Check for updates

ACM DIGITAL LIBRARY

Association for Computing Machinery

acm open

Latest updates: https://dl.acm.org/doi/10.1145/3757676

PDF Download
3757676.pdf
30 March 2026
Total Citations: 0
Total Downloads: 268

Published: 16 October 2025

Citation in BibTeX format

RESEARCH-ARTICLE

ThingMoji: User-Captured Cut-Outs For In-Stream Visual Communication

ERZHEN HU, University of Virginia, Charlottesville, VA, United States

QIAN WAN, City University of Hong Kong, Hong Kong, Hong Kong

CHANGKONG ZHOU, Aalto University, Espoo, Uusimaa, Finland

MD AASHIKUR RAHMAN AZIM, University of Virginia, Charlottesville, VA, United States

PIAOHONG WANG, City University of Hong Kong, Hong Kong, Hong Kong

XINGYI HU, The Hong Kong Polytechnic University, Hong Kong, Hong Kong, Hong Kong

View all

Open Access Support provided by:

City University of Hong Kong

University of Virginia

Aalto University

George Mason University

The Hong Kong Polytechnic University

Proceedings of the ACM on Human-Computer Interaction, Volume 9, Issue 7 (October 2025)
https://doi.org/10.1145/3757676
EISSN: 2573-0142

ThingMoji: User-Captured Cut-Outs For In-Stream Visual Communication

ERZHEN HU, University of Virginia, USA
QIAN WAN, City University of Hong Kong, China
CHANGKONG ZHOU, Aalto University, Finland
MD AASHIKUR RAHMAN AZIM, University of Virginia, USA
PIAOHONG WANG, City University of Hong Kong, China
XINGYI HU, The Hong Kong Polytechnic University, China
YUHAN ZENG, City University of Hong Kong, China
ZHICONG LU, George Mason University, USA
SEONGKOOK HEO*, University of Virginia, USA

Live streaming has become increasingly popular, driven by the desire for direct and real-time interactions between streamers and viewers. However, current text-based interactions and pre-defined emojis limit expressiveness, especially when referring to specific stream moments. We propose ThingMoji, a type of user-captured cut-outs to enhance user expression and foster more effective communication between streamers and their audience in the comment section. ThingMojis are unique digital icons created by users by capturing snapshots and annotating specific areas at any point during the stream. We developed StreamThing, a live-streaming platform integrated with ThingMojis, to explore their use during object-focused live streaming contexts. In a user study with three in-the-wild deployments reveals the expressive use of ThingMojis in diverse live-streaming scenarios with rich visual contents. Our findings show that ThingMojis enable viewers to reference specific objects, express emotions, and create shared visual narratives. Streamers found ThingMojis valuable for facilitating on-the-fly communication around visual content and fostering playful interactions. The study also uncovered challenges in ThingMoji comprehension, issues for long-term uses of ThingMojis, and potential concerns regarding misuse. Based on these insights, we discussed new opportunities for supporting object-focused communication during live streaming environments.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**.

Additional Key Words and Phrases: Live-Streaming, Human-AI, Video-Mediated Communication, One-To-Many Communication, Shared Narrative

ACM Reference Format:
Erzhen Hu, Qian Wan, Changkong Zhou, Md Aashikur Rahman Azim, PiaoHong Wang, Xingyi Hu, Yuhan Zeng, Zhicong Lu, and Seongkook Heo. 2025. ThingMoji: User-Captured Cut-Outs For In-Stream Visual

*Corresponding author

Authors' Contact Information: Erzhen Hu, University of Virginia, Charlottesville, VA, USA, eh2qs@virginia.edu; Qian Wan, City University of Hong Kong, Hong Kong, China, ahin.qianwan@gmail.com; Changkong Zhou, Aalto University, ESPOO, Finland, changkong.zhou@aalto.fi; Md Aashikur Rahman Azim, University of Virginia, Charlottesville, VA, USA, ma6zp@virginia.edu; PiaoHong Wang, City University of Hong Kong, Hong Kong, China, piaohwang2-c@my.cityu.edu.hk; Xingyi Hu, The Hong Kong Polytechnic University, Hong Kong, China, xingyi.hu@connect.polyu.hk; Yuhan Zeng, City University of Hong Kong, Hong Kong, China, yhzeng3-c@my.cityu.edu.hk; Zhicong Lu, George Mason University, Fairfax, VA, USA, zlu6@gmu.edu; Seongkook Heo, University of Virginia, Charlottesville, VA, USA, seongkook@virginia.edu.

This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.
© 2025 Copyright held by the owner/author(s).
ACM 2573-0142/2025/11-ARTCSCW495
https://doi.org/10.1145/3757676

Communication . *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW495 (November 2025), 29 pages. https://doi.org/10.1145/3757676

1 Introduction

Live streaming has become an integral part of online communication and is used across various domains such as gaming, crafting, and e-commerce [51]. What sets live streaming apart from other digital media is its real-time, dynamic nature, where streamers and audiences engage in a real-time exchange, fostering a sense of community and participation [4, 12, 19, 24, 25]. Viewers can comment and ask questions during the stream, allowing streamers gauge viewers' interests and dynamically adjust the content and direction. Consequently, viewer-streamer communication has become a key aspect of the streaming experience, moving beyond passive video consumption.

A critical challenge in live streaming emerges from the inadequacy of current communication tools — predominantly text chat and pre-defined emojis [4] — in supporting precise object-focused discourse. While viewers frequently need to reference specific objects, moments, and spatial relationships within streams [17, 50, 62, 64, 72, 81], existing tools often fall short in facilitating these nuanced interactions. These communication challenges are especially evident in streams with complex visual content. For instance, when a jewelry maker demonstrates intricate wire-wrapping techniques or when a model builder showcases various assembly components, viewers struggle to precisely indicate which elements they're discussing. Similarly, in e-commerce streams where multiple products are displayed simultaneously [72], viewers often resort to ambiguous temporal references ("*the necklace you showed earlier*") or imprecise spatial descriptions ("*the blue piece on the left*"). Such descriptions not only consume time to compose but frequently lead to miscommunication and reduced engagement between streamers and viewers. Such descriptions lack the details and can be time-consuming to write and difficult for other viewers and the streamer to quickly understand, resulting in communication gaps and reduced engagement [73].

Previous research has attempted to address these referencing challenges through various visual annotation approaches [49, 73]. These systems [49, 73] have integrated pre-defined and free-form annotation tools for users to attach snapshots and sketches with chat-based comments, aiming to provide a more direct and visual form of communication. However, these snapshot-based methods with screenshot attachment can clutter the comment section and disrupt focus. These methods often result in coarse-grained area of interest identification and disjointed interactions, as they lack integration with textual communication. Moreover, while these approaches have shown promise in creative live streams like visual arts [49, 73], they struggle with *object-focused communication tasks* [41, 50, 56], *e.g.*, hands-on activities like maker projects, crafting, and model building, e-commerce interactions involving shared physical objects, and specific gaming streams that emphasize tracking objects of interest rather than general areas of interest.

We believe that addressing these challenges and facilitating fluid visual communication of objects in live streams is critical to ensuring effective communication between viewers and the streamer. To achieve this goal, our work draws inspiration from how digital icons and emojis have successfully enhanced expressivity in online communication [16, 63]. While emojis excel at conveying emotions and simple concepts succinctly, current live streaming platforms limit viewers to pre-defined sets of emojis that cannot capture or reference the dynamic visual content being shared [49]. This suggests an opportunity: combining the communicative efficiency of emojis with the ability to reference specific objects and moments from the stream.

Our work investigates the use of **user-captured cut-outs**, named ThingMojis, designed for object-focused communication. Unlike typical pre-defined emojis, ThingMojis allow users to efficiently communicate context in text messages while preserving the natural flow of conversation. ThingMoji offers a unique way to enhance live-stream interaction by representing key objects

of interest. It is *object-based*, capturing items identified by viewers; *context-based*, preserving the whole imagery from the stream and allowing users to create narratives enriched with contextual meanings; and *time-based*, turning selected live-stream moments into expressive cut-outs aligned with the video timeline.

With design principles derived from prior literature, we developed a live streaming system, StreamThing, as a prototype to investigate the use of ThingMojis during object-focused live streaming contexts. StreamThing enables viewers to efficiently create ThingMojis by brushing over objects they want to reference in the live video. These ThingMojis can then be seamlessly embedded into chat messages like regular emojis, carrying with them the rich context of when and where they were captured in the stream. The system also helps viewers and streamers keep track of these visual references through a timeline that shows when different ThingMojis were used throughout the stream. We conducted a deployment study with three streamers and 29 viewers using StreamThing. Our investigation focused on understanding how viewers create and use ThingMojis, how these visual elements integrate into stream communication, and how their meaning and utility evolve over time. The findings demonstrated that viewers used ThingMojis as an expressive visual language channel for referencing and self-expression, while also revealing important challenges in comprehension and organization that inform future design of similar systems.

In summary, our contributions include:

- The design and implementation of ThingMoji in a live stream platform that streamlines referencing and communicating spatial and temporal contexts in live video.
- Results from the deployment study about the use of ThingMoji for in-stream communication during object-focused live-streaming contexts.

2 Related Work

In this section, we review the previous literature on rich in-stream communication, video summarization, and user-generated emojis.

2.1 Rich In-Stream Communication

Previous work supporting viewer inputs has proposed to enrich the predominantly text-based communication, *e.g.*, by text summarization [48, 49]. However, referencing a specific object within the visual context remains challenging. Several studies further proposed to involve multi-modal interactions beyond text inputs [4, 49, 73], and tools like SnapStream [73], LiveMâché [26], StreamSketch [49] enabled users to reference streaming content for communication via sketch-based interactions. However, these design risks overwhelming users when overloaded with context-rich snapshots, leading to scalability issues. In addition to viewer-generated content, VisPoll [8] offers an alternative approach with visual inputs such as symbols and shapes for polling. However, it primarily caters to streamer-initiated interactions and does not focus on images of physical objects that contain more complex information. Furthermore, another line of research focuses on the direct representation of social signals and incorporates physiological data such as heart rate [60] to represent audience signals in games and streaming.

In summary, most of these works require efforts for either the streamer to design and interact with the user interface for viewer input or viewer-driven methods that require the viewer to archive notes. As a result, several works that aim to support such "viewer input" techniques found some time delays or the streamer's split of attention between the streamer-viewer interaction and their main interaction. This problem can be even more severe for the sharing of physical space, as this type of activity may cause the split of attention between interactions in the streamer's physical

space, and the interactions in the digital space. Our system enables user-captured cut-outs that can be embedded in the text, and can then be visualized on the timeline for later review.

2.2 Custom Digital Icons, Emojis, and Stickers for Digital Communication

In computer-mediated communication, people use emojis, icons, and stickers to help self-expression and reinforce the intent of a message [11], which is a different visual channel from text messages. Previous work has found that emojis can be used for self-expression, which can be essential in video-mediated communication, such as video conferencing settings, to compliment the lack of non-verbal cues [36, 39, 46, 83]. These emojis and icons were found to be useful for clarification of the message intent [46] and can serve as symbols for jokes, enable the creation of visual narratives, and demonstrate or express interest in relationships [11, 67, 70, 83]. Emojis can not only supplement text, but also convey concepts without textual accompaniment [10, 33, 34]. One line of research focused on the design of emojis and icons, as preferences may vary depending on their applications [43]. For example, Snapchat's Scissors feature allows users to create personalized stickers from videos in chat interfaces. However, this feature lacks contextual information and does not preserve the liveness of the video during video-mediated communication. Emojis were also used during video-mediated platforms such as Zoom for providing reactions to remote users, which has also been explored during video-mediated communications such as video-conferencing [7, 36]. Streamers also use some predefined graphic symbols for self-expression so that viewers can request visual effects to be overlain on the live streams. However, there is limited flexibility for viewers to generate graphic symbols. Live-streaming platforms such as Twitch also have their bank of emojis, developed by the company and available for use while using the platform. Furthermore, streamers can create their own Twitch emotes [38].

Our work, instead, focuses on supporting the user-captured cut-outs to improve streamer-viewer communication by enabling extracting items of interest from the shared visual spaces during live-streaming. We aim to investigate the use of ThingMojis, which enable extracting items of interest from the shared visual spaces during live-streaming, for text-based communication. Unlike the platform-provided or streamer-created emotes, ThingMojis are user-captured and allow for a more flexible and personalized way of expressing ideas and referencing shared content within the live-streaming context.

