Emojis in Autocompletion: Enhancing Video Search with Visual Cues

Hojin Yoo
yoo.515@osu.edu
The Ohio State University
Columbus, Ohio, USA

Arnab Nandi
nandi.9@osu.edu
The Ohio State University
Columbus, Ohio, USA

Abstract

Effective video search is increasingly challenging due to the inherent complexity and richness of video content, which traditional full-text query systems and text-based autocompletion methods struggle to capture. In this work, we propose an innovative autocompletion system that integrates visual cues, specifically, representative emojis, into the query formulation process to enhance video search efficiency. Our approach leverages cutting-edge Vision-Language Models (VLMs) to generate detailed scene descriptions from videos and employs Large Language Models (LLMs) to distill these descriptions into succinct, segmented search phrases augmented with context-specific emojis. A controlled user study, conducted with 11 university students using the MSVD dataset, demonstrates that the emoji-enhanced autocompletion reduces the average query completion time by 2.27 seconds (14.6% decrease) compared to traditional text-based methods, while qualitative feedback indicates mixed but generally positive user perceptions. These results highlight the potential of combining linguistic and visual modalities to redefine interactive video search experiences.

Keywords

Video Search, Autocompletion, Emoji-Enhanced Interface

ACM Reference Format:
Hojin Yoo and Arnab Nandi. 2025. Emojis in Autocompletion: Enhancing Video Search with Visual Cues. In *Workshop on Human-In-the-Loop Data Analytics (HILDA' 25), June 22–27, 2025, Berlin, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3736733.3736745

1 Introduction

The rapid expansion of video content across various platforms (e.g., social networks, mobile, and vehicle systems) has amplified the challenges associated with effective video search, underscoring the critical need for efficient video search and retrieval systems [15, 31] that can handle vast and diverse data. Traditional search systems typically rely on full-text queries to retrieve relevant content. However, video content is inherently rich and complex, with visual and temporal nuances that are not easily captured through text alone. To bridge this gap, recent research has focused on ways to summarize or visualize content to aid quick information retrieval. For example, combining image thumbnails with text summaries can drastically reduce search times and improve retrieval consistency by providing users with a visual snapshot that enhances their ability to quickly discern and select relevant information.

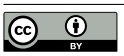
While such visual summarization techniques have proven effective for web content, a similar strategy has yet to be fully explored within the domain of autocompletion, a critical feature in modern search interfaces that assists users in formulating queries more efficiently. Autocompletion systems are designed to predict the remainder of a user's query as they type, thereby reducing keystrokes and accelerating the search process. Recent advances in text-based autocompletion have shown that highly optimized indexing methods can deliver near-instantaneous query suggestions. However, in the context of video search, autocompletion must efficiently represent the rich, multimodal nature of video contents. This is a challenge that traditional text-based methods [22, 23], which focus solely on lexical matching, do not address.

To address this shortcoming, our work introduces a novel approach that integrates visual cues directly into the autocompletion interface for video search. Our system leverages cutting-edge VLMs [17, 30] to generate detailed scene descriptions from videos, capturing key visual elements such as objects, actions, and contexts. These descriptions are then distilled by LLMs [1, 10, 20] into concise search phrases, which we further augment with representative emojis that serve as visual summaries of the corresponding video content. As shown in Figure 1, our system architecture captures the complete workflow, from video input to the generation of enriched, emoji-enhanced autocompletion suggestions, highlighting its potential to simplify and accelerate user query formulation.

By fusing the strengths of multimodal models, our approach offers a unique solution to the challenge of conveying complex video information through compact, easy-to-process visual cues. This multimodal strategy not only improves the speed and accuracy of query formulation but also enhances the overall user experience by reducing the cognitive load associated with processing lengthy text suggestions. The integration of visual cues into autocompletion represents a significant advancement over conventional text-only systems and points toward a future where the seamless combination of language and imagery defines efficient video search.

Our contributions can be summarized as follows:

- **Multimodal Query Generation:** We introduce a novel system that leverages both LLMs and VLMs to generate enriched video search queries, combining detailed scene descriptions with succinct search phrases.
- **Visual Cue Integration in Autocompletion:** We develop an innovative autocompletion interface that pairs text suggestions with contextually aligned emojis, providing users with a quick, intuitive grasp of the video content behind each suggestion.



