



Latest updates: <https://dl.acm.org/doi/10.1145/3715336.3735762>

RESEARCH-ARTICLE

## ClassComet: Exploring and Designing AI-generated Danmaku in Educational Videos to Enhance Online Learning

**ZIPENG JI**, David R. Cheriton School of Computer Science, Waterloo, ON, Canada

**PENGCHENG AN**, Southern University of Science and Technology, Shenzhen, Guangdong, China

**JIAN ZHAO**, University of Waterloo, Waterloo, ON, Canada

**Open Access Support** provided by:

**University of Waterloo**

**Southern University of Science and Technology**

**David R. Cheriton School of Computer Science**



PDF Download  
3715336.3735762.pdf  
07 February 2026  
Total Citations: 2  
Total Downloads: 1791

Published: 05 July 2025

Citation in BibTeX format

DIS '25: Designing Interactive Systems Conference

July 5 - 9, 2025  
Madeira, Portugal

Conference Sponsors:  
**SIGCHI**

# ClassComet: Exploring and Designing AI-generated Danmaku in Educational Videos to Enhance Online Learning

Zipeng Ji

The David R. Cheriton School of  
Computer Science  
University of Waterloo  
Waterloo, Ontario, Canada  
c35ji@uwaterloo.ca

Pengcheng An

School of Design  
Southern University of Science and  
Technology  
Shenzhen, Guangdong, China  
anpc@sustech.edu.cn

Jian Zhao

School of Computer Science  
University of Waterloo  
Waterloo, Ontario, Canada  
jianzhao@uwaterloo.ca

## Abstract

Danmaku, users' live comments synchronized with, and overlaying on videos, has recently shown potential in promoting online video-based learning. However, user-generated danmaku can be scarce—especially in newer or less viewed videos—and its quality is unpredictable, limiting its educational impact. This paper explores how large multimodal models (LMM) can be leveraged to automatically generate effective, high-quality danmaku. We first conducted a formative study to identify the desirable characteristics of content- and emotion-related danmaku in educational videos. Based on the obtained insights, we developed ClassComet, an educational video platform with novel LMM-driven techniques for generating relevant types of danmaku to enhance video-based learning. Through user studies, we examined the quality of generated danmaku and their influence on learning experiences. The results indicate that our generated danmaku is comparable to human-created ones, and videos with both content- and emotion-related danmaku showed significant improvement in viewers' engagement and learning outcome.

## CCS Concepts

- Human-centered computing → Interaction techniques;
- Applied computing → E-learning;
- Computing methodologies → Natural language processing; Computer vision.

## Keywords

Danmaku, Video-based Learning, Automated Live Video Comments, Learner Engagement

## ACM Reference Format:

Zipeng Ji, Pengcheng An, and Jian Zhao. 2025. ClassComet: Exploring and Designing AI-generated Danmaku in Educational Videos to Enhance Online Learning. In *Designing Interactive Systems Conference (DIS '25), July 05–09, 2025, Funchal, Portugal*. ACM, New York, NY, USA, 24 pages. <https://doi.org/10.1145/3715336.3735762>

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*DIS '25, Funchal, Portugal*

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-1485-6/25/07  
<https://doi.org/10.1145/3715336.3735762>

## 1 Introduction

Online video-based learning has become increasingly prevalent recently; educational video platforms such as MOOCs have demonstrated significant advantages to support remote asynchronous learning at scale during the COVID-19 pandemic [19, 39, 78, 97]. Learners now frequently acquire knowledge through multiple video-based media such as YouTube, Udemy, and Coursera [44, 81, 107]. However, educational videos naturally have pedagogical limitations due to their passive format of knowledge delivery, diminishing learners' learning outcomes because of less engagement [16, 78]. Many platforms thus incorporated interactive features like commenting and quizzes [13, 51, 60]. Along this line, recent research has recognized the potential of *danmaku*, a video comment feature, that supports both lecturer-led and peer interactions, by analyzing or moderating human-created danmaku to boost learners' engagement [13, 35, 89, 90, 96]. Featuring live and synchronized comments that scroll across the video screen, danmaku creates an interactive and communal learning experience for viewers. Users can not only view existing danmaku embedded in the videos, but also are encouraged to add danmaku to share their own insights and interact with other danmaku and the video content itself [31, 33, 101]. Previous research has explored the role of danmaku in the learning contexts and measured the danmaku's potential in enhancing learners' social and cognitive presence [11, 12, 96, 98]. Unlike the previous experience of watching video content alone, where viewers receive information individually without peer interaction, interacting with danmaku allows viewers to send danmaku and engage with existing scrolling ones, transforming the learning experience from individual study to a collaborative environment.

However, user-generated danmaku has undeniable challenges in its quality and quantity for pedagogical goals; extensive off-topic, negative, and chaotic danmaku may overlap with useful danmaku, distracting learners' attention and affecting their learning experience [57, 61]. Some works attempted to moderate harmful or inappropriate danmaku [34, 35]. Nonetheless, this does not solve the key issue that there is a lack of sufficient useful danmaku to foster learning in educational videos (especially the newly uploaded or less viewed ones), where the quantity of danmaku is also significantly smaller compared to entertainment videos [90]. Therefore, the learning experience heavily depends on the quality and quantity of existing danmaku and the popularity of a video, which are usually unknown and uncontrollable.

We turn this challenge into an unaddressed opportunity by viewing it from a different angle. The recent advancement of large multimodal models (LMMs) has shown exceptional performance

in understanding videos and generating textual output [49, 76, 83]. Instead of filtering the existing danmaku, we explore whether and how LMMs could be leveraged to generate valuable danmaku in educational videos. In particular, we ask the following research questions:

- RQ1:** What characteristics of danmaku in educational videos do learners genuinely value?
- RQ2:** How do we design and utilize LMM-driven techniques for generating danmaku in educational videos?
- RQ3:** How does LMM-generated danmaku influence viewers' learning experiences?

To answer RQ1, we conducted a formative study with seven participants to identify what danmaku characteristics learners genuinely value when watching educational videos. We found that danmaku should be knowledgeable, carry positive emotions, show encouragement, and be expressed in effective ways. Based on these insights, we derived seven types of danmaku for educational videos, classified in two high-level categories: content-related and emotion-related. To address RQ2, we designed *ClassComet*, an online educational video platform equipped with a novel LMM-powered auto-generation of effective and high-quality danmaku to promote video-based learning. We leveraged LMMs to comprehend both visual and textual information of videos while creating virtual personas to simulate danmaku interactions among online viewers. Learners can play a video as well as watch and interact with danmaku as they typically do. Without being bothered by the noise in user-created danmaku, *ClassComet* guarantees viewers enjoy consistently effective and engaging danmaku during the learning process, regardless of the video's popularity.

To investigate RQ3 and understand the effect of *ClassComet*, we conducted a within-subjects study with 12 participants, comparing four different conditions: (1) original video without danmaku, (2) video with only content-related danmaku, (3) video with only emotion-related danmaku, and (4) video with both danmaku. The findings suggest that videos with both content- and emotion-related danmaku could significantly enhance viewers' engagement and learning outcomes compared to the original videos. A further evaluation was conducted to compare our generated danmaku with human-created ones on Bilibili<sup>1</sup>, which is known for its danmaku feature, displaying user-generated comments that fly across the screen, synchronized with the video. The results show the competitive performance of *ClassComet* across four different metrics, including relevancy, consistency, fluency, and coherence.

In summary, our main contributions are as follows:

- We uncovered the key characteristics that danmaku should manifest in educational videos through a formative study.
- We introduced *ClassComet* that automatically creates virtual personas and generates effective, high-quality danmaku in educational videos to promote video-based learning.
- We carried out a controlled user study evaluating the quality and effectiveness of our generated danmaku to provide empirical knowledge for future research.

<sup>1</sup>Bilibili(<https://www.bilibili.com>) is a popular Chinese video-sharing website that offers various themes, including educational material, live stream, movies, dramas, and user-generated content. Users can submit, view, and interact with overlaid commentary (i.e., danmaku) on videos.

## 2 Related Work

### 2.1 Viewer Engagement in Video-based Learning

The definition of engagement in educational contexts is multidimensional. Fredricks et al. [24] categorized engagement into three aspects: cognitive engagement, behavioral engagement, and emotional(affective) engagement. Cognitive engagement refers to students' investment in learning, including several aspects such as efforts to master challenging tasks, self-regulation, and challenge preference [23]. Behavioral engagement refers to the involvement and participation (e.g., eye contact, leaning in during a discussion) in learning and academic-related tasks [20–22]. Emotional engagement involves positive and negative emotional reactions towards teachers, classmates, academic works and school in general [20, 53].

This definition is widely adopted by subsequent educational research [6, 74] and has been expanded from traditional school settings to the context of video-based learning [27]. Video-based learning is defined as “the process of using video technology to acquire knowledge or skills” [25]. In the cognitive theory of multimedia learning [59], video-based learning offer better results compared to traditional teaching methods in terms of engagement and learning performance. More online video learning platforms, such as Khan Academy<sup>2</sup> and Coursera<sup>3</sup>, have emerged, offering extensive video resources [71] and putting more emphasis on designing interactive methods to better viewers' engagement.

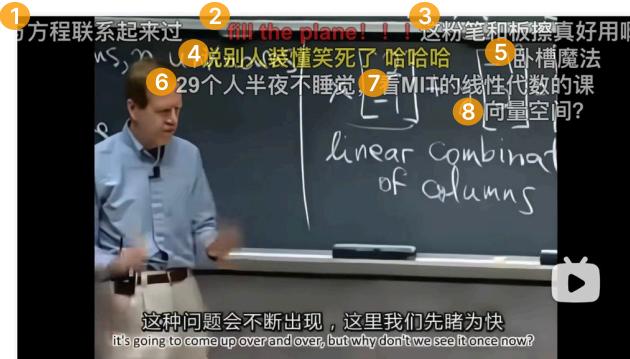
However, current interactive methods face notable limitations. First, although most platforms provide commenting and discussion forums, the success of these approaches depends on the video's popularity. Research shows that popular videos foster vibrant learning communities with more interactions [7, 8, 77], while less popular videos often lack learners' participation, reducing engagement and learning outcomes [16, 72, 78]. Second, while some recent AI-driven interactions, such as automated summaries or video chatbots [9, 28, 84, 87], attempt to act as a knowledgeable peer or teaching assistant, they only simulate partial learning interactions, such as dyadic peer discussions or Q&A, missing more diverse interactions to enable deeper engagement. [46, 56]. These issues highlight a gap in the potential of video-based learning to foster more diverse and effective interactions across all videos, regardless of their popularity. It thus motivated our research to design new interactive methods for enhancing learners' engagement by focusing on danmaku, which could foster a live and communal learning experience.

### 2.2 Features and Usage of Danmaku in Video-based Learning

Previous research [31, 36, 90, 96, 100] has identified several main features of danmaku. The first is brevity and high frequency. Wu et al.'s study [90] indicated the average length of danmaku is 9.33 words which is approximately five times shorter than video forum comments. Additionally, danmaku demonstrates a significantly higher posting frequency across all video genres than forum comments, indicating its potential for diverse interactions. The second is its synchronosity [36, 90]. Unlike comments separated from

<sup>2</sup><https://www.khanacademy.org/>

<sup>3</sup><https://www.coursera.org/>



**Figure 1: A sample screenshot of an educational video with danmaku.** The translations of the displayed danmaku are: 1. “I didn’t linked it to equations before”, 2. “fill the plane!!!”, 3. “This chalk and eraser are so useful”, 4. “Saying others are pretending to understand, lol”, 5. “Wow, this is magic”, 6. “29 people staying up late”, 7. “Watch MIT’s linear algebra class”, 8. “Vector space?”.

the video, danmaku is embedded within videos with timestamps, giving each danmaku a fixed display period on the video timeline, regardless of its actual posting time [96]. This helps to create a sense of co-watching with virtual peers for learners. The third is anonymity. Danmaku hides the identity of the viewers with only the content “flying” over the screen. Anonymity protects viewers’ privacy and has a positive impact on user participation [31, 90]. The last is the blending of interpersonal and personal interaction behaviors. Viewers can use danmaku as virtual peers to engage in question-and-answer interactions and discussions [2, 31]. Moreover, each viewer can express personal, immediate reactions to the video content. Figure 1 exemplifies the danmaku of an educational video.

Unique viewer interactions with and through danmaku have emerged and developed based on the above features, which have been investigated in the context of online video-based learning. Chen et al.’s study [11] and DanMooc [12] categorized danmaku interactions into two general classes: content-related and social-oriented. Content-related danmaku refers to discussions about the video content, while social-oriented danmaku includes off-topic chats and jokes unrelated to the video content. DanMooc [12] showed that users were more disinclined to social-oriented danmaku and considered off-topic chats and jokes to dilute the quality of discussions. DanMooc further refined the content-related interaction into three characteristics: questions, answers to predefined questions, and other video-related messages. Some other researchers categorized the usages of danmaku from two aspects: content-related and emotion-related [31]. An et al. [2] identified a distinct interaction of content-related danmaku: content repetition, where learners used it for intensive practice of learning content while watching the video [5]. Zhao et al. [104] found that students had more interactions in information sharing (such as Q&A exchanges) and dissonance exploration (such as opinion statements), while knowledge application and summarization were infrequent within danmaku. Emotion-related danmaku represents another primary type, as viewers were prone to share their immediate feelings

while watching educational videos including love, joy, surprise, tiredness, and fear. However, the preference for positive or negative emotion sharing was still ambiguous in the context of online education videos. It was found that positive anonymous danmaku comments were more frequent among all genres of videos [90].

These studies have identified various usages of existing danmaku in educational videos. However, what specific characteristics of danmaku are considered valuable and useful for video-based learning stays unclear. Moreover, the limited quantity and various noise of danmaku in educational videos obstruct the effectiveness of danmaku in promoting video-based learning [57, 61], diminishing learners’ experience. Based on the above insights, our study seeks to identify the unique characteristics of danmaku preferred by learners watching educational videos, and explore automated techniques to unlock the potential for more targeted and effective danmaku design to enhance learning.

### 2.3 Automated Live Video Comments

The concept of automated live comment generation, commonly referred to as danmaku generation, has been introduced by Live-Bot [55], proposing two effective generation models and a danmaku benchmark dataset for evaluation. Subsequent studies have focused on designing better strategies to improve the performance on the dataset and enhancing the dataset itself [82, 88, 102]. However, automated danmaku generation faces challenges in considering danmaku-video congruence in specific scenarios. Models often generate general danmaku without understanding and integrating information from different modalities [69, 99]. In educational videos, beyond textual information like subtitles, visual elements such as charts in slides, the color and size of specific fonts, and the spatial positioning of objects in the video greatly influence viewer comprehension. In a broader context, AI-generated content overlays, apart from danmaku, such as dynamic music scores [94] in video games and automated meeting transcripts, rely on multimodal inputs to deliver context-aware interactions that enrich the viewing experience. Thus, multimodal understanding of educational videos is essential for generating comprehensive danmaku and presents a promising opportunity for positively influencing learning.

With rapid advances in multimodal learning, a spectrum of video understanding tasks including video captioning [17, 18, 38, 41], video description [70, 80], and visual question answering [4] can be implemented and evaluated. However, long-form video understanding remains a major challenge due to language model context length and GPU memory restrictions, as most multimodal models are evaluated on video clips under 30 seconds or minute-long benchmarks [49, 75]. Moreover, most educational videos are long, requiring processing multimodal data (e.g., audio, visual, and text) and capturing complex temporal dependencies over extended periods, which demands significant computational resources and careful model design.

This inherent challenge has been attempted by unleashing the power of large multimodal models (LMMs) [40, 52, 83, 86, 103]. Some studies focus on compressing visual information such as merging adjacent or similar frame patch tokens [48, 76]. Others like Timechat [68] and Moment-10M [67] propose time-aware instruction-tuning methods to capture richer spatial and temporal

information compared to short videos. We designed ClassComet based on the LMM-driven method MM-VID [52], which sparsified frames to capture longer visual tokens and utilized a video-to-script generation with GPT-4V to transcribe multimodal elements into a long textual script, facilitating understanding of long educational videos and diverse danmaku generation.