2.3 Gesturing, and Remote Referencing of Visual Content on the Shared Visual Space

Earlier work in CSCW systems focused on the support of remote referencing and gestures to enhance non-verbal cues during video-mediated communication and collaboration. Different referencing methods such as the use of cursor [56], laser pointers [28], annotations [3, 20], visual overlays of bodies and arms [65], or malleable videos and screens [22, 30], and force feedback and Vibrotactile approaches [55, 71] were applied and explored to provide non-verbal cues and remote gestures.

One line of work in *shared visual space* [40] focused on augmenting shared video streams [15, 29, 32, 61, 76, 77]. Recent work, such as ThingShare [29], has implemented real-time segmentation using COCO data sets to identify objects in video conferences to augment the video stream of users. However, this line of research has primarily focused on one-on-one or small-group communication [5, 15, 29, 31, 32, 59, 76, 77], which can be different from one-to-many communication. Additionally, it emphasizes the expressiveness and malleability of sharing the snapshots within the user's personal video window.

Another approach is to enhance the comment section for reference purposes during video-mediated communication, which is more prevalent in one-to-many scenarios that rely heavily on text-based interactions. This has been explored in both collocated and remote communication contexts as digital backchannels, forming a secondary layer that augments the primary focus of

attention [14, 52, 54, 69]. Such backchannels are particularly relevant in settings such as educational lectures, presentations, conferences, and live streaming events. Very few work investigated the referencing of things in real-time comment space. For example, SlideSpecs [69] references the chats with the slides during the feedback phase of the presentation. For asynchronous comment section such as Youtube videos, Yamand *et al.* [75] investigated the pattern of time-based referencing, and identifies visual entities that include actors, objects, and events.

Extending these works, our work investigates the referencing patterns with the use of ThingMojis in the text-based comment section, and whether that augments or impedes in-stream communication for object-focused sharing. Unlike ThingShare [29], which augments the video feed to highlight physical objects controlled by the presenter, ThingMoji introduces an emoji-like referencing system that allows remote viewers to actively and collectively participate in discourse using cut-outs from the streamer's video. This shift enables distributed agency in one-to-many settings, where text-based visual referencing can enhance engagement beyond the limitations of video-embedded approaches.

3 Designing User-Captured Cut-outs for Object-Focused Live Streaming Scenarios

In this section, we discuss challenges we faced when iteratively building user-captured cut-outs, which we coined as **ThingMojis**, for live streaming communication. We highlight what makes such a process particularly challenging and emphasize important aspects that interactive systems should support or alleviate for object-focused communication. We reflected on these challenges through the lens of prior work and identified a list of design considerations (**DG1-DG6**) that informed the design of ThingMoji and the complete system we describe in Section 4. These design considerations also align with our research questions (**RQ1-RQ4**), which we explored through ThingMoji as a prototype and later revisited in our user study.

Our workflow for integrating user-captured cut-outs into real-time video contexts was shaped by insights from previous studies on live-streaming systems, particularly those focusing on integrating visuals into chat interfaces. A summary of these workflows is presented in Figure 1.

3.1 Supporting Seamless and Precise Object-Focused Communication During Live Streams

Conveying expressions and questions demand a clear vocabulary, and simple words might not suffice in all situations. For example, when trying to communicate a specific object, a vibration sequence, or the spatial movement, it maybe easier to gesture the object to the remote streamer than use words. Using words might be limited to provide clear questions, comments, or emotional expressions towards an object, during live streaming with rich visual content.

Previous research has explored various approaches to this challenge. Systems like SnapStream [73] and StreamSketch [49] introduced annotation capabilities through predefined shapes (*e.g.*, circles and boxes) and freeform sketching, enabling viewers to reference specific areas during streams. These annotation methods have proven effective in art-focused streams, fostering engagement through creative visual referencing and contributing to the attention directed to different *visual representations* of objects of interest within shared visual spaces [40]. However, the focus on visual art streams (Figure 1 - C1), where content changes are gradual, makes them less suitable for scenarios requiring rapid interaction or precise object referencing.

Nevertheless, we see opportunities to support additional interaction needs in object-focused contexts, where referencing must be both swift and precise. While SnapStream and StreamSketch allow screenshot attachments and inline annotations, frequent or large-scale image insertions can occupy substantial space in chat windows. While instant messaging platforms often handle images and text without major disruption, they typically do so in smaller group or one-to-one settings, and

usually without a simultaneous live video feed demanding the audience's attention. In contrast, multi-viewer live streams involve a large, concurrent audience that must track both real-time video content and a scrolling chat, making large or frequent images more likely to clutter the conversation and disrupt viewer focus [49, 73] (Figure 1 - C2). Though object-focused referencing has been well-studied in video-mediated communication [44, 56] - particularly for remote collaboration, assistance, and hands-on activities [41] - its application in one-to-many streaming scenarios with text-based communication remains under-explored.

These observations motivate our goal of designing a live streaming system that enables more precise object-focused communication. We propose addressing these challenges through mark-based cut-outs (Figure 1a), supporting *fine-grained identification* for object referencing. This approach would allow users to precisely highlight object contours within the video stream while maintaining conversational flow. By enabling more accurate object identification, the system can facilitate detailed commenting and feedback without cluttering the interface. However, the implementation must carefully balance precision with usability - complex workflows or delays found by prior work [73] could disrupt the natural flow of communication.

> **DG1 - Support seamless and precise creation of ThingMoji:** The system should minimize interruption during ThingMoji creation through quick, intuitive object selection that integrates seamlessly with the chat interface.
> **DG2 - Support instant previews and adjustments:** Users should be able to preview and refine ThingMojis rapidly within the chat interface to maintain message clarity and reduce miscommunication.
> **DG3 - Highlight with context on demand:** The system should maintain a balance between highlighting objects of interest yet preserving their surrounding context [2, 9]. This approach enables users to focus on specific elements, with the option to access broader contextual relationships as needed.

These design goals address our first research question:

> **RQ1:** How do viewers employ ThingMojis to reference objects and contexts during live streams, and what impact does this process have on engagement and distraction?

3.2 Enhancing Expressive Communication by Embedding Visuals in Live Chats

During live streaming, streamers and viewers communicate through text-based messages in the comment section. ThingMojis function as *spontaneous visual elements* embedded within chat messages, making communication more expressive and context-aware, similar to the role of emojis and digital icons. The difference between these impromptu, user-captured cut-outs and pre-defined emojis or stickers lies in the ad-hoc nature of ThingMoji creation and the condensed information that they contain. This spontaneity introduces challenges in categorization and retrieval, echoing prior research on emoji categorization and recommendation [23, 35], and the use of emojis and hashtags to annotate user's nuanced emotions such as confusion and curiosity [78, 79]. Grouping or tagging have also been used in conventional computer-mediated communication systems (*e.g.*, [78]).

Beyond proactive creation and grouping of these emojis, we aim to understand how people comprehend chat messages that combine text and user-captured cut-outs identified by other viewers. This differs from prior work [49, 73], where images and text were separated in the chat.

> **DG4 - Enhance expressive messaging:** Embed ThingMojis seamlessly into text messages to enrich communication, similar to how emojis enhance emotional expression and referencing.

Embedding these cut-outs directly into the text helps bridge the gap between textual and visual communication, allowing users to convey complex ideas quickly and with more nuance. However,

this integration also presents challenges, such as ensuring that messages containing ThingMojis remain clear and easily interpretable, even in fast-paced conversations. The seamless blending of ThingMojis within text must support emotion, humor, and reference creation, enriching the expressiveness of both viewers and streamers.

By construing the "enhance expressive messaging" goal, we aim to investigate how viewers may use ThingMojis for object-centric communication **(RQ2)** and how users (both viewers and streamers) comprehend messages with ThingMojis **(RQ3)**.

> **RQ2:** In what ways do viewers integrate ThingMojis into chat messages to reference content, comment, or express emotions during live streams?
> **RQ3:** How do streamers and viewers interpret messages enriched with ThingMojis, and what challenges arise in comprehension?

3.3 Maintaining Engagement Through Reusable Cut-Outs and Time-Based Navigation

Prior work has identified the loss of context when joining a live stream in the middle [48, 74] and the challenges of understanding the contexts based from piles of chat messages [8]. Beyond augmenting chat interfaces, interactive overlay extensions[1] allow streamers to track and visualize viewer interactions based on click locations [21]. However, these interactions are often ephemeral and lack persistent context. User-captured cut-outs were not only aimed for ad-hoc and impromptu chat-based communication but also preserve the timestamps of those messages sent with these cut-outs, which can be post-hocly pinned on the timeline (Figure 1c) for viewers to catch up with the overview of the stream and navigate through the long, chat-based comments.

User-captured cut-outs support both *immediate engagement and long-term narrative building* in live streams. ThingMojis should be reusable, offering users the flexibility to construct their own narratives—unlike prior work [49, 73], where interaction is limited to tagging or replying. Prior work used textual notes to annotate time-based media such as videos where color-coded rectangles were displayed on the timeline for navigation (*e.g.*, [45, 57, 58]). Video streaming platforms like Bilibili[2] enabled time-based anchors with emojis to highlight key moments in live streams. Expanding on these approaches, the system should provide a structured yet flexible navigation method that helps users quickly locate and visually categorize content, bridging past and present interactions.

> **DG5 - Create a reusable visual backchannel:** The system offers a gallery as a portal that collects all ThingMojis, making it easy to quickly find, reuse, reconstruct their own messages and narrative.
> **DG6 - Provide time-based navigation:** The system provides a timeline that allows users to revisit key moments efficiently using ThingMojis.

With the reusable and time-based navigation capabilities of ThingMojis, our aim is to investigate the following research question **(RQ4)** around user's perception of ad-hoc versus longer-term engagement of ThingMojis:

> **RQ4:** How do viewers and streamers perceive the evolving meaning of ThingMojis over time, from ad-hoc use to longer-term engagement?

4 ThingMoji

We developed StreamThing as a prototype to explore how live-streaming systems can better support *object-focused communication* (**RQ1**), how the *use of user-captured cut-outs* affect streamer-viewer

[1] http://heat.j38.net/
[2] https://www.bilibili.com/

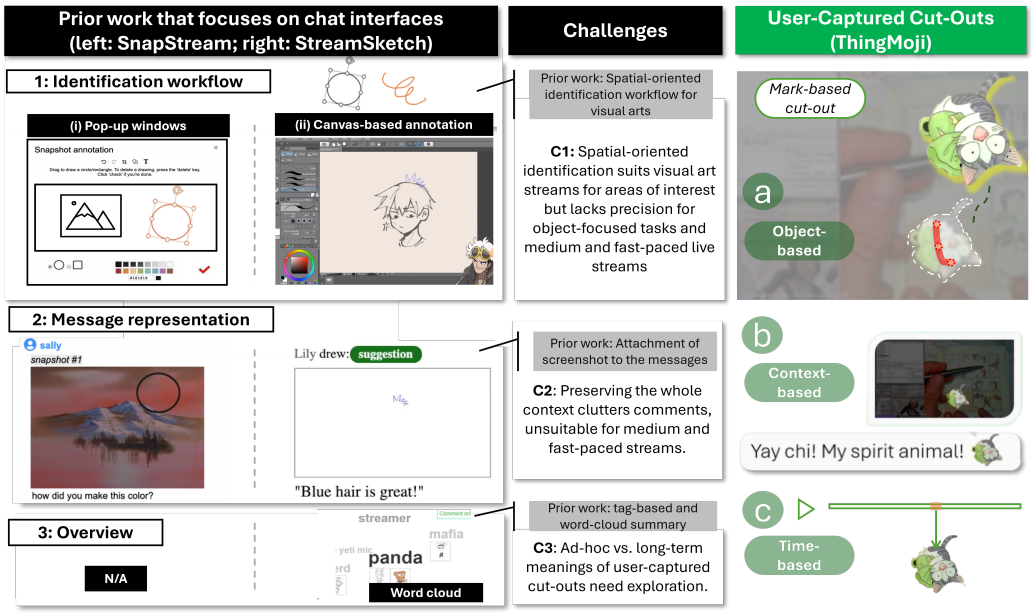

Fig. 1. Summary of prior systems based on 1) *Identification workflow* that describes the method used to identify or annotate objects from live video, which can be grouped as pre-defined annotation [73], freeform annotation [49], and mark-based cut-out; 2) *Message representation* that describes how the messages are represented, which can be separating images and text versus embedding cut-out objects into text; 3) *Overview*: provides an overview or summary of the captured visual content for navigation.

communication(**RQ2-RQ3**), and how the *short-term vs. long-term* meanings of user-captured cut-outs be perceived by viewers and streamers (**RQ4**)?