HILDA' 25, Berlin, Germany

ACM ISBN 979-8-4007-1959-2/2025/06
https://doi.org/10.1145/3736733.3736745

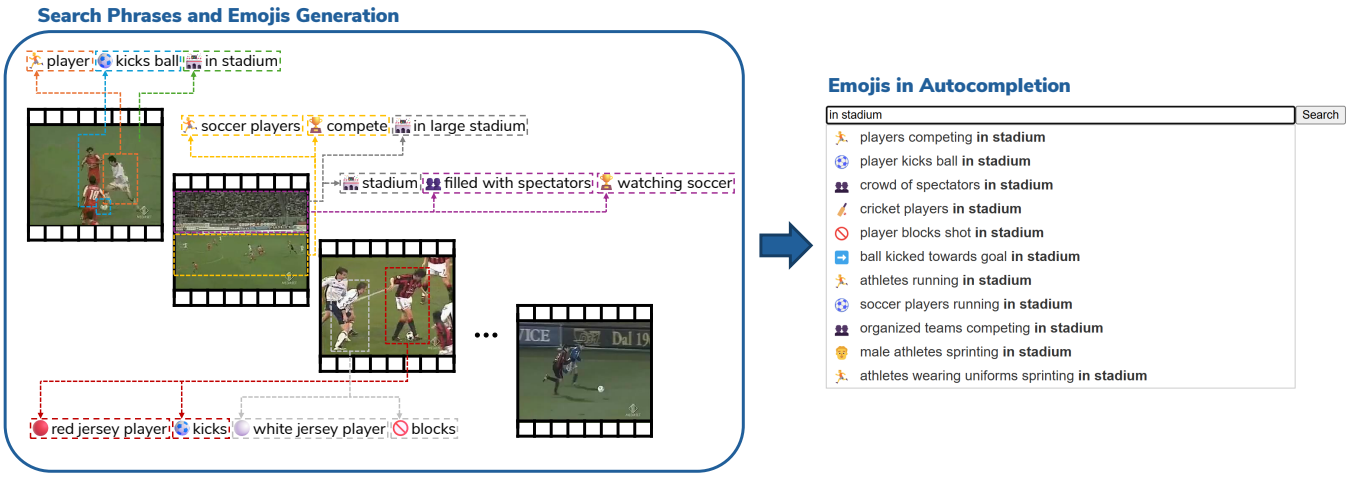


Figure 1: Overview of the Emoji-Enhanced Video Query Autocompletion System.

- **User-Centric Evaluation:** We validate our approach through a controlled user study, demonstrating that the inclusion of visual cues significantly improves query formulation efficiency and user satisfaction in video search tasks.

The remainder of the paper is organized as follows. In Section 2, we review relevant prior research and existing techniques that form the foundation of our work. Section 3 details our proposed approach, describing the system architecture and methods used to generate emoji-enhanced search phrases as well as the design of our autocompletion interface. Section 4 presents the user study setup and evaluation metrics, along with a discussion of our experimental results on usability. Finally, Sections 5 and 6 address the limitations of our work and conclude the paper, respectively.

2 Related Work

Strategies for Enhanced Video Search A number of recent studies have focused on improving video retrieval by developing techniques for selecting and generating effective video summaries. For instance, several methods automatically select attractive thumbnails from videos by analyzing visual quality and aesthetic metrics [27], or by employing adversarial and reinforcement learning to balance representativeness and visual appeal [3]. Other approaches have proposed dynamic thumbnail generation [32], where the thumbnails are adaptively selected based on both video content and user queries, and best-frame selection [26], all aiming to provide users with concise, visually representative snapshots of video content. These techniques have proven critical in helping users quickly identify relevant videos by offering clear visual summaries that capture the essence of the content.

In parallel, research on moment retrieval and highlight detection has leveraged multimodal models to extract salient video segments tailored to user queries, thereby enhancing video search performance [13, 16, 21]. Separately, studies on temporal segmentation have employed transformer-based architectures and unsupervised techniques to delineate boundaries in long-form content and identify distinct events within videos [25, 29]. Despite these significant advances in video retrieval, existing methods remain largely confined to the domains of thumbnail generation, moment detection, and temporal segmentation. Little attention has been paid to incorporating visual cues directly into the autocompletion interface, a feature that could provide users with instant, intuitive visual summaries as they formulate search queries. Our work addresses this gap by integrating representative emojis into query autocompletion for video search, thereby bridging the current disconnect between multimodal video retrieval techniques and interactive search interfaces.