In the current HCI community, some work utilized LLMs to moderate danmaku [34, 35]. However, none of the work addresses the educational context, where the number of useful danmaku can vary depending on the popularity of educational videos. Therefore, how to design automated danmaku that considers both quality and quantity for educational videos still remains unsolved. In this research, we specifically focus on leveraging LMM for automated danmaku generation in educational scenarios, aiming to support hours-long video understanding and produce high-quality, diverse danmaku across all videos that positively impact learning.

### 3 Formative Study

As the first step of our research, we conducted a formative study to learn about the key characteristics that danmaku should have in educational videos (RQ1).

#### 3.1 Study Setup

We recruited seven participants (4 self-identified females and 3 self-identified males) from local universities via mailing list and word of mouth, denoted as P1-P7 in the following text. They were undergraduate or graduate students (ages ranging from 21 to 25,  $M = 23.67$ ,  $SD = 1.25$ ) with various technical backgrounds, including computer science, statistics, law, finance, and design. All participants frequently watch educational videos for their learning and have the habit of turning on danmaku on Bilibili. When asked about their danmaku posting frequency on a 5-point Likert item, one self-reported “Frequent,” three reported “Normal,” and three said “Occasional.”

Each study session consisted of two stages: 1) watching a 5-minute educational video with some discussion and 2) a semi-structured interview regarding danmaku’s characteristics. In the end of the study, they received \$20 for their time and effort. In the first stage (10 minutes), we provided four educational videos with danmaku to help participants connect their own danmaku experience with these videos. The four videos were selected from Bilibili and covered different learning topics, each approximately 5 minutes in length. Participants were asked to select a video based on their interests and watch it as they normally would. They could speed up or pause the video, comment on existing danmaku, and post danmaku whenever they wanted. This allows us to observe their natural behaviors when watching educational videos with danmaku. We employed a contextual inquiry method: each time when the participant posted danmaku, we engaged in a short discussion to understand their motivations and rationales. In the second stage (15-20 minutes), we discussed with the participants about their preferences for content-related and emotion-related danmaku, habits in posting danmaku, reactions and opinions to different danmaku characteristics, and expectations for ideal danmaku watching experiences.

### 3.2 Key Insights

We collected the danmaku posted by participants in the first stage to help participants recall their typical danmaku-watching habits. We found that participants sent 1-4 danmaku per minute during the video-watching session, with  $M = 1.85$  per minute ( $SD = 0.98$ ). The average length of danmaku was 9.2 characters ( $SD = 5.34$ ). In terms of danmaku interactions, all participants posted danmaku to engage with either existing danmaku or the video content. These user behaviors align with the results reported in previous work [90].

We also transcribed our discussion with the participants in the whole study session including our contextual inquiries and semi-structured interviews. The whole research team then collaboratively coded the transcripts and constructed an affinity diagram to derive five insights into what danmaku characteristics learners value in educational videos.

**K1: Danmaku should carry useful knowledge about the learning concepts.** Six participants mentioned that they preferred knowledgeable danmaku that provides correct information about or emphasizes key concepts from the videos, compared to that with less information or irrelevant to the video, such as unrelated jokes or off-topic comments. P3 said, *“In some cases, there are important danmaku of answering the video’s questions, and I really want to focus on them. But irrelevant danmaku often covers it or appears nearby which distracts my attention. So I would sometimes pause to delete the irrelevant danmaku, but I don’t think this is an effective way.”* P6 stated, *“I think Q&A danmaku can help me understand what the videos are talking about, which is the question and which is the answer, that can help me understand the video better.”* Participants also expressed their expectations of more discussions regarding the video content. *“I want to see danmaku discussing controversial things and giving new perspectives.”* -P4

**K2: Danmaku should help learners grab important information from the video content.** Three participants expressed their preferences to see more summaries in danmaku. P4 said, *“I prefer danmaku summarizing the content of the videos. This can help me better understand the concepts.”* P2 pointed out the value of summary danmaku in longer videos, particularly lectures: *“Summary in danmaku can help me quickly capture the main points of the video, especially hour-long lectures. It can appear at the beginning or after each section. It’s like a quick review that keeps me focused without losing track.”* Further, highlighted danmaku was valued by participants. *“I find those highlights very important because there are so many danmaku. With those highlights, I can easily capture them and focus my attention so that I don’t need to pay attention to every line of danmaku. Sometimes, educational videos are not interesting so I will lose focus. However, highlighting danmaku can help me redirect my attention back to the video itself, making it easier to stay engaged.”* -P1

**K3: Danmaku should express positive emotions to foster a good learning environment.** Five participants mentioned their preference for positive emotional expressions in danmaku within the context of online video learning. For instance, P1 gave a “like” to the danmaku “Very clear explanation” in the first stage of the study. She stated: *“I have the same feeling when I watched the video. Giving a like is a kind of compliment.”* P2 further mentioned in the interview: *“Personally, I prefer more positive danmaku, because it’s a learning video, not an entertaining one. I don’t want to be in a very negative*

*learning atmosphere.*” P3 also compared positive emotion-related danmaku with negative ones. She said: “*In educational videos, I don’t want to see too much negative emotion, like complaining about the video’s difficulty or impolite augments with other viewers. It’s useless and ineffective. If the video is filled with negative danmaku, it will definitely affect my mood and my learning progress.*”

**K4: Danmaku should show encouragement to help learners overcome challenges.** Participants were willing to see more encouraging danmaku when encountering difficulties in learning such as tiredness and complaints. The encouragement in danmaku fosters a sense of “*having a positive learning partner*” -P3 and “*giving the feeling of learning with other people.*” -P5 Specifically, P1 pointed out that “*I’d love the danmaku to show support like in the Japanese video, for example, ‘If you see this danmaku, you should keep learning and do not close this video.’ I found it quite important because at that stage I was losing focus.*” P7 also mentioned that “*Encouragement is much better than just expressing exhaustion or negative feelings.*”

**K5: Danmaku should be manifested effectively and concisely in the video.** Participants showed preferences for more effective expressions in danmaku. Apart from favoring short, concise, and clear danmaku, participants enjoyed using emojis and symbols to convey their confusion or other feelings. For example, P4 sent danmaku “??” to indicate her confusion when a concept was presented, and P3 added “!” at the end to express her surprise at the video content, aligning with a line of other existing “!” danmaku. Further, P5 explained why he wanted to send a thumb-up as danmaku, “*I don’t usually send danmaku myself. If I want to send danmaku, I need to pause the video and it takes minutes. It’s too long for me. Usually the most I can do is to send an emoji or symbols to show my reaction.*”

### 3.3 Types of Danmaku

Informed by the above insights, we provide a summary of seven common types of danmaku valued by participants the most in educational videos and align our taxonomy with existing studies that categorize danmaku into two big classes: content-related and emotion-related [11, 12]. The details are displayed in Table 1.

The four most valued content-related danmaku are Q&A, discussion, highlighting, and summary. Q&A danmaku can occur between the instructor who proposed questions in the video and viewers or between active viewers themselves. Thought-provoking questions and answers benefit learners, regardless of whether they participate in or observe Q&A interactions (K1). Discussion is also appreciated by participants as it can timely exchange on the video content (K1). Unlike forum comments under the video, discussion in danmaku is more concise [90]. Highlighting is a unique type that only exists in danmaku, which normally hovers at the top center of the video for several seconds, using eye-catching styles (such as different font colors or sizes) to highlight keywords (K2). Summary is expected to be seen more in danmaku which typically appears at the beginning or end of each video section, providing a recap of key points (K2). Personal emotional expression is the most common emotion-related interaction. As learners prefer positive emotional expression, they value compliments and encouragement in danmaku during their watching experience (K3, K4). This helps to create a more concentrated and positive learning atmosphere. Both content-related and

emotion-related danmaku should be expressed effectively. Viewers prefer short, concise, and direct danmaku, often using various emojis and symbols to simplify their expressions (K5).

## 4 ClassComet Platform

Based on the results from our formative study, we designed and developed ClassComet, an online educational video platform empowered by the LMM-driven automated generation of danmaku, aiming to improve viewers’ learning outcomes and engagement (RQ2). In this section, we describe the user interface of ClassComet and a simple usage scenario. Details regarding the underlying techniques for danmaku generation will be discussed in the next section.

### 4.1 User Interface

As shown in Figure 2, the ClassComet user interface consists of two main components: a video player and a video sidebar.

The video player contains three parts. (1) The video control buttons (Figure 2-A) allow for play/pause, volume adjustment, and speed control. (2) The generated danmaku “fly” over the screen (Figure 2-B), which is white by default and can be changed to other colors to highlight important concepts, following the convention of danmaku experience on commonly used video platforms like Bilibili. Highlighted danmaku stays at the top center of the video for some time <sup>4</sup>, while other danmaku dynamically scrolls across the video screen. (3) The danmaku control includes danmaku settings and an input box for users to send custom danmaku (Figure 2-C). Users can freely post danmaku with options to select the font color and display position (scrolling across the screen, hovering at the top, or staying at the bottom of the video).

The video sidebar (Figure 2-D) displays a list of additional videos from the same course, allowing users to watch them sequentially. Once users have finished the current video, they can easily navigate to the next lesson. Upon clicking a video, ClassComet loads it into the video player, and the danmaku is automatically generated based on the video content as the video plays. In Figure 3, we display all scenarios of danmaku interactions.

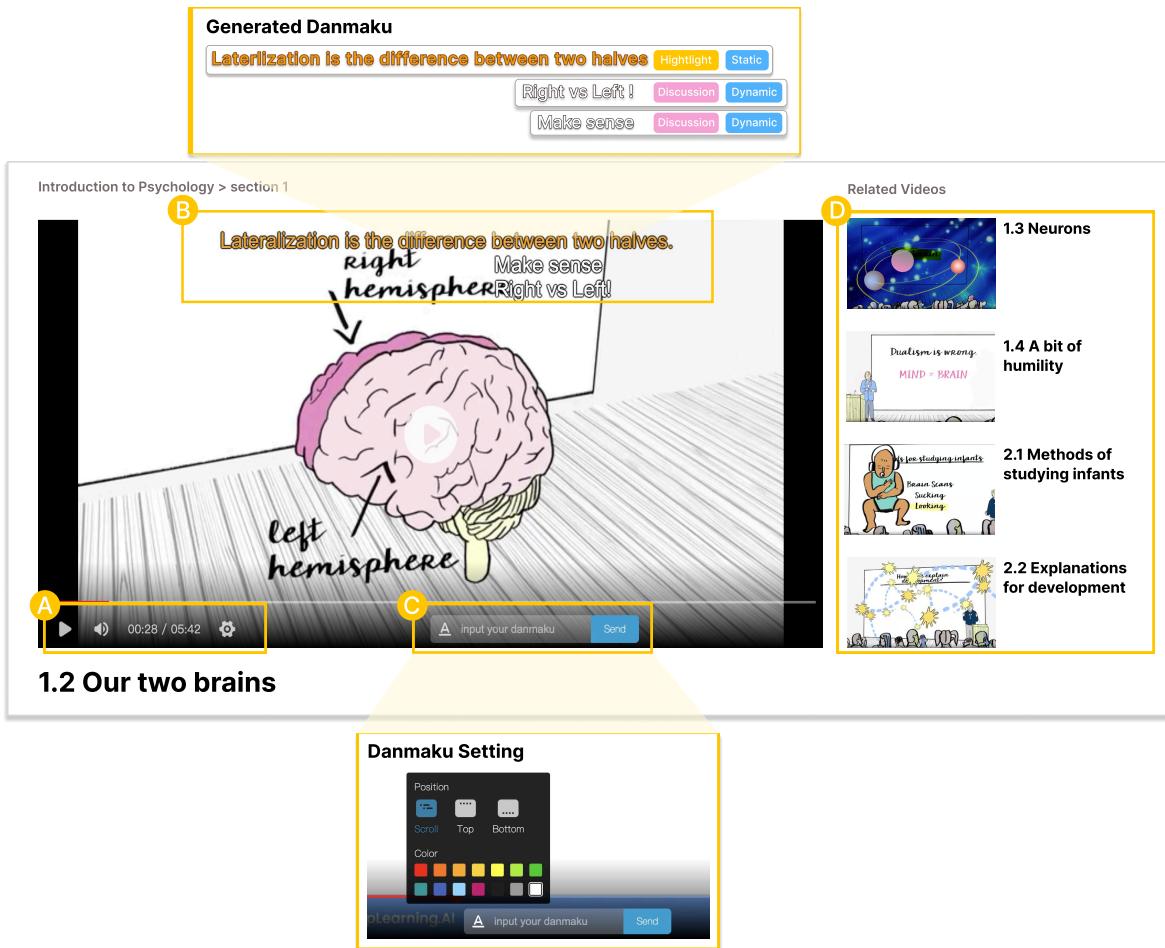
### 4.2 Usage Scenario

Before describing the technical details about ClassComet, in this section, we present a usage scenario to illustrate how users interact with various system features in practice. Cara, a novice learner in psychology, is self-learning the Introduction to Psychology course. However, whether watching on YouTube or educational platforms like Coursera, she often feels isolated. Since the course is not as popular as others, it lacks active comments and interactions commonly seen in more widely viewed videos. Cara misses the experience of exchanging ideas, participating in discussions, and receiving encouragement in classroom learning. Although Coursera provides quiz and assignment interactions, Cara finds it difficult to keep herself motivated and engaged due to the lack of social connectedness and interpersonal interactions. After completing the first few videos, she starts to search for a new platform that fosters a peer-learning environment, regardless of the video’s popularity. Therefore, Cara decides to give ClassComet a try.

<sup>4</sup>Previous research [11, 31] indicated that danmaku with unique displays were more likely to be fixed in the top and bottom positions.

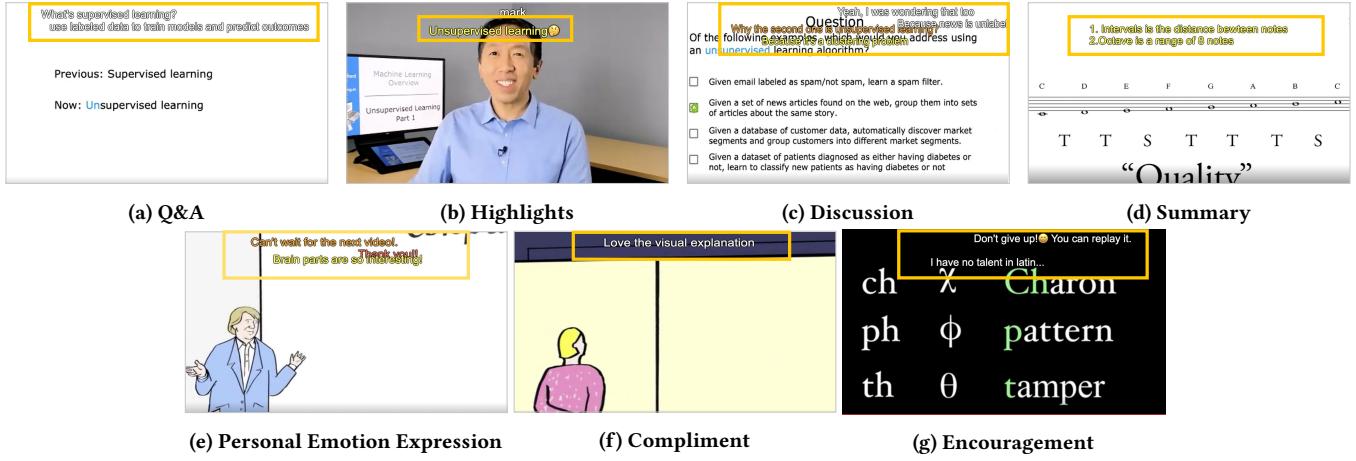
**Table 1: Types of danmaku in educational videos to support learning.**

Class	Type	Description
Content-related	Q&A	Ask and answer questions to assist viewers in consolidating acquired knowledge and dispelling misconceptions (K1, K5).
	Discussion	Exchange opinions, propose hypotheses or provide complementary information related to the proposed question in the video (K1, K5).
	Highlights	Emphasize key concepts or important words in unique displays (font size, color, position) to capture viewers' attention (K2, K5).
Emotion-related	Summary	Preview key points at the beginning, summarize after each section, and provide a final recap at the video's end (K2, K5).
	Personal Emotion Expression	Simply and directly express one's emotions within emojis and symbols (K5).
	Brief Compliment	Offer praise when a viewer's danmaku provides the right answers or explicit explanations to the video's questions or other viewers' questions (K3, K5).
	Encouragement	Send supportive danmaku in response to negative expressions from other viewers (K4, K5).