ThingMoji is unique in being *object-based* (Figure 1a), directly representing objects of interest identified by viewers; *context-based* (Figure 1b), preserving imagery from the video stream and enabling flexible combinations of user narratives with contextual meanings; and *time-based* (Figure 1c), transforming fragmented live-stream moments into expressive cut-outs distributed across the video timeline.

4.1 StreamThing: UI and Workflow

StreamThing consists of four main components - the video stream that shows the streamer's live video scene (Figure 2a), a comment box (Figure 2c) that displays all the viewers' messages, the ThingMoji Gallery (Figure 2e) that stores all the ThingMojis, and the ThingMoji Timeline (Figure 2h) that organizes all the ThingMojis based on the message timing.

4.2 Streamlined ThingMoji Creation

Prior work [73] highlighted the inefficiency of the creation and annotation of snapshots, with frustration with the back-and-forth nature of the process involving the pop-up window, brushing, and confirmation. Building on these insights, we designed ThingMoji's creation process to prioritize efficiency and seamless interaction **(DG1)**.

4.2.1 Direct Creation Process. We eliminated the pop-up window to minimize any potential friction during iteration; we shifted our focus towards creating a more direct ThingMoji creation workflow,

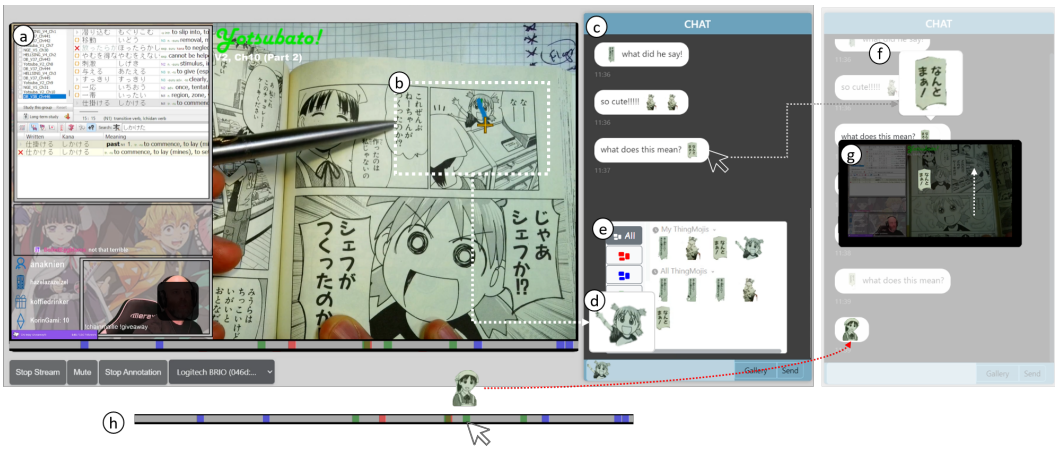

Fig. 2. System Overview: When watching the (a) video stream, viewers can (b) brush on a specific area of interest, which will create (d) **ThingMoji** that can be (c) embedded in the text-based messages, stored and shared in the (e) ThingMoji gallery. Users can hover to see (f) an enlarged view and click to see the (g) original image. (h) The colored segments on the timeline represent moments ThingMojis were used in the chat, and hovering on them shows the used ThingMoji. Clicking on the segments will direct the user to the related comments.

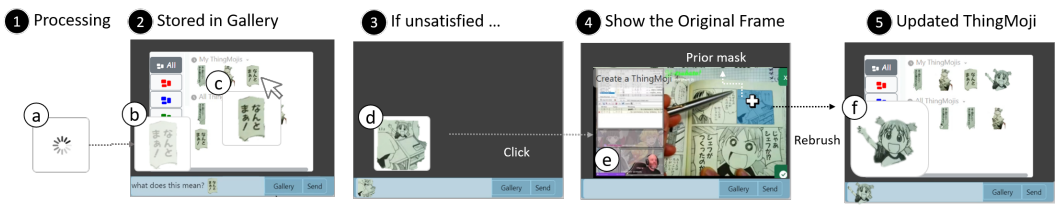

Fig. 3. The Workflow of Modifying ThingMojis: (1) The created ThingMojis will be updated immediately in the Gallery, with the preview window vanishing in six seconds. (2) Once the user creates a new ThingMoji, the preview window will be shown as (b) animation to notify the user. (3) If the user is not satisfied with the ThingMoji, they can click on the thumbnail, which shows the (4) peek window of the original image, with the prior segmentation mask highlighted in light blue. The user can brush on the peek window, (5) and the system will re-estimate and update the (e) ThingMoji in the preview window and Gallery.

allowing users to annotate videos directly rather than following the "brush and confirm" procedure. The annotation mode was enabled by default. Users can disable it by toggling the "Annotation" Switch. Upon the user annotation on the video (Figure 2b), the system generates ThingMojis, which are small visual representations of the highlighted areas of interest. To facilitate precise selection during rapid visual changes, the system automatically pauses the video when users begin their annotation, resuming only after they complete their selection.

4.2.2 Instant Preview and Refinement. In video streams with rapid visual changes, users often lose the opportunity to modify their ThingMojis once the previous image frame has passed.

The system provides immediate visual feedback through a preview window (Figure 2d) that appears next to the text input area **(DG2)**. The ThingMoji also appears in the text input box and is saved in the Gallery. This preview window displays an enlarged version of the newly created

ThingMoji and remains visible for six seconds before fading away (Figure 3b). During this time, the system automatically adds the ThingMoji to both the text input box and the Gallery. Once this preview window disappears, further modifications to the ThingMoji are not available.

4.2.3 Modification Workflow. If a user was unsatisfied with the most recent ThingMoji, they can access a refinement interface by clicking the preview window (Figure 3d). This opens a peek window showing the original frame with the current selection highlighted in light blue (Figure 3e). Users can make new annotations on the original image and see the updated ThingMoji instantly (Figure 3f). This process can be repeated until users are satisfied with the result.

This streamlined process ensures accurate ThingMoji creation while maintaining fluid interaction between viewers and streamers. The ephemeral nature of the preview window encourages spontaneous creation while still providing opportunities for refinement when needed.

4.3 Context-Rich Message Representation

4.3.1 Interacting with ThingMoji Messages in the Chat. A comment box (Figure 2c) is integrated into the system, allowing users to exchange messages and engage in conversations. To enable the expressive messaging (**DG4**), the user can seamlessly insert ThingMojis anywhere in the message just like a typical emoji (See Figure 2c). Furthermore, the system should preserve the broader context of referenced objects **(DG3)** to support further exploration and meaningful engagement, and avoid loss of context on demand. This balance between precision, usability, and minimal interference is essential for effective communication in live streams. Hence, the design is focused on balancing the focus and contexts and allows the users to not only examine ThingMojis in detail but also to preview their original image contexts. Once a message containing ThingMoji(s) is sent, both streamers and viewers can hover over the ThingMoji in the chat messages to view an enlarged version of it (Figure 2f). Clicking on the ThingMoji on the message expands it to the center of the chatbox with the original image. The surrounding elements in the image were dimmed to draw focus to the ThingMoji (Figure 2g). Once the original image is shown, users can scroll up and down to adjust the size of the original image and explore its finer details.

4.3.2 Sending Messages from Gallery. The ThingMoji Gallery provides personalized content organization through two main sections: "My ThingMojis" for user-created content and "All ThingMojis" for community-wide sharing. Furthermore, to enable the users to view in detail the ThingMojis created by any user (**DG5**), Users can preview enlarged versions of any ThingMoji through hover interactions (Figure 3c), and insert them into messages like traditional emojis, enabling dynamic communication that combines text and visual elements. This functionality allows for a more dynamic and expressive communication style, enabling users to flexibly convey complex ideas and emotions through a combination of text and visual elements.

4.4 Automated Time-Based Context Overview

ThingMoji is not only used for short-term and ad-hoc chat-based communication but also can be used for the high-level visual summary of the streaming context, where it supports users to quickly navigate and explore the viewer input during the live stream on a timeline such that they can jump back to specific moments based on the ThingMoji timestamps. Previous work used tagging [78] or word cloud visualization [49] to filter and group multi-modal messages. Some work also investigated comments on Youtube videos with time-based referencing [75]. However, these methods often required users to explicitly mention or describe images for effective filtering and tagging. Additionally, ThingMoji was seen as a potential substitution for text descriptions in conveying complex objects, where naming for filtering purposes might be impractical because of its spontaneous nature.

To enable the interaction between the ThingMoji timeline (Figure 2h) and the chat interface (Figure 2c), ThingMojis are visually organized on a timeline as colored segments (Figure 2h), supporting time-based cross-referencing to quickly navigate to the prior context via visually identified ThingMojis **(DG6)**. Users can hover over segments to preview associated ThingMojis and click to access related chat messages. This design facilitates quick navigation to previous contexts through visual references.

To simplify content organization without increasing user burden, the system automatically groups ThingMojis using deep learning-based analysis of visual attributes and semantic meanings. The timeline displays these groupings through three distinct color categories, which are also reflected in the Gallery's sidebar (Figure 2e). This automated categorization helps viewers distinguish between different types of ThingMojis and more easily locate specific content or moments from the stream.

5 StreamThing Implementation

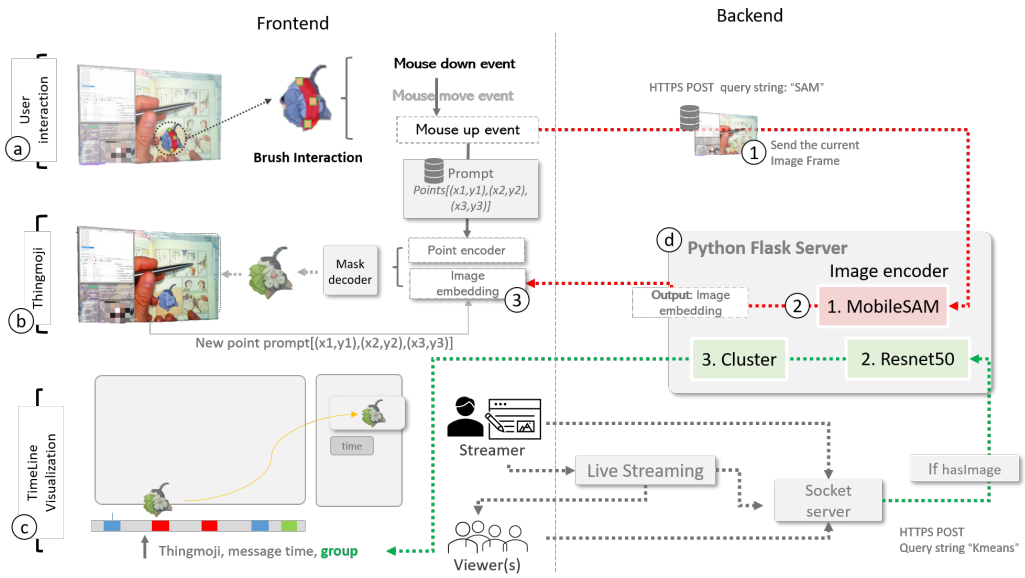

Fig. 4. StreamThing System Architecture Diagram.

The system consists of a front-end web interface and back-end signaling servers. The front-end UI was built using React, Javascript, HTML, and CSS. The back-end server was built using Node.js with socket.io for chat comments, with an ION-SFU server [1] for live-streaming between the streamer and viewers (See Figure 4).

The annotation interaction of the front-end (Figure 4a) consists of three events: mouse-down, mouse-move, and mouse-up events. Upon the triggering of mouse-up event on the video (by releasing their cursor), the current image frame is sent to the back-end python flask server (Figure 4d) via HTTPS POST with the query string to call the MobileSAM [80] function. For our cases, we tested the performance of MobileSAM image encoding on both a cloud server equipped with a Nvidia A10G GPU and a laptop with a Nvidia RTX 2060 GPU to provide image embedding within 0.1-0.5 seconds, which is significantly faster than the original SAM. The MobileSAM then runs the image encoder with a GPU that outputs the image embedding of the given image frame, returning the image frame to the front-end via HTTPS POST. Since more points will not lead to better model

performance according to [80], the continuous annotation interaction was downsampled to three discrete points. Once the embedding comes back, the three points were fed into the prompt encoder. Once the point prompt and image embedding are ready, the output of the prompt encoder and image embedding is then fed into the lightweight mask decoder that can be run on the front-end web browser CPU. The output is the final mask composed with the original image to get the area of interest and created the digital icon, named ThingMoji. To enable quick refinement, the system caches image embeddings, allowing users to modify their ThingMojis through additional point prompts if they are unsatisfied with initial results.

