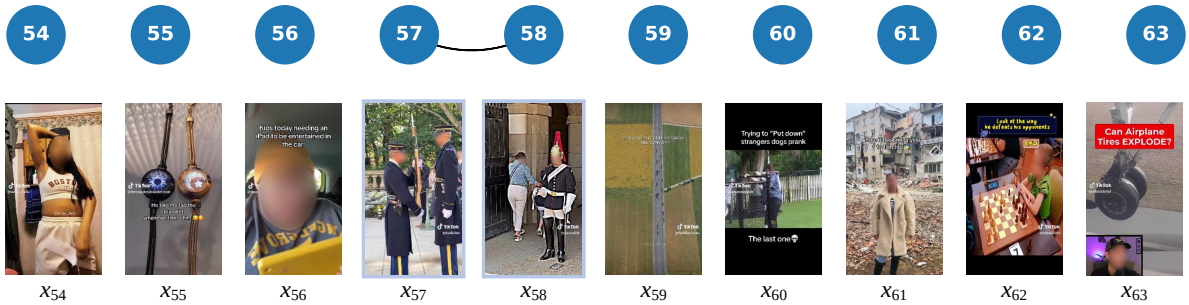


Figure 8: The visualization for the window $w_{70,10}$ in the Mental Health trace.

assessment. They can further visualize generated recommendations, as shown in Figs. 6 to 9, to validate the alignment between their interests and platforms' recommendations. Such insights are paramount, as they will allow end-users to assess the extent of personalization and help them in adjusting their personalization settings and user experience.

7 Conclusion

This paper presented ClipMind, a framework for auditing AI-powered recommendation systems on short-format video platforms. ClipMind is highly scalable because it eliminates the need for manual annotations to analyze video content. It leverages SOTA AI models to generate embeddings for short-format videos, requiring only a few seconds to produce



x_{54} x_{55} x_{56} x_{57} x_{58} x_{59} x_{60} x_{61} x_{62} x_{63}

Figure 9: The visualization for the window $w_{54,10}$ in the War/Military trace.

Trace 1 \ Trace 2	**Random**	**Food**	**Beauty Care**	**Mental Health**	**War/Military**
Random	–	0.4629	0.5030	0.9183	0.9943
Food	2.1602	–	1.0865	1.9836	2.1478
Beauty Care	1.9881	0.9204	–	1.8256	1.9768
Mental Health	1.0890	0.5041	0.5478	–	1.0828
War/Military	1.0058	0.4656	0.5059	0.9235	–

Table 3: The pairwise balance measurements for all considered traces, where each value represents the ratio of trace 1 to trace 2.

all necessary embeddings for a single video. Additionally, it incorporates an automated system with temporal and graph analysis to understand video recommendations. While ClipMind partially relies on closed-source AI models chosen for their superior performance (as of January 2024), it stands to grow more cost-effective and powerful with ongoing AI developments. ClipMind is also highly generalizable as it is capable of adapting to other short-format video platforms through simple parameter adjustments. Moreover, it can be customized for diverse applications by tailoring the keyword pool or integrating fine-tuned AI models to align with specific user preferences. Its modular design further ensures smooth integration with evolving AI technologies, making it well-suited for future enhancements. We demonstrated the applicability of ClipMind on TikTok, finding little evidence of the algorithm steering users towards information rabbit holes of extreme content. We believe ClipMind can contribute to efforts to use AI for social good. We argue that our framework can be used by the research community, regulators, social media platforms, and users to understand and audit algorithmic video recommendations on short-format video platforms like TikTok, YouTube Shorts, and Instagram Reels.