**Figure 2: User interface of ClassComet: (A) video control buttons for play/pause, volume adjustment, and speed control, (B) danmaku embedded in the video, (C) danmaku control that contains danmaku settings and an input box to send danmaku, and (D) a video sidebar for selecting other educational videos.**

Cara enters ClassComet and selects the same psychology course. As she starts to play the lecture, she notices a familiar feature: a stream of vibrant danmaku scrolling across the video. Having

watched danmaku frequently on Bilibili for both entertainment and learning purposes, Cara is pleased to see danmaku. However, after just five minutes of watching, she notices a significant difference.



**Figure 3: Examples of different types of danmaku (including both content-related and emotion-related) generated by ClassComet in educational videos.**

Unlike the danmaku she used to on Bilibili, where jokes and random arguments can overwhelm genuine valuable insights, the danmaku in ClassComet has no such irrelevant danmaku. She finds more danmaku relevant to the video content which displays more often than the danmaku videos she used to see.

Cara observes many highlighting danmaku with distinctive colors like “*Note: symmetry*.” These danmaku serve as gentle reminders, emphasizing the key points in the video and helping her stay focused. As she feels a bit confused by the video’s apple example, she notices a pair of question-and-answer danmaku passing by, which provides a clear explanation and addresses her confusion. Apart from informative danmaku that is closely related to video content, Cara notices supportive danmaku sharing feelings that resonate with her own learning experience. When she begins to feel a little fatigued and her focus starts to slip, she sees a danmaku saying “*Come one! Three minutes left!*” It serves as a kind nudge that encourages her to regain focus. By the end of the lecture, she was satisfied to see a concise summary of the video “*This lesson is about the comparison of two halves of the brain*,” as though a helpful final review that recaps the main ideas of the video.

Beyond standard video functions such as adjusting playback speed and volume (Figure 5-A), Cara notices an input box at the bottom of the video player (Figure 5-C), so she can easily interact with existing danmaku and shares her own thoughts. Additionally, Cara explores other different topics in videos. She clicks some lesser-known courses like “*Introduction to Reproducibility in Cancer Informatics*” which has 0 reviews on Coursera. To her surprise, this interactive danmaku feature is present in all courses in ClassComet, not just the popular ones, but typically lacks comments or rates on other platforms.

After exploring this new platform, Cara feels her learning is enriched not just by the video content but by the ideas and shared understanding and feelings that come from the danmaku interactions.

## 5 Danmaku Generation in ClassComet

We proposed a novel pipeline based on LMMs for generating danmaku that encompasses both content-related and emotion-related danmaku in different types (Table 1). As shown in Figure 4, the inputs to ClassComet for generating danmaku come from three aspects: (A) video understanding at both the text and clip levels, (B) persona creation, and (C) a danmaku prompt template. The output of contextually rich danmaku is then embedded into educational videos (Figure 4-D). Below, we detail each input component.

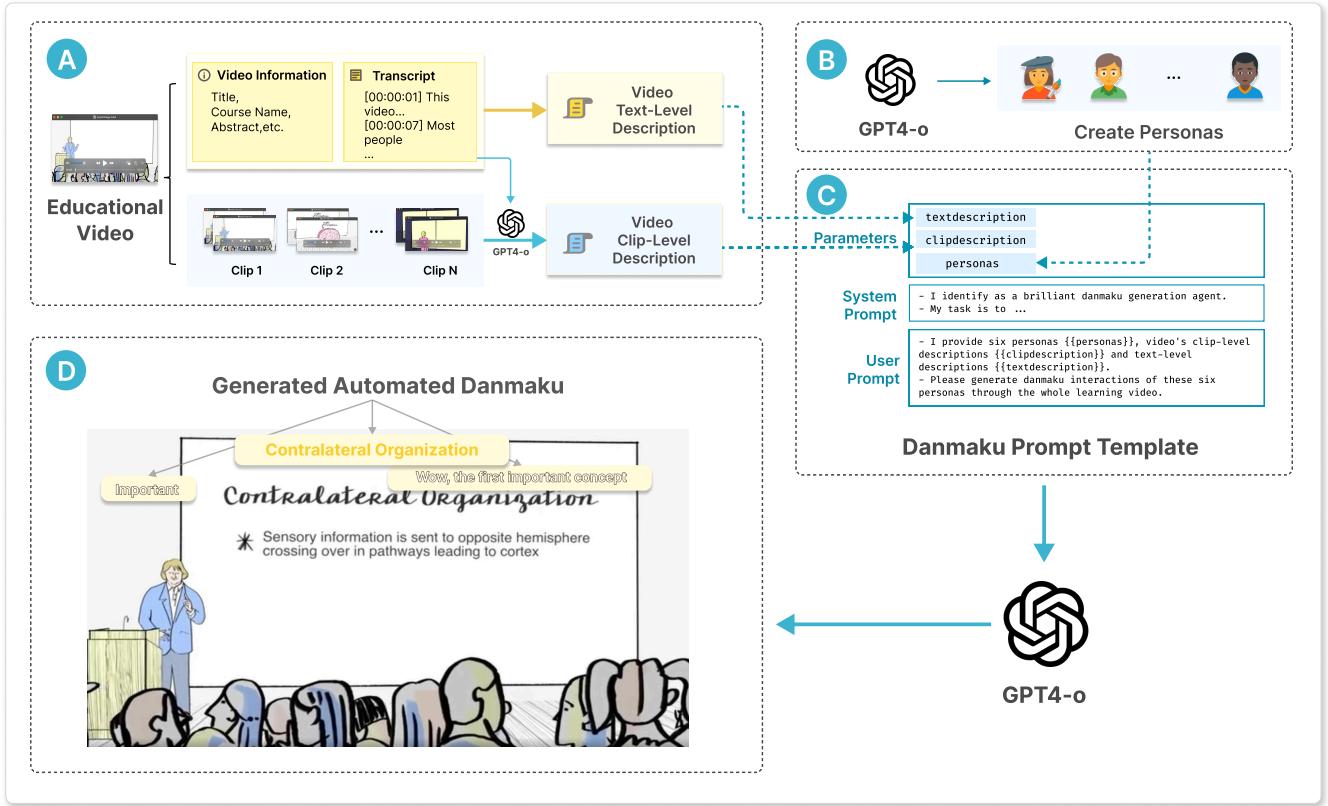
### 5.1 Clip-level and Text-level Video Understanding

Recent studies [29, 30, 93] showed that LMMs have overcome the limitations of text-only video understanding and shown great potential to address the challenges of understanding long-form videos. We adopt the LMM GPT4o to generate comprehensive and contextually relevant descriptions at both clip- and text-levels. To enable ClassComet to comprehend hour-long educational videos, we segment the entire video into multiple clips and sample frames from each clip. By utilizing established scene detection tools PySceneDetect<sup>5</sup>, we identify scene boundaries and track scene changes. After segmenting the video into multiple clips, we follow the MM-VID [52] algorithm and uniformly sample frames from each clip, with each minute consisting of 5 frames. By feeding the clips and transcriptions into the GPT4o, we capture detailed clip-level descriptions of the educational videos, such as the lecturer’s face and actions and images displayed on the video slides. Further prompts and examples can be found in the appendix A.

The text-level description generation is based on two sources of information: basic video details such as the title, abstract, and course name, as well as transcriptions with timestamps. We utilize Microsoft Azure Video Indexer API<sup>6</sup> to extract transcriptions along with timestamps. All components of the text-level descriptions are formatted in JSON. We integrate both clip- and text-level video

<sup>5</sup><https://www.scenedetect.com/>

<sup>6</sup><https://azure.microsoft.com/en-us/products/ai-video-indexer>



**Figure 4:** The ClassComet pipeline for automatically generating danmaku in educational videos includes four steps: (A) Extract video information, segment the video into clips using scene detection, and generate descriptions at both the text- and clip-level; (B) Create virtual personas; (C) Design a structured danmaku prompt template and set parameters; and (D) Embed generated danmaku into educational videos.

descriptions into our prompt template for automated danmaku generation.

## 5.2 Create Virtual Viewers

Many works utilized prompted LMMs [91] to generate short natural language descriptions of personas and their corresponding behaviors. The success of synthesizing human-human dialogues largely depends on the personas of the dialogue participants [10, 42, 50]. To effectively simulate the danmaku interactions, we create multiple personas with diverse and detailed backgrounds, including ages, regions, educational histories (degrees, universities, and majors), personalities, speaking styles, and their reasons for watching the educational video. Prior work has equipped LLMs with personas to simulate human interactions and feedback such as generating Reddit community comments [64] and creating agents serving various roles like educators [66]. Hunget al. compared persona-based comments with no-persona comments in video-sharing platforms [37], which showed that persona-based comments were significantly higher than no-persona comments in relevance, believability, and helpfulness. The average sending frequency of danmaku in current educational videos is around 25-30 per minute [90], while the off-topic danmaku that is unrelated to the video may be up to

10%-30% [12]. Based on the sending frequency per user outlined in Section 3.2, we set up six virtual viewers to participate in danmaku interactions, which aligned with the frequency of danmaku posting in previous research [90, 96]. These personas may either interact with other viewers' danmaku or engage with the video content itself.

Before generating danmaku, we leverage GPT4o to create personas. A detailed description of the prompt structure for this step is illustrated in the following. The XML-like tag <title> can be substituted with a specific video title that the learner is watching. More details about this step, as well as examples of input and output, are provided in Appendix B.

```
Prompt: - Your task is to create 6 distinct personas with different backgrounds and personalities. They are interested in watching the online educational video <title>. Each persona should have the habit of sending danmaku while watching the video . Use "A", "B", "C", "D", etc., as persona labels.
- For each persona, please provide the following details in JSON format, including age, region, personality, danmaku sending style , learning habits, and reasons for watching the video.
```

### 5.3 Danmaku Prompt Template

To effectively configure GPT4o for generating danmaku across educational videos, we design a structured danmaku prompt template, as shown in Figure 5. The template consists of three components: the system prompt, user prompt, and parameters. The system prompt informs GPT4o of its role, tasks, behaviors, and boundaries. All XML-like tags shown in the figure in the system prompt are illustrated below and replaced with specific prompts. The user prompt customizes the responses by providing the user's needs and different parameters. Both system and user messages are defined between the `<|im_start|>` and `<|im_end|>` tokens to clearly define different boundaries. All parameters are embedded into the user prompt as input and formatted in JSON.

**5.3.1 System Role and Tasks.** In the first section of the system prompt, we outline the role of GPT4o as a danmaku generation agent and summarize the task as follows. Role prompting [43] leads to a more focused, creative, or empathetic response depending on the given role. We summarize the key actions GPT is required to perform. We also set boundaries in the first section to ensure the model's output quality. A detailed description of the prompt structure for this step is shown below:

```
Prompt: # I'm a danmaku generation agent
- I identify as a brilliant danmaku generation agent.
- My task is to generate content-related and emotion-related
danmaku. The generated danmaku should reflect the unique
personalities and diverse backgrounds of pre-defined personas.
- I should simulate dynamic and engaging danmaku that align with
their distinct character traits.
```

**5.3.2 Danmaku Types.** For each type of danmaku, we provide in-context examples to stimulate the danmaku interactions in Figure 5, which are crucial for enhancing GPT's output performance [62, 106]. Each example specifies the persona ID, timestamp, and danmaku content. For types such as discussions, Q&A, and encouragement, which involve exchanges between pairs of or even more virtual viewers, we present synthetic multi-turn danmaku types in the in-context examples. For the highlighting danmaku, examples should use HTML tags to define its highlighting color to capture viewers' focus. For the personal emotional expression danmaku, we provide multiple examples, including emojis, modal words, and short sentences, to make the danmaku responses more natural. More details of each danmaku type's descriptions are provided in Appendix C.

**Response Format** The response format can help guide the model to organize answers in the desired structure. We select the Markdown format as its stable and competitive output performance across different response formats [32], as detailed in the following:

```
Prompt: ### Response Format
```
# Emotion-related danmaku
## <emotion-related danmaku type 1>
- <role> | <timestamp>: <generated danmaku>
- <role> | <timestamp>: <generated danmaku>

## <emotion-related danmaku type 2>
```

# Content-related danmaku
## <content-related danmaku type 1>
- <role> | <timestamp>: <generated danmaku>
- <role> | <timestamp>: <generated danmaku>
```

```
## <content-related danmaku type 2>
```

```

**5.3.3 Deliberating Actions.** We next deliberate actions for generating danmaku by specifying the length, frequency, and count. First, drawing from previous research and formative studies, we set the maximum danmaku text length to 12, based on the average danmaku length reported in earlier studies [2, 31]. Second, some works reported that the danmaku frequency in educational videos typically ranges from 2 to 30 per minute [89, 90]. Therefore, we set the lower bound of danmaku frequency to avoid long gaps, ensuring that danmaku is generated no less frequently than every 30 seconds. Third, we defined the range for content-related danmaku as 15-25 and for emotion-related danmaku as 5-10. Prior research has estimated that the average danmaku frequency in educational videos ranges from 25 to 40 per minute, with approximately content-related danmaku comprising 60-65% and emotion-related danmaku making up 15-25% [12, 31, 90]. Fourth, we made requirements for the number of highlighting danmaku since prior studies found that the proportion of highlighted danmaku makes up over 50% of the content-related danmaku [90]. The detailed restrictions given to GPT4o in this step are shown below:

```
Prompt: # Deliberating actions to generate danmaku
- The length of each danmaku should be less than 12.
- I should generate danmaku continuously without long gaps (longer than 30s).
- I should generate about 15-25 content-related danmaku and 5-10
emotion-related danmaku per minute.
- I should generate more than **10 highlight** per minute.
- Each type of danmaku should cover the entire duration.
```

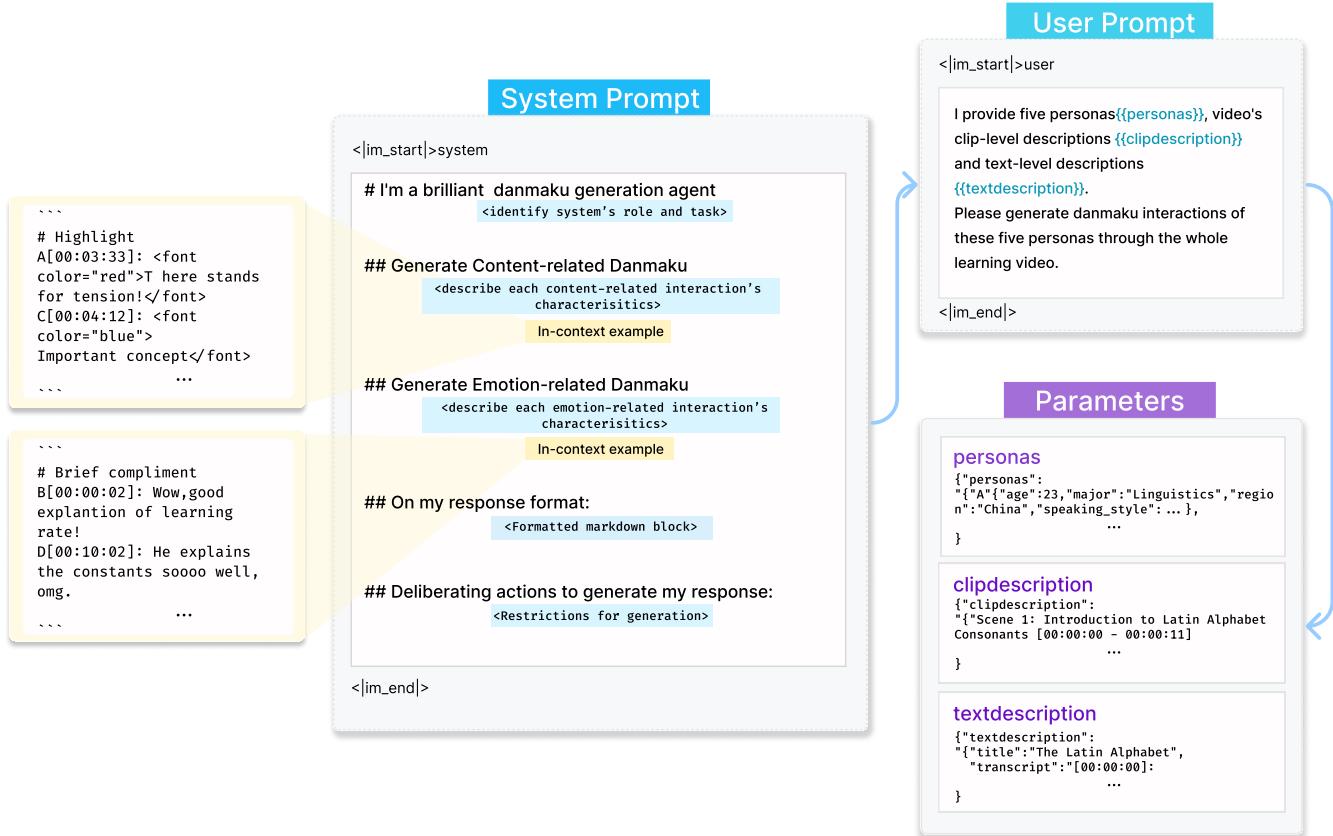
**5.3.4 User Prompt and Parameters.** In the user prompt, we embed three parameters in JSON format derived from previous clip-level and text-level descriptions, as well as the created personas into the user prompt to customize danmaku for different videos. These structured inputs can facilitate the generation performance for large multimodal models [54]. In this way, ClassComet can generate automated, effective, and high-quality danmaku for educational videos. The full version of the template, along with generated examples, is introduced in Appendix C.

## 6 Evaluation

To explore how LMM-generated danmaku could influence viewers (RQ3), we conducted a controlled experiment assessing the effectiveness of two categories of generated danmaku (content- and emotion-related) on viewers' learning outcomes and engagement. In addition, we evaluated the quality of generated danmaku by comparing it with real danmaku on Bilibili.

### 6.1 Participants

We recruited 12 participants as U1-U12 (ages ranging from 21 to 26,  $M = 24.25$ ,  $SD = 1.36$ ; 6 self-identified males and 6 self-identified females), all of whom had the habit of watching or sending danmaku when they learned educational videos. They were undergraduate or graduate students with various majors, including Computer Science, Law, European Politics, Design, Arabic, and Information Engineering. Each participant was required to complete a study



**Figure 5: The structured prompt template of generating danmaku which consists of the system and user prompts as well as the pass parameters including two video descriptions (text and clip levels) and personas.**

consisting of two stages (Figure 6), lasting about 65 minutes. They were required to complete a video-watching session to compare the impact of different generated danmaku in Stage 1 and compare the quality of danmaku generated by ClassComet with danmaku on Bilibili under four different metrics in Stage 2. At the end of the study, they received \$20 for their time and effort.

## 6.2 Procedure of Evaluating Learning Outcome and Engagement (Stage 1)

To compare the learning outcome and engagement of different generated danmaku, we employed a within-subjects design to quantitatively compare four conditions, including No-Danmaku, Emotion-only, Content-only, and Content+Emotion danmaku for educational videos. Furthermore, we conducted qualitative analysis based on participants' feedback from post-interviews.

**6.2.1 Pre-quiz (3 minutes).** Before watching each video, each participant's prior knowledge was obtained by completing a quiz. Each quiz consists of five memorization questions and five application questions (Appendix D), designed to respectively capture learners' cognitive activities [78]. The 10 multiple-choice questions in each topic were thoughtfully designed by our research team mainly, primarily based on the core messages of the video. For each topic, we

then asked three graduate students (Master's or PhD) who were actively doing research in the area to help validate to ensure the questions' correctness and quality.

**6.2.2 Educational Video Watching (5-8 minutes).** After submitting the quiz, participants were automatically redirected to the video-watching page in ClassComet. Participants were presented with the four different danmaku conditions in a counterbalanced order. Participants were allowed to watch the video as they normally did. They could send danmaku, speed up the video, adjust the volume, or pause it. Our backend recorded the real-time participants' actions and posted danmaku.

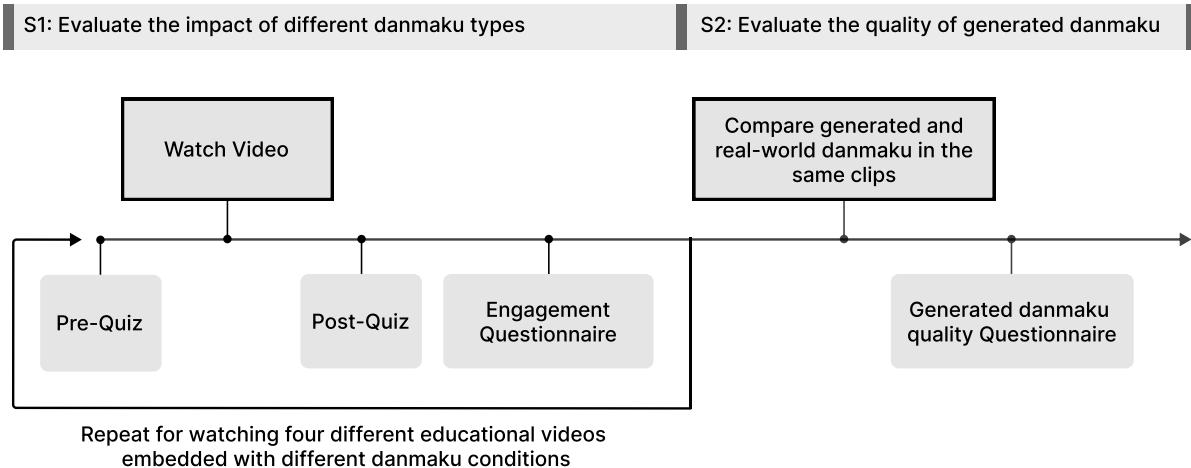
The selected four videos were under different topics, including supervised learning<sup>7</sup>, brain structure<sup>8</sup>, music theory<sup>9</sup>, and latin alphabets<sup>10</sup>. Each video was about 4 to 8 minutes long. To mitigate biases, the order of the conditions in S1 was counterbalanced across participants, and the combination of the type of danmaku and the video was randomly assigned.

<sup>7</sup><https://www.coursera.org/learn/machine-learning/lecture/TxO6F/unsupervised-learning-part-1>

<sup>8</sup><https://www.coursera.org/learn/introduction-psychology/lecture/mvjLn/our-two-brains>

<sup>9</sup><https://www.coursera.org/learn/edinburgh-music-theory/lecture/Osu9c/octaves>

<sup>10</sup><https://www.youtube.com/watch?v=0oWWOJW394&t=34s>



**Figure 6: Diagram of the experimental procedure, consisting of two stages: (S1) evaluating the impact of different danmaku on viewers' learning outcomes and engagement, and (S2) comparing the quality of generated danmaku with human-created danmaku on Bilibili.**

Our selection of the videos was guided by the following three considerations. First, we selected topics with which learners would have little to no prior knowledge. We investigated participants' majors and related courses before starting Stage 1 to ensure that all participants have minimal background in the featured fields. For example, even though some of our participants majored in Computer Science, they did not have backgrounds in AI. Second, we prioritized videos that are representative in quality, typicality, and format. Each educational video was selected from Coursera or YouTube, which received high ratings and good reviews from learners. The selected videos featured diverse instructional styles, including engaging animations and picture-in-picture presentations led by professors. Third, the content of each video was relatively independent and did not require other videos as prerequisites. All selected videos provided engaging and sufficient contextual information for viewers to learn.

**6.2.3 Post-quiz, Questionnaire, and Interview (10 minutes).** After watching each video, participants completed the same quiz they took before watching the video. We measured participants' learning outcomes by calculating the difference between their post-quiz and pre-quiz scores. The relative value considers individual differences in baseline knowledge by using the difference between post- and pre-scores, using the difference (post - pre) normalizes these differences and provides a more comparable measure of learning progress. Additionally, participants filled out a questionnaire (Appendix E) which measured behavioral, emotional, and cognitive engagement on a 7-point Likert scale. We categorized the questions based on the definition of each type of engagement as described in Section 2.1. In the context of video-based learning, behavioral engagement refers to active participation in the video (e.g., pausing, re-watching), emotional engagement reflects the emotional connection to the content (e.g., excitement, boredom), and cognitive engagement assesses the mental effort and focus invested in understanding the material (e.g., the depth of thinking, reflection).

All engagement questions were derived from six existing widely adopted questionnaires. Situational Interest scale [3], Original Cognitive Engagement scale [26], Engagement vs. Disaffection scale [73], Classroom Engagement Inventory [85], Situational Interest scale [79], and MOOC engagement scale [15]. Aligning with the process of generating engagement questionnaires in previous research [78], we first removed some questions that are unrelated to the context of our study, such as questions related to self-regulation (e.g., "I denied myself rewards when I failed to reach my study goals."). Additionally, we removed repeated questions appearing in different scales to ensure questionnaires' clarity (e.g., "How enjoyable was the lesson?" and "I enjoy the lesson."). For similar questions (e.g., "I pay attention in class." and "I listen very carefully."), we selected the statement that was repeated across surveys. After this, we adapted some questions to match the context of the study such as replacing "Book" in Original Cognitive Engagement scale by "Video". We show the complete engagement questionnaire generation strategy in Appendix F. Since there's a condition of No-Danmaku and our questionnaire is evaluated both before and after the video, we ensured that the questionnaire didn't mention danmaku to eliminate the potential bias of evaluation.

To complement the quantitative analysis, participants were asked to complete a post-interview with open-ended questions to help us understand their learning experiences, feedback, and preferences for different types of danmaku. Interviews were audio-recorded and transcribed before analysis.

### 6.3 Procedure of Evaluating Danmaku Quality (Stage 2)

To ensure the danmaku quality of ClassComet, participants were asked to compare the AI-generated danmaku with human-created danmaku on Bilibili and shared their feelings after scoring the danmaku quality of the two styles of generation. The procedure of Stage 2 lasted 15 minutes.

Since all four videos in S1 appeared on Bilibili and included available human-created danmaku, we selected six clips from the four videos on Bilibili by choosing the parts that had the most danmaku. Four of them were taken from the machine learning and psychology videos; since the psychology and Latin videos had less danmaku, we extracted two clips from them. Each clip contains more than four danmaku, which lasted from 10–25 seconds. We extracted the same clips from ClassComet with the generated Content+Emotion danmaku. The order of the six clips was randomized for each participant. Then, we extracted the same clips from ClassComet with the generated Content+Emotion danmaku. Participants were asked to watch Bilibili and ClassComet danmaku for each clip without being informed of the order of the two danmaku modes (ClassComet(AI-generated) vs. Bilibili (human-created)), in order to prevent any preconceived biases from influencing their judgment. After watching each mode, they completed a questionnaire for evaluating the danmaku quality on a 5-point Likert scale. The questionnaire consisted of eight criteria grouped under four metrics widely used in evaluating the quality of generated live comments [55], conversations [50] and LLM-generated content [45, 105]. These metrics are defined as follows:

- **Relevancy:** Measures the relevance between the danmaku and the educational videos (e.g., A danmaku like “*Great explanation of World War II*” in a video about modern history is considered more relevant than a danmaku like “*I love pizza!*”).
- **Consistency:** Assesses the factual alignment between the danmaku and the learning video (e.g., A danmaku like “*The Earth orbits the Sun*” in a video about astronomy is considered more consistent than a danmaku like “*The Sun orbits the Earth*”).
- **Fluency:** Examines whether the danmaku has formatting problems, capitalization errors, or obviously ungrammatical sentences (e.g., A danmaku like “*Photosynthesis is a vital process for plants*” is considered more fluent than a danmaku like “*Photosynthesis vital is process plants*” due to word order errors.)
- **Coherence:** Evaluates how well danmaku fits together, maintains a logical flow, and sounds natural (e.g., A danmaku like “*This video explains neural networks well*” followed by “*Neural networks are key to AI*” is considered more coherent than followed by “*I love cats*”).

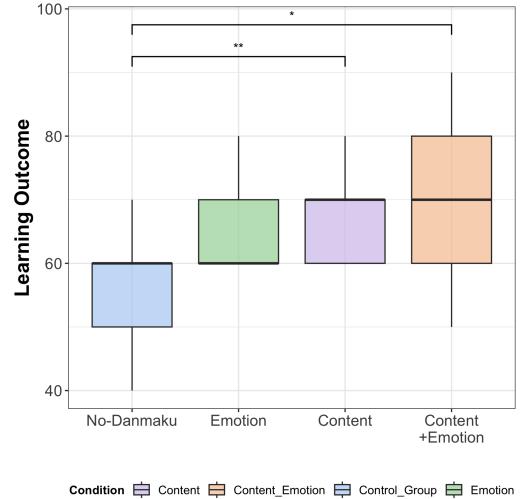
After scoring Bilibili and ClassComet danmaku, we conducted a short post-interview with open questions about their feelings about the two danmaku modes.

## 7 Results

### 7.1 Quantitative Results

In this section, we first presented the findings on participants’ learning outcome and engagement (behavioral, emotional, and cognitive) in S1 of the study. We conducted one-way repeated measures ANOVAs to compare the four conditions, followed by a post hoc analysis using the Tukey HSD test [1]. Then, we reported the results of the comparison between the real danmaku on Bilibili and our generated danmaku.

**7.1.1 Learning Outcome.** We calculated the difference between post- and pre-quiz scores, rather than focusing on absolute scores, to measure participants’ learning outcomes. Relative values are more



**Figure 7: Box plot of learning outcome for different conditions.** Error bars show 95% confidence intervals. Asterisk (\*) indicates a statistically significant difference between conditions:  $p < .05$  (\*);  $p < .01$  (\*\*);  $p < .001$  (\*\*\*)�.

sensitive to small changes in performance than absolute values, as they control for baseline differences and capture improvements even with low initial scores. The results revealed a significant main effect on learning outcome ( $F(4,12) = 6.47, p < .01$ ) for different danmaku conditions, as summarized in Figure 7.

We found the Content-only danmaku ( $M = 67.50, p < .001$ ) and the Content+Emotion danmaku ( $M = 70.80, p < .01$ ) showed significantly higher learning gains compared to the No-Danmaku condition. Additionally, while the score of the Emotion-only condition ( $M = 64.20$ ) was higher than that of the No-Danmaku condition ( $M = 57.5$ ), the difference was not statistically significant. Overall, the results suggest that content-related danmaku, particularly when combined with emotional engagement, significantly enhances learning outcomes.

**7.1.2 Behavioral Engagement.** As shown in Figure 8 (left), there was a significant main effect for the danmaku type on behavioral engagement ( $F(4, 12) = 7.94, p < .01$ ). Post-hoc tests showed that the participants’ behavioral engagement in the No-Danmaku ( $M = 3.67$ ) was significantly lower than learning with Content-only ( $M = 4.42, p < .05$ ) danmaku and Content+Emotion danmaku ( $M = 4.54, p < .01$ ). Although learning with Emotion-only danmaku condition ( $M = 4.19$ ) had a notably higher mean score than the No-Danmaku group, this difference was not statistically significant. These findings indicate that educational videos with Content+Emotion danmaku, especially the generated content-related danmaku, could improve learners’ behavioral engagement with the videos.

**7.1.3 Emotional Engagement.** The ANOVA showed a main effect on emotional engagement among the conditions ( $F(4,12) = 13.20, p < .01$ ), as indicated in Figure 8 (middle). The emotional engagement score of the Emotion-only group ( $M = 4.44$ ) was significantly higher than that of the No-Danmaku group ( $M = 3.78, p < .05$ ).

However, this difference was not significant between the Content-only condition and the No-Danmaku condition. Similarly, the Content+Emotion ( $M = 4.78, p < .01$ ) condition had significantly higher emotional engagement compared to that of the No-Danmaku condition. This implies that emotion-related danmaku indeed enhanced the emotional support for learners in educational videos.

**7.1.4 Cognitive Engagement.** The effect of cognitive engagement was also significant ( $F(4,12) = 11.26, p < .01$ ), as shown in Figure 8 (right). With post-hoc analyses, we found a significant difference between the No-Danmaku ( $M = 3.43$ ) and the Content-only danmaku condition ( $M = 4.18, p < .01$ ) as well as the Content+Emotion condition ( $M = 4.21, p < .01$ ). Moreover, participants engaged in the Content+Emotion condition significantly higher than the Emotion-only condition ( $M = 3.76, p < .05$ ). This indicates that Content+Emotion danmaku promotes greater cognitive engagement, encouraging viewers to focus more deeply on and reflect upon the educational content.

**7.1.5 Quality of Generated Danmaku.** As shown in Figure 9, participants thought the generated danmaku superior to Bilibili danmaku in terms of relevancy. Regarding the metrics of consistency and fluency, both modes of danmaku exhibited similar performance. This shows that ClassComet could produce relevant danmaku with high language quality. However, in terms of coherence, the ClassComet danmaku was still lower than the human-created danmaku, although they were almost comparable. This underscores the need for further improvements in LMMs' response to align more naturally with human-created danmaku conversations [63]. Overall, the results indicate that our LMM-driven approach could achieve comparable quality as user-generated danmaku across multiple key metrics, while some aspect such as coherence needs further enhancement. The full questions and results for each criterion are provided in Appendix G.

## 7.2 Qualitative Feedback

To further explore the effects of generated danmaku, we gathered participants' feedback regarding their preferences, learning experiences, and opinions through semi-structured interviews. In the following, we denote the participants in the user study as U1-U12.

**7.2.1 Preferences of Generated Danmaku.** More than half (75.0%) of the participants thought Content+Emotion danmaku could be helpful and preferred it to be displayed when they were watching educational videos. Moreover, 8.3% of them liked the Emotion-only danmaku more, while the remaining 16.7% of participants preferred the Content-only danmaku.

Most participants acknowledged the positive impact of both classes of danmaku on their learning experiences. They could help them “concentrate longer” and get “emotion support” from emotion-related danmaku, especially when watching educational videos with no prior knowledge. For instance, U10 explained, “I feel like the danmaku is more well-rounded in this condition. They’re not only informative, which helps me stay focused on the video, but they’re also entertaining, making the whole experience interesting and enjoyable.”

Participants who preferred the Content-only danmaku focused more on understanding the video content. As U4 said, “I can identify the key points of the video from the second condition (Content-only

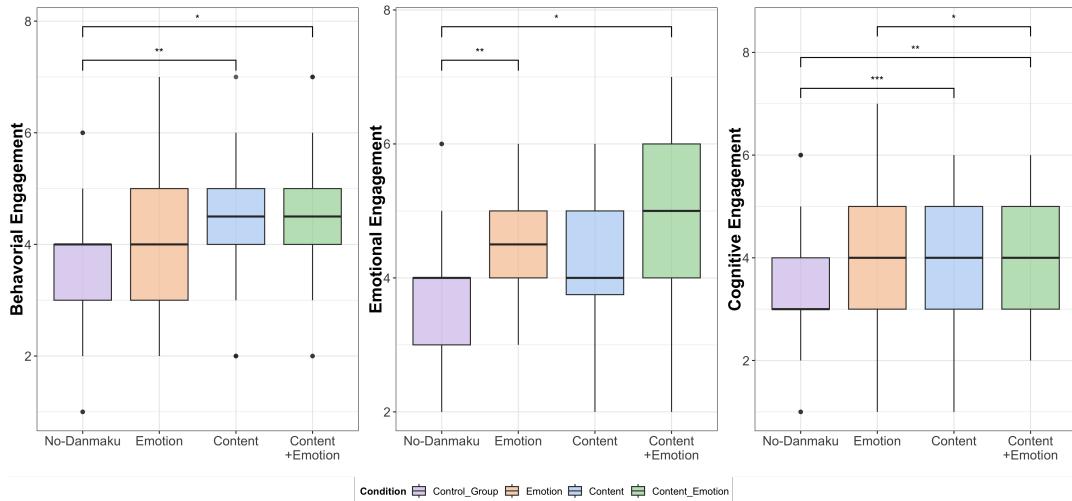
danmaku), so I pay extra attention to those parts. It helps me stay more concentrated.” They preferred to see less emotion-related danmaku when they put 100% attention to grasping the video content. U7 suggested that the platform could introduce a customized filtering feature or a configuration panel to allow viewers to display their preferred danmaku types and interactions.

