TalkDirector: Real-time Multimodal Slide Augmentation via Adaptive Presenter Integration

GEONSUN LEE, University of Maryland, College Park, USA and Google, USA
YURAN DING, University of Maryland, College Park, USA and Max-Planck Institue for Informatics, Germany
VRUSHANK PHADNIS, Google, USA
DINESH MANOCHA, University of Maryland, College Park, USA
RUOFEI DU, Google, USA

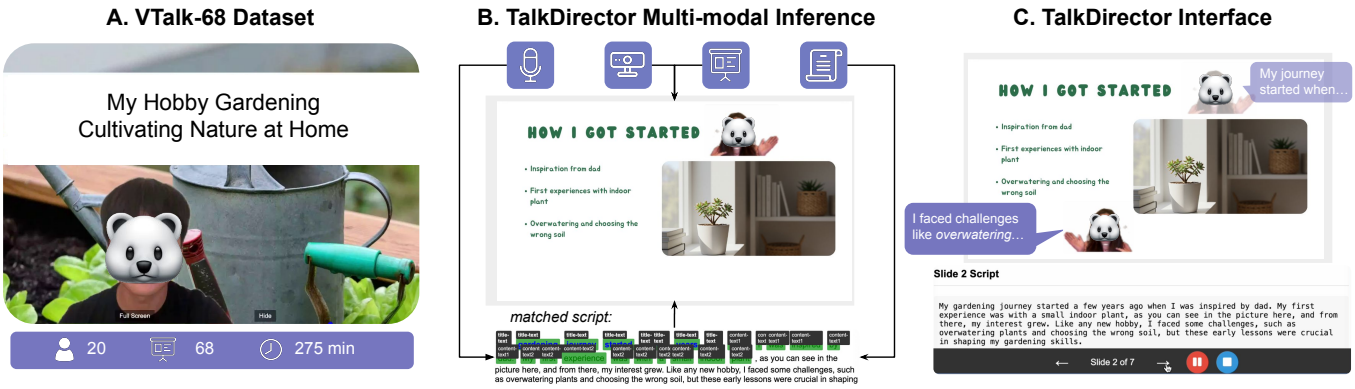


Fig. 1. TalkDirector is an interactive system that dynamically integrates a presenter's video feed into slides during online presentations. We contribute: (A) VTalk-68, the first dataset of 68 online presentations with presenter-driven, personalized video integration, designed to support analysis of dynamic positioning behaviors; (B) a novel multi-modal inference pipeline that determines video placement and size based on speech, gesture, slide content, and scripts; and (C) an interface that operationalizes this pipeline, enabling real-time background segmentation, gesture detection, ASR, and adaptive video rendering within slides.

Online presentation tools limit expressivity by enforcing static video layouts, preventing presenters from fluidly coordinating gestures and speech with slide content. This disrupts the natural alignment of modalities crucial for engagement in face-to-face talks. We present **TalkDirector**, an interactive system that dynamically integrates the presenter's video into slides using multimodal input. TalkDirector operates in two stages: a preprocessing phase where a vision-language model aligns the presentation script with the contents of the slides; and a second, live phase that detects gestures and transcribes speech to identify currently referenced content through semantic matching against the pre-aligned elements. The system then adaptively positions and resizes the video feed in relation to the referenced content, enabling layout-aware and context-aware video integration.

Additionally, we contribute **VTalk-68**, a dataset of 68 annotated presentation recordings that informed the system's design. A two-part user study (n=12×2) demonstrates that TalkDirector reduces presenter effort, increases expressiveness, and enhances audience connection compared to conventional presentations.

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

Additional Key Words and Phrases: online presentation, video-mediated communication, multi-modal models, collaborative work, video conferencing, augmented communication

ACM Reference Format:
Geonsun Lee, Yuran Ding, Vrushank Phadnis, Dinesh Manocha, and Ruofei Du. 2025. TalkDirector: Real-time Multimodal Slide Augmentation via Adaptive Presenter Integration. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), Sep 28–Oct 1, 2025, Busan, Korea.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Online presentations often disrupt the fluid, expressive delivery; characteristic of face-to-face talks. In person, presenters naturally align their speech, gestures, and body orientation with visual content, effortlessly directing audience attention and emphasizing key elements with diectic references [2, 25]. In contrast, most online presentation tools confine presenter's video to fixed, manually positioned frames, disconnected from the specific content being discussed or gestured toward. This spatial and temporal misalignment reduces both audience engagement and comprehension [24, 35], while limiting presenter expressiveness [16].

Authors' Contact Information: Geonsun Lee, gsunlee@umd.edu, University of Maryland, College Park, College Park, MD, USA and Google, San Francisco, CA, USA; Yuran Ding, yurand@umd.edu, University of Maryland, College Park, College Park, MD, USA and Max-Planck Institue for Informatics, College Park, MD, Germany; Vrushank Phadnis, Google, , USA, vrushank@google.com; Dinesh Manocha, University of Maryland, College Park, , USA, dmanocha@umd.edu; Ruofei Du, Google, , , USA, ruofei@google.com.

UIST '25, Busan, Korea

While commercial platforms like Zoom and Microsoft Teams offer basic presenter overlay features, they rely on static, manual configurations. These limitations hinder effective communication, particularly in education or technical talks, where presenters frequently reference specific elements on complex, information-dense slides. Inspired by professional TED-style talks—where human directors actively adjust framing to align with the speaker's intent—we explore whether automated, multimodal presenter integration can enhance the usability, expressiveness, and engagement of online presentations, fostering a stronger sense of connection with the speaker. Our aim is to bring back some of the natural coordination between modalities that makes in-person talks effective.

Research across communication and cognitive science shows that viewers process verbal and nonverbal signals—the intricate interplay between slides, verbal communication, gestures and body movements—as an integrated whole [2, 10, 25]. Spatial alignment between presenter gestures and content playing a crucial role in directing audience attention and enhancing comprehension [13, 39].

However, current presentor-video integration approaches fall into two categories: (1) manual overlays (Zoom Virtual Background [62], Microsoft Cameo [37]), and (2) content-driven augmentations (RealityTalk [32], ChalkTalk [42]). Manual overlays require presenters to pre-configure or manually position their video, increasing cognitive burden sweller1994cognitive, ellis1992perceived and often resulting in static layouts [16]. In contrast, content-driven augmentations allow dynamic manipulation of contextual overlays, but fail in traditional information-dense slide decks, where content clarity is paramount [60].