Limitations and Future Work. Naturally, our work has some limitations. First, we demonstrated the applicability of our framework only on TikTok. However, we anticipate that our framework is widely applicable to all platforms, assuming that the content can be modeled as a sequence of video recommendations. As part of our future work, we plan to expand our research on other short-format video platforms like YouTube Shorts and Instagram Reels. This is mainly because of the use of powerful, generalizable, and SOTA multimodal AI models that are able to capture the semantics of video content. Second, our analysis focuses on a small number of topics. In our future work, we plan to investigate the prevalence of information rabbit holes on a larger set of topics, such as political content and entertainment. Also, our analysis focuses only on a single snapshot of the TikTok algorithm. Future work is needed to conduct longitudinal audits of recommendations online, as these algorithms are dynamic and evolve over time. Third, our analysis did not account for the measurement of biases or the evaluation of diversity in recommended content. In our future work, we plan to customize the keyword pool by incorporating detailed categories and topics within a specific domain. For example, analyzing various cuisines and culinary traditions in the Food domain. Finally, we acknowledge that ClipMind partially relies on closed-source AI models, incurring a cost of $30 to analyze all the traces in this paper. Such models might not be an option (e.g., due to the cost). In our future work, we plan to develop OpenClipMind, a variant of our framework that will exclusively rely on open-source AI models.

Ethics Statement

Our work relies solely on publicly available datasets obtained from the TikTok platform. We do not anticipate any potential harm arising from our work. At the same time, we believe that our work has a broader impact and can benefit our society in multiple ways. First, our framework can assist researchers and regulators in auditing algorithmic recommendations online. Such audits can pressure social media operators to improve their algorithms to ensure that their algorithms and services have no adverse effects on people. Second, our framework can be used by end-users to analyze and understand their social media feeds. In particular, users can obtain their social media activity by requesting their data from the platform and then using our framework to analyze what kind

of videos are recommended online. This can assist the users in identifying signs of potentially harmful recommendations. Overall, we believe that our work's benefits outweigh any potential harm that may arise from this work (the authors do not foresee any potential harm).

Acknowledgements

This work was partially funded by an unrestricted gift from Google.

References

Ali, M.; Goetzen, A.; Mislove, A.; Redmiles, E. M.; and Sapiezynski, P. 2023. Problematic advertising and its disparate exposure on Facebook. *arXiv preprint arXiv:2306.06052*.

Badola, P. 2023. *Russia and Ukraine: A content analysis of "the world's first TikTok war"*. Ph.D. thesis.

Bandy, J. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *CSCW*.

Barde, B. V.; and Bainwad, A. M. 2017. An overview of topic modeling methods and tools. In *IEEE ICICCS*.

Boeker, M.; and Urman, A. 2022. An empirical investigation of personalization factors on TikTok. In *ACM Web Conference 2022*, 2298–2309.

Braghieri, L.; Levy, R.; and Makarin, A. 2022. Social media and mental health. *American Economic Review*.

Burgess, S. 2023. Andrew Tate: Controversial influencer pushed on to 'teen's' YouTube Shorts and Instagram video feeds. https://news.sky.com/story/andrew-tate-controversial-influencer-pushed-onto-teens-youtube-shorts-and-instagram-video-feeds-12849572. Accessed: 2025-04-03.

European Commission. 2023. The Digital Services Act package. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package. Accessed: 2025-04-03.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. ImageBind: One Embedding Space To Bind Them All. In *CVPR*.

Greene, R.; Sanders, T.; Weng, L.; and Neelakantan, A. 2022. New and improved embedding model. https://openai.com/blog/new-and-improved-embedding-model. Accessed: 2025-04-03.

Hussein, E.; Juneja, P.; and Mitra, T. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. *CSCW*.

Jasser, G.; McSwiney, J.; Pertwee, E.; and Zannettou, S. 2023. 'Welcome to #GabFam': Far-right virtual community on Gab. *New Media & Society*, 25(7): 1728–1745.

Klug, D.; Qin, Y.; Evans, M.; and Kaufman, G. 2021. Trick and please. A mixed-method study on user assumptions about the TikTok algorithm. In *WebSci*, 84–92.

Ledwich, M.; and Zaitsev, A. 2019. Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*.

Liaropoulos, A. N. 2023. Victory and virality: War in the age of social media. *Georgetown Journal of International Affairs*.

Ling, C.; Gummadi, K. P.; and Zannettou, S. 2023. "Learn the facts about COVID-19": Analyzing the use of warning labels on TikTok videos. In *ICWSM*, volume 17, 554–565.