For automatic organization (Figure 4c), the system processes ThingMojis through a pre-trained ResNet50 [27] for feature extraction, capturing rich representations of visual content. These features feed into a K-means clustering algorithm that organizes ThingMojis into three prominent groups. The system maintains these groupings dynamically, updating the clusters as users create new ThingMojis. These organizations are reflected in both the timeline and gallery interfaces through color-coded segments, helping users navigate and access related content.

Deployment. We deployed StreamThing across two cloud servers: one with an Nvidia A10G GPU for image processing, and another for hosting the front-end, Node.js, and live-streaming services. The system was accessible through modern web browsers (*e.g.*, Chrome, Firefox, and Edge) and supported Open Broadcaster Software (OBS) virtual camera integration, allowing streamers to maintain their existing multi-camera setups common in content-rich streams [13].

6 Evaluation

Our IRB-approved study investigates how using ThingMojis enhances streamer-viewer communication through three in-the-wild deployments.

6.1 Participants

We recruited three streamers with prior experience in object-focused live streaming scenarios with rich visual content. Three Twitch streamers participated in the study. S1 is an American streamer with 2 years of experience and streamed about once a day, with 544 followers on Twitch. His content includes live readings of various manga series in Japanese, which he explains in English. Additionally, he covers maker and craft topics like cross-stitch, chain maille, incense making, and occasionally video games. S2 is a Taiwanese streamer with 9 years of experience and streamed about once a day, with 278 followers on Twitch. His content primarily revolves around model creation, gaming, and cooking. S3 is an American streamer with over 235,000 followers on Twitch. With eight years of streaming experience, including four years focused on 3D painting streams, his content includes maker and crafting topics such as 3D/resin printing, assembling plastic/metal model kits, painting miniatures, and building Lego kits. Each live stream lasted for around 55-60 minutes.

For each streamer, we recruited viewers who were interested in the live-streaming topic and regularly watch live streams. We followed prior research findings about the importance of engaging regular community members rather than paid participants [66]. We recruited through Discord channels and university mailing lists, targeting viewers in the US/EU regions for S1 and S3's English sessions, and Hong Kong/Taiwan regions for S2's Mandarin session.

In total, we recruited 29 unique viewers (ages 19-53; 12 female, 17 male): 13 viewers for S1, 13 for S2, and 12 for S3. Two regular viewers from S1's community participated, with one also attending S3's session. The viewers were active stream consumers: 27.6% watched daily, 37.9% watched multiple times weekly, and the remainder watched monthly. Their interests spanned

gaming, sports, makers, art, e-commerce, and anime. This experimental setup was widely used during the evaluation of live streaming systems [4, 73].

6.2 Procedure

We conducted a pre-session setup with each streamer via video call one day before their stream. Streamers tested StreamThing with their OBS virtual camera setup and were encouraged to develop strategies for incorporating ThingMojis into their typical streaming content. Each session was planned for approximately one hour (See Figure 5 for their streaming screenshots). Viewers joined the Discord voice channel 20 minutes before each session for orientation. This included a 5-minute instruction period followed by a 10-minute test stream where viewers could freely explore StreamThing using their desktop or laptop computers.

We video-recorded the sessions as well as chat logs including ThingMojis. Following the sessions, we collected feedback through viewer surveys and conducted semi-structured interviews with streamers. Streamers received $100 compensation and viewers received $10 per session.

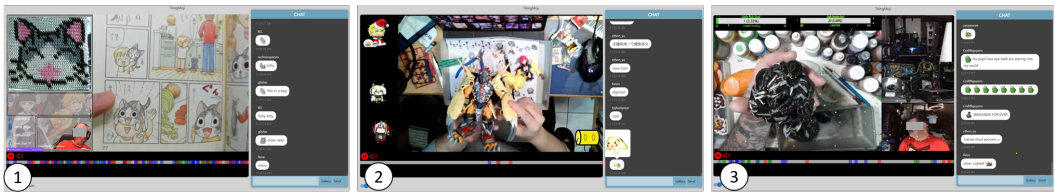

Fig. 5. Screenshot of three sessions. 1) S1's session, featuring Japanese woodblock prints, chainmaille, and manga reading with three camera angles; 2) S2's model making session with a single camera view; and 3) S3's session on 3D/Resin printing and painting, using four camera streams for comprehensive coverage.

6.3 Data Collection and Analysis

6.3.1 Data Collection and Interview Analysis. Survey responses were collected around viewers' usability ratings about the system features with 5-point Likert scales (1-Strongly Disagree, 5-Strongly Agree), and open-ended feedback about the system features compared to tranditional live streaming platforms and the use of pre-defined emojis, such as how StreamThing influenced their watching experience and comprehension of the text-based chat. Survey questions were based on prior work examining live streaming interface usability [42, 48, 49, 73]. We conducted 30-minute one-on-one semi-structured interviews with each streamer after their sessions. All interviews were recorded, transcribed, and iteratively coded by the research team. One interview, conducted in Mandarin, was translated by the interviewing researcher before analysis.

6.3.2 ThingMoji Analysis. Our analysis of ThingMoji usage followed a three-step process. First, we conducted thematic analysis of ThingMojis in both chat messages and video contexts, focusing on how they were used for object referencing and playful engagement. Second, researchers analyzed recorded videos and chat logs to identify and categorize different types of ThingMoji usage. Finally, we cross-referenced these categories with interview findings, using key episodes from the streams to validate and refine our understanding of emerging usage patterns.

7 Findings

This section presents the results of our investigation into understanding the behaviors and experiences of streamer-viewer communication.

7.1 RQ1: Experiences with ThingMojis: Engagement and Distractions

Viewer-Specific Experience with ThingMoji. Viewers reported their experiences when creating, sending and perceiving the ThingMojis (see Figure 6). Out of 29 unique viewers, 25 responded to the survey. In cases where viewers attended sessions with multiple streamers, only their response from the first session was included in the analysis.

Overall, ThingMoji enabled viewers to interact with streamers and viewers in an engaging way, and supported better understanding during live streaming. Viewers reported that using ThingMoji was fun and enjoyable (Median=4, IQR=1), and they felt more engaged in live streams using ThingMoji (Median=5, IQR=1). They also reported that ThingMoji was helpful to communicate during live streams (Median=5, IQR=1) and increased their understanding of other viewer's comments (Median=4, IQR=1), and the context of the stream (Median=4, IQR=1). Although a few viewers found that creating ThingMoji could potentially be distractive (Mediam=4, IQR=3), viewers reported that they felt ThingMoji was easy to use (Median=4, IQR=1), easy to learn how to use (Median=5, IQR=1), and quick to use (Median=4, IQR=2). They would like to use ThingMoji when watching live streams in the future (Median=5, IQR=1).

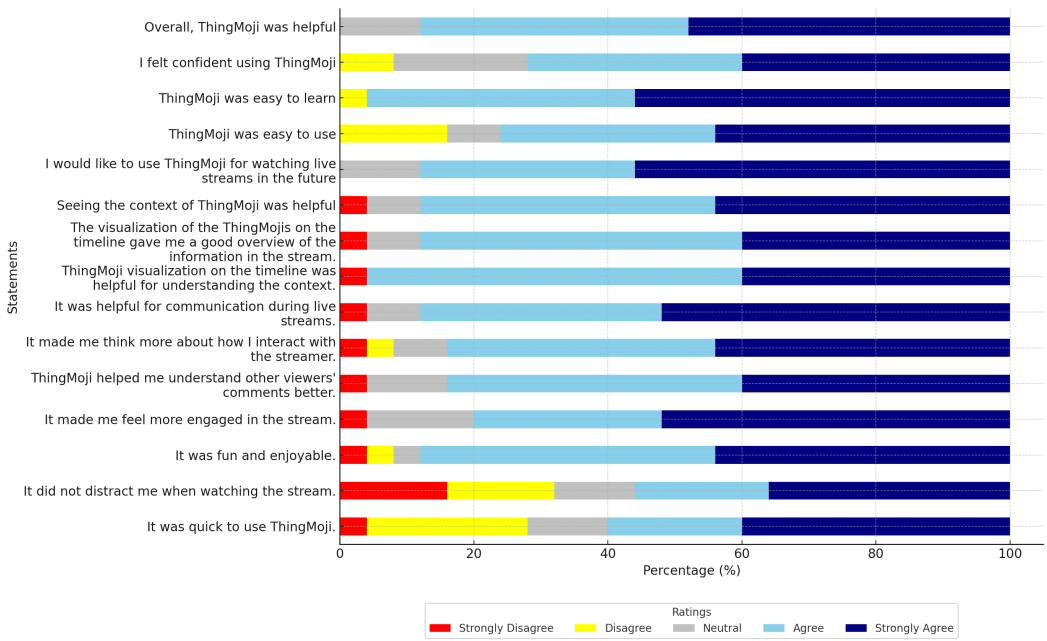

Fig. 6. Survey responses about the experience of using ThingMoji.

Here, we report viewer-specific findings on the creation process of ThingMojis.

7.1.1 Personalization and Emotional Connection with ThingMojis Compared to Pre-Defined Emojis. Viewers perceived ThingMoji creation as more unique and contextually relevant compared to traditional emojis. For instance, V3 expressed, "*The aspect I enjoy the most is creating ThingMojis. I take pleasure in capturing screenshots and including objects I like within the image. It's like crafting my own emojis that you can't find anywhere else...interacting with 'my ThingMoji' is quite intriguing, particularly when I notice other viewers sending similar emojis.*" This personalization fostered shared emotional experiences when viewers saw others using similar ThingMojis. Viewers

appreciated the flexibility of capturing and segmenting content with a single stroke. As V9 noted, "*I love how flexible and applicable it is to capture all interesting images and make them into my ThingMoji. I frequently use emojis and memes, and I find that a combination of both related pictures and text is an efficient expression when communicating with others.*" Five viewers highlighted how this flexibility enhanced meme creation through capturing funny moments, exemplified by V10's comment, *e.g.*, "*When there is a funny moment (the streamer makes some funny thing), the audience can use ThingMoji immediately and save the sticker so as to recall this moment in the future. This could build a good memory for both the streamer and the audience.*" This suggests the potential for long-term use of ThingMoji across different streaming sessions. The above findings align with our design goals of supporting seamless ThingMoji creation (DG1) and enhancing expressive messaging (DG4). The ability to quickly capture and personalize visual elements supports fluid interaction, while the emotional and social aspects of sharing ThingMojis enable richer forms of expression beyond what is possible with pre-defined emojis.

7.1.2 Distraction of ThingMoji Creation and Usage in the Chat. Despite ThingMoji's benefits, viewers reported two main sources of distraction. First, the creation process could be frustrating when captures did not align with intentions. For example, V7 commented "*It can be slightly distracting trying to get a perfect picture of the element...*" Furthermore, V9 reported "*Sometimes the capture is a little off, and editing the graphics in a short time is difficult.*" Second, a few viewers reported distractions from excessive ThingMoji usage. For instance, V22 noted, "*It may provide a lot of meaningless messages in the chat and flood out the meaningful message.*" This echoes findings of previous work about how increased interactivity can lead to unusable chat experiences and increased distraction [18].