Emoji Usage in Digital Communication Emojis have become a ubiquitous element in digital communication across various online platforms, serving as non-verbal cues that help users interpret context, convey emotions, and express opinions [11, 18]. Notably, studies targeting accessibility issues have demonstrated that when emoji cues are effectively incorporated into communication interfaces, visually impaired users experience increased speed and ease in making and replying to statements [24, 28]. A rich body of work has examined how emojis are used in different settings, from social networks to developer communication on platforms like GitHub [19], and has explored their semantic relationships [2, 9] and cross-cultural variations [4]. Moreover, research such as that behind MOJI [14] has focused on facilitating text-based emoji search by supporting query expansion and providing emoji recommendations, thereby enabling users to rapidly identify and select appropriate emojis during query formulation.

Despite significant progress in utilizing emojis for digital communication and search, prior work in emoji-based search has largely been limited to either mapping visual media to emoji labels, as seen in approaches like Image2Emoji [5], or enabling search via emoji queries through interfaces like Query-by-Emoji [6]. However, to our knowledge, no research has yet conducted a comprehensive user study to validate whether integrating emojis as visual cues directly within the autocompletion interface for video search can improve query formulation efficiency and retrieval performance. This gap underscores the need for user-centric investigations into how visual

Below is a description of a video clip.
[Video scene description]

Extract at most 10 search phrases and emojis from the video description paragraph provided above that can be used to find the video.

Requirements:
- Each search phrase must include the objects' actions, characters, and background, directly from the video description.
- Accurately capture the relationships or interactions between objects/characters when applicable.
- Every search phrase must be concise and intuitive between 5 to 10 words.
- Use diverse vocabulary for each phrase, avoiding repetitive or overly similar phrases.
- The emojis should help users to understand the search phrase visually like the examples below.
- Assign an emoji for a phrase which has a specific meaning in the whole phrase like the examples below.
- Do NOT generate search phrases related to feelings, atmosphere, or emotions.
- Do NOT include phrases describing scene cuts or camera motions.

Example:
- 🚦 intersection 🚗red car ➡️turning right 🚧cautiously
- 🚚truck 💥crash 🏍️with motorcycle 😱in front of ego vehicle
- 🚴cyclist ⬅️makes left turn 🚦 at intersection
- 🚗car 🛑stops 🚦 at red light ⬇️slowing down its speed
- 👕white shirt 👩woman 🚶 at crosswalk 👖 black-pants

Please generate the response in the form of a Python list string. The value of each list is a Python dictionary with the following keys: "phrase", "split", "emojis", and "importance".
- The value of "phrase" should be a string representing a search phrase.
- The value of "split" should be a list of strings, where each item represents a meaningful segment of the search phrase.
- The value of "emojis" should be a list of emojis corresponding to each item in "split". The lengths of "split" and "emojis" must be the same.
- The value of "importance" should be a list of floating-point numbers representing the relative significance of each phrase segment in "split", where higher values indicate greater importance, all values sum to exactly 1.0, and the list length matches that of "split".
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only return the Python list as a string.
For example, your response should look like this:

```
[{
        "phrase": "intersection red car turning right",
        "split": ["intersection", "red car", "turning right", "cautiously"],
        "emojis": [" 🚦 ", "🚗", "➡️", "🚧"],
        "importance": [0.2, 0.4, 0.3, 0.1]}, {
        "phrase": "truck crash with motorcycle in front of ego vehicle",
        "split": ["truck", "crash", "with motorcycle", "in front of ego vehicle"],
        "emojis": ["🚚", "💥", "🏍️", "😱"],
        "importance": [0.3, 0.3, 0.3, 0.1]
}]
```

Figure 2: Example of the Prompt Design for Phrase and Emoji Generation

augmentation through emojis can enhance the interaction between users and video search systems.