Meta. 2019. Powered by AI: Instagram's Explore recommender system. https://ai.meta.com/blog/powered-by-ai-instagrams-explore-recommender-system/. Accessed: 2025-04-03.

Microsoft. 2024. Playwright: A framework for Web Testing and Automation. https://github.com/microsoft/playwright. Accessed: 2025-04-03.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Papadamou, K.; Papasavva, A.; Zannettou, S.; Blackburn, J.; Kourtellis, N.; Leontiadis, I.; Stringhini, G.; and Sirivianos, M. 2020. Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. In *ICWSM*.

Papadamou, K.; Zannettou, S.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Sirivianos, M. 2022. "It is just a flu": Assessing the effect of watch history on YouTube's pseudoscientific video recommendations. In *ICWSM*.

Paudel, P.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2023. LAMBRETTA: Learning to rank for Twitter soft moderation. In *IEEE S&P*.

Pilař, L.; Kvasničková Stanislavská, L.; Kvasnička, R.; Bouda, P.; and Pitrova, J. 2021. Framework for social media analysis based on hashtag research. *Applied Sciences*.

Pretorius, C.; McCashin, D.; and Coyle, D. 2022. Mental health professionals as influencers on TikTok and Instagram: What role do they play in mental health literacy and help-seeking? *Internet interventions*, 30: 100591.

Ribeiro, M. H.; Ottoni, R.; West, R.; Almeida, V. A.; and Meira Jr, W. 2020. Auditing radicalization pathways on YouTube. In *FAccT*, 131–141.

Shahbaznezhad, H.; Dolan, R.; and Rashidirad, M. 2021. The role of social media content format and platform in users' engagement behavior. *Journal of Interactive Marketing*.

Shang, L.; Kou, Z.; Zhang, Y.; and Wang, D. 2021. A multimodal misinformation detector for COVID-19 short videos on TikTok. In *IEEE Big Data*.

Shen, X.; He, X.; Backes, M.; Blackburn, J.; Zannettou, S.; and Zhang, Y. 2022. On Xing Tian and the perseverance of anti-China sentiment online. In *ICWSM*.

Song, S.; Xue, X.; Zhao, Y. C.; Li, J.; Zhu, Q.; and Zhao, M. 2021. Short-video apps as a health information source for chronic obstructive pulmonary disease: Information quality assessment of TikTok videos. *Journal of medical Internet research*, 23(12): e28318.

TikTok. 2020. How TikTok recommends videos #ForYou. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you. Accessed: 2025-04-03.

TikTok. 2024. TikTok Explore page. https://www.tiktok.com/explore. Accessed: 2025-04-03.

Weimann, G.; and Masri, N. 2023. Research note: Spreading hate on TikTok. *Studies in conflict & terrorism*.

WSJ Staff. 2021. Inside TikTok's algorithm: A WSJ video investigation. https://www.wsj.com/articles/tiktok-algorithm-video-investigation-11626877477. Accessed: 2025-04-03.

Zannettou, S.; Nemes-Nemeth, O.; Ayalon, O.; Goetzen, A.; Gummadi, K. P.; Redmiles, E. M.; and Roesner, F. 2024. Analyzing User Engagement with TikTok's Short Format Video Recommendations using Data Donations. In *CHI*.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858*.

A Topic-Specific Traces Hashtags

This section contains the definitions of the four topics and details on how the hashtag pool of each topic-specific trace was constructed. Table 4 contains the hashtag pool of each trace. In each pool, the hashtags that were used for the "bootstrapping" phase are written in the boldface.

A.1 Topic Definitions

The definitions of the four topics are presented as follows:

- The Food trace involves content related to cooking, dining, and culinary culture, including meal preparation, food reviews, recipes, and festive or cultural food traditions.
- The Beauty Care trace focuses on personal grooming, makeup, skincare, haircare, and nail art, emphasizing beauty trends, techniques, products, and self-care routines to enhance personal aesthetics.
- The Mental Health trace includes emotional well-being, mental health challenges, and coping strategies, highlighting topics such as anxiety, depression, trauma, and the importance of awareness, support, therapy.
- The War/Military trace covers content related to armed conflicts, military and police forces, and defense, emphasizing historical wars, modern warfare, military personnel, and geopolitical dynamics.