Streamer-Specific Experience with ThingMoji. Here we reported the streamer-specific experience, especially in reflecting the experience in their type of live streaming contexts. Post-session interviews revealed streamers' positive experiences with ThingMoji across various streaming types. S1 noted its general benefit for viewer engagement: "*...this will make people more inquisitive, and gets people a tool where they can ask questions more easily, that's exciting.*"

For manga and maker content streams, ThingMoji enabled precise referencing of visual elements. S1 explained, "*It's hard for the chatter to point out what their interest on the page of the manga. ThingMoji is really powerful, for the ability to ask questions, to talk about a thing on the screen, such as it can capture the characters from the Kunji. For maker and crafting... it's also easier for me to understand what's the thing on the table that the chatter pointed out.*"

In educational and gaming streams, ThingMoji allowed viewers to highlight objects or characters, creating universally understandable visual references. S3, who streams model-making, educational, and gaming content, observed, "*Many elements in the game are challenging to describe, especially for new and inexperienced viewers. Even if some expert viewers use the correct terminology, it may not be universally understandable. However, if you capture a specific element as a ThingMoji, it creates a visual reference that everyone can immediately grasp.*"

In particular, streamers valued ThingMoji for streams with frequent visual changes, preferring it over full-screen captures. S3 mentioned, "*For live-streaming with more visual changes, I would prefer ThingMoji as it shows the most recent status and user's real-time expressions of the live streaming.*" S2 compared it to Danmaku: "*ThingMojis can serve a similar purpose to Danmaku comments but can be compressed like an emoji, making them more versatile in fast-paced streams.*"

While acknowledging challenges in fast-paced gaming scenarios, streamers noted that ThingMoji-embedded comments provide valuable context both in real-time and in retrospect. They can review these comments during breaks, post-game, or post-stream. S1 pointed out a potential area for

improvement: "*It would be helpful to have a way to organize or categorize ThingMojis, especially in longer streams with many created.*"

7.2 RQ2: Use of ThingMojis

During the three streaming sessions, 29 unique viewers contributed to in total 400 ThingMojis were mostly used in the messages (Table 1). In total, 286 comments with ThingMojis were shared in the chat. In the Session 1, viewers created 239 ThingMojis and shared 184 messages containing ThingMojis (61.5% of the total 299 messages). Notably, 79 of the 184 were ThingMoji-only messages. In Session 2, 106 ThingMojis were created. Viewers shared 57 messages with ThingMojis (52.3% of the 109 total messages). Within this, 17 comments has ThingMoji only without any text-based description, which accounts for 29.8% of the messages with ThingMojis. In Session 3, 115 ThingMojis were created, and viewers shared 45 messages with ThingMojis, making up 39.1% of the total 115 messages. Within this, 19 comments contained only ThingMoji(s), which is 42.2% of the total 45 messages with ThingMojis. This indicates that overall, more than half of the comments with ThingMojis mix the language and ThingMojis. Post-hoc analysis revealed that out of the eight viewers who did not create or share ThingMojis, one provided only text-based messages, while the remaining seven watched passively. This pattern is consistent with previous live-streaming studies [49], where a subset of viewers opt not to engage in chat.

	Session 1	**Session 2**	**Session 3**	**Total**
# of viewers watching the stream	13 viewers	13 viewers	12 viewers	29 (unique)
# of viewers who has used ThingMoji	10 viewers	9 viewers	9 viewers	21 (unique)
# of created ThingMojis	239	106	55	400
# of comments	299	109	115	523
# (%) of comments with ThingMojis	184 (61.5%)	57 (52.3%)	45 (39.1%)	286 (54.7%)
# (%) of ThingMoji-only comments	79 (42.9%)	17 (29.8%)	19 (42.2%)	115 (40.2%)

Table 1. The number of interactions for each session

The diverse usage of ThingMojis can be categorized into two main aspects: usage patterns and purposes.

7.2.1 Usage Patterns of ThingMojis in a Message. Usage patterns refer to the ways in which ThingMojis are utilized to construct messages and communicate within the stream. These patterns focus on the form of incorporating ThingMojis into the text.

Replacing Textual Words. Viewers often used ThingMojis to replace words, making their messages more visual and immediate. This pattern involves substituting parts of or entire textual messages with ThingMojis. For instance, a viewer cropped the streamer's mouse to use it as a playful response, bypassing the need for a textual explanation - " a mouse for a "

Repetitive Use of a Single ThingMoji: ThingMojis evolved beyond ad-hoc references to become recurring thematic elements in streams. Some ThingMojis were used repeatedly to establish a theme or continuity within the stream.

> An example is the pondering cat " " being used multiple times (20 times during session 1) to create a consistent theme during the session, even when it was not shown on the shared stream anymore.

S1 found it very useful to set a theme for specific events, and he used it twice during the start and the end of the streaming to warm up and responded to the viewers. He observed this emerging behavior and recognized its potential for community building: "*...You could almost at the beginning of a stream... have little images like that, where people can create like a happy cat or whatever and then let that be a theme for a stream... 'hey, when you like things or whatever, use the "happy cat", when you don't you know, use the frowny cat' or whatever...there are potentially lots of silly and fun ways to theme an event or even just like for a stream, in general, to always have like 'hey, this is our thing.'*" This spontaneous co-creation of visual themes offered advantages over standard Twitch emotes, as S1 noted: "*We kind of do that through Twitch with emotes, but you're kind of limited on Twitch. This (ThingMoji) lets you very dynamically create things based on what you're doing.*"

Using Multiple Same ThingMojis to Express Intensity. Viewers used multiple instances of the same ThingMoji to convey strong emotions or reactions. For example, sending several "hangry cat" ThingMojis to emphasize a humorous frustration.

7.2.2 Purposes of Using ThingMojis. Purposes refer to the reasons why ThingMojis are used within the stream. These purposes focus on the intentions and results of using ThingMojis in communication.

Facilitating Questions and Answers. Viewers primarily used ThingMojis to reference specific objects and ask questions about stream content.

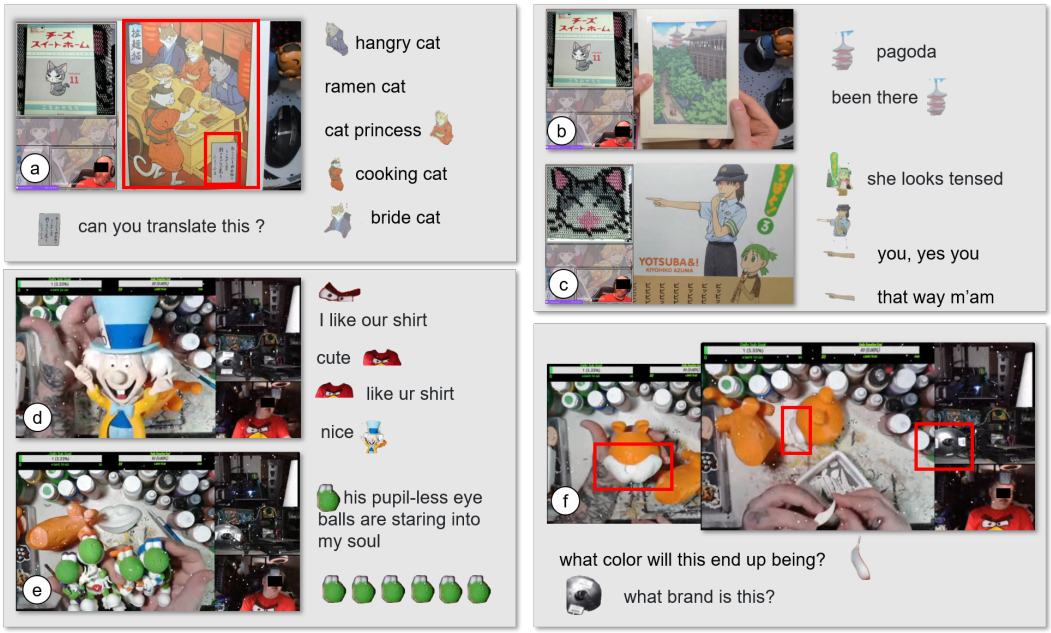

Fig. 7. Use of ThingMojis

In Session 1 (Figure 7a), a viewer asked the streamer "*can you translate this?*" Similarly, during S3's 3D painting stream (Figure 7f), a viewer inquired "*what color will this end up being?* " while referencing an uncolored piece.

ThingMojis can also be used to express and answer the questions for the streamer, for example,

> When the streamer introduced a place in the art pieces (Figure 7b), a viewer responded " pagoda" by pointing to the place and responding the name.

Expressing Emotions and Personal Connections. Viewers used ThingMojis to convey emotions about specific objects and share personal connections.

> During Session 1, a viewer captured the moon of an artwork entitled "seizing the moon" and expresses it in Japanese " *suki ga kirei des ne!*" to express to the streamer that "the moon is beautiful." Another viewer captured the cat from the manga being shared, and commented " *adorable...* "

Viewers also related stream content to their experiences, for example,

> In Session 1 (Figure 7b), a viewer commented "been there " about a location shown in an artwork.

ThingMojis were also used for playful interactions and expressions, adding a fun and creative layer to the communication. Viewers gave ThingMojis humorous names or used them to create jokes and playful comments. This also added a layer of fun or irony to a conversation:

> For example, when S3 showed a 3D-printed turtle (Figure 7e), a viewer commented " *his pupil-less eye balls are staring into my soul*", and then shared a spawning of the turtle in the next comment.

In sessions where streamers' video feeds were visible, viewers creatively engage with stream content about the streamer by utilizing ThingMoji in a humorous and interactive manner.

> In the sessions where the streamer's person video was shared, the streamer's image was cropped and labeled with their Twitch username, referred to as "Big [Streamer's Twitch name + person stream] ."Specifically in Session 3 (Figure 7d), the streamer's clothing, which featured an angry bird design, was cropped multiple times. (*e.g.*, " *like ur shirt*")

This action highlights the interactive and playful aspects of streaming. S3 had a positive session but raised some moderation concerns. As S3 noted: "*When you're streaming, you try to keep out all the bad stuff... if someone did that and said you're fat or ugly, that's where the concern comes in.*" S1 agreed, stating that inappropriate behavior was not unique to ThingMoji and could be managed with warnings or bans. These reflections emphasize ThingMoji's positive impact on engagement while recognizing the need for vigilance against misuse. Though ThingMoji did not directly cause these concerns, it could facilitate negative comments, highlighting the importance of moderating or blocking ThingMoji use in some parts of the stream.

Viewers also demonstrated clever spatial awareness:

> During Session 1, when S1 was showcasing a manga cover featuring a policewoman (Figure 7c) pointing at someone with a record, viewers playfully interacted by cropping the hand of the policewomen. They cleverly used the spatial arrangement of the streamer's stream on the left and chat messages on the right: " you, yes you", followed by another viewer commenting " that way m'am".

These interactions often reflected deeper community connections. S1 described how a long-time viewer created emojis from his hand gestures: "*It was interesting to see what people were going to decide to do. Like, [a regular viewer's username], he got, at some point, he liked to start making emojis of my fingers when they appeared on the screen, but we've known each other for a long time. I think he felt a little bit more freedom to just be kind of silly.*" Such moments highlight ThingMoji's potential for fostering playful, personal interactions between streamers and their community.

Creating Shared Visual Narratives. ThingMojis enabled shared visual narratives and collective storytelling among viewers. This was evident in the creation of copypasta chains where viewers would build upon each other's contributions to form a cohesive and creative theme.

> During Session 1 (Figure 7a), viewers built a narrative around a "ramen cats" artwork, sharing sequences of related ThingMojis:“ hangry cat”, “ramen cat”, “cat princess ”, “ cooking cat”, and “ bride cat”, with cropped related Thingmojis, when the streamer was sharing and describing an art piece called “ramen cats”.

This collaborative creation process added unpredictability to streams, as S1 observed: “*...it kind of like if you've ever played D&D (Dungeons & Dragons)...where someone is trying to tell a story, but sometimes the players don't quite go in the direction that the storyteller wants...That interface allows people to key in on anything...*” S2 also elaborated: “*...chat ends up, the group of people ends up being its own sort of entity. And they're just going to pick what kind of makes them amused. And you never know quite what that's going to be.*”

7.3 RQ3: Communicational Experiences and Comprehension of Messages with ThingMojis

Here, we report some shared feedback for both viewers and streamers, particularly regarding the communication and comprehension of ThingMojis in the chat interface.

7.3.1 The Use of ThingMojis Enhanced Communication and Comprehension. Our findings reveal how ThingMojis improved communication and comprehension for both streamers and viewers during live streams.

Viewer Feedback. Eight of 25 surveyed viewers highlighted ThingMoji's effectiveness in referencing specific objects and improving stream comprehension. As V3 noted, "*I believe this is beneficial for grasping the content of conversations among viewers, especially during live streams, where typos are common and can make it difficult to understand what the viewer is mentioning about certain objects.*" Most viewers emphasized ThingMoji's precision advantage over text, with V25 observing that “*Audiences can directly point out objects in the live broadcast footage, which is more precise than describing object by text.* ”

ThingMojis also facilitated better responses to streamer questions. V19 explained: “*... when the streamer asks which part of the art does the audience like the most, it is difficult to answer with only words. But with ThingMoji, the audience can reply the streamer with the part that they like.*” In addition, viewers appreciated the fact that ThingMojis allowed them to notice elements they might have missed during the live stream, both from other viewers' references and even things that the streamer themselves overlooked, saying, "*I can see things that I didn't notice at the moment and create content that others didn't notice.*"

Streamer Feedback. All streamers reported that ThingMoji facilitated better communication and comprehension of viewer comments in their specific streaming contexts.