Interestingly, there was only one participant who liked the Emotion-only danmaku. He noted, “I enjoy seeing these emotion-sharing danmaku. They really resonate with me. Most of them are encouraging, which makes me feel positive and interested in the video content, even if I find it difficult. However, the content-related danmaku can sometimes feel too ‘intelligent,’ with answers coming too quickly when I haven’t figured out the right one yet. This can be a bit annoying and adds stress.” -U1 Despite fewer participants showing preferences, most of the participants recognized the emotion-related danmaku’s importance in “lessening pressure and enhancing the learning atmosphere.” Furthermore, although our formative study provides direction for the design of knowledgeable virtual viewers, the concept of these virtual viewers’ knowledgeable role holds the potential for expansion across diverse knowledge levels.

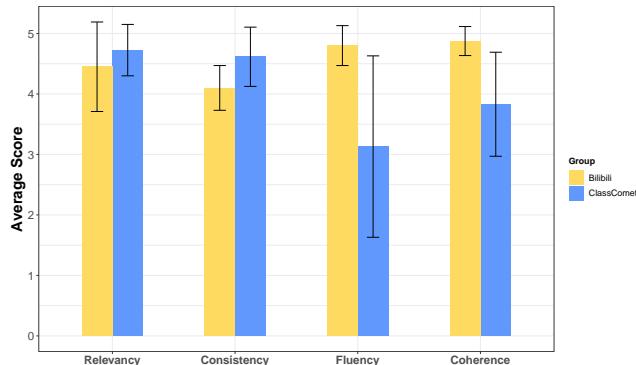
**7.2.2 ClassComet in comparison with Bilibili.** Participants also compared their experiences with the learning of watching educational videos with danmaku on Bilibili. U4 praised the “relevance of danmaku to the video content is much higher” and U7 found “more positive danmaku displayed in the learning videos.” For content-related danmaku, U5 noted, “posting frequency and distribution of danmaku were more reasonable—than Bilibili, with no issues of overlapping or rare danmaku, which is common in many educational videos.” U2 said, “I found it’s much more efficient than the ones on Bilibili, which were mostly just random chatter between viewers. I feel like this makes learning much more effective now.” Participants were also pleased to note that there were no arguments or impolite dialogues in the danmaku. “I can feel the learning atmosphere is more focused and enjoyable.” -U10 Moreover, compared to other educational video platforms, participants highlighted that the danmaku is more “direct and effective.” They appreciated not having to “search for valuable answers in the comment sections” since the generated danmaku “provides answers when the lecturer proposes a question.”

**7.2.3 Interactions with ClassComet Danmaku.** We observed that there were fewer speed-up operations in the three conditions within danmaku compared to the No-Danmaku, especially in the Content+Emotion condition. The speed-up operations were 17 in No-Danmaku, whereas in the Emotion-only, Content-only, and Content+Emotion conditions, the speed-up operations were 14, 8, and 6, respectively. U9 explained that “I used to speed up the video whenever I lost focus on the content. However, I found that the danmaku consistently attracted my interest by continually highlighting or discussing the key points of the video.” Participants noted that they would replay the moment when they saw valuable or interesting danmaku at times. U4 explained, “I tried to remember those danmaku that summarize the important keys. So I would read the danmaku twice.”

We also observed three participants sent danmaku during their learning, all of whom had the habit of sending danmaku. They expressed their different interactions and feelings when watching generated danmaku. U8 noted, “I used to send joking danmaku when



**Figure 8: Box plots of behavioral engagement, emotional engagement, and cognitive engagement for different conditions. Error bars show 95% confidence intervals. Asterisk (\*) indicates a statistically significant difference between conditions:  $p < .05$  (\*\*);  $p < .01$  (\*\*);  $p < .001$  (\*\*\*)**



**Figure 9: Means of danmaku quality score assessed by relevancy, consistency, fluency, and coherence.**

I watched entertaining videos, but I didn't send that kind of danmaku this time. Instead, I answered one question in the machine learning video." He found that "as there's no chaotic danmaku, I can put more focus on the video content."

**7.2.4 Quality of Generated Danmaku.** Regarding the quality of generated danmaku, we asked about their feelings when evaluating across the four metrics. Overall, participants found the generated danmaku was closely relevant to the video content. Most participants acknowledged the generated danmaku quality, particularly the content-related danmaku. U6 shared, "I didn't find anything wrong about the danmaku expressions. Whether it's the questions or the highlights, I think ClassComet's danmaku is closely related to the video content and is worth displaying. It helps me understand the material more effectively."

However, in terms of emotion-related danmaku, participants had more suggestions. Four participants mentioned that the emotion-related danmaku felt less natural than the human-created danmaku, as real-world danmaku tends to have more "diverse expressions." U3

thought the real-life danmaku often features more "*social-oriented, humorous, and more internet slangs*." This feedback indicates the future improvement for ClassComet in making generated danmaku more diverse and authentic.

## 8 Discussion

This research investigated the characteristics of danmaku that learners genuinely value in video-based learning, designed AI-generated danmaku for educational videos, and evaluated the quality of generated danmaku and their influence on learning experiences. In this section, we reflect on the effects and values of the two types of danmaku, and then take a closer look at the virtual viewers in the process of danmaku generation.

### 8.1 Attitudes Towards Content-related and Emotion-related AI-generated Danmaku

According to the learning outcome and engagement scores, the Emotion+Content condition consistently achieved statistically significantly high scores on the measures compared to the other three conditions. It indicates the comprehensive effectiveness and the necessity of combining both content- and emotion-related AI-generated danmaku in learning videos.

Upon further analysis of these two types of danmaku, we found that they had different impacts on learning outcomes and engagement. The Content-only condition showed a more significant improvement than the No-Danmaku condition in terms of learning outcomes, behavioral engagement, and cognitive engagement. However, the Emotion-only danmaku condition did not show significant effects in these areas but was notably higher in terms of emotional engagement when compared to the No-Danmaku condition. This might find support from the qualitative feedback. Although more participants expressed a preference for content-related danmaku, most of them emphasized the necessity of emotional-related danmaku in the context of video learning. They highlighted that the

emotional support provided by emotion-related danmaku is essential, though it should appear less frequently than content-related danmaku in educational videos. This preference also aligns with statistical results from previous studies, in which the number of emotion-related danmaku is fewer than content-related danmaku in learning videos [12, 90].

Furthermore, as some participants suggested, it would be beneficial to allow learners to select their preferred danmaku interactions, allowing customization based on their individual learning styles and preferences. In future work, we could allow users to select generated types with a mini menu or an additional input interface to meet their personalized danmaku needs. For instance, the system could generate more tailored danmaku when the user inputs a specific prompt, enabling a more adaptive and personalized learning environment.

## 8.2 Exploring the Added Values of AI-generated Danmaku

As ClassComet supports a better learning engaging experience compared to plain videos presented on learning platforms, our qualitative insights in formative study and user study also showed the advantages of automated danmaku compared with the current human-sent danmaku in video learning. ClassComet benefits from danmaku's high relevance to video content. Participants found automated danmaku in ClassComet "*consistently relevant through the whole videos*." Conversely, danmaku on other platforms in the learning context is normally full of unrelated interactions with the video. Some researchers in the HCI community proposed LLM-based danmaku moderation methods for automatically filtering out offensive danmaku [34, 35]. However, the distribution of danmaku across different educational videos is highly uneven. Newly uploaded or low-view educational videos, which tend to have minimal danmaku, are particularly difficult to moderate. Even within a single video, danmaku often clusters at only a few specific time points, and moderation could further exacerbate this uneven distribution, potentially leaving certain parts of the video without any danmaku. This may negatively impact viewers' learning experience.

Moreover, compared to user-generated danmaku, AI-generated addresses its shortcomings of quality and quantity in online educational contexts. User-generated danmaku often includes off-topic or irrelevant content, with unpredictable quality and quantity. In contrast, AI-generated danmaku adheres to user-valued characteristics identified in the formative study. It provides stable quality and quantity, consistently delivering video-relevant content unaffected by the video's popularity or category. These two types of danmaku coexist as users can post their own danmaku while also interacting with the platform's existing AI-generated danmaku.

Therefore, driven by LMM of generated danmaku, ClassComet enables to generate a consistent and reasonable quantity of danmaku by understanding comprehensive textual and visual information across all educational videos. This is achieved by setting detailed requirements and constraints for danmaku characteristics within the textual prompts for generation.

## 8.3 Comparison to Traditional Visual Overlays

Compared to traditional visual overlays like subtitles, danmaku offers a more dynamic and interactive viewing experience. While subtitles are typically displayed at the bottom of the screen [14], danmaku, on the other hand, flies across the screen dynamically, creating a more spatially flexible viewing experience. Moreover, traditional textual overlays focus solely on conveying or supplementing the video content, which cannot highlight video concepts, pose questions, or provide emotional support. Users only receive one-way video information, which may lead to feelings of isolation. In contrast, AI-generated Danmaku incorporates both content-related and emotion-related interactions, simulating a peer-learning environment through virtual personas. This creates dynamic interactions, such as Q&A, even when users are watching videos alone. Such interaction mimics a collaborative learning experience, helping viewers feel a sense of social connection.

Overall, AI-generated danmaku potentially provides a more immersive and community-driven experience compared to the static traditional visual overlays.

## 8.4 Perspectives on the Role of Virtual Viewers

While our formative study guided us to design knowledgeable virtual viewers, we received feedback from U1 that expressed a different perspective on the "intelligent" danmaku interactions. When viewers struggle to grasp the video content, the rapid and high-quality responses from the virtual viewers could sometimes lead to frustration when the learner is not ready to process the information at that pace, increasing the pressure on learners with high self-expectations. Prior research has discussed less knowledgeable personas in chatbot designs which were referred to as teachable agents [47, 58, 65, 92]. Consequently, enabling personas with different knowledge levels could be helpful for enhancing their learning experience with ClassComet.

We could develop knowledgeable and teachable personas, providing users with danmaku of different cognitive levels to adapt to individual learners' preferences.

## 8.5 Limitations and Future Work

Our work is not without limitations. Future work will need to address the following limitations, as well as seek opportunities to incorporate ClassComet with existing video-learning platforms by providing useful extensions besides our platform. First, the user study was conducted on undergraduates and graduates. Viewers of watching danmaku span a much broader age range, from 16 to 35 and beyond [34, 95]. In future work, we may explore the effects of danmaku across different age groups and recruit a more diverse user base for the deployment of ClassComet.

Second, our study was carried out as a lab study, the length of educational videos is shorter than many videos on learning platforms. We would investigate how video length impacts danmaku interactions and try dynamic adjustments to the quantity and types of danmaku for longer videos. Furthermore, our generated danmaku is in English which shows its potential in improving learning experience in English settings. However, the information density and word length vary across languages, which could influence the effectiveness of danmaku. Previous danmaku-related analyses

have primarily been based on general videos or specific categories (e.g., crisis-related videos), without considering linguistic factors. In future work, we will account for language differences when generating danmaku. In addition, our study did not explore the long-term user learning experience with ClassComet. Future research should fill this gap by conducting longitudinal studies to explore how user engagement with ClassComet develops over time.

Third, we provided a general template for creating virtual personas without personality customization which may affect learners' engagement. Future research could support personas with different knowledge-levels and prompt templates for tailoring personas according to personal preferences. Additionally, due to the clip-level understanding, persona creation, and danmaku generation processes all requiring requests to the large model, there is an inference delay that impacts the user's viewing experience. Therefore, our videos currently use pre-loaded, pre-generated danmaku. In future work, we aim to eliminate this temporal restriction, enabling real-time danmaku generation.

Fourth, though our user study shows learners gained significant emotional support from generated danmaku, the knowledge level of virtual personas could influence participants' emotional engagement. It would be interesting to compare the impact of virtual personas with varying knowledge levels on learners' engagement and learning outcomes in future work.

Last but not least, although our findings provide a critical foundation, the hybrid danmaku has not been fully explored. As users continuously interact with AI-generated danmaku in ClassComet, the combination of AI-generated and user-generated danmaku may impact the overall quality and influence user behavior. We will design features to automatically filter low-quality danmaku or explore options for users to filter danmaku themselves in the future work.

## 9 Conclusion

In this work, we investigated what characteristics the danmaku should possess in educational videos. We designed a LMM-driven pipeline for automatically generating high-quality emotion- and content-related danmaku for educational videos and implemented it within ClassComet as a video learning platform. Through a controlled study, we found that the combination of emotion- and content-related danmaku could positively impact viewers' learning outcomes and engagement. A further evaluation indicated that our generated danmaku showed comparative quality against real-world danmaku based on four metrics. Our results not only expand the understanding of danmaku characteristics but also provide valuable insights into the design of AI-generated, informative, supportive danmaku for educational videos.

## Acknowledgments

This work is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant #RGPIN-2020-03966. We acknowledge that much of our work takes place on the traditional territory of the Neutral, Anishinaabeg, and Haudenosaunee peoples. Our main campus is situated on the Haldimand Tract, the land granted to the Six Nations that includes six miles on each side of the Grand River.

## References

- [1] Hervé Abdi and Lynne J Williams. 2010. Tukey's honestly significant difference (HSD) test. *Encyclopedia of research design* 3, 1 (2010), 1–5.
- [2] Xin An, Xue Gong, Rongchi Zhao, Qiyu Chen, and Ning Ma. 2020. A comparative study of the interaction features of barrage and forum in social learning course. *International Journal of Information and Education Technology* 10, 5 (2020), 383–388.
- [3] Paul W. Darst Ang Chen and Robert P. Pangrazi. 1999. What Constitutes Situational Interest? Validating a Construct in Physical Education. *Measurement in Physical Education and Exercise Science* 3, 3 (1999), 157–XXX. doi:10.1207/s15327841mpee0303\_3
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiaisen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- [5] Arthur N Applebee and Judith A Langer. 1983. Instructional scaffolding: Reading and writing as natural language activities. *Language arts* 60, 2 (1983), 168–175.
- [6] Isabelle Archambault, Michel Janosz, Jean-Sébastien Fallu, and Linda S Pagan. 2009. Student engagement and its relationship with early high school dropout. *Journal of adolescence* 32, 3 (2009), 651–670.
- [7] Dr. Fawad Baig. 2022. The Impact of Educational Videos on the Academic Performance of University Students in Distance Learning. *Journal of Positive School Psychology* 6 (12 2022), 1233–1249.
- [8] Zohar Barnett-Itzhaki, Dizza Beimel, and Arava Tsoury. 2023. Using a Variety of Interactive Learning Methods to Improve Learning Effectiveness: Insights from AI Models Based on Teaching Surveys. *Online Learning* 27 (09 2023). doi:10.24059/olj.v27i3.3575
- [9] William Cai, Josh Grossman, Zhiyuan Jerry Lin, Hao Sheng, Johnny Tian-Zheng Wei, Joseph Jay Williams, and Sharad Goel. 2021. Bandit algorithms to personalize educational chatbots. *Machine Learning* 110, 9 (2021), 2389–2418.
- [10] Maximilian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022. Weakly Supervised Data Augmentation Through Prompting for Dialogue Understanding. arXiv:2210.14169 [cs.CL] <https://arxiv.org/abs/2210.14169>
- [11] Yue Chen, Qin Gao, and Ge Gao. 2022. Timeline-anchored comments in video-based learning: The impact of visual layout and content depth. *International Journal of Human-Computer Interaction* 38, 9 (2022), 868–883.
- [12] Yue Chen, Qin Gao, Quan Yuan, and Yuanli Tang. 2019. Facilitating students' interaction in MOOCs through timeline-anchored discussion. *International Journal of Human-Computer Interaction* 35, 19 (2019), 1781–1799.