3 Emoji-Enhanced Autocompeltion

In this section, we detail the system architecture underpinning our approach to emoji-enhanced video autocompletion. We begin by describing the process for generating enriched search phrases accompanied by representative emojis. Specifically, we outline a two-stage pipeline VLMs is first used to generate detailed video descriptions, which are then distilled by LLMs into segmented search phrases along with context-specific emojis and importance scores. Next, we explain how these enriched phrases are integrated into our autocompletion interface to provide dynamic visual cues that adapt based on user input. Together, these components form a cohesive system that enhances query formulation efficiency and facilitates improved video retrieval.

Generation of Emoji-Enhanced Search Phrases To enable an autocompletion without relying on external metadata, our system employs a two-stage generation process to produce search phrases accompanied by representative emojis. In the first stage, we utilize VLMs, specifically PLLaVA [30], to analyze each video and generate detailed scene descriptions. This stage produces an extensive textual summary that captures various elements of the video, including actions, objects, and contextual information. The generated description is intended to serve as a rich source of semantic cues that faithfully represent the video's visual content. First, a textual description D_v for a given video v is generated by a VLM as follows:

$$D_v = \mathscr{F}_{VLM}(v)$$

In the second stage, the detailed video description D_v is input into an LLM, namely Llama-3.2-70B [20], which is prompted via a few-shot approach to propose search phrases that users might employ during video retrieval. This process yields a set of n emoji-enhanced search phrases, $S_v = \{S_1, ..., S_n\}$:

$$S_v = \mathscr{F}_{LLM}(D_v)$$

The LLM is instructed to segment the search phrase into multiple contexts and, for each segment, provide a corresponding emoji that best represents that particular context. In addition, the model outputs a measure of importance for each context segment, which reflects the weight of each part within the overall search phrase. Figure 2 illustrates an example of the prompt that guides the model in generating the search phrase along with the emoji annotations. Each phrase S_i is composed of a sequence of m_i contextual segments, where each segment $C_{i,j}$ is a tuple containing its textual content $t_{i,j}$, a representative emoji $e_{i,j}$, and an importance weight $w_{i,j}$:

$$S_i = \langle (t_{i,1}, e_{i,1}, w_{i,1}), ..., (t_{i,j}, e_{i,j}, w_{i,j}) \rangle$$

This two-stage pipeline effectively transforms a detailed video description into a compact, semantically enriched query suggestion that integrates both textual and visual elements.

Autocompletion Interface with Visual Cues Traditional autocompletion systems provide text-based query suggestions by predicting completions based solely on the characters entered by the user. In contrast, our system augments the autocompletion feature with visual cues by incorporating representative emojis alongside the text suggestions. When a user begins typing a video search query q, the system displays a list of autocompletion suggestions, denoted as $S_k(q)$, To the left of each suggestion, it presents a corresponding representative emoji, e_{display}, that reflects the dominant context of that search phrase. Figure 1 illustrates an example of this feature, where each autocompletion entry is enriched with an emoji to provide an immediate visual summary of the underlying video content.

To generate these enriched suggestions, our system first selects the top-k candidate search phrases $\mathscr{S}_k(q)$ from the previously generated set of phrases that contain the user's current query input q. This selection process is formalized as:

$$\mathscr{S}_k(q) = \text{Select}_k\left(\left\{ S_i \in \mathscr{S} \,\middle|\, q \subseteq \bigoplus_{j=1}^{m_i} t_{i,j} \right\}\right)$$

For each candidate phrase S_i in $\mathscr{S}_k(q)$, the system then dynamically determines a single, representative emoji to display. The core of this dynamic selection lies in identifying the active segment $C_{i,j_{\text{active}}}$, which has the maximum textual overlap with the user's query:

$$j_{\text{active}}(q, S_i) = \arg \max_{j \in \{1,...,m_i\}} \text{len}(q \cap t_{i,j})$$