S1 highlighted how ThingMoji solved the challenge of referencing complex visual elements:

> “*... When people highlighted some of the Kanji... it's something complicated and it's not something that they can describe... they would normally have to say like, 'hey, what's that character that looks like a box with a squiggle on the top...*”

S3 emphasized how ThingMoji fostered more meaningful interactions around his 3D printing work: “*I was using it [3D printing] to pick some parts... it even brings up the conversation of 'what are you printing?' or 'Why are you printing that?'*” He particularly valued how viewers could directly

reference objects: "*...they were able to, like draw that circle, colored in and bring it over and put it in chat and say, 'what's this?'*" anticipating that his community "*would have a blast with that.'*"

7.3.2 Concerns About Comprehension of ThingMojis. We asked streamers and viewers to recall and reflect on different moments where they found it could be hard to understand ThingMojis. Here, we summarized the possible reasons attributed to the (mis)comprehension of ThingMojis.

Lack of textual context: Streamers, in particular, encountered challenges in understanding ThingMojis lacking accompanying text descriptions. A significant portion of messages (42.9% in Session 1 and 29.8% in Session 2) contained only ThingMojis without text, creating interpretation challenges. S2 expressed difficulty in interpreting the intended meaning without any text descriptions, stating, "*.. I couldn't discern the significance when they didn't include text... It's challenging to grasp the message behind Thingmojis without additional text.* " S1 elaborated that while some ThingMojis were self-explanatory (like the pondering cat), others required context: "*.... some (ThingMojis) are clearer than others, right? Like...they grab the picture of the pondering cat...but when it's just getting my mouse, Um, interesting, but I don't know what to say about that...*"

Clarity of captures: The effectiveness of ThingMojis can also be compromised by unclear captures, leading to confusion and reduced utility. S1 described struggling with a "chainsaw dog sculpture" capture - "*...I couldn't quite initially tell what it was and it took me some time...*" In such cases, streamers often defaulted to ignoring unclear ThingMojis unless they had time to seek clarification. As S1 mentioned "*If someone has not necessarily gotten a good grab of something... unless I've got time to ask, 'Hey, what was that?' I'd probably just ignore it, and just move on to something that was more clear...*"

Interaction Overhead for ThingMoji Preview. While the hover feature for enlarging ThingMojis improved message comprehension for all users, its implementation created different challenges for streamers and viewers. S3 appreciated the enlarged view functionality: "*...because when you see the object in the chat, you're able to go and bring it up and it brings it up bigger. So that makes it easier...* " They also liked the options of click and see the whole context. While viewers can easily hover and click on ThingMojis, streamers face interaction overhead, particularly those engaged in creative crafting and gaming. They found the manual hover interaction burdensome and suggested an automatic preview system that would temporarily display enlarged ThingMojis without requiring explicit interaction, allowing them to maintain focus on their primary streaming activities while passively monitoring chat content.

This disparity between viewer and streamer experiences points to broader challenges in managing multimodal chat interactions during active streaming. Along with the previously noted issues of textual context and capture quality, these interaction challenges suggest a need for more streamlined approaches to ThingMoji display and management.

7.4 RQ4: Spontaneous VS. Long-Term Usage of ThingMojis

7.4.1 Varied Utilization and Perceptions of the ThingMoji Gallery. Among 29 viewers, ThingMoji Gallery usage patterns revealed different preferences for spontaneous versus archived content. Ten viewers used the Gallery multiple times, while others emphasized the ephemeral nature of ThingMojis. Some of them would not be interested in using Gallery once a ThingMoji was used, *e.g.*, V11 noted - "*No, because the topics in stream is changing and not going back.*" Some viewers preferred creating new ThingMojis over reusing existing ones, with V2 noting: "*I used them (those in the gallery) a few times. It was much more interesting to create your own.* " "*No, If I want to show something to the chat room, I would rather capture myself than search the Gallery* (V21)"

However, the Gallery served different purposes for different users. About half the viewers valued its convenience, as V9 noting "*it is quicker (to access the Gallery) than making a thingmoji...*" Five viewers regularly used it to catch up with stream content, as V16 described - "*Nice to go back and see what others clipped so I could ask similar questions or see what I might have missed.*"

However, V22 expressed mixed feelings about the use of Gallery due to the potential clutter compared to traditional galleries of pre-defined emojis, noting, "*Yes, since many people have already make some better stickers of which I want to capture, it is more convenient to use the gallery. But it may be flooded by a lot of meaningless ThingMojis and hard to find some well-captured pieces.*"

This indicates a need for balancing accessibility with organization to ensure the Gallery remains a useful tool for viewers. This also suggests that the introduction of ThingMojis may create barriers for users in locating newly generated ThingMojis within the gallery. Even with the three colors categorizing the visual categories of the ThingMojis in the Gallery, most viewers were mainly paying attention to the "All ThingMojis" section to check the ThingMojis created by other users.

7.4.2 Balancing Detail and Overview in Time-Based Visualization. The timeline visualization feature received mixed feedback from both viewers and streamers.

Viewer Feedback. Viewers found timeline visualization useful for catching up with the context, as V2 noted, "*I really like that I can use ThingMoji to look back at the streaming content...*" V5 added, "*If I need to scroll back to find out what happened and why they are laughing, I have to go back to the specific area, but ThingMoji visualization helps me get a sense and jump to the correct moment.*"

Furthermore, a few viewers expressed mixed feelings about the timeline visualization, noting that it can become cluttered with too many colored boxes, particularly during Session 1 with a high volume of ThingMoji-embedded messages, suggesting automatic filtering based on real-time user interaction data. While the time-based attributes of ThingMojis allowed for sequential browsing of ThingMoji-related messages, some viewers expressed a preference for textual descriptions of chat events mediated by ThingMoji. V23 commented, "*...I think it depends on the time range. For example, I preferred seeing more detailed, time-based information for the past 10 minutes, yet a more high-level textual summary embedded with ThingMojis, for the past hour...*" This highlights how the perception of time-based information varies with different time frames. Immediate contexts may benefit from detailed, high-fidelity information, facilitating attention to and interaction with individual messages, whereas longer contexts may benefit from a more summarized, high-level overview of events.

Additionally, viewers commented on the colored groupings on the timeline, which many had overlooked during the study. Some viewers (3/29) suggested topic-based segmentation over purely chronological display, *e.g.*, V3 noted - "*...it can still contain the high-level sections summarizing each time segment, and then within each time segment, there could be these groups of thingmojis...*" This highlights a macro level that divides the timeline into larger segments with summaries, and then a micro level that organizes ThingMojis within each segment into meaningful groups.

Streamer Feedback. The need for ThingMoji timeline can be diverse based on the streaming contexts. By working on different aspects of a project in each stream (like the dragon model mentioned), S3 maintains viewer interest over time by showing the project progression of objects. "*...Like the dragon I'm going to be doing, the dragon I show, that's going to be one of... we're almost done with it now. So when I start working on the wings and start assembling it...more people are gonna probably come in because they've been waiting for the dragon...*" Viewers return to see the progress and eventual completion of these projects. This approach of showing a work in progress and the anticipation of the final outcome keeps the audience engaged and coming back for more, and echoes the importance of visualizing the ThingMoji timeline for viewers who joined later. The importance

of the timeline was perceived differently. For instance, in manga-sharing sessions, its use may shift towards separating apart specific elements like characters and Kanji in speech bubbles, as noted by S1, contrasting with the project-based streams where the focus is on the progression of objects.

8 Discussion

The results from study highlight the numerous benefits and challenges faced by streamers and viewers when engaging in different types of live streaming.

8.1 Reflecting ThingMoji Characteristics for Communication

In our study, we explored the use of ThingMojis as a means of interaction during live streams. While slower-paced or art-focused streams may benefit from full screenshots or sketches to capture broad areas [73], ThingMojis enable swift, emoji-like referencing of specific items directly in chat. This agility is especially valuable for tasks requiring precise identification (e.g., model-building or product demos). However, heavily cropped images risk ambiguity, and their frequent use may distract viewers during fast-paced discussions. Meanwhile, comprehensive screenshots can clutter the interface and slow the flow of conversation. By allowing both fine-grained cut-outs and larger annotated screenshots to coexist, streamers and viewers can choose the most suitable method for each context, thereby retaining greater control over communication and collaboration. Future work should further investigate how different interaction tasks affect users' preferences for these complementary techniques.

8.2 Design Implications

Here we reflect on some design implications for future design of user-captured cut-outs to address the challenges of comprehending, customizing, and categorizing ThingMojis. It is essential to recognize that user needs and preferences vary, and not all recommendations will be applicable to every user or live streaming category.

8.2.1 Comprehension Challenges with ThingMoji. Some pre-designed emojis are used as a sequence [34] as a powerful vocabulary without any text-based accompaniment. However, when used alone and without text, ThingMojis were perceived to be difficult to understand. In addition, some types of ThingMojis were found to be more challenging to understand than others, mainly due to the depth of information it can encapsulate.

Furthermore, similar to the ambiguities frequently found in text-based messages for communication, the use of ThingMoji can contribute to misunderstandings, especially when their meaning is not immediately clear or their quality is poor. ThingMojis, just like some language nuances and emojis usage that can lead to different interpretations, may carry varied meanings based on context.

One possible way is to incorporate automated and suggestive text-based descriptions using large language models that aligns with the chosen ThingMoji and the overall video context to aid in reducing ambiguity. This approach would provide immediate contextual backing, and human-in-the-loop to help viewers re-configure their communication in a timely manner, making the intent and meaning of ThingMojis more accessible and explicit.

Customization and Style Generation of ThingMojis. As identified by [75] as the taxonomy of referent types for visuals, ThingMojis, as static emojis, are effective in referencing characters and objects but less so for dynamic events or occurrences in the sequential video frames. Furthermore, current ThingMoji supports direct subtraction from the scene which did not explore the expressiveness spectrum. This limitation opens up opportunities for enhancing ThingMojis through customization, blending, and style generation, possibly leveraging diffusion models. This approach could align with prior techniques used in VisiBlends [6] or methods for creating compound icons [82], offering

a richer, more nuanced way to convey complex ideas through ThingMoji. Such advancements would expand the utility and expressiveness of ThingMojis in digital communication.

8.2.2 Organizing ThingMojis with Different Time Frames. Viewers and streamers have expressed varied opinions on the grouping aspect of ThingMojis on timeline and gallery.

The viewer feedback regarding the timeline visualization suggests that the perception of time-based information varies significantly with different time frames. For immediate contexts, viewers benefit from detailed time-based information from the timeline that facilitates attention to and interaction with individual messages. In contrast, for longer contexts, viewers reported the preference towards less detailed but more summarized information to provide a clear overview of events. To enhance the user experience, future designs could incorporate a hybrid approach that combines detailed and summarized information [68]. For example, providing detailed visualizations for the last few minutes and high-level summaries for older messages could cater to varying viewer preferences and improve overall satisfaction.

The findings from streamer feedback emphasized the context-dependent nature of the time-based ThingMojis organization. Streamers' preferences for organizing ThingMojis vary based on the type of content they produce. Streamers, particularly in maker and crafting streams, emphasize the value of progressive presentation, where showcasing the development of projects is key to attracting viewer interest. For them, a chronological organization of ThingMojis might be most beneficial. Conversely, for the streamer who shared manga-related streams, the focus shifts to effectively organizing different main components within the video stream like characters and Kunji into different groups. This indicates that a more categorical organization might be preferred in this context, and demonstrates the need for adaptable grouping strategies tailored to the specific type of stream.

8.2.3 Shared Creative Narratives with ThingMojis. Viewers creatively and meaningfully engage with ThingMojis, often coining unique terms that reflect both individual and collective narratives. The use of ThingMoji chains further enriches viewer narratives, adding a layer of creativity to the streaming experience. This practice distinguishes ThingMojis from traditional or streamer-created emotes, which usually come pre-named. The process of naming ThingMojis can be challenging yet rewarding, as it encapsulates diverse social, personal, and cultural contexts. This leads to each ThingMoji potentially symbolizing different entities, imbued with varied attributes as perceived by different viewer communities.