These definitions were based on either the types of content mainly featured in TikTok's Explore page or the key aspects informed by recent research concerns.

A.2 Food and Beauty Care traces

For each of the Food and Beauty Care traces, we create the hashtag pool by collecting 40 most-trending hashtags of that specific topic in the USA over the last 120 days timeframe. We consider these hashtags because they reflect current user engagement trends and audience interests, providing the latest relevant representation of popular content within that topic. These trending hashtags, as measured by TikTok, are available on https://ads.tiktok.com/business/creativecenter/inspiration/popular/hashtag/. We collected these trending hashtags in January 2024. Next, we choose the 5 most-trending hashtags among the hashtag pool to use in the "bootstrapping" phase.

A.3 Mental Health and War/Military traces

For each of the Mental Health and War/Military traces, we first choose 5 common hashtags related to that specific topic and use them to initialize a hashtag pool, which are written in boldface in Table 4. We determine these 5 common hashtags by beginning with the hashtags representing the definition of the topic itself, such as #mentalhealth and #war, because they are foundational to the topic and are likely to co-exist with other related hashtags during subsequent search-and-expansion procedure. Then, we search the related hashtags of each existing hashtag in the pool and expand the pool by adding such related hashtags. To search the related hashtags of a specific hashtag, we scrape its corresponding webpage on https://ads.tiktok.com/. For instance, the webpage of #burger is https://ads.tiktok.com/business/creativecenter/hashtag/burger and its related hashtags can be found in "Recommended for you". This approach provides a data-driven way to identify algorithmically related hashtags, ensuring relevance and alignment with current TikTok trends. We iterate this procedure four times and, in the end, we manually refine the pool and keep 40 collected hashtags. We collected the hashtags in January 2024. Next, we choose the 5 common hashtags to use in the "bootstrapping" phase.

B Filtering Hashtags

For the five traces, the frequencies of the top 1% hashtags are shown in Fig. 10. For the five traces, we further report the pools of meaningless hashtags in Table 5.

C Examples of Generating Metadata

Two examples of generating metadata are shown in Fig. 12.

D Examples of Generating Keywords

Two examples of generating keywords are shown in Fig. 13.

E Video Similarity

The CDF of the cosine similarity of all video pairs in the Random trace is shown in Fig. 11.

F Connected Components

Detailed information about the connected components in the randomly sampled windows for the four topic-specific traces is provided below.

Food trace. As shown in Fig. 6, the window $w_{17,10}$ has one connected component $C_1 = (\mathcal{V}_1, \mathcal{E}_1)$ where

$$\begin{aligned}\mathcal{V}_1 &= \{x_{19}, x_{20}, x_{21}, x_{23}, x_{24}, x_{25}\},\\ \mathcal{E}_1 &= \{(x_{19}, x_{21}), (x_{19}, x_{23}), (x_{20}, x_{23}),\\ &\qquad (x_{20}, x_{25}), (x_{21}, x_{23}), (x_{21}, x_{24}), (x_{23}, x_{25})\}.\end{aligned}$$

The component keyword of C_1 is "Food".

Beauty Care trace. As shown in Fig. 7, the window $w_{163,10}$ has one connected component $C_1 = (\mathcal{V}_1, \mathcal{E}_1)$ where

$$\begin{aligned}\mathcal{V}_1 &= \{x_{164}, x_{165}, x_{166}, x_{168}, x_{169}, x_{172}\},\\ \mathcal{E}_1 &= \{(x_{164}, x_{165}), (x_{164}, x_{166}), (x_{164}, x_{169}),\\ &\qquad (x_{165}, x_{166}), (x_{165}, x_{168}), (x_{165}, x_{172}),\\ &\qquad (x_{166}, x_{168}), (x_{166}, x_{172}), (x_{168}, x_{172})\}.\end{aligned}$$

The component keywords of C_1 are "Beauty Care" and "Daily Life".