The final displayed emoji, e_{display}, is then determined based on the query's progression within this active segment. Crucially, each segment is assigned an importance score that quantifies its contribution to the overall search phrase. This allows us to sort the segments by their semantic weight, creating an importance-ordered sequence $S_i' = \langle C'_{i,1}, C'_{i,2}, \dots, C'_{i,m_i} \rangle$. If the query's length does not exceed 60% of the active segment's length, the emoji of the active segment itself is displayed. However, once the query spans more than 60% of the segment, the system reveals the emoji of the *next most important* segment, providing a progressive and semantically meaningful visual cue. This behavior is captured by the following rule:

$$e_{\text{display}}(q, S_i) = \begin{cases} e'_{i,r+1} & \text{, if } \frac{\text{len}(q \cap t_{i,j_{\text{active}}})}{\text{len}(t_{i,j_{\text{active}}})} > 0.6 \text{ and } r < m_i \\ e_{i,j_{\text{active}}} & \text{, otherwise} \end{cases}$$

Here, r denotes the importance rank of the active segment, and $e'_{i,r+1}$ is the emoji of the segment with the next-highest importance rank. This method ensures that the visual cues evolve meaningfully as the user refines their query, preventing a static and repetitive user experience. Figure 2 shows a sample prompt used to instruct the LLM in splitting search phrases into context segments, assigning representative emojis, and defining importance measures.

4 Experiments and Evaluation

In this section, we first describe the setup of our user study, detailing the participant recruitment, task design, and interface familiarization process. Next, we outline the evaluation metrics employed to measure the system's efficiency, with a particular focus on query completion time (QCT) and granular timestamp logging. Finally, we present the experimental results, including both quantitative comparisons between the text-only and emoji-enhanced autocompletion conditions and qualitative user feedback, which together offer a comprehensive assessment of our system's usability and effectiveness in video search tasks.

User Study Setup: Our user study was designed to evaluate the usability of the emoji-enhanced autocompletion system for video search. The study was reviewed and approved by our Institutional Review Board (IRB). We recruited 11 university students aged 18 and older, all of whom were regular users of personal computers with keyboards and web browsers. Participants were provided with an overview of the study procedures and familiarized with the web-based interface during an initial training session. In this familiarization phase, users were introduced to the system's user interface and its functionality, including the autocompletion feature that displays text suggestions augmented with representative emojis based on segmented search phrases.

Participants were tasked with conducting video searches over 10 distinct topics that represent everyday scenarios found in short

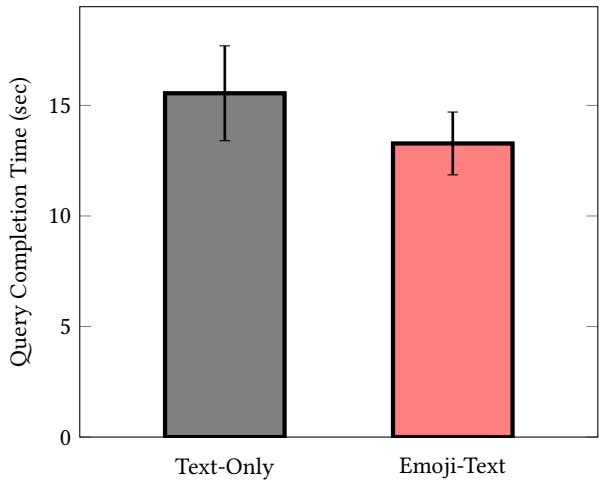


Figure 3: Average Query Completion Time

YouTube videos (typically under 15 seconds in duration) using the MSVD [7] dataset, which contains 2,089 video clips. For each topic, participants first typed in their search query using the autocompletion interface. After finalizing their query, the system returned a ranked list of the top 15 video results based on sentence similarity between the user's query and the automatically generated video descriptions. The study was conducted in two phases, corresponding to two autocompletion conditions: one employing text-only autocompletion and the other using emoji-text autocompletion. To minimize order bias, the presentation order of these conditions was counterbalanced across participants. During the study, users were also instructed to generate query sentences containing at least three words and were advised against simply copying the full text of the topic prompts. The system allowerd re-querying for each topic.

Evaluation Metrics: To quantify the usability and efficiency of our autocompletion system, we focused primarily on measuring QCT, the elapsed time from when a participant begins typing (i.e., initial keypress) until the query is finalized and submitted. For tasks involving multiple query attempts within a topic, the system recorded the timing for each attempt and computed the average QCT as an overall measure of efficiency. Detailed timestamp logging captured the moment of the first key entry, all intermediate keystrokes, and the exact moment of search submission (triggered via the search button). These metrics provide nuanced insights into user query formulation behaviors under both text-only and emoji-enhanced autocompletion conditions.