Such viewer-driven interactions lead to dynamic and unpredictable narratives, with ThingMojis acting as tools for inclusive and versatile communication. The interpretation of ThingMojis may vary across different live-streaming contexts and communities, indicating potential cultural differences in their design and perception. This is akin to the cultural variances observed in the use of conventional emojis [37, 47, 53]. User-captured ThingMojis may have their unique advantages, especially in active streaming communities. They play a crucial role in reducing misunderstandings and miscommunications [67], fostering a more inclusive and connected environment.

8.3 Limitation and Future Work

8.3.1 Need for A Real-World Study. A notable challenge we faced with the current MobileSAM model was the potential delay in receiving ThingMojis from the server, which could take several seconds based on the network conditions. While our deployment study involved direct viewer-streamer interactions with an in-the-wild deployment, inviting streamers and viewers to use our system, it is important to acknowledge potential novelty effects. Furthermore, the study was conducted on a relatively small scale, with only 12 to 13 viewers per experiment. This is far from the actual scale of live-streaming audiences, which can reach hundreds or even thousands

of viewers. A high volume of ThingMoji usage may lead to clutter, potentially overwhelming both users and the system. This raises concerns about the effectiveness of key features such as timelines, clustering results, and the ThingMoji gallery in high-traffic scenarios. Future work should explore scalable solutions, such as automatic filtering mechanisms or adaptive display techniques, to manage the potential clutter. Moreover, large-scale evaluation can benefit from log analysis to track user interactions, such as timeline navigation, click patterns, and hover behaviors. Such an analysis could provide information on how viewers engage with the system and inform design improvements for better usability in real-world live streaming environments. Additionally, while our study highlights the spontaneous and playful nature of ThingMoji use, further research should investigate how ThingMojis can be systematically integrated into long-term streaming practices and existing platforms like Twitch. For instance, streamers and communities may develop and identify recurring themes or structured use cases for ThingMojis, such as dedicated visual references for specific topics or interactive overlays that evolve over multiple sessions. Evaluating how streamers and audiences adopt ThingMojis over time could offer deeper insights into their sustained utility and potential impact on audience engagement.

8.3.2 Exploring ThingMojis in Other Live-Streaming Contexts. The in-the-wild deployments were conducted in object-focused scenarios, *i.e.*, manga sharing, model making, and maker and crafting type of live streams. Beyond its current use, ThingMoji can be integrated into gaming streams and e-commerce platforms. In gaming, it can provide viewers with a more interactive way to engage with gameplays and strategies. For e-commerce, ThingMoji can facilitate more dynamic interactions between sellers and potential buyers, enhancing the shopping experience. Streamers have also shared their insights regarding the applicability of ThingMojis in different types of live streams.

8.3.3 Beyond Twitch And Asynchronous Video Interactions. ThingMoji's potential extends beyond Twitch to other platforms and one-to-many communication scenarios where text-based communications are predominant for audiences, like online courses, seminars, and presentation in video conferencing platforms. ThingMoji could be used to highlight specific parts of educational content or to express reactions and questions in a visually engaging manner. Additionally, ThingMoji can be adapted for asynchronous interactions, such as YouTube videos, to link reactions to specific video moments, offering precise and context-rich viewer engagement. In videos with rich visual content, ThingMoji can enhance the viewer's ability to interact with specific visual elements, making discussions more precise and engaging. For video creators, this provides valuable insights into which parts of their content are resonating most with the audience.

9 Conclusion

We introduce ThingMoji, a user-captured cut-out designed to enhance streamer-viewer communication. ThingMojis serve as a powerful tool for spotlighting, condensing, and summarizing crucial contextual information related to object-oriented elements across various application scenarios, including creative crafting, e-commerce product presentations, and gaming. Our user study with streamers and viewers revealed promising use cases for ThingMojis across diverse live-streaming contexts, with design implications for the continued exploration of the ThingMoji concept. We anticipate that ThingMoji will pave the way for novel opportunities in object-focused sharing within the realm of live streaming, sparking further exploration of such design principles within the HCI community.

Acknowledgments

We thank the invited streamers and viewers for their invaluable insight. Erzhen Hu was supported by the Google PhD Fellowship. This work was supported in part by the UVA White Ruffin Byron Center for Real Estate research grant.

References

[1] 2021. Ion-SFU. https://github.com/pion/ion-sfu. Accessed: 2021-08-15.

[2] Patrick Baudisch, Nathaniel Good, and Paul Stewart. 2001. Focus plus context screens: combining display technology with visualization techniques. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology* (Orlando, Florida) *(UIST '01)*. Association for Computing Machinery, New York, NY, USA, 31–40. doi:10.1145/502348.502354

[3] Yuan-Chia Chang, Hao-Chuan Wang, Hung-kuo Chu, Shung-Ying Lin, and Shuo-Ping Wang. 2017. AlphaRead: Support Unambiguous Referencing in Remote Collaboration with Readable Object Annotation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2246–2259. doi:10.1145/2998181.2998258

[4] Di (Laura) Chen, Dustin Freeman, and Ravin Balakrishnan. 2019. Integrating Multimedia Tools to Enrich Interactions in Live Streaming for Language Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300668

[5] Xinyue Chen, Shuo Li, Shipeng Liu, Robin Fowler, and Xu Wang. 2023. MeetScript: Designing Transcript-based Interactions to Support Active Participation in Group Video Meetings. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 347 (Oct. 2023), 32 pages. doi:10.1145/3610196

[6] Lydia B. Chilton, Savvas Petridis, and Maneesh Agrawala. 2019. VisiBlends: A Flexible Workflow for Visual Blends. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300402

[7] Haena Cho, Hyeonjeong Im, Sunok Lee, and Sangsu Lee. 2021. "I Want More than " User-Generated Icons for Better Video-Mediated Communications on the Collaborative Design Process. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 348, 6 pages. doi:10.1145/3411763.3453655

[8] John Joon Young Chung, Hijung Valentina Shin, Haijun Xia, Li-yi Wei, and Rubaiat Habib Kazi. 2021. Beyond Show of Hands: Engaging Viewers via Expressive and Scalable Visual Communication in Live Streaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 109, 14 pages. doi:10.1145/3411764.3445419

[9] Andy Cockburn, Amy Karlson, and Benjamin B. Bederson. 2009. A review of overview+detail, zooming, and focus+context interfaces. *ACM Comput. Surv.* 41, 1, Article 2 (Jan. 2009), 31 pages. doi:10.1145/1456650.1456652

[10] Neil Cohn, Jan Engelen, and Joost Schilperoord. 2019. The grammar of emoji? Constraints on communicative pictorial sequencing. *Cognitive research: principles and implications* 4 (2019), 1–18.

[11] 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* (Florence, Italy) *(MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 504–509. doi:10.1145/2935334.2935370

[12] Audubon Dougherty. 2011. Live-Streaming Mobile Video: Production as Civic Engagement. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (Stockholm, Sweden) *(MobileHCI '11)*. Association for Computing Machinery, New York, NY, USA, 425–434. doi:10.1145/2037373.2037437

[13] Ian Drosos and Philip J. Guo. 2022. The Design Space of Livestreaming Equipment Setups: Tradeoffs, Challenges, and Opportunities. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 835–848. doi:10.1145/3532106.3533489

[14] Honglu Du, Mary Beth Rosson, and John M. Carroll. 2012. Augmenting classroom participation through public digital backchannels. In *Proceedings of the 2012 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '12)*. Association for Computing Machinery, New York, NY, USA, 155–164. doi:10.1145/2389176.2389201

[15] Florian Echtler, Vitus Maierhöfer, Nicolai Brodersen Hansen, and Raphael Wimmer. 2023. SurfaceCast: Ubiquitous, Cross-Device Surface Sharing. *Proceedings of the ACM on Human-Computer Interaction* 7, ISS (2023), 286–308.

[16] Thorsten M Erle, Karoline Schmid, Simon H Goslar, and Jared D Martin. 2022. Emojis as social information in digital communication. *Emotion* 22, 7 (2022), 1529.

[17] Travis Faas, Lynn Dombrowski, Alyson Young, and Andrew D. Miller. 2018. Watch Me Code: Programming Mentorship Communities on Twitch.Tv. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 50 (nov 2018), 18 pages. doi:10.1145/

3274319

[18] Colin Ford, Dan Gardner, Leah Elaine Horgan, Calvin Liu, a. m. tsaasan, Bonnie Nardi, and Jordan Rickman. 2017. Chat Speed OP PogChamp: Practices of Coherence in Massive Twitch Chat. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 858–871. doi:10.1145/3027063.3052765

[19] C. Ailie Fraser, Joy O. Kim, Alison Thornsberry, Scott Klemmer, and Mira Dontcheva. 2019. Sharing the Studio: How Creative Livestreaming Can Inspire, Educate, and Engage. In *Proceedings of the 2019 Conference on Creativity and Cognition* (San Diego, CA, USA) *(C&C '19)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3325480.3325485

[20] Steffen Gauglitz, Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. 2014. World-stabilized annotations and virtual scene navigation for remote collaboration. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 449–459. doi:10.1145/2642918.2647372

[21] Elena L. Glassman, Juho Kim, Andrés Monroy-Hernández, and Meredith Ringel Morris. 2015. Mudslide: A Spatially Anchored Census of Student Confusion for Online Lecture Videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1555–1564. doi:10.1145/2702123.2702304

[22] Jens Emil Sloth Grønbæk, Marcel Borowski, Eve Hoggan, Wendy E Mackay, Michel Beaudouin-Lafon, and Clemens Nylandsted Klokmose. 2023. Mirrorverse: Live tailoring of video conferencing interfaces. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[23] Gaël Guibon, Magalie Ochs, and Patrice Bellot. 2018. From emoji usage to categorical emoji prediction. In *International Conference on Computational Linguistics and Intelligent Text Processing*. Springer, 329–338.

[24] Oliver L. Haimson and John C. Tang. 2017. What Makes Live Events Engaging on Facebook Live, Periscope, and Snapchat. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 48–60. doi:10.1145/3025453.3025642

[25] William A. Hamilton, Oliver Garretson, and Andruid Kerne. 2014. Streaming on Twitch: Fostering Participatory Communities of Play within Live Mixed Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1315–1324. doi:10.1145/2556288.2557048

[26] William A. Hamilton, Nic Lupfer, Nicolas Botello, Tyler Tesch, Alex Stacy, Jeremy Merrill, Blake Williford, Frank R. Bentley, and Andruid Kerne. 2018. Collaborative Live Media Curation: Shared Context for Participation in Online Learning. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3174129

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[28] Jon Hindmarsh, Mike Fraser, Christian Heath, Steve Benford, and Chris Greenhalgh. 1998. Fragmented interaction: establishing mutual orientation in virtual environments. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '98)*. Association for Computing Machinery, New York, NY, USA, 217–226. doi:10.1145/289444.289496

[29] Erzhen Hu, Jens Emil Sloth Grønbæk, Wen Ying, Ruofei Du, and Seongkook Heo. 2023. ThingShare: Ad-Hoc Digital Copies of Physical Objects for Sharing Things in Video Meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 365, 22 pages. doi:10.1145/3544548.3581148

[30] Erzhen Hu, Jens Emil Grønbæk, Austin Houck, and Seongkook Heo. 2023. OpenMic: Utilizing Proxemic Metaphors for Conversational Floor Transitions in Multiparty Video Meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany, USA) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 511, 17 pages. doi:10.1145/3544548.3581013

[31] Erzhen Hu, Mingyi Li, Xun Qian, Alex Olwal, David Kim, Seongkook Heo, and Ruofei Du. 2024. Experiencing Thing2Reality: Transforming 2D Content into Conditioned Multiviews and 3D Gaussian Objects for XR Communication. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST Adjunct '24)*. Association for Computing Machinery, New York, NY, USA, Article 23, 3 pages. doi:10.1145/3672539.3686740

[32] Xincheng Huang, Michael Yin, Ziyi Xia, and Robert Xiao. 2024. VirtualNexus: Enhancing 360-Degree Video AR/VR Collaboration with Environment Cutouts and Virtual Replicas. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 55, 12 pages. doi:10.1145/3654777.3676377

[33] Minal Jain, Sarita Seshagiri, and Simran Chopra. 2016. How Do I Communicate My Emotions on SNS and IMs?. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Florence, Italy) *(MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 767–774. doi:10.1145/2957265.2961862