Mental Health trace. As shown in Fig. 8, the window $w_{70,10}$ has two connected components $C_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $C_2 = (\mathcal{V}_2, \mathcal{E}_2)$ where

$$\begin{aligned}\mathcal{V}_1 &= \{x_{72}, x_{73}, x_{77}, x_{79}\},\\ \mathcal{E}_1 &= \{(x_{72}, x_{73}), (x_{72}, x_{77}), (x_{72}, x_{79}), (x_{73}, x_{77})\},\\ \mathcal{V}_2 &= \{x_{74}, x_{76}\},\ \mathcal{E}_2 = \{(x_{74}, x_{76})\}.\end{aligned}$$

The component keywords of C_1 are "Society" and "Drama", while those of C_2 are "Society" and "Education".

War/Military trace. As shown in Fig. 9, the window $w_{54,10}$ has one connected component $C_1 = (\mathcal{V}_1, \mathcal{E}_1)$ where

$$\mathcal{V}_1 = \{x_{57}, x_{58}\},\ \mathcal{E} = \{(x_{57}, x_{58})\}.$$

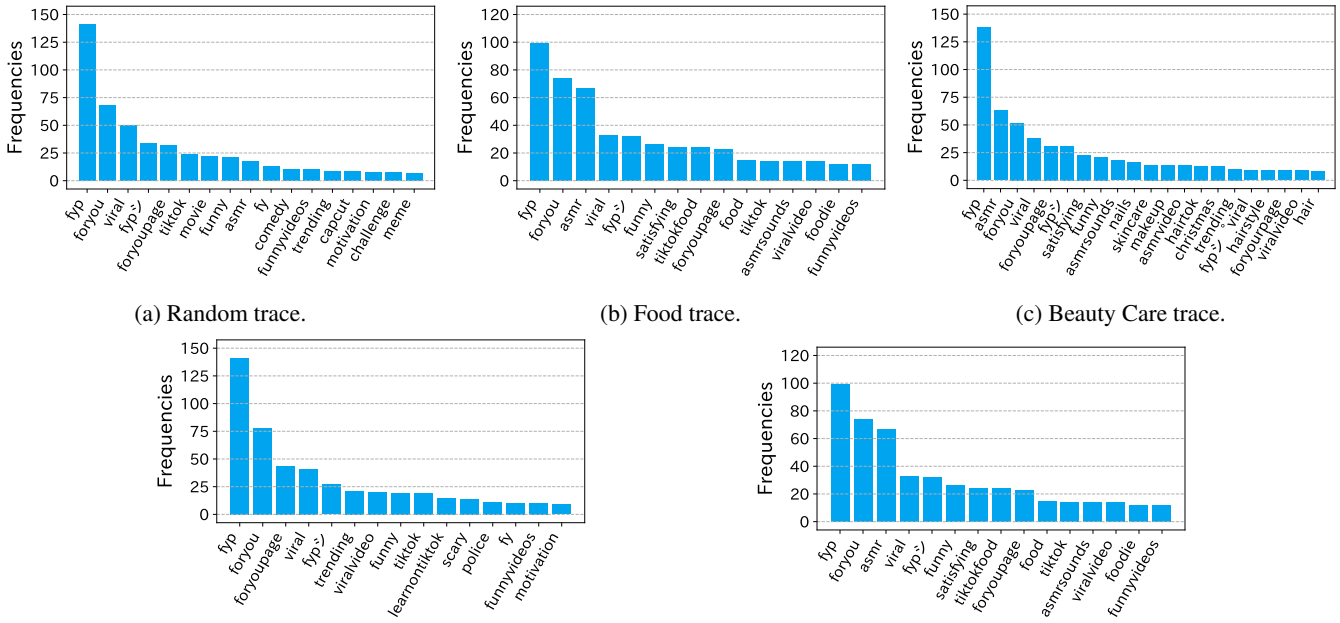
The component keywords of C_1 are "Society" and "Daily Life".