This gap matters: studies show that thoughtful video placement—such as on the right side of slides [60] or using larger video feeds [9]—can enhance learning, audience engagement and perceived speaker presence. However, manually optimizing such factors during *live* presentations imposes significant cognitive load on presenters, often leading to fixed, suboptimal layouts that fail to leverage the full potential of their non-verbal cues. As a result, many presenters opt to keep their video feed fixed for convenience, sacrificing opportunities for more expressive and engaging delivery.

We introduce **TalkDirector**, an interactive system enabling real-time, multimodal presenter-slide integration. TalkDirector dynamically adjusts the position and size of the presenter's video feed in real-time based on speech, gestures, and slide content (Figure 1C).

Our approach integrates automatic speech recognition, gesture detection, and structured slide-layout analysis through a two-stage inference pipeline: (1) it first semantically aligns presentation scripts with visual elements on each slide, then continuously matches real-time speech to these aligned components using sliding window algorithm; (2) during the live presentation, when a pointing gesture is detected—indicating a presenter's intent to emphasize a specific element—the system repositions and resizes the video feed next to the referenced content. This ensures dynamic, context-aware transitions that maintain content visibility and highlight presenter intent.

Informed by the VTalk-68 dataset—68 annotated online presentations recording user behavior in dynamically adjusting their video feed based on context—we designed TalkDirector to support expressive delivery, engagement while minimizing presenter effort. To evaluate TalkDirector's effectiveness, we conducted two-part user studies from both presenter and audience perspectives. The presenter study (n=12) compared TalkDirector against a manual baseline (Zoom's *Slides as Virtual Background* feature [62]), to assess preparation effort and perceived expressiveness, measured through logging preparation time and subjective ratings. The audience study (n=12) compared presentations viewed via TalkDirector to a conventional side-by-side video-slide layout, evaluating perceived engagement, speaker connection, and potential distraction with subject questionnaire results. Results indicate that TalkDirector significantly reduced preparation time and cognitive load for presenters while enhancing their perceived expressiveness. Audiences reported a stronger sense of connection to the speaker, without finding the dynamic video integration significantly more distracting than the static baseline layout.

Our contributions are threefold:

- **TalkDirector**: An interactive system demonstrating real-time, multimodal presenter video integration driven by speech, gesture, and slide layout analysis, leveraging LMMs to drive adaptive video placement.
- **Dual-perspective evaluation**: Presenter ($n = 12$) and audience ($n = 12$) studies, demonstrating the effectiveness of TalkDirector in reducing presenter effort and enhancing perceived expressiveness, engagement, and clarity.
- **Design insights and dataset**: Findings from two formative workshops ($n = 5$) and release of **VTalk-68**, an annotated multimodal presentation dataset[1] to facilitate future research on presenter behavior and video placement strategies, as well as AI-powered presentation tools.

2 Related Works

Our work is inspired by prior literature in gestures, language, and contents in presentations, as well as recent advances in body-content interaction in augmented presentations.

2.1 Gestures, Language, and Contents in Presentation

Effective presentations are inherently multimodal, relying on presentation slides, verbal communication, and nonverbal cues such as gestures, facial expressions, and body movements [25]. Studies have shown that humans interpret these multimodal signals as a unified whole in communication [10], with spatial alignment between gestures and content playing a crucial role in directing audience attention [39]. When presenters can physically reference and interact with content through body movements and gestures, they create stronger cognitive associations for their audience [13]. Research has demonstrated that the spatial positioning of presenter video relative to content significantly impacts learning outcomes and audience engagement. For instance, Zhang et al. found that placing the instructor's video on the right side of the screen increases learning performance and satisfaction [60], while Ellis et al. showed that the size of presenter video affects perceived instructor presence [9]. The alignment of body movements with content spatial layout allows presenters to create a more interactive and engaging environment [45], adaptable to both in-person and virtual

[1] GitHub link omitted for anonymity

Project	POC	COP	Adaptive	Input Modality
Tutor In-sight [53]	✓ (avatar)		✓	mouse
Microsoft Teams, Microsoft Cameo, Google Meet [15, 37, 38]				mouse & real-world contents
Zoom 'Slides as Virtual Background' [62]	✓			mouse
OpenMic [22]				mouse
ChatDirector [44]			✓	speech
Charade [1]				hand (glove-tracking)
Bringing physics to the surface [57]	✓		✓	hand
Chalktalk [42]		✓		mouse
RealitySketch [51]		✓	✓	mouse & real-world objects
RealityTalk [32]		✓	✓	speech & gestures
Elastica [4]		✓	✓	pre-defined animations, speech & gestures
Interactive Body-Driven Graphics [48]		✓	✓	gestures & postures
Augmented Chironomia [17]		✓	✓	gestures
ARCADE [50]		✓	✓	hand
ThingShare [23]				mouse
Matulic et al., [34]	✓		✓	gestures & slide elements
Our Work (TalkDirector)	✓		✓	gestures, speech, slide content (layout, elements, sequence, contexts)

Table 1. This table categorizes related works on online-presentations based on the following criteria: whether they integrate the presenter's video over content (**Presenter Over Content**—**POC**), whether they overlay content on the presenter's video (**Content Over Presenter**—**COP**), whether they are **adaptive**, meaning they can dynamically adjust layout, content placement, or presentation style in response to the presenter's actions or content context without manual input, and lastly, the **input modality** that determines UI placements in the online presentation system. TalkDirector is the only presenter-video-over-content system to enable real-time, multimodal adaptive presenter placement by combining speech, gesture, and slide content analysis.

settings [8]. These findings highlight the importance of thoughtful integration between presenter video and content in digital presentations.

2.2 Presenter Video-Content Integration in Digital Presentations

Digital presentation systems have evolved to bridge the gap between in-person and remote communication by integrating presenter video with content in various ways. These approaches have been widely adopted across educational settings [7, 14, 42, 53], public presentations [46, 47], and online explanatory videos [54]. While many early augmented presentations relied on post-production [32], the rise of livestreaming and video conferencing has sparked new research into real-time presenter-content integration [1, 5, 14, 17, 27, 34]. These systems follow two distinct approaches: Presenter-Video-Over-Content (POC) Integration Systems, which render the presenter's video feed over slides to maintain physical presence, and Content-Over-Presenter-Video (COP) Integration Systems, which overlay interactive content over the presenter's video to enable dynamic content manipulation. Here, we examine these approaches and their implications for presentation effectiveness.