[34] Sujay Khandekar, Joseph Higg, Yuanzhe Bian, Chae Won Ryu, Jerry O. Talton Iii, and Ranjitha Kumar. 2019. Opico: A Study of Emoji-first Communication in a Mobile Social App. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 450–458. doi:10.1145/3308560.3316547

[35] Joongyum Kim, Taesik Gong, Bogoan Kim, Jaeyeon Park, Woojeong Kim, Evey Huang, Kyungsik Han, Juho Kim, Jeonggil Ko, and Sung-Ju Lee. 2020. No More One Liners: Bringing Context into Emoji Recommendations. *Trans. Soc. Comput.* 3, 2, Article 9 (apr 2020), 25 pages. doi:10.1145/3373146

[36] Yeon Soo Kim, Hyeonjeong Im, Sunok Lee, Haena Cho, and Sangsu Lee. 2023. "We Speak Visually": User-Generated Icons for Better Video-Mediated Mixed-Group Communications Between Deaf and Hearing Participants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 610, 16 pages. doi:10.1145/3544548.3581151

[37] Philippe Kimura-Thollander and Neha Kumar. 2019. Examining the "Global" Language of Emojis: Designing for Cultural Representation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300725

[38] Konstantin Kobs, Albin Zehe, Armin Bernstetter, Julian Chibane, Jan Pfister, Julian Tritscher, and Andreas Hotho. 2020. Emote-Controlled: Obtaining Implicit Viewer Feedback Through Emote-Based Sentiment Analysis on Comments of Popular Twitch.Tv Channels. *Trans. Soc. Comput.* 3, 2, Article 7 (apr 2020), 34 pages. doi:10.1145/3365523

[39] Lucas Kohnke and Benjamin Luke Moorhouse. 2022. Facilitating synchronous online language learning through Zoom. *Relc Journal* 53, 1 (2022), 296–301.

[40] Robert E. Kraut, Darren Gergle, and Susan R. Fussell. 2002. The use of visual information in shared visual spaces: informing the development of virtual co-presence. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work* (New Orleans, Louisiana, USA) *(CSCW '02)*. Association for Computing Machinery, New York, NY, USA, 31–40. doi:10.1145/587078.587084

[41] Audrey Labrie, Terrance Mok, Anthony Tang, Michelle Lui, Lora Oehlberg, and Lev Poretski. 2022. Toward Video-Conferencing Tools for Hands-On Activities in Online Teaching. *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 10 (Jan. 2022), 22 pages. doi:10.1145/3492829

[42] Pascal Lessel, Alexander Vielhauer, and Antonio Krüger. 2017. Expanding Video Game Live-Streams with Enhanced Communication Channels: A Case Study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1571–1576. doi:10.1145/3025453.3025708

[43] Weijian Li, Yuxiao Chen, Tianran Hu, and Jiebo Luo. 2018. Mining the relationship between emoji usage patterns and personality. In *Proceedings of the international AAAI conference on web and social media*, Vol. 12.

[44] Christian Licoppe, Paul K Luff, Christian Heath, Hideaki Kuzuoka, Naomi Yamashita, and Sylvaine Tuncer. 2017. Showing objects: Holding and manipulating artefacts in video-mediated collaborative settings. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 5295–5306.

[45] Xingyu "Bruce" Liu, Ruolin Wang, Dingzeyu Li, Xiang Anthony Chen, and Amy Pavel. 2022. CrossA11y: Identifying Video Accessibility Issues via Cross-modal Grounding. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 43, 14 pages. doi:10.1145/3526113.3545703

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

[47] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the Ubiquitous Language: An Empirical Analysis of Emoji Usage of Smartphone Users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 770–780. doi:10.1145/2971648.2971724

[48] Zhicong Lu, Seongkook Heo, and Daniel J. Wigdor. 2018. StreamWiki: Enabling Viewers of Knowledge Sharing Live Streams to Collaboratively Generate Archival Documentation for Effective In-Stream and Post Hoc Learning. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 112 (nov 2018), 26 pages. doi:10.1145/3274381

[49] Zhicong Lu, Rubaiat Habib Kazi, Li-yi Wei, Mira Dontcheva, and Karrie Karahalios. 2021. StreamSketch: Exploring Multi-Modal Interactions in Creative Live Streams. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 58 (apr 2021), 26 pages. doi:10.1145/3449132

[50] Zhicong Lu, Peng Tan, Yi Ji, and Xiaojuan Ma. 2022. The Crafts+Fabrication Workshop: Engaging Students with Intangible Cultural Heritage-Oriented Creative Design. In *Proceedings of the 2022 ACM Designing Interactive Systems*

Conference (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1071–1084. doi:10.1145/3532106.3533525

[51] Zhicong Lu, Haijun Xia, Seongkook Heo, and Daniel Wigdor. 2018. You watch, you give, and you engage: a study of live streaming practices in China. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[52] Joseph F. McCarthy, danah boyd, Elizabeth F. Churchill, William G. Griswold, Elizabeth Lawley, and Melora Zaner. 2004. Digital backchannels in shared physical spaces: attention, intention and contention. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work* (Chicago, Illinois, USA) *(CSCW '04)*. Association for Computing Machinery, New York, NY, USA, 550–553. doi:10.1145/1031607.1031700

[53] Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. "Blissfully happy" or "ready tofight": Varying interpretations of emoji. In *Proceedings of the international AAAI conference on web and social media*, Vol. 10. 259–268.

[54] Qianqian Mu, Marcel Borowski, Jens Emil Sloth Grønbæk, Susanne Bødker, and Eve Hoggan. 2024. Whispering Through Walls: Towards Inclusive Backchannel Communication in Hybrid Meetings. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1032, 16 pages. doi:10.1145/3613904.3642419

[55] Archana Narayanan, Erzhen Hu, and Seongkook Heo. 2022. Enabling Remote Hand Guidance in Video Calls Using Directional Force Illusion. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, Taiwan) *(CSCW'22 Companion)*. Association for Computing Machinery, New York, NY, USA, 135–139. doi:10.1145/3500868.3559470

[56] James Norris, Holger Schnädelbach, and Guoping Qiu. 2012. CamBlend: an object focused collaboration tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 627–636. doi:10.1145/2207676.2207765

[57] Pavel Okopnyi, Oskar Juhlin, and Frode Guribye. 2022. Designing for Collaborative Video Editing. In *Nordic Human-Computer Interaction Conference*. 1–11.

[58] Amy Pavel, Dan B. Goldman, Björn Hartmann, and Maneesh Agrawala. 2016. VidCrit: Video-based Asynchronous Video Review. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 517–528. doi:10.1145/2984511.2984552

[59] Vitaliy Popov, Xinyue Chen, Jingying Wang, Michael Kemp, Gurjit Sandhu, Taylor Kantor, Natalie Mateju, and Xu Wang. 2024. Looking Together ≠ Seeing the Same Thing: Understanding Surgeons' Visual Needs During Intra-operative Coordination and Instruction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 459, 12 pages. doi:10.1145/3613904.3641929

[60] Raquel Breejon Robinson, Ricardo Rheeder, Madison Klarkowski, and Regan L Mandryk. 2022. " Chat Has No Chill": A Novel Physiological Interaction For Engaging Live Streaming Audiences. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[61] Mose Sakashita, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew David Wilson. 2024. SharedNeRF: Leveraging Photorealistic and View-dependent Rendering for Real-time and Remote Collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 675, 14 pages. doi:10.1145/3613904.3642945

[62] Beejoli Shah. [n. d.]. Facebook Live Is the New QVC | WIRED. arXiv:https://www.wired.com/story/facebook-live-qvc-pearl-parties/

[63] Luke Stark and Kate Crawford. 2015. The conservatism of emoji: Work, affect, and communication. *Social Media+ Society* 1, 2 (2015), 2056305115604853.

[64] Xiumei Su. 2019. An empirical study on the influencing factors of e-commerce live streaming. In *2019 International Conference on Economic Management and Model Engineering (ICEMME)*. IEEE, 492–496.

[65] Anthony Tang, Carman Neustaedter, and Saul Greenberg. 2007. Videoarms: embodiments for mixed presence groupware. In *People and Computers XX—Engage: Proceedings of HCI 2006*. Springer, 85–102.

[66] John Tang, Gina Venolia, Kori Inkpen, Charles Parker, Robert Gruen, and Alicia Pelton. 2017. Crowdcasting: Remotely Participating in Live Events Through Multiple Live Streams. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 98 (dec 2017), 18 pages. doi:10.1145/3134733

[67] Garreth W. Tigwell and David R. Flatla. 2016. Oh That's What You Meant! Reducing Emoji Misunderstanding. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Florence, Italy) *(MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 859–866. doi:10.1145/2957265.2961844

[68] Ruotong Wang, Lin Qiu, Justin Cranshaw, and Amy X Zhang. 2024. Meeting Bridges: Designing Information Artifacts that Bridge from Synchronous Meetings to Asynchronous Collaboration. *Proceedings of the ACM on Human-Computer*

Interaction 8, CSCW1 (2024), 1–29.

[69] Jeremy Warner, Amy Pavel, Tonya Nguyen, Maneesh Agrawala, and Bjoern Hartmann. 2023. SlideSpecs: Automatic and Interactive Presentation Feedback Collation. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 695–709. doi:10.1145/3581641.3584035

[70] Sarah Wiseman and Sandy J. J. Gould. 2018. Repurposing Emoji for Personalised Communication: Why Pizza Means "I Love You". In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3173574.3173726

[71] Dennis Wittchen, Alexander Ramian, Nihar Sabnis, Richard Böhme, Christopher Chlebowski, Georg Freitag, Bruno Fruchard, and Donald Degraen. 2025. CollabJam: Studying Collaborative Haptic Experience Design for On-Body Vibrotactile Patterns. In *2025 ACM Conference on Human Factors in Computing Systems (CHI 2025)*.

[72] Qunfang Wu, Yisi Sang, Dakuo Wang, and Zhicong Lu. 2023. Malicious Selling Strategies in Livestream E-Commerce: A Case Study of Alibaba's Taobao and ByteDance's TikTok. *ACM Trans. Comput.-Hum. Interact.* 30, 3, Article 35 (jun 2023), 29 pages. doi:10.1145/3577199

[73] Saelyne Yang, Changyoon Lee, Hijung Valentina Shin, and Juho Kim. 2020. Snapstream: Snapshot-Based Interaction in Live Streaming for Visual Art. 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, 1–12. doi:10.1145/3313831.3376390

[74] Saelyne Yang, Jisu Yim, Juho Kim, and Hijung Valentina Shin. 2022. CatchLive: Real-Time Summarization of Live Streams with Stream Content and Interaction Data. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 500, 20 pages. doi:10.1145/3491102.3517461

[75] Matin Yarmand, Dongwook Yoon, Samuel Dodson, Ido Roll, and Sidney S. Fels. 2019. "Can You Believe [1:21]?!": Content and Time-Based Reference Patterns in Video Comments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300719

[76] Ye Yuan, Peter Genatempo, Qiao Jin, and Svetlana Yarosh. 2024. Field Trial of a Tablet-based AR System for Intergenerational Connections through Remote Reading. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–28.

[77] Ye Yuan, Qiao Jin, Chelsea Mills, Svetlana Yarosh, and Carman Neustaedter. 2024. Designing Collaborative Technology for Intergenerational Social Play over Distance. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–26.

[78] Amy X Zhang and Justin Cranshaw. 2018. Making sense of group chat through collaborative tagging and summarization. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–27.

[79] Amy X Zhang, Michele Igo, Marc Facciotti, and David Karger. 2017. Using student annotated hashtags and emojis to collect nuanced affective states. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. 319–322.

[80] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. 2023. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. *arXiv preprint arXiv:2306.14289* (2023).

[81] 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.

[82] Nanxuan Zhao, Nam Wook Kim, Laura Mariah Herman, Hanspeter Pfister, Rynson W.H. Lau, Jose Echevarria, and Zoya Bylinskii. 2020. ICONATE: Automatic Compound Icon Generation and Ideation. 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, 1–13. doi:10.1145/3313831.3376618

[83] Rui Zhou, Jasmine Hentschel, and Neha Kumar. 2017. Goodbye Text, Hello Emoji: Mobile Communication on WeChat in China. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 748–759. doi:10.1145/3025453.3025800

Received October 2024; revised April 2025; accepted August 2025