Topic-Specific Trace	Hashtag Pool
Food	**#chef**, **#foodreview**, **#restaurant**, **#orderpacking**, **#mealprep**, #candyboxsubscription, #welovecandy, #thanksgivingvibes, #sushi, #healthyfood, #netherlands, #tastetest, #thanksgivingdinner, #homecooking, #drinking, #ramen, #drpepper, #tiktokmademebuylt, #coffeetok, #thanksgiving2023, #canadia, #chipotle, #baker, #whiskey, #groceryhaul, #birthdaycake, #recipesoftiktok, #rollerskates, #coffeetime, #sydneyaustralia, #hotchocolate, #coffeelover, #keithlee, #tamales, #gingerbreadhouse, #koreanfood, #streetfood, #foodietiktok, #snackreview, #rollerskater
Beauty Care	**#lashes**, **#barber**, **#nailart**, **#nailtech**, **#sephora**, #lashextensions, #nailinspo, #mua, #haircare, #perfume, #acrylicnails, #knotlessbraids, #nailsartvideos, #fragrance, #pressonnails, #glitter, #lipgloss, #lashtech, #perfumetiktok, #christmasnails, #beginnernailtech, #wiginfluencer, #barberlife, #ulta, #lipstick, #hairextensions, #silkpress, #fragrancetiktok, #protectivestyles, #gelnails, #nailtutorial, #drunkelephant, #facial, #cleangirl, #selfcareroutine, #braidstyles, #blowout, #glowrecipe, #sephorahaul, #naturalhairstyles,
Mental Health	**#mentalhealth**, **#cry**, **#anxiety**, **#mentalhealthmatters**, **#mentalhealthillness**, #depressed, #believe, #adhd, #addiction, #stress, #consciousness, #anxietyrelief, #heart, #ocd, #die, #think, #hurt, #heartbroken, #sadquotes, #awakening, #ptsd, #stressrelief, #sadedits, #trauma, #mentalhealthawareness, #anxietydisorder, #sadtok, #sadtiktok, #grief, #grieftok, #griefandloss, #griefjourney, #chidloss, #babyloss, #depressionanxiety, #grievingmom, #therapy, #therapist, #painhub, #crying,
War/Military	**#war**, **#ww1**, **#ww2**, **#military**, **#soldier**, #putin, #recon, #f22, #russie, #russia, #nato, #zelensky, #usaf, #moscow, #fighterjet, #marines, #soldiers, #usairforce, #usmilitary, #ussoldier, #armyedit, #marinedrillinstructor, #airforce, #f35, #f16, #spaceforce, #militarygf, #marinedrill, #usarmy, #ww3, #infantry, #specialforces, #navy, #ucrania, #militarytiktok, #ukraine, #ukrainewar, #militarywife, #troops, #army,

Table 4: Sets of hashtags that were used to collect topic-specific traces. The hashtags in bold were used for the "bootstrapping" phase.

Trace	Hashtag Pool
Random	#foryoupage, #trending, #foryou, #tiktok, #capcut, #viral, #fypシ, #fyp, #fy
Food	#foryoupage, #viralvideo, #foryou, #tiktok, #viral, #fypシ, #fyp
Beauty Care	#foryourpage, #foryoupage, #fypシviral, #viralvideo, #trending, #foryou, #viral, #fypシ, #fyp
Mental Health	#foryoupage, #viralvideo, #trending, #foryou, #tiktok, #viral, #fypシ, #fyp, #fy
War/Military	#foryourpage, #foryoupage, #viralvideo, #fypシviral, #trending, #foryou, #tiktok, #viral, #fypシ, #fyp, #fy

Table 5: The pools of meaningless hashtags for the five traces.



(a) Random trace.

(b) Food trace.

(c) Beauty Care trace.

(d) Mental Health trace.

(e) War/Military trace.

Figure 10: The frequencies of top 1% hashtags for the five traces.