2.2.1 Presenter-Video-Over-Content Integration Systems. Current video conferencing platforms like Zoom, Microsoft Teams, and Google Meet provide basic integration of presenter video over slide contents but struggle to support natural presenter-content interactions [15, 21, 38, 62]. Research has shown that effective integration of presenter video with content is crucial: Friedland et al. [12] demonstrated that separating presenter video from content creates a split-attention effect, while strategic video placement can enhance learning outcomes. Ellis et al. [9] found that larger presenter video sizes increase perceived instructor presence, and Zhang et al. [60] showed that right-side video placement improves learning performance and satisfaction. Recent commercial solutions have attempted to address these challenges. Zoom's 'Slides as Virtual Background' [63] and Microsoft Cameo [37] allow video integration with slides, but require manual positioning and lack context awareness. While Microsoft Teams' 'Dynamic View' optimizes general content layout [38], it doesn't specifically address presenter video placement. In the research community, several systems have explored more sophisticated approaches. Some focus on video manipulation in conferencing contexts, such as proxemic-based resizing [22] and space-aware avatar transitions [44]. Others have investigated gesture-based interactions [20, 57] and presenter avatar integration [34]. However, these systems typically focus on single modalities rather than combining multiple inputs for video placement. TalkDirector advances prior work by integrating speech, gestures, and slide content to dynamically adjust presenter video in real time. Unlike manual commercial tools or research systems limited to single modalities, our multimodal approach reduces presenter cognitive load, increases expressiveness and ease of use, and enhances audience engagement.

2.2.2 Content-Over-Presenter-Video Integration Systems. In contrast to Presenter-Video-Over-Content Integration Systems which place video feeds over slides, this approach overlays interactive content on the presenter's video feed. These systems can be categorized by their input modalities and content types. Basic systems like Chalktalk [42] and RealitySketch [51] enable mouse-based sketching and animation over video. More advanced systems incorporate gesture-based interactions: Interactive Body-Driven Graphics [48] allows real-time manipulation of graphical elements through gestures and postures, while Hall et al. [17] developed bimanual interactions with dynamic charts overlaid on presenter video. Some systems have explored multimodal inputs for content overlay. RealityTalk [32] combines speech and gestures to control text and

image overlays, while Elastica [4] synchronizes predefined animations with speech and gestures. Others have investigated specialized approaches, such as ThingShare's [23] gesture-based object manipulation, Tutor-Insight's [53] MR avatars for attention direction, and ARCADE's [50] 3D holographic integration. Some researchers have even explored physical space integration through holography [27, 28]. While these systems focus on enriching the presenter's visual space with overlays, TalkDirector takes the complementary approach—enhancing presenter-content alignment by dynamically repositioning the presenter within slides using multimodal cues, enabling more expressive, usable and adaptive communication in traditional slide-based online presentation formats.

2.3 LLM Augmentation Systems in HCI

Recent HCI research has explored how LLMs can augment human capabilities and act as engines within interactive systems—goals directly aligned with TalkDirector's multimodal inference pipeline for real-time presenter integration.

In augmenting human performance, LLMs have been used to reduce effort and enhance user expression. Visual Captions [33] predicts user intent in video conferencing, while LLMR [6] enables context-aware adaptation in mixed reality. Marco [11] supports sensemaking in business document collections, and Kobiella et al. [29] examines how ChatGPT affects workplace productivity. Tools like PaperWeaver [31] and EvaluatingLLM [55] help researchers generate ideas and synthesize literature—parallel to how TalkDirector assists presenters by offloading layout and content reasoning to an intelligent backend.

LLMs also serve as core components within interactive systems. Farsight [56] generates speculative scenarios, GenLine [26] translates natural language to code, and MindTalker [58] supports users with dementia through conversation. LLMs have been used for insight extraction, as seen in Memoro [64] and PaperWeaver, which process conversational and bibliographic histories. Visual Captions fine-tunes LLMs to interpret intent from live speech, and InstructPipe [61] coordinates multiple LLM modules for pipeline generation—similar in spirit to TalkDirector's orchestration of ASR, gesture detection, and layout analysis for adaptive presenter positioning.

Across this space, LLMs act as prompt-based agents [52], components within larger workflows [18, 30], or engines for end-to-end systems [56, 59]. TalkDirector builds on this theme, demonstrating how LLMs can enable expressive, real-time human-computer interaction through multimodal reasoning and context-aware layout adaptation.

3 Design Rationale

To inform the design of TalkDirector and ground our system in real user needs, we conducted two formative studies that explored how presenters integrate video feeds in slide-based presentations. The first was an expert workshop aimed at surfacing presenter-centered and information-centered preferences for video feed control and automation in online presentations. The second was an interaction pattern exploration study, where users gave presentations using a prototype that logged detailed video feed behavior. Together, these studies helped us uncover key placement strategies, gesture integration desires, and pain points with current workflows—directly shaping TalkDirector's multimodal pipeline and adaptive placement algorithm.

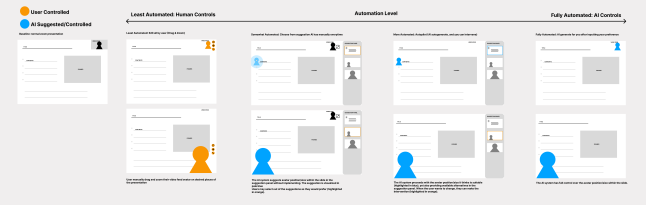


Fig. 2. Mock-ups exploring users' decision-making processes and workflows in dynamic presenter video feed generation, based on adaptation level. These mockups serve as an ideation point for the iterative development of a presenter-slide integrated dynamic presentation system.

3.1 Workshop Study

We conducted an expert workshop to understand how presenters currently integrate video feeds into online presentations and to derive design implications for automated video feed management.