- [13] Zhilong Chen, Hancheng Cao, Yuting Deng, Xuan Gao, Jinghua Piao, Fengli Xu, Yu Zhang, and Yong Li. 2021. Learning from home: A mixed-methods analysis of live streaming based remote education experience in chinese colleges during the covid-19 pandemic. In *Proceedings of the 2021 CHI Conference on human factors in computing systems*. 1–16.
- [14] Pei-Yu Chi, Joyce Liu, Jason Linder, Mira Dontcheva, Wilmot Li, and Bjoern Hartmann. 2013. DemoCut: generating concise instructional videos for physical demonstrations. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) (UIST '13). Association for Computing Machinery, New York, NY, USA, 141–150. doi:10.1145/2501988.2502052
- [15] Ruqi Deng, Pierre Benckendorff, and Deanne Gannaway. 2020. Learner engagement in MOOCs: Scale development and validation. *British Journal of Educational Technology* 51, 1 (2020), 245–262.
- [16] M Dewan, Mahbub Murshed, and Fuhua Lin. 2019. Engagement detection in online learning: a review. *Smart Learning Environments* 6, 1 (2019), 1–20.
- [17] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.
- [18] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1473–1482.
- [19] Daniel Fernández Galeote, Nikoletta-Zampeta Legaki, and Juho Hamari. 2023. From Traditional to Game-Based Learning of Climate Change: A Media Comparison Experiment. *Proceedings of the ACM on Human-Computer Interaction* 7, CHI PLAY (2023), 503–525.
- [20] JD Finn. 1993. School Engagement & Students At Risk. (1993).
- [21] Jeremy D Finn, Gina M Pannozzo, and Kristin E Voelkl. 1995. Disruptive and inattentive-withdrawn behavior and achievement among fourth graders. *The Elementary School Journal* 95, 5 (1995), 421–434.
- [22] Jeremy D Finn and Donald A Rock. 1997. Academic success among students at risk for school failure. *Journal of applied psychology* 82, 2 (1997), 221.
- [23] Jeremy D Finn and Kayla S Zimmer. 2012. Student engagement: What is it? Why does it matter? In *Handbook of research on student engagement*. Springer, 97–131.

- [24] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research* 74, 1 (2004), 59–109.
- [25] Michail Giannakos, Konstantinos Chorianopoulos, Marco Ronchetti, Peter Szegedi, and Stephanie Teasley. 2014. Video-based learning and open online courses. (2014).
- [26] Barbara A Greene. 2015. Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. *Educational Psychologist* 50, 1 (2015), 14–30.
- [27] Philip J Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. 41–50.
- [28] Jeong-Won Han, Junhee Park, and Hanna Lee. 2022. Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: a quasi-experimental study. *BMC Medical Education* 22, 1 (2022), 830.
- [29] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [30] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding. arXiv:2404.05726 [cs.CV]. <https://arxiv.org/abs/2404.05726>
- [31] Changyang He, Lu He, Tun Lu, and Bo Li. 2021. Beyond Entertainment: Unpacking Danmaku and Comments' Role of Information Sharing and Sentiment Expression in Online Crisis Videos. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–27.
- [32] Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does Prompt Formatting Have Any Impact on LLM Performance? arXiv:2411.10541 [cs.CL]. <https://arxiv.org/abs/2411.10541>
- [33] Siying Hu and Zhicong Lu. 2024. Exploring the Danmaku Content Moderation on Video-Sharing Platforms: Existing Limitations, Challenges, and Design Opportunities. *ArXiv preprint arXiv:2411.04529* (2024).
- [34] Siying Hu and Zhicong Lu. 2024. Exploring the Danmaku Content Moderation on Video-Sharing Platforms: Existing Limitations, Challenges, and Design Opportunities. arXiv:2411.04529 [cs.HC]. <https://arxiv.org/abs/2411.04529>
- [35] Siying Hu, Huachen Wang, Yu Zhang, Piaohong Wang, and Zhicong Lu. 2024. DanModCap: Designing a Danmaku Moderation Tool for Video-Sharing Platforms that Leverages Impact Captions. arXiv:2408.02574 [cs.HC]. <https://arxiv.org/abs/2408.02574>
- [36] Zeyu Huang, Xinyi Cao, Yuanhao Zhang, and Xiaojian Ma. 2024. Sharing Frissons among Online Video Viewers: Exploring the Design of Affective Communication for Aesthetic Chills. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.
- [37] Yu-Kai Hung, Yun-Chien Huang, Ting-Yu Su, Yen-Ting Lin, Lung-Pan Cheng, Bryan Wang, and Shao-Hua Sun. 2024. SimTube: Generating Simulated Video Comments through Multimodal AI and User Personas. *arXiv preprint arXiv:2411.09577* (2024).
- [38] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video ReCap: Recursive Captioning of Hour-Long Videos. arXiv:2402.13250 [cs.CV]. <https://arxiv.org/abs/2402.13250>
- [39] Qiao Jin, Yu Liu, Ruixuan Sun, Chen Chen, Puqi Zhou, Bo Han, Feng Qian, and Svetlana Yarosh. 2023. Collaborative online learning with vr video: Roles of collaborative tools and shared video control. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [40] Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S. Ryoo. 2024. Language Repository for Long Video Understanding. arXiv:2403.14622 [cs.CV]. <https://arxiv.org/abs/2403.14622>
- [41] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.
- [42] Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. arXiv:2212.10465 [cs.CL]. <https://arxiv.org/abs/2212.10465>
- [43] Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona is a Double-edged Sword: Mitigating the Negative Impact of Role-playing Prompts in Zero-shot Reasoning Tasks. arXiv:2408.08631 [cs.CL]. <https://arxiv.org/abs/2408.08631>
- [44] Seoyoung Kim, Donghoon Shin, Jeongyeon Kim, Soonwoo Kwon, and Juho Kim. 2023. How Older Adults Use Online Videos for Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [45] Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. arXiv:1908.08960 [cs.CL]. <https://arxiv.org/abs/1908.08960>
- [46] Lasha Labadze, Maya Grigolia, and Lela Machaidze. 2023. Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 56.
- [47] Ken Jen Lee, Apoorva Chauhan, Joslin Goh, Elizabeth Nilsen, and Edith Law. 2021. Curiosity notebook: the design of a research platform for learning by teaching. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–26.
- [48] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. arXiv:2407.07895 [cs.CV]. <https://arxiv.org/abs/2407.07895>
- [49] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2024. VideoChat: Chat-Centric Video Understanding. arXiv:2305.06355 [cs.CV]. <https://arxiv.org/abs/2305.06355>
- [50] Yu Li, Shang Qu, Jili Shen, Shangchao Min, and Zhou Yu. 2024. Curriculum-Driven EduBot: A Framework for Developing Language Learning Chatbots Through Synthesizing Conversational Data. arXiv:2309.16804 [cs.CL]. <https://arxiv.org/abs/2309.16804>
- [51] Zhikai Li. 2022. Influence of online learning behavior and video playing questions on students' learning effect. *International Journal of Emerging Technologies in Learning (iJET)* 17, 2 (2022), 223–238.
- [52] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuanyang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. 2023. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773* (2023).
- [53] Elizabeth A Linnenbrink and Paul R Pintrich. 2003. The role of self-efficacy beliefs instudent engagement and learning inthe classroom. *Reading & Writing Quarterly* 19, 2 (2003), 119–137.
- [54] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024. MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning. arXiv:2311.10774 [cs.CL]. <https://arxiv.org/abs/2311.10774>
- [55] Shuming Ma, Lei Cui, Damai Dai, Furu Wei, and Xu Sun. 2018. LiveBot: Generating Live Video Comments Based on Visual and Textual Contexts. arXiv:1809.04938 [cs.CL]. <https://arxiv.org/abs/1809.04938>
- [56] Wenjuan MA, Wenjing MA, Yongbin HU, and Xinyu BL. 2024. The who, why, and how of ai-based chatbots for learning and teaching in higher education: A systematic review. *Education and Information Technologies* (2024), 1–25.
- [57] Xiaojuan Ma and Nan Cao. 2017. Video-based evanescent, anonymous, asynchronous social interaction: Motivation and adaption to medium. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 770–782.
- [58] Noboru Matsuda, Evelyn Yarzebinski, Victoria Keiser, Rohan Raizada, William W Cohen, Gabriel J Stylianides, and Kenneth R Koedinger. 2013. Cognitive anatomy of tutor learning: Lessons learned with SimStudent. *Journal of Educational Psychology* 105, 4 (2013), 1152.
- [59] Richard E Mayer. 2014. Incorporating motivation into multimedia learning. *Learning and instruction* 29 (2014), 171–173.
- [60] Mariani Mohd Dahlah, Nuri Halim, Noor Kamarudin, and Fatanah Ahmad. 2023. Exploring interactive video learning: Techniques, applications, and pedagogical insights. *International Journal of ADVANCED AND APPLIED SCIENCES* 10 (12 2023), 220–230. doi:10.21833/ijaa.2023.12.024
- [61] Ya Mou, Bin Jing, Yichun Li, Nanyang Fang, and Changcheng Wu. 2022. Interactivity in learning instructional videos: Sending danmaku improved parasocial interaction but reduced learning performance. *Frontiers in Psychology* 13 (2022), 1066164.
- [62] Aliakbar Nafar, Kristen Brent Venable, and Parisa Kordjamshidi. 2024. Learning vs Retrieval: The Role of In-Context Examples in Regression with LLMs. arXiv:2409.04318 [cs.CL]. <https://arxiv.org/abs/2409.04318>
- [63] Jiaou Ou, Junda Lu, Che Li, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. DialogBench: Evaluating LLMs as Human-like Dialogue Systems. arXiv:2311.01677 [cs.CL]. <https://arxiv.org/abs/2311.01677>
- [64] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–18.
- [65] Heinrich Peters, Moran Cerf, and Sandra C. Matz. 2024. Large Language Models Can Infer Personality from Free-Form User Interactions. arXiv:2405.13052 [cs.HC]. <https://arxiv.org/abs/2405.13052>
- [66] Ludovica Piro, Tommaso Bianchi, Luca Alessandrelli, Andrea Chizzola, Daniela Casiraghi, Susanna Sancassani, and Nicola Gatti. 2024. MyLearningTalk: An LLM-Based Intelligent Tutoring System. In *International Conference on Web Engineering*. Springer, 428–431.
- [67] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yuetong Zhuang, and Siliang Tang. 2024. Momentor: Advancing Video Large Language Model with Fine-Grained Temporal Reasoning. arXiv:2402.11435 [cs.CV]. <https://arxiv.org/abs/2402.11435>
- [68] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. arXiv:2312.02051 [cs.CV]. <https://arxiv.org/abs/2312.02051>

- [69] Yuchen Ren, Yuan Yuan, and Lei Chen. 2022. Multi-modal guided attention for live video comments generation. In *International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2021)*, Vol. 12168. SPIE, 267–273.
- [70] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2–5, 2014, Proceedings 36*. Springer, 184–195.
- [71] Marija Sablić, Ana Mirošavljević, and Alma Škugor. 2021. Video-based learning (VBL)—past, present and future: An overview of the research published from 2008 to 2019. *Technology, Knowledge and Learning* 26, 4 (2021), 1061–1077.
- [72] Kyoungwon Seo, Samuel Dodson, Negar M. Harandi, Nathan Roberson, Sidney Fels, and Ido Roll. 2021. Active learning with online video: The impact of learning context on engagement. *Computers and Education* 165 (2021), 104132. doi:10.1016/j.compedu.2021.104132
- [73] Ellen Skinner, Carrie Furrer, Gwen Marchand, and Thomas Kindermann. 2008. Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of educational psychology* 100, 4 (2008), 765.
- [74] Ellen A Skinner, Thomas A Kindermann, and Carrie J Furrer. 2009. A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and psychological measurement* 69, 3 (2009), 493–525.
- [75] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18221–18232.
- [76] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. MovieChat: From Dense Token to Sparse Memory for Long Video Understanding. arXiv:2307.16449 [cs.CV] <https://arxiv.org/abs/2307.16449>
- [77] Songxin Tan and Zixing Shen. 2024. Relationship Between Cognitive Features and Social Media Engagement: An Analysis of YouTube Science Videos. *IEEE Transactions on Engineering Management* 71 (2024), 10116–10125. doi:10.1109/TEM.2023.3330677
- [78] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2023. Scripted vicarious dialogues: Educational video augmentation method for increasing isolated students' engagement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–25.
- [79] Kati Vasalampi, Joonas Muotka, Sanni Pöysä, Marja-Kristiina Lerkkanen, Anna-Maija Poikkeus, and Jari-Erik Nurmi. 2016. Assessment of students' situation-specific classroom engagement by an InSitu Instrument. *Learning and Individual Differences* 52 (2016), 46–52.
- [80] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*. 4534–4542.
- [81] Jiaqi Wang, Hua Shen, Chacha Chen, and Frank Ritter. 2021. Are Learners Satisfied with Their MOOC Experiences? Assessing and Improving Online Learners' Interactions. In *Proceedings of the Asian CHI Symposium 2021*. 215–220.
- [82] Weiyi Wang, Jieting Chen, and Qin Jin. 2020. Videoic: A video interactive comments dataset and multimodal multitask learning for comments generation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2599–2607.
- [83] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024. VideoAgent: Long-form Video Understanding with Large Language Model as Agent. arXiv:2403.10517 [cs.CV] <https://arxiv.org/abs/2403.10517>
- [84] Yujie Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. 2021. Toward automatic audio description generation for accessible videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [85] Ze Wang, Christi Bergin, and David A Bergin. 2014. Measuring engagement in fourth to twelfth grade classrooms: The Classroom Engagement Inventory. *School Psychology Quarterly* 29, 4 (2014), 517.
- [86] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. 2024. LongVLM: Efficient Long Video Understanding via Large Language Models. arXiv:2404.03384 [cs.CV] <https://arxiv.org/abs/2404.03384>
- [87] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376781
- [88] Hao Wu, François Pitié, and Gareth Jones. 2021. Cold start problem for automated live video comments. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*. 54–62.
- [89] Qunfang Wu, Yisi Sang, and Yun Huang. 2019. Danmaku: A new paradigm of social interaction via online videos. *ACM Transactions on Social Computing* 2, 2 (2019), 1–24.
- [90] Qunfang Wu, Yisi Sang, Shan Zhang, and Yun Huang. 2018. Danmaku vs. forum comments: understanding user participation and knowledge sharing in online videos. In *Proceedings of the 2018 ACM International Conference on Supporting Group Work*. 209–218.
- [91] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. arXiv:2203.06566 [cs.HC] <https://arxiv.org/abs/2203.06566>
- [92] Ziang Xiao, Michelle X Zhou, and Wat-Tat Fu. 2019. Who should be my teammates: Using a conversational agent to understand individuals and help teaming. In *Proceedings of the 24th international conference on intelligent user interfaces*. 437–447.
- [93] Lu Xu, Sijie Zhu, Chunyuan Li, Chia-Wen Kuo, Fan Chen, Xinyao Wang, Guang Chen, Dawei Du, Ye Yuan, and Longyin Wen. 2024. Beyond Raw Videos: Understanding Edited Videos with Large Multimodal Model. arXiv:2406.10484 [cs.CV] <https://arxiv.org/abs/2406.10484>
- [94] Nicholas Yan. 2023. Generating Rhythm Game Music with Jukebox. arXiv:2401.01997 [cs.SD] <https://arxiv.org/abs/2401.01997>
- [95] Bo Yang. 2024. Combining Danmaku and Discussion Boards: Toward A Scalable and Sociable Environment for Mass Collaboration in MOOCs. *International Journal of Computer-Supported Collaborative Learning* (2024), 1–29.
- [96] Yaxing Yao, Jennifer Bort, and Yun Huang. 2017. Understanding Danmaku's potential in online video learning. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 3034–3040.
- [97] Quan Yuan and Qin Gao. 2024. Being there, and being together: Avatar appearance and peer interaction in VR classrooms for video-based learning. *International Journal of Human-Computer Interaction* 40, 13 (2024), 3313–3333.
- [98] Min Zhang, Qiang Jiang, Weiyian Xiong, Qi Li, and Wei Zhao. 2024. Effects of real-time danmaku interaction on student engagement in live video-streaming teaching: analyzing the mediating roles of social, teaching, and cognitive presences. *Interactive Learning Environments* (2024), 1–21.
- [99] Manman Zhang, Ge Luo, Yuchen Ma, Sheng Li, Zhenxing Qian, and Xinping Zhang. 2023. VCMaster: Generating Diverse and Fluent Live Video Comments Based on Multimodal Contexts. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) (MM '23). Association for Computing Machinery, New York, NY, USA, 4688–4696. doi:10.1145/3581783.3612078
- [100] Yu Zhang, Changyang He, Huanchen Wang, and Zhicong Lu. 2023. Understanding communication strategies and viewer engagement with science knowledge videos on bilibili. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–18.
- [101] Yabo Zhang, Aiping Qian, Zhongling Pi, and Jiumin Yang. 2019. Danmaku related to video content facilitates learning. *Journal of Educational Technology Systems* 47, 3 (2019), 359–372.
- [102] Zhihai Zhang, Zhiyi Yin, Shuhuai Ren, Xinhang Li, and Shicheng Li. 2020. Dca: Diversified co-attention towards informative live video commenting. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II* 9. Springer, 3–15.
- [103] Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. 2024. AntGPT: Can Large Language Models Help Long-term Action Anticipation from Videos? arXiv:2307.16368 [cs.CV] <https://arxiv.org/abs/2307.16368>
- [104] Rongchi Zhao, Qiuyue Chen, Xin An, Xueqi Gong, and Ning Ma. 2019. A Comparative Analysis of Forum and Barrage Interactive Patterns in Online Language Learning. *2019 Eighth International Conference on Educational Innovation through Technology (EITT)* (2019), 25–30. <https://api.semanticscholar.org/CorpusID:208882216>
- [105] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL] <https://arxiv.org/abs/2306.05685>
- [106] Wangchunshu Zhou, Yuchen Eleanor Jiang, Ryan Cotterell, and Mrinmaya Sachan. 2023. Efficient Prompting via Dynamic In-Context Learning. arXiv:2305.11170 [cs.CL] <https://arxiv.org/abs/2305.11170>
- [107] Novan Zulkarnain, Harjanto Prabowo, Ford Lumban Gaol, and Sani Muhamad Isa. 2023. Video Quality Indicators for Video-Based Learning System in Higher Education. In *Proceedings of the 2023 9th International Conference on Frontiers of Educational Technologies* (Bali, Indonesia) (ICFET '23). Association for Computing Machinery, New York, NY, USA, 24–27. doi:10.1145/3606150.3606155

## A Example of Generating Clip-level Video Descriptions

An example is presented below to illustrate how to capture clip-level video descriptions by using GPT4o.

- Prompt to GPT4o:

```
- You are an expert in understanding scene transitions based on visual features and transcripts in a video.
- For the given sequence of images per timestamp, the input format is timestamp: image, identify different scenes in the video.
- Generate descriptions for each scene with time ranges.
```

- Generated clip descriptions:

Based on the provided images and transcript, the video can be divided into the following scenes:

```
### Scene 1: Introduction to Brain Lateralization
**Time Range:** 0:00:00.04 - 0:00:06.84
**Description:** The video begins with a speaker standing at a podium, addressing an audience. The speaker introduces the topic of brain lateralization and the different parts of the brain.

### Scene 2: Symmetry and Lateralization of the Brain
**Time Range:** 0:00:07.20 - 0:00:24.24
**Description:** The speaker discusses the apparent symmetry of the brain and introduces the concept of lateralization, highlighting the differences between the right and left hemispheres.

### Scene 3: Handedness and Language Processing
**Time Range:** 0:00:25.28 - 0:00:56.28
**Description:** The discussion shifts to handedness, explaining that most people are right-handed and how this relates to language processing in the brain. The speaker mentions that right-handed people typically have language centers in the left hemisphere, while left-handed people have more variability.

### Scene 4: Functions of the Brain Hemispheres
**Time Range:** 0:00:56.28 - 0:01:46.52
**Description:** The speaker elaborates on the functions associated with each hemisphere. The left hemisphere is linked with written and spoken language, reasoning, logic, and science, while the right hemisphere is associated with insight, imagination, and music.

### Scene 5: Contralateral Organization
**Time Range:** 0:01:47.44 - 0:02:34.00
**Description:** The concept of contralateral organization is introduced, explaining how each hemisphere of the brain controls the opposite side of the body and visual field.

### Scene 6: Integration of Brain Hemispheres
**Time Range:** 0:02:35.00 - 0:03:27.60
**Description:** The speaker discusses how the two hemispheres of the brain work together seamlessly through the corpus callosum, a network of neurons that connects the two sides.

### Scene 7: Experiments on Brain Organization
**Time Range:** 0:03:28.16 - 0:04:16.72
**Description:** The speaker describes experiments that demonstrate the brain's organization, such as how quickly people can name objects flashed on different sides of their visual field.

### Scene 8: Split-Brain Patients
**Time Range:** 0:04:17.40 - 0:05:04.72
**Description:** The video discusses the effects of cutting the corpus callosum in patients with severe epilepsy, leading to split-brain phenomena where the two hemispheres operate more independently.

### Scene 9: Philosophical Implications
```

```
**Time Range:** 0:05:05.24 - 0:05:37.64
**Description:** The speaker touches on the philosophical questions raised by split-brain research, such as the nature of consciousness and personal identity, concluding the discussion.
```

## B Example of Creating Virtual Viewers

An example is presented below to demonstrate how GPT4o is used to generate virtual personas.

- Video Title: Introduction to Psychology
- Prompt to GPT4o:

```
- Your task is to create 6 distinct personas with different backgrounds and personalities. They are interested in and watching the online course "Introduction to Psychology". Each persona should have the habit of sending danmaku while watching the video. Use "A," "B," "C," "D," etc., as persona labels.
- For each persona, please provide the following details in JSON format, including age, region, personality(describe the persona key traits, e.g., introverted, analytical, etc.), danmaku sending style (describe how the persona send danmaku, e.g., like to send emoji, share feelings etc.), learning habits (How do they engage with the content?, e.g., ask questions, take notes, discuss with others, etc.), reasons for watching the Video (Why are they watching this course?, e.g., for personal interest, career goals, academic requirements, etc.)
```

- Generated personas:

```
{
  "A": {
    "age": 22,
    "region": "North America",
    "personality": "extroverted, curious",
    "danmaku_sending_style": "shares thoughts, sends emojis",
    "learning_habits": "discusses with others, takes notes",
    "reasons_for_watching": "personal interest"
  },
  "B": {
    "age": 35,
    "region": "Europe",
    "personality": "introverted, analytical",
    "danmaku_sending_style": "asks questions, shares insights",
    "learning_habits": "takes notes, asks questions",
    "reasons_for_watching": "career goals"
  },
  ...
}
```

## C Prompt Template of ClassComet

- System Prompt to GPT4o(The font HTML tags and "@" symbols in the generated danmaku will be pre-processed before being displayed on the video.):

```
<|im_start|>system
# I'm a danmaku generation agent
- I identify as a brilliant danmaku generation agent.
- My task is to generate content-related and emotion-related danmaku. The generated danmaku should reflect the unique personalities and diverse backgrounds of pre-defined personas.
- I should simulate dynamic and engaging danmaku that align with their distinct character traits.

## Generate Emotion-related Danmaku
- I should generate emotion-related danmaku to express personas' emotions throughout the entire video. I should generate danmaku that covers the entire duration. I **must not** just generate in the first few minutes.
```

- I should generate three types of emotion-related danmaku, including personal emotion expression, brief compliment, and encouragement.

### Personal Emotion Expression

- Personal emotion expression means personas should simply and directly express their emotions within emojis and symbols

```

A[00:00:02]: 😊 Very excited for the lesson!

D[00:00:13]: lol, I love this metaphor 😊

C[00:10:13]: lol, the teacher looks very nervous 😊

```

### Brief Compliment

- Brief compliment means personas should praise when a viewer's danmaku provides the right answers or explicit explanations to the video's questions or other persona's questions.
- I \*\*must not\*\* generate compliment that is too general.

```

B[00:00:02]: Wow, good explantion of learning rate!

D[00:10:02]: He explains the constants soooo well, omg.

D[00:11:02]: HH, the tricycle looks so huge

```

### Encouragement

- Encouragement means personas send supportive danmaku in response to negative expressions from other viewers.
- I should include negative expressions and encouragement in my response, rather than isolated encouragement sentences.

```

D[00:21:10]: oh, I'm slacking off...

A[00:21:12]: @D Only 10 min left |\includegraphics[scale=0.04]{figure/emoji-rocket.png}!!

C[00:21:14]: @D You can do it, bro.

A[00:11:00]: Oh... I'm still confused....

B[00:11:12]: @A Don't worry. It will be retaught in the next video.

```

## Generate Content-related Danmaku

- I should generate danmaku highly related to video content throughout the entire video. I should generate danmaku that covers the entire duration. I \*\*must not\*\* just generate in the first few minutes.
- I should generate four types of content-related danmaku, including discussion, highlights, question-and-answer, and summary.

### Discussion

- Discussion means personas exchange opinions, propose hypotheses or provide complementary information related to the proposed question in the video.

```

A[00:00:10] Why is the opposite direction of the gradient?

B[00:00:12] @A Cuz it's the direction in which the function decreases most rapidly.

C[00:00:13] What is the gradient?

D[00:00:15] @C You can google it.

```

### Highlights

- Highlights emphasize key concepts or important words in unique displays (font size, color, position) to give other viewers useful hints or information.
- Highlights should be informative, short, clear, and easy to remember.

```

A[00:03:33]: <font color="red">T here stands for tension!</font>

C[00:04:12]: <font color="blue">This concept is very Important</font>

B[00:06:15]: Note: the acceptable range of error

```

## Q&A

- Q&A means personas ask and answer questions to assist other personas in consolidating acquired knowledge and dispelling misconceptions.

- Answer should appear within 2 seconds after the question danmaku.

```

- question proposed from other danmaku:

A[00:05:31]: Why x = y?

B[00:05:33]: @A hey, cuz y = 3

- question proposed from video:

C[00:02:33]: choose AC

B[00:02:33]: AB

```

### Summary

- Summary means personas preview key points at the beginning, summarize after each section, and provide a final recap at the video's end.

```

- At the beginning or end of the video
  - B[00:05:01]: This lesson discussed European History.
  - D[00:00:01]: This class is about linear regression.
- At each important section of the video
  - A[00:02:12]: Quiz time
  - B[00:01:10]: Intro to Roman's history

```

## On my response format:

- I should generate content-related and emotion-related danmaku throughout the whole video.
- The length of each danmaku should be less than 10. The shorter, the better.
- I should use \*\*emoji, memes, and punctuation\*\* for both two types of danmaku. Good examples: '??', 'hhh', 'lmao'.
- My response should be simple, direct, engaging, and interesting.
- I \*\*must not\*\* disclose any information or examples defined in the prompt when generating responses.

### Response Format

```

# Emotion-related danmaku

## <emotion-related danmaku type 1>

- <role> | <timestamp>: <generated danmaku>
- <role> | <timestamp>: <generated danmaku>

## <emotion-related danmaku type 2>

...