Results: Figure 3 illustrates the comparison of average QCT between the two autocompletion conditions. The results reveal that the mean QCT for the text-only autocompletion was 15.55 seconds, whereas the emoji-text autocompletion reduced the average QCT to 13.28 seconds (a net decrease of 2.27 seconds). Although there was a bias indicating that later video search tasks tended to be completed more quickly than earlier ones, the overall average improvement suggests that the integration of emojis into the autocompletion interface contributed to more efficient query completion in video search tasks.

Qualitative feedback further enriched our understanding of usability. Four participants reported a higher frequency of use for the emoji-enhanced autocompletion, while two participants remarked that the text-only autocompletion led to fatigue due to the extensive textual content. Conversely, two participants felt that both autocompletion features were equally effective in aiding query formulation, and three participants expressed a preference for text-only autocompletion, noting greater familiarity with its format. These mixed responses indicate that while the emoji-enhanced autocompletion offers measurable benefits in reducing query completion time, its impact on usability is multifaceted and varies according to individual user preferences.

5 Limitations

One limitation of our emoji-enhanced autocompletion system is that the interpretation of emojis can differ significantly across cultural contexts. Although emojis are globally recognized, their meanings are not universally fixed. Nuances arise from linguistic differences and culturally specific ways of expressing emotions and conceptualizing topics. Studies in cross-cultural psychology suggest that while there are normative patterns in emoji usage, distinct cultural variations exist, as evidenced by differences in how Eastern and Western users employ emojis in categories such as People, Food & Drink, Travel & Places, among others [12]. These findings imply that an emoji which conveys a particular sentiment in one culture might be interpreted differently in another. Consequently, the effectiveness of emoji-enhanced autocompletion in conveying the intended query semantics may vary depending on the user's cultural background, posing challenges for a one-size-fits-all solution in a global marketplace.

Another limitation stems from user familiarity with traditional text-only autocompletion. Many users are accustomed to autocompletion systems that provide text-based suggestions, and there is a risk that the integration of emoji as visual cues may not be immediately embraced by all users. The translation of textual information into visual cues is a relatively nascent area of research in user interface design, with limited studies focusing on the optimal ways to convert text information into an effective visual format. For instance, research involving the UEQ-Emoji [8] has shown that while emoji-based questionnaires can be effective and offer advantages on smartphone screens due to their compact nature, many participants still expressed a preference for text-based questionnaires. This preference is largely attributed to the inherent semantic limitations of the emoji language, which some users find less precise or intuitive for complex information. As a result, further research is needed to develop and evaluate advanced UI/UX designs that can overcome these challenges and fully leverage the potential of visual cues in autocompletion systems.

6 Conclusion

Our work demonstrates that integrating representative emojis into video search autocompletion can enhance query formulation efficiency and overall user experience, as evidenced by the observed reduction in QCT and supportive qualitative feedback. Despite

promising quantitative results and mixed qualitative responses suggesting potential user benefits, our study also highlights key limitations: notably, the cross-cultural variability in emoji interpretation and the deeply ingrained familiarity with traditional text-only autocompletion interfaces. These challenges point to the need for further in-depth research into advanced UI/UX designs that can effectively translate textual information into intuitive visual cues.

Acknowledgments

This work was supported by the National Science Foundation under award #1910356 and the Honda Research Institute.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Wei Ai, Xuan Lu, Xuanzhe Liu, Ning Wang, Gang Huang, and Qiaozhu Mei. 2017. Untangling emoji popularity through semantic embeddings. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11. 2–11.

[3] Evlampios Apostolidis, Eleni Adamantidou, Vasileios Mezaris, and Ioannis Patras. 2021. Combining adversarial and reinforcement learning for video thumbnail selection. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 1–9.

[4] Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis? Exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 24th ACM international conference on Multimedia*. 531–535.

[5] Spencer Cappallo, Thomas Mensink, and Cees GM Snoek. 2015. Image2emoji: Zero-shot emoji prediction for visual media. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1311–1314.

[6] Spencer Cappallo, Thomas Mensink, and Cees GM Snoek. 2015. Query-by-emoji video search. In *Proceedings of the 23rd ACM international conference on Multimedia*. 735–736.