3.1.1 Participants and Procedure. The study consisted of two sessions of semi-structured interviews with five participants (1 female, 4 males), experienced in creating presentation videos with embedded speaker feeds. Participants were recruited based on their experience with creating online presentation videos that incorporate speaker video feeds. The semi-structured interviews lasted 60-90 minutes via Zoom, with participants receiving $10 compensation. The first session (40-50 minutes) explored participants' current practices and challenges, while the second session (20-30 minutes) gathered feedback on mock-ups representing different levels of automation in video feed management.

We presented four mock-ups with increasing levels of automation: manual control (drag/zoom), AI suggestions with user selection, single AI suggestion with an override option, and fully automated placement. Participants provided feedback on using these variations for both live and post-editing scenarios. (Figure 2

3.1.2 Results and Design Implications. We generalized and analyzed findings from the semi-structured workshop using affinity diagramming. Our following key findings directly informed TalkDirector's design:

(1) *Content-Aware Placement:* Participants emphasized that video placement should adapt to slide content and presentation goals. P5 preferred large, centered video during introductions but minimized corner placement for dense content like "Related Work." P3 suggested a "one-third rule" to maintain content visibility. Participants also noted that presentation types vary—information-centric talks (e.g., tutorials) might benefit from minimal presenter presence, while presenter-centric ones (e.g., personal intros) prioritize visibility to enhance memorability. These insights shaped our layout-aware algorithm, which dynamically adjusts video size and position based on content type, complexity, and transitions.

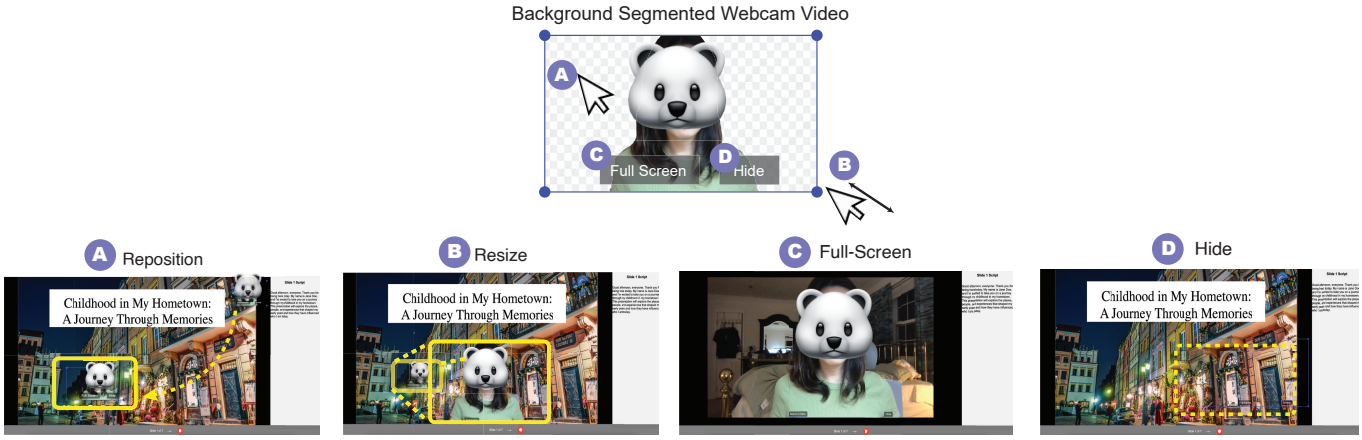


Fig. 3. Screenshots of the prototype used in our data collection study. Users can dynamically edit their webcam video in their presentations with four features; (a) repositioning (b) resizing, (c) turning into full-screen, and (d) hiding mode. The presenter can edit their video and hit the record button on the bottom to record a presentation video.

(2) *Gesture-Responsive Integration:* Preserving natural gestures was seen as essential to expression. P5 noted a strong desire for gesture-aware placement: "if there are ways of deriving how my gestures are in an actual presentation... I would definitely rely on those." This directly informed our real-time gesture detection and positioning system.
(3) *Unified, Real-Time Workflow:* Participants found existing multi-tool workflows tedious. P1 described the burden of switching between video and slide editors, while P2 requested an integrated overlay preview, where camera feed, slides, and audio could be synchronized in real-time. This led us to design TalkDirector as a single unified tool that synchronizes slides, camera feed, and script for live presentation control.
(4) *Balanced Automation with Control:* While automation was welcomed, participants preferred maintaining final control. P4 suggested "a first pass automated, then tweak from there," supporting our pipeline's semi-automated mode with override flexibility. This hybrid design supports both presenter-centric and information-centric workflows.

3.2 Interaction Pattern Exploration Study

Given the lack of data on how presenters naturally integrate video feeds for expressive communication, we conducted an exploratory study to uncover user behaviors and preferences in video-enhanced presentations. Our goal was to identify common interaction patterns to inform the design of automated presenter integration systems.

3.2.1 Presentation Tool Prototype. We developed a web-based prototype to support online PowerPoint presentations with enhanced control over webcam video placement. Built with the PPTX2HTML library [43], the system renders slides and integrates a live webcam feed. Users upload a slide deck and a corresponding script (in .txt format).

Webcam background was removed using MediaPipe [36], enabling seamless video-slide integration similar to Zoom's "Slide as Virtual Background."