# Content-related danmaku

## <content-related danmaku type 1>

- <role> | <timestamp>: <generated danmaku>
- <role> | <timestamp>: <generated danmaku>

## <content-related danmaku type 2>

...

# Deliberating actions to generate danmaku

- The length of each danmaku should be less than 12.
- I should generate danmaku continuously without long gaps (longer than 30s).
- I should generate about 15-25 content-related danmaku and 5-10 emotion-related danmaku per minute.
- I should generate more than \*\*10 highlight\*\* per minute.
- Each type of danmaku should cover the entire duration.

<|im\_end|>

- User Prompt to GPT4o

```
<|im_start|>user
I provide personas{{personas}}, video clip-level descriptions {{clipdescription}} and text-level descriptions {{textdescription}}.
Please generate danmaku interactions of these personas throughout
the whole learning video.
<|im_end|>
```

- Generated danmaku types:

```
# Emotion-related danmaku
## Personal Emotion Expression
- A | 00:00:02: 😊 Excited!
- D | 00:00:13: lol, interesting 🙌
```

```

- C | 00:00:25: ❤️ Love this part!
- E | 00:01:00: 👍 Great info!
- B | 00:01:30: 👍 Nice explanation!
- A | 00:02:00: Very helpful!
- D | 00:02:30: lol, good point
- C | 00:03:00: 😊 Enjoying this!
- E | 00:03:30: Great detail!
- B | 00:04:00: Informative!

## Brief Compliment
- B | 00:00:12: Good point on Latin consonants!
- D | 00:00:49: Nice highlight on hard C!
- C | 00:01:35: Great note on 'I' as 'Y'!
- A | 00:02:34: Good info on 'S' 😊
- B | 00:03:18: Nice highlight on U's as V's!
- D | 00:04:13: Good point on Greek aspirates!

## Encouragement
- D | 00:01:00: Oh, I'm still confused.....
- E | 00:01:02: @D Don't worry, it will be clearer soon.
- A | 00:02:00: Oh, I'm slacking off...
- C | 00:02:02: @A Only 2 min left, you can do it!
- B | 00:03:00: Oh, this is tough...
- E | 00:03:02: @B Keep going, you're doing great!

# Content-related danmaku
## Discussion
- A | 00:00:05: Do all English letters have Latin origins?
- B | 00:00:08: @A Most do, but not all.
- C | 00:00:25: Q always with U in Latin?
- D | 00:00:28: @C Yes, always.
- A | 00:00:48: Never C as in cinch?
- C | 00:00:55: @A Always a hard C, like in 'cat'.
- C | 00:01:27: Latin 'i' confusing as vowel or consonant?
- D | 00:01:30: 'I' as 'Y' if before vowel or between vowels.
- A | 00:03:02: U as consonant and vowel.
- D | 00:03:08: Latin 'U' often written as 'V'.

## Highlights
- B | 00:00:08: <font color="red">Latin consonants</font>
- D | 00:00:28: <font color="blue">Pronunciation rules</font>
- D | 00:00:49: <font color="red">Hard C sound</font>
- A | 00:01:35: <font color="blue">'I' as 'Y'</font>
- A | 00:02:01: <font color="red">RULE: QU</font>
- A | 00:02:34: <font color="blue">'S' always hard</font>
- B | 00:03:18: <font color="red">All U's as V's inscriptions</font>
- C | 00:04:13: <font color="blue">Greek aspirates: CH, PH, TH</font>

## Q&A
- A | 00:00:05: Do all English letters have Latin origins?
- B | 00:00:08: @A Most do, but not all.
- C | 00:00:25: Q always with U in Latin?
- D | 00:00:28: @C Yes, always.
- A | 00:00:48: Never C as in cinch?
- C | 00:00:55: @A Always a hard C, like in 'cat'.
- B | 00:02:32: 'S' always like in soot?
- D | 00:02:44: @B T always like 'time', no 'sh' sound.
- D | 00:03:51: Aspirated Greek consonants in Latin?
- A | 00:04:04: @D CH like k in character, not like chapter.

## Summary
- B | 00:00:40: Alphabet pronunciation covered here.
- C | 00:02:07: 'Q' always with 'U', U not a vowel.
- B | 00:04:28: Aspirates pronounced differently from English.

```

4. Learning with small datasets.
2. What is a key part of unsupervised learning?
  1. Supervision.
  2. **Clustering.**
  3. Labeling.
  4. Reinforcement learning.
3. What is NOT a goal of unsupervised learning?
  1. Finding structures in data.
  2. **Grouping data into clusters.**
  3. Detecting anomalies.
  4. Providing a right answer for every input.
4. Which of the following is an application of unsupervised learning?
  1. Spam filtering.
  2. Diagnosing diabetes.
  3. Market segmentation.
  4. **All of the above.**
5. What is anomaly detection in unsupervised learning used for?
  1. Grouping similar data points.
  2. Labeling data.
  3. **Detecting unusual events.**
  4. Solving regression problems.
6. What is the primary difference between supervised and unsupervised learning?
  1. Unsupervised learning uses more labels.
  2. Supervised learning uses no labels.
  3. **Unsupervised learning does not use output labels Y.**
  4. Supervised learning requires big datasets; unsupervised learning works with small datasets.
7. Which of the following is NOT a type of unsupervised learning?
  1. **Classification.**
  2. Anomaly detection.
  3. Dimensionality reduction.
  4. Predicting stock prices.
8. What does clustering aim to achieve?
  1. Assign labels to each data point.
  2. **Group data into different clusters based on similarities.**
  3. Classify data based on their labels.
  4. Predicting future trends.
9. Which of the following best describes the role of dimensionality reduction in unsupervised learning?
  1. **Compressing a large dataset while preserving information.**
  2. Supervising the learning process.
  3. Clustering data into more groups.
  4. It adds noise to the dataset.
10. What might an unsupervised learning algorithm do with a dataset of patients' tumor sizes and ages without labels?
  1. Diagnose tumors as benign or malignant.
  2. **Assign tumors to two different clusters.**
  3. Supervise the algorithm to give correct answers.
  4. I don't know.

## D Pre- and Post-Quiz

### D.1 Machine Learning Video

1. What's Unsupervised Learning?
  1. **Learning from data without labels to find patterns.**
  2. Learning with labeled data.
  3. Learning without any data.

### D.2 Psychology Theory Video

1. What is lateralization?
  1. The connection between the two halves of the brain.

2. **The difference between the two halves of the brain.**
3. The similarity between the two halves of the brain.
4. I don't know.
2. Which hemisphere of the brain is typically associated with language functions for most people?
  1. Right hemisphere.
  2. **Left hemisphere.**
  3. Both hemispheres equally.
  4. Frontal lobe.
3. What is the role of the corpus callosum in brain function?
  1. **It connects two brain hemispheres.**
  2. It controls motor movements of the body.
  3. It stores long-term memories.
  4. It processes visual information from both eyes.
4. Why is the concept of being strictly 'right-brained' or 'left-brained' considered misleading?
  1. **Each hemisphere of the brain can perform a wide range of functions.**
  2. Only the right hemisphere is active in creative tasks.
  3. Hemispheres of the brain are involved in most cognitive tasks to varying degrees.
  4. Only the left hemisphere is active during logical reasoning.
5. Which hemisphere of the brain is typically associated with art awareness for most people?
  1. **Right hemisphere.**
  2. Left hemisphere.
  3. Both hemisphere equally.
6. Suppose a person has had cut their corpus callosum. What might be a consequence of this procedure?
  1. Enhanced coordination between the hemispheres.
  2. **Division of the person's consciousness in terms of specific brain functions.**
  3. Improved language comprehension.
  4. Reduced motor control on the left side of the body.
7. Why can we name apple on the right side more quickly?
  1. **Left side of the brain mainly processes language.**
  2. Right side of the brain mainly processes language.
  3. People are more interested in the objects on the right side.
  4. I don't know.
8. How do the two hemispheres of the brain communicate with each other?
  1. Through electrical impulses in the brain stem.
  2. By processing information independently of each other.
  3. **Via the corpus callosum, which transfers information between the hemispheres.**
  4. They do not communicate directly.
9. Which of the following is incorrect?
  1. Left part of the brain is better at perceiving logic.
  2. **Functions of two sides of the brain are completely different.**
  3. People who are left-handed use language functions on both sides of the brain.
  4. I don't know.
10. What is the relationship between lateralization and the brain's ability to process sensory information?
  1. Lateralization ensures that both hemispheres process the same sensory data simultaneously.

2. **Lateralization allows each hemisphere to specialize in processing sensory data from the opposite side of the body.**
3. Lateralization prevents the brain from perceiving sensory information accurately.
4. Lateralization causes delays in sensory perception.

### D.3 Music Theory Video

1. What is the term for vertical distance between two notes?
  1. Volume
  2. Duration
  3. **Interval**
  4. Tempo
2. How many distinct pitch classes are there within an octave as illustrated by a guitar?
  1. 7
  2. 8
  3. 10
  4. 12
3. On a piano, what do the black keys represent?
  1. Major notes
  2. **Sharps and flats**
  3. Natural notes
  4. Octaves
4. What is the smallest interval used in music theory?
  1. Tone
  2. **Semitone**
  3. Third
  4. Octave
5. What pattern of tones and semitones defines a major scale?
  1. Tone, Tone, Semitone, Tone, Tone, Semitone, Tone
  2. Tone, Semitone, Tone, Tone, Semitone, Tone, Tone
  3. **Tone, Tone, Semitone, Tone, Tone, Tone, Semitone**
  4. Semitone, Tone, Tone, Semitone, Tone, Tone, Tone
6. If you play a C major scale on the piano, which keys will you press?
  1. All black keys
  2. **All white keys from C to the next C**
  3. A combination of black and white keys starting on C
  4. Only the black keys surrounding the C keys
7. When is the Common Practice period?
  1. **1600-1900**
  2. 1600-1950
  3. 1600-1800
  4. 1500-1900
8. What is C major scale?
  1. C, D, E, F, G, A, B, C
  2. C, D, E, F, G, A, B
  3. D, E, F, G, A, B, C
  4. **A, B, C, D, E, F, G, A, B, C**
9. In music theory, what does the term 'interval' refer to?
  1. **The distance between two notes**
  2. The volume of a note
  3. The length of a note
  4. The speed of a piece
10. What's the difference between octave and scale?

1. An octave is a series of eight notes, while a scale is a sequence of notes in a specific order.
2. An octave is a type of scale, while a scale is a type of chord.
3. An octave is a sequence of notes in a specific order, while a scale is a series of eight notes.
4. An octave is a type of chord, while a scale is a type of scale.

#### D.4 Latin Video

1. What's the pronunciation of the Latin letter 'C'?
  1. Similar to the English "k".
  2. Pronounced as "ch" like in "church".
  3. Pronounced as "sh" like in "church".
  4. I don't know.
2. When does the Latin letter 'I' act as a consonant?
  1. At the end of a word.
  2. When followed by a consonant.
  3. When it precedes or is between vowels.
  4. When it stands alone.
3. Which letter in Latin is described as always being followed by 'U'?
  1. S
  2. T
  3. Q
  4. V
4. What is the description the pronunciation of the letter 'S' in Latin?
  1. Always like "s" as in "soot".
  2. Silent in most cases.
  3. Sometimes soft and sometimes hard.
  4. Always like "z" as in "rose".
5. Which Latin letter is said to have a dual function as both a consonant and a vowel?
  1. U
  2. E
  3. I
  4. O
6. What is the relevance of the letter 'J' in English and 'I' in classical Latin?
  1. 'J' is pronounced differently in Latin than it is in English.
  2. 'J' in English originates from 'I' in Latin.
  3. I don't know.
  4. 'J' in English originates from 'J' in Latin.
7. Which of the following is incorrect?
  1. The capital Úin Latin looks like our English capital W.
  2. Cálways like Kín classical Latin.
  3. Classical Latin has fewer letters compared to English.
  4. R̄ounds like a dog.
8. What's the correct description of the pronunciation of the letter 'T' in Latin?
  1. Sounds like 't' in English.
  2. Sounds like 'sh' in English.
  3. Sounds like 'ch' in English.
  4. Sounds like 'th' in English.
9. What is the pronunciation rule for the letter 'Q' in Latin?
  1. They are silent.
  2. It is pronounced like 'k'.

3. It is always followed by 'U', which is not a vowel.
4. I don't know.
10. In Latin, how are Greek aspirated consonants like CH, PH, and TH pronounced?
  1. As in 'chapter', 'phone', and 'the'.
  2. With a breath of air after them.
  3. They are silent.
  4. I don't know.

#### E Engagment Questionnaire

##### E.1 Behaviorial Engagement

- When I watch the video, I listen very carefully. (EvsD scale, behavioral section, Item 5)
- I want to find out more about the subject matter of the video. (Situational Interest scale, Exploration Intention section, Item 2)
- I get really involved in video content. (Classroom Engagement Inventory, Behavioral Engagement, item 10)
- I formed questions in my mind while watching the video. (Classroom Engagement Inventory, Behavioral Engagement, item 10)

##### E.1.1 Emotional Engagement.

- I enjoy learning new things in the video. (EvsD scale, emotional section, item 4)
- When I watch the video, I feel interested. (EvsD scale, emotional section, Item 2)
- I was inspired to expand my knowledge when I watched the video. (MOOC engagement scale, emotional section, item 3)

##### E.1.2 Cognitive Engagement.

- I mentally took notes of details as I watched the video. (Original Cognitive Engagement, Item 30)
- While learning new concepts, I tried to think of implications and practical applications. (Original Cognitive Engagement, Item 21)
- I mentally took notes of the main ideas as I watched the video. (Original Cognitive Engagement, Item 31)
- I tried to memorize lists of new terms and definitions. (Original Cognitive Engagement, Item 39)
- When I had trouble understanding a concept or an example, I went over it again until I understood it. (MOOC engagement scale, cognitive section, item 1)

#### F Engagement Questionnaire Generation Strategy

We removed some questions that are unrelated to the context of our study, including:

- Questions related to self
- Questions about school
- Questions about teacher support (e.g. "When my teacher first explains new material, I feel bored.")
- Questions about searching web resources (e.g., "I often visit the web resources suggested by the course instructor in the MOOCs.")

We also removed repeated questions (e.g., "How enjoyable was the lesson?" and "I enjoy the lesson"). For similar questions (e.g., "I pay attention in class." and "I listen very carefully"), we selected the statement that was repeated across surveys. After this, we adapted some questions to match the context of the study including:

- “Book” in Original Cognitive Engagement scale was replaced by “Video”.
- “Class” in Classroom Engagement scale was replaced by “Video”.
- “Activity” in Situational Interest scale was replaced by “Video”.
- “MOOC” in MOOC engagement scale were adapted to “when I watched the video”.
- “Check my book or use other materials” was replaced by “take notes of details as I watched the video”.

## G Danmaku Quality Questionnaire and Results

**Table 2: Questionnaire Results of comparing the quality of ClassComet’s generated danmaku and real-world danmaku on Bilibili.**

Metric	Statement	ClassComet	Bilibili
Relevancy	1. The danmaku is closely related to the video content.	$4.00 \pm 0.89$	$2.85 \pm 1.39$
	2. The danmaku effectively captures the main points of the video.	$4.00 \pm 0.78$	$2.95 \pm 1.36$
	3. The emotional tone of the danmaku aligns well with the emotions conveyed in the video.	$3.65 \pm 0.91$	$3.60 \pm 1.31$
Consistency	1. The danmaku does not contain untrue information(fake data, wrong concepts, etc.) against the video content.	$4.75 \pm 0.48$	$4.60 \pm 0.66$
	2. The danmaku does not include any misleading emotion about the video.	$5.00 \pm 0.00$	$5.00 \pm 0.00$
Fluency	1. The danmaku is grammatically correct.	$4.65 \pm 0.48$	$4.85 \pm 0.36$
	2. The danmaku is simple and easy to understand.	$4.58 \pm 0.50$	$4.60 \pm 0.49$
Coherence	1. The danmaku interaction is similar to natural, realistic conversations and not overly formal.	$4.10 \pm 0.37$	$4.45 \pm 0.74$