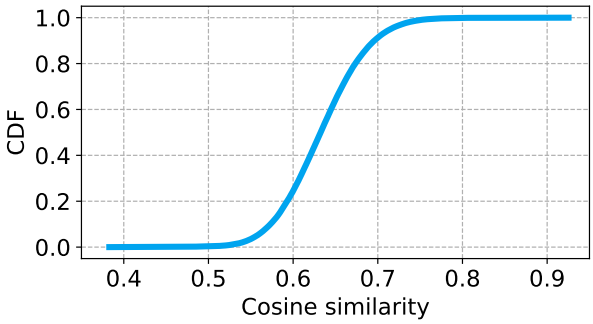
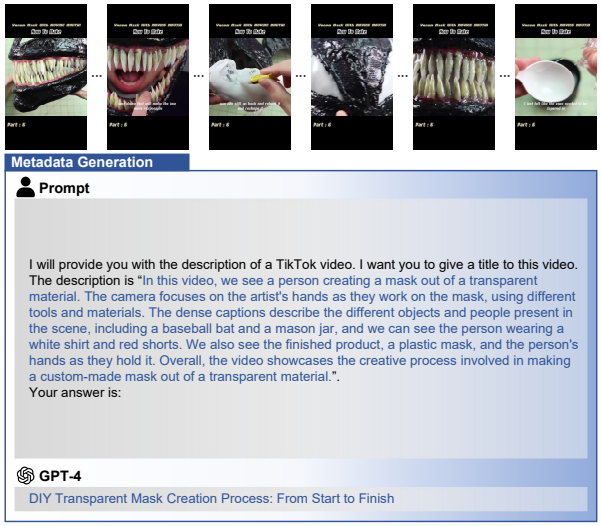
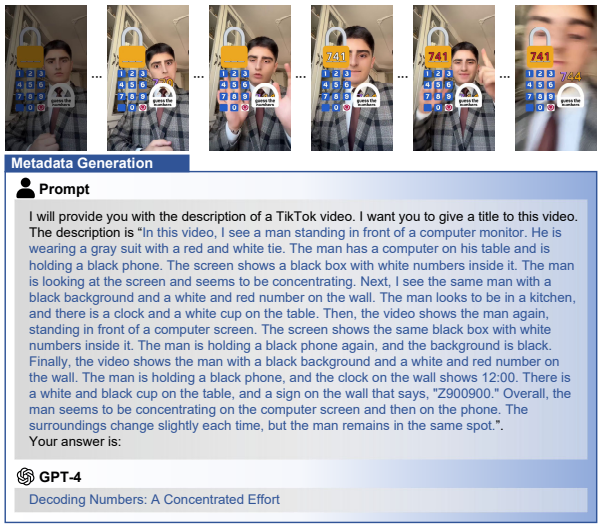


Figure 11: The CDF of the cosine similarity of all video pairs in the Random trace.

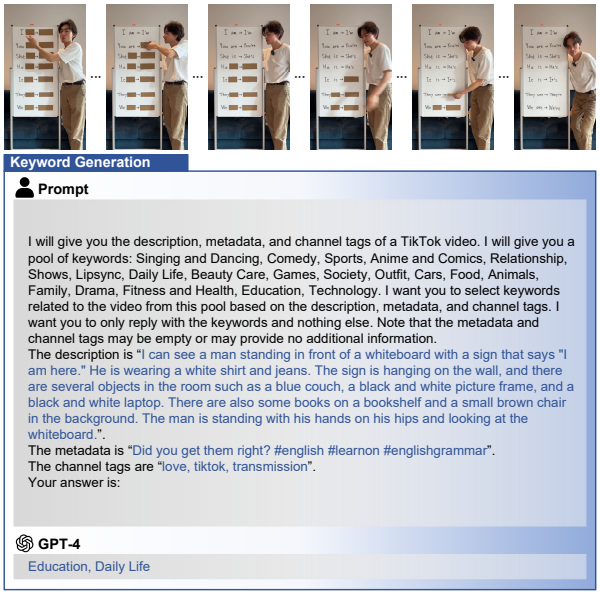


(a)



(b)

Figure 12: Examples of generating user-defined metadata using GPT-4 and LLM-generated descriptions.



(a)

(b)

Figure 13: Examples of generating keywords using GPT-4, LLM-generated descriptions, user-defined metadata, and channel tags.