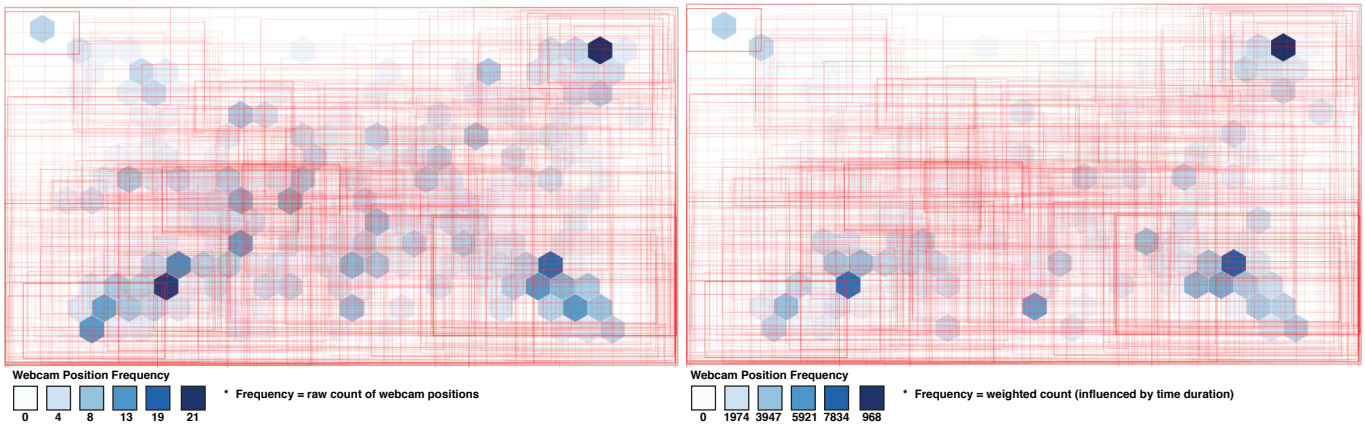
The prototype supports four main video manipulation features (Figure 3): drag-and-drop repositioning, corner-based resizing, full-screen mode, and visibility toggling. A control panel allows slide navigation and recording with pause/resume functionality. During sessions, the system logs webcam position, size and mode states (e.g., full-screen, hidden); enabling detailed post-hoc interaction analysis.

3.2.2 Study Design. Guided by insights from our formative study, we created three types of mock-up slide decks to represent diverse presentation contexts: personal stories (presenter-centric), tutorials (information-centric), and professional presentations that balance presenter presence with information delivery. Each type included two variants, and every deck consisted of seven slides designed to test different layout scenarios—such as a title slide, a full-text slide, a full-image/video slide, and slides combining text and visuals in varied configurations.

The study was conducted remotely via Zoom. Participants were tasked with preparing a monologue presentation with an emphasis on expressivity. They were randomly assigned to one of two slide variants per presentation category. After reviewing the content, participants used our platform to record their talk while dynamically controlling webcam placement. This setup enabled us to observe natural video feed integration behaviors across diverse presentation styles.

3.2.3 Participants. Twenty participants (12 female, 8 male, M age=26.6, SD=3.23) were recruited from a university sample, receiving $10 compensation. On 5-point Likert Scales, participants reported high proficiency with presentation software (M = 4.05, SD=0.60) and moderate to high presentation skills (M = 3.73, SD=0.79). Sixteen participants consented to release their video feeds as part of the dataset.

3.2.4 Identified Patterns Quantitative Analysis. To identify common webcam placement and sizing patterns, we analyzed timestamped JSON logs from 20 participants across three presentation types



(a) Heatmap of presenter video positions and sizes (raw count)

(b) Heatmap of presenter video positions and sizes (weighted)

Fig. 4. The heatmaps display the spatial distribution of presenter video positions across all participants, with (a) showing raw occurrence counts (each position counted once) and (b) showing time-weighted distributions where positions are weighted proportionally to their duration (calculated by multiplying each position by the time elapsed until the next logged change). The visualization uses hexagonal binning to aggregate position data into discrete spatial units, where each blue hexagon represents a region of the slide. Darker blue indicates higher position frequency within that region, while red outlines show the actual size and placement of individual webcams. Analysis revealed two dominant positioning strategies: bottom-left (18%), and bottom-right (17%), suggesting presenters prefer corner placement to maximize slide visibility. Full-screen and hidden modes were excluded from this analysis per experimental protocol.

(personal, tutorial, professional). We excluded entries where the webcam was in fullscreen or hidden mode (`isFullScreen=true`, `isHidden=true`) as these reflected intentional mode changes rather than typical positioning behavior. The entries where webcam was in fullscreen or hidden mode were analyzed later with counted instances, instead of plotting them in the heatmap.

To reflect user preferences, we used a time-weighted analysis: the longer a position was held, the more it contributed to the overall distribution. This helped highlight intentional choices over brief adjustments.

Positions were normalized relative to the slide container, using the webcam's center point (x, y) as a proportion of the slide area. We then applied D3.js's hexbin algorithm to generate a spatial density heatmap, where darker regions indicate more frequent placements.

From the analyzed data ($N = 692$ positions), we identified that the most common webcam placements occurred in the *bottom-left (18%)* and *bottom-right (17%)* regions of a standardized 3×3 grid layout. The average webcam size was consistent across categories, with slight variations: tutorials (321 × 180 pixels), personal stories (367 × 207 pixels), and professional presentations (356 × 200 pixels). Furthermore, we observed that fullscreen and hidden modes were predominantly activated on the first and last slides, primarily for personal introductions and conclusions (full-screen mode: 35 total instances), and during visually dense slides featuring large figures or images (hidden mode: 12 total instances). Interestingly, 8 participants never engaged either mode, preferring consistent visibility throughout their presentations.

3.2.5 Identified Patterns Qualitative Analysis. We conducted post-trial semi-structured interviews to explore participants' motivations and experiences. We created an affinity diagram using thematic analysis to systematically categorize participant responses.

Most participants (14 of 20) were pleasantly surprised by the level of control over webcam integration, valuing the ability to reposition and resize dynamically. However, nearly half (9 of 20) expressed a preference for automated adjustments, noting that manual control added cognitive effort. One participant summarized: *"I found myself just repositioning into the corner blank space whenever possible, and making my webcam as big as the available space allowed."*

Participants also noted that integrated webcam positioning enhanced their expressiveness. They became more aware of their gestures and facial expressions as central to the presentation rather than peripheral. As one participant put it: *"I felt more deliberate in my gestures and how I expressed myself because I felt like I was a part of the slide, not just on the side."*

Participants often adjusted webcam placement to emphasize key slide content. Five specifically mentioned repositioning and resizing to highlight important points or visual-text associations, especially in personal narratives. In contrast, those giving tutorial-style presentations (5 of 20) made fewer adjustments, though three valued the ability to use gestures for pointing. Additionally, five participants noted that their use of webcam integration varied depending on the presentation's purpose and content style.

3.2.6 Design Considerations. Based on our quantitative and qualitative findings, we derived design considerations that directly link presenter behaviors and preferences to the design of our automated multimodal integration system.

DC1: Automated Blank-Space Utilization Presenters often moved their webcam to bottom corners to avoid occluding content. Automated systems should detect and prioritize these blank areas—especially bottom-left and bottom-right—for dynamic, non-intrusive placement.

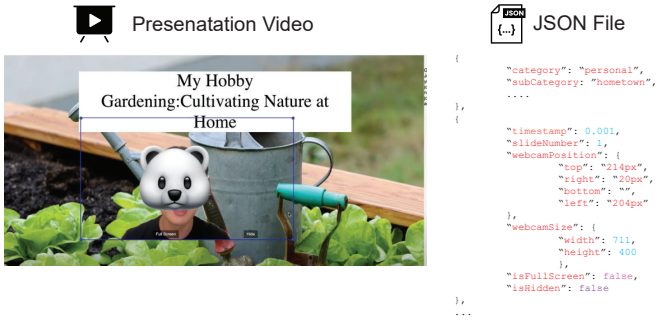


Fig. 5. VTalk-68 dataset, that collects information on user video feed placements across various slide layouts and contents.

DC2: Contextual Mode Adaptation Fullscreen mode was typically used during introductions and conclusions, while webcams were hidden on visually dense slides. Systems should recognize these contextual cues and automatically adjust visibility (e.g., show, hide, fullscreen) accordingly.

DC3: Gesture-Aligned Placement Participants felt more expressive when their gestures aligned with slide content. Automated systems should incorporate gesture recognition to position the presenter video near referenced elements, enhancing clarity and engagement.

DC4: Seamless Presentation Flow Manual adjustments were seen as disruptive. To maintain flow, systems should continuously adapt video placement and size using multimodal inputs, minimizing the need for presenter intervention.

3.2.7 Dataset. We are releasing our dataset, ***VTalk-68 Dataset*** Figure 5, which includes both a webcam video file and a screen recording file for each participant, capturing how users decided to move and resize their webcam video based on the content of the presentation. This dataset reflects the participants' interaction with the presentation system, particularly their decisions regarding the placement and resizing of the video feed.

Additionally, the dataset includes a JSON file for each recording, logging parameters such as the size and position of the webcam video, the activation of fullscreen or hide modes, and timestamps of any changes made by the user. This allows for a detailed analysis of participants' interaction patterns with the system.

We share this dataset with the goal of encouraging further research into understanding user behavior during interactive presentations. We hope it will support future work that aims to investigate how presenters utilize interactive features when presenting slides. *The dataset URL (7.38 GB) will be provided in the camera-ready version for anonymity.*

4 TalkDirector

Based on insights from Section 3.1 and 3.2, we developed ***TalkDirector***, an intelligent online presentation interface that dynamically integrates presenter video feeds with slides. The system optimizes the position and size of the presenter's video in real-time by analyzing slide content, spoken words, and presenter gestures, enhancing visual engagement while minimizing cognitive load. Our approach introduces a novel multimodal inference technique that uniquely combines layout analysis, speech recognition, and gesture detection to position presenter video adaptively.

TalkDirector employs a two-stage architecture: (1) preprocessing and (2) real-time processing. This separation serves three critical purposes: it front-loads computationally intensive tasks such as layout analysis and content alignment to minimize latency during live presentations; establishes precise relationships between scripts and slide elements for accurate real-time decisions; and robustly adapts to spontaneous changes in the presenter's speech and actions. Figure 6 illustrates the system architecture.

4.1 Preprocessing Pipeline

The preprocessing pipeline aligns slide layouts with presentation scripts, preparing structured references for real-time positioning.

4.1.1 Layout Analysis. Slides are initially converted into PNG images and processed through Tesseract OCR [49] to extract text and preliminary bounding boxes. To improve accuracy and robustness, particularly in complex graphical layouts, we refine these initial bounding boxes using GPT-4o [41]. GPT-4o semantically tags slide elements with hierarchical labels such as `<title-text>` and `<figure-image1>`, clearly distinguishing titles, content blocks, and visual elements. Coordinates are normalized to maintain consistency across slides, and results are stored in structured JSON. The resulting labeled bounding boxes are preserved for each slide and later used to construct the occupancy grid required for blank space detection during real-time positioning.

4.1.2 Script-Layout Alignment. Presentation scripts are segmented by explicit slide markers (e.g., "[Slide 1]"). GPT-4o then matches these segments to labeled slide elements using XML tags corresponding exactly to component labels. We employ a two-step robust matching strategy: initial candidate matches via TF-IDF vectorization and refined semantic matching using All-MiniLM-L6-v2 embeddings. Based on Section 3.2 insights, we also label special semantic tags to accommodate full-screen and hidden transitions:

- `<full>`: Indicates sections such as self-introduction or Q&A invitations, prompting the presenter's video to occupy the full screen.
- `<hide>`: Marks slides containing full-screen visuals or complex figures where presenter visibility is less critical, hiding the video entirely.

4.2 Real-time Processing

The real-time pipeline is designed to interpret multimodal input from the presenter and adapt the video positioning accordingly. It integrates three components in parallel: (1) speech recognition and matching with scripted content, (2) gesture recognition and spatial interpretation, and (3) blank space detection. With all combined, speaker's video is dynamically repositoning and resized during the presentation.

4.2.1 Speech Recognition and Script Matching. We combine the Web Speech API for responsiveness and OpenAI's Whisper API for accuracy under variable audio conditions. The Web Speech API

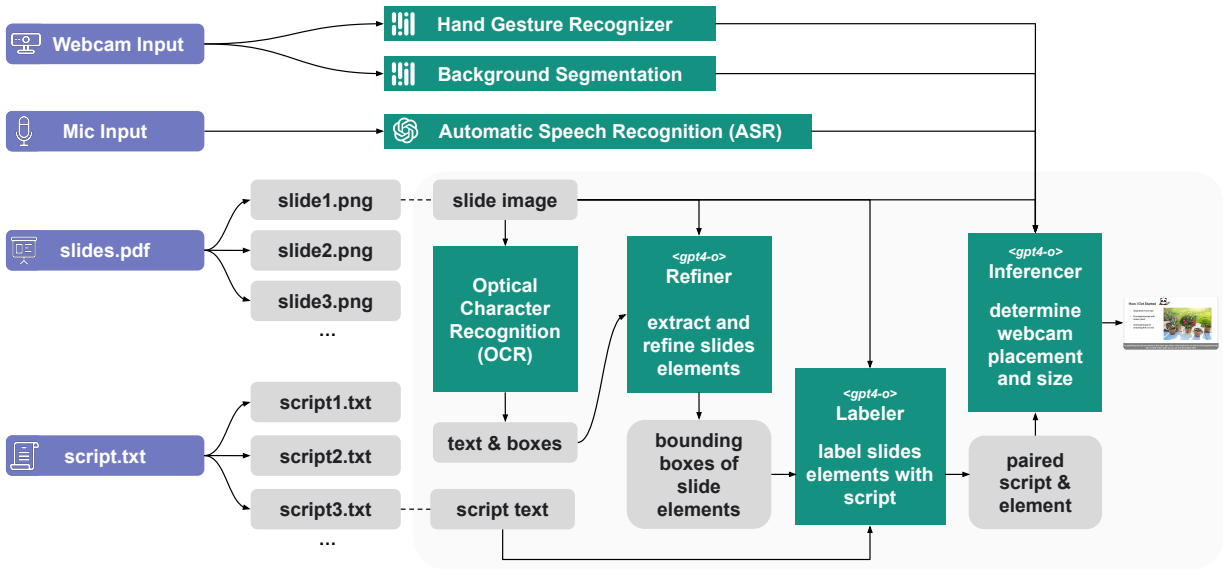


Fig. 6. TalkDirector's multi-modal layout inference pipeline. We first process inputs of slides presentation and scripts with TesseractOCR [49] and Large Generative Multimodal Modal (GPT4-o) with a refiner and a labeler. This pipeline effectively reduces hallucinations of GPT4-o and generate paired script and element of each slide. During the presentation, we run a real-time pipeline to segment webcam background, recognize hand gestures, and parse speech with WhisperAI [40], an Automatic Speech Recognition (ASR) service. Finally, we leverage GPT4-o to determine determine webcam placement and size, hence rendering the recommended segmented video onto the slides.

provides immediate transcription updates, while Whisper periodically refines transcriptions. A sliding-window algorithm matches recent transcribed speech segments to script segments using fuzzy matching, accommodating spontaneous speech variations and minor transcription errors.

4.2.2 Gesture Recognition. Based on DC3, which emphasizes presenter intent to spatially associate themselves with content, we interpret pointing gestures as an explicit signal to be placed adjacent to the referenced slide component for emphasis and contextual grounding. For gesture recognition, we use MediaPipe Hands [36] to detect stable pointing gestures (index finger extended, others curled) held for at least 500ms. Detected gestures are geometrically mapped to slide coordinates, accurately aligning gestures with slide elements for precise positioning. Gesture processing occurs in parallel with speech recognition to maintain responsiveness and minimize latency.

4.2.3 Blank Space Detection and Default Positioning. To determine suitable regions for placing the presenter's video, TalkDirector analyzes blank areas on each slide that do not contain important content. Each slide is discretized into a $W \times H$ binary occupancy gird G with a resolution of $R = 5$ pixels per cell. Occupied cells corresponding to text and visual elements are marked as 1, and unoccupied cells are marked as 0. Given an empty cell (x, y), its largest continuous empty region can be computed in Algorithm 1:

Each candidate region (x, y, w, h) is evaluated based on size constraints. To ensure webcam sizes are within a desirable size range, we enforce a size constraint of the webcam relative to the slide ($w_{min} = 0.15 \times$ slide_width, $w_{max} = 0.4 \times$ slide_width and corresponding height based on a fixed aspect ratio). For each qualifying region, the system calculates the latest possible webcam dimension within bounds and ranks candidates based on proximity to slide content and user-preferred placement locations identified in the formative study. Following the order of: bottom-right → bottom-left → top-right → top-left.

Algorithm 1 Blank Space Detection Algorithm

```
procedure ExpandBlankSpace(G, x, y, W, H)
    Initialize blank space's width w ← 0, height h ← 0
    while x + w < W ∧ G[y][x + w] = 0 do
        w ← w + 1
    end while
    while y + h < H ∧ IsRowEmpty(G, y + h, x, w) do
        h ← h + 1
    end while
    return blank space (x, y, w, h)
end procedure
```

Blank space detection is performed once per slide transition and cached, avoiding recomputation during live interactions. This maintains real-time responsiveness while adapting to layout changes between slides.

4.2.4 Gesture-Speech Triggered Repositioning. When a pointing gesture is detected with a valid speech-to-script match, the system repositions the video feed adjacent to the referenced slide component.

Given the bounding box (x, y, w, h) of the matched component, TalkDirector attempts to place the video feed near the element using

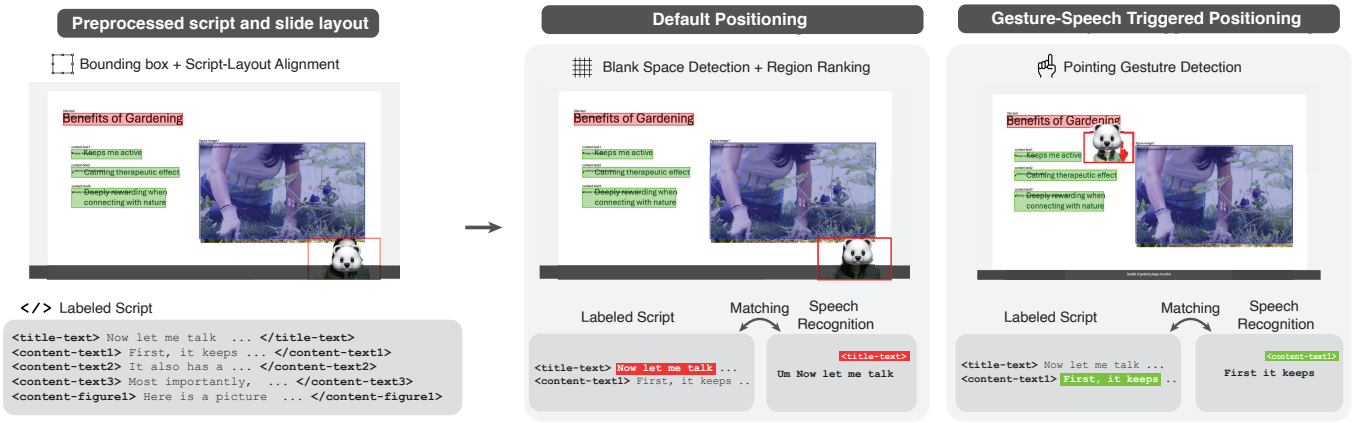


Fig. 7. Visualization of *TalkDirector*'s context-aware video positioning: default positioning and sizing and real-time positioning and sizing. In the default mode, the system detects blank spaces using slide layout analysis, placing the presenter's video in a non-obstructive area with an appropriately sized feed. In the real-time mode, the system dynamically adjusts the position and size of the presenter's video based on gesture recognition and speech-to-script matching, aligning the video feed with the relevant content being discussed on the slide. This approach ensures an adaptive and engaging presentation flow.

the following anchor positions:

$$p_{right} = (x + w + M,\ y + h - h_{webcam}), \tag{1}$$

$$p_{left} = (x - w_{webcam} - M,\ y + h - h_{webcam}), \tag{2}$$

$$p_{below} = (x,\ y + h + M), \tag{3}$$

where M is a fixed margin (10 pixels) from the component boundary. Each candidate position is validated to ensure it does not overlap existing slide content or exceed slide boundaries. If all adjacent placements are infeasible, the system falls back to the highest-ranked default corner.

Figure 7 visually summarizes default and gesture-triggered positioning based on multimodal inputs.

4.3 Implementation and Performance Evaluation

We implemented TalkDirector as a web-based application, using a Node.js backend for preprocessing tasks (interactions and layout analysis via GPT-4o API) and browser-side JavaScript for real-time processing (speech recognition, gesture detection, and webcam positioning). The real-time frontend runs entirely within a Chrome browser (version 123) on a commodity laptop (Intel Core i7, 16 GB RAM), simulating typical end-user presentation scenarios.

To evaluate system responsiveness, we measured the latency of real-time components in TalkDirector: gesture recognition, speech-to-script matching, blank-space detection, and webcam repositioning. We conducted latency measurements using prerecorded presentation segments derived from all six slide decks used in Section 3.2, ensuring content diversity and layout differences.

Each operation was evaluated across 30 runs per slide deck (total $N = 180$ per component). We varied the spatial target and gesture direction (left, right, upward) across runs to reflect natural pointing diversity observed in the user study. Gesture recognition latency was recorded from the onset of the pointing gesture to successful recognition by the system. Speech-script matching latency was measured from the moment a spoken phrase ended to the identification of the corresponding script segment. This measurement includes both the real-time transcription update (via Web Speech API) and the periodic refinement using Whisper, followed by fuzzy script alignment. While Web Speech provides immediate feedback, Whisper introduces additional overhead due to neural inference, contributing to increased—but still acceptable—latency. Blank-space detection latency was recorded from initiation of the positioning query to the successful return of candidate regions. Webcam repositioning latency was measured from the moment of detection confirmation to visual update on the slide interface.

Table 2 presents the latency results for each component, reporting mean and standard deviation values:

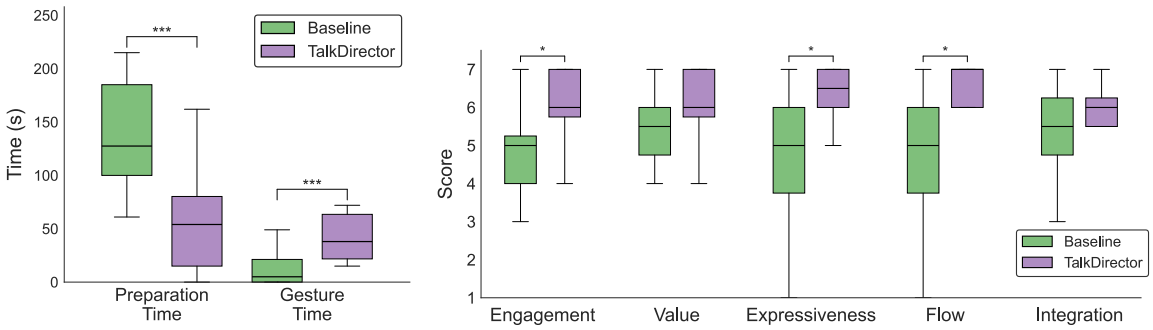
Operation	Mean Latency (ms)	Std. Dev. (ms)
Gesture Recognition	22	7
Speech-script Matching	132	24
Blank-space Detection	49	9
Webcam Repositioning	28	5

Table 2. Latency evaluation results of TalkDirector's real-time operations (N=180).

With the exception of speech-to-script matching—which includes transcription refinement and alignment with scripted segments—all latency values remained well below 100 ms. Although speech matching introduces slightly higher latency, the overall timing still falls within a range that supports smooth and responsive live presentations. These latency profiles collectively support the use of TalkDirector in live presentation settings, where timely feedback is critical to maintaining flow and audience engagement.

5 Evaluation

We evaluated TalkDirector from both presenter and audience perspectives by conducting a *two-part user study*. Our system, which dynamically integrates the presenter's video feed into the slide layouts based on content and gestures, was compared to industry-standard commercial video conferencing tools. Specifically, we examined whether (1) the automatic video control and repositioning features



(a) Talk preparation and gesture visible time.

(b) 7-point likert scale ratings for subjective questions.

Fig. 8. Quantitative analysis of time spent in preparation, with gestures, and talk presentation, as well as count of pointing gesture activation. The statistic significance is annotated with *, **, or *** (representing p<.05, p<.01, and p<.001, respectively). With comparable total presentation time, TalkDirector significantly reduced presenter's preparation time and increased both duration and activation of hand gestures.

help presenters feel more expressive, improve usability, and reduce cognitive load, and (2) whether these features enhance audience engagement and perceived connection with the speaker.

5.1 Presenter-Side Evaluation

Based on our formative insights, we hypothesize that TalkDirector's content and gesture driven video integration enhances presenters' expressiveness, improves perceived usability, and reduces cognitive load associated with manual configuration. To test this, we conducted a within-subjects comparison between TalkDirector and a baseline modeled after Zoom's *"Slide as virtual background"* feature, which allows only basic manual toggles of video feed (e.g., repositioning, fullscreen, and hide). Participants prepared and delivered slide presentations using both systems. We collected subjective ratings on usability, expressiveness, and cognitive load, along with interaction logs tracking how participants repositioned or resized their video—used as a behavioral indicator of expressive control over