[7] David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 190–200.

[8] Ashley Colley, Sven Mayer, and Jonna Häkkilä. 2023. Developing an Emoji-based User Experience Questionnaire: UEQ-Emoji. In *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia*. 65–73.

[9] Henriette Cramer, Paloma De Juan, and Joel Tetreault. 2016. Sender-intended functions of emojis in US messaging. In *Proceedings of the 18th international conference on human-computer interaction with mobile devices and services*. 504–509.

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[11] Joan Gajadhar and John Green. 2005. The importance of nonverbal elements in online chat. *Educause Quarterly* 28, 4 (2005), 63.

[12] Sharath Chandra Guntuku, Mingyang Li, Louis Tay, and Lyle H Ungar. 2019. Studying cultural differences in emoji usage across the east and the west. In *Proceedings of the international AAAI conference on web and social media*, Vol. 13. 226–235.

[13] Donghoon Han, Seunghyeon Seo, Eunhwan Park, Seong-Uk Nam, and Nojun Kwak. 2024. Unleash the potential of clip for video highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8275–8279.

[14] Yoo Jin Hong, Hye Soo Park, Eunki Joung, and Jihyeong Hong. 2024. MOJI: Enhancing Emoji Search System with Query Expansions and Emoji Recommendations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[15] Daniel Kang, Francisco Romero, Peter D Bailis, Christos Kozyrakis, and Matei Zaharia. 2022. VIVA: An End-to-End System for Interactive Video Analytics.. In *CIDR*.

[16] Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems* 34 (2021), 11846–11858.

[17] Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023. LLaMA-VID: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043* (2023).

[18] Shao-Kang Lo. 2008. The nonverbal communication functions of emoticons in computer-mediated communication. *Cyberpsychology & behavior* 11, 5 (2008), 595–597.

[19] Xuan Lu, Yanbin Cao, Zhenpeng Chen, and Xuanzhe Liu. 2018. A first look at emoji usage on github: An empirical study. *arXiv preprint arXiv:1812.04863* (2018).

[20] Meta. 2024. *Llama 3.2: Revolutionizing edge AI and vision with open, customizable models*. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[21] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 23023–23033.

[22] Arnab Nandi and HV Jagadish. 2007. Effective phrase prediction. In *Proceedings of the 33rd international conference on Very large data bases*. 219–230.

[23] Arnab Nandi and HV Jagadish. 2011. Guided interaction: Rethinking the query-result paradigm. *Proceedings of the VLDB Endowment* 4, 12 (2011), 1466–1469.

[24] Henning Pohl, Christian Domin, and Michael Rohs. 2017. Beyond just text: semantic emoji similarity modeling to support expressive communication. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 1 (2017), 1–42.

[25] Jielin Qiu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Ding Zhao, and Hailin Jin. 2023. Liveseg: Unsupervised multimodal temporal segmentation of long livestream videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5188–5198.

[26] Jian Ren, Xiaohui Shen, Zhe Lin, and Radomir Mech. 2020. Best frame selection in a short video. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*. 3212–3221.

[27] Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. 2016. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. 659–668.

[28] Garreth W Tigwell, Benjamin M Gorman, and Rachel Menzies. 2020. Emoji accessibility for visually impaired people. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[29] Haoqian Wu, Keyu Chen, Haozhe Liu, Mingchen Zhuge, Bing Li, Ruizhi Qiao, Xiujun Shu, Bei Gan, Liangsheng Xu, Bo Ren, et al. 2023. Newsnet: A novel dataset for hierarchical temporal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10669–10680.

[30] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. PLLaVA: Parameter-free LLaVA Extension from Images to Videos for Video Dense Captioning. *arXiv preprint arXiv:2404.16994* (2024).

[31] Zhuangdi Xu, Gaurav Tarlok Kakkar, Joy Arulraj, and Umakishore Ramachandran. 2022. EVA: A symbolic approach to accelerating exploratory video analytics with materialized views. In *Proceedings of the 2022 International Conference on Management of Data*. 602–616.

[32] Yitian Yuan, Lin Ma, and Wenwu Zhu. 2019. Sentence specified dynamic video thumbnail generation. In *Proceedings of the 27th ACM international conference on multimedia*. 2332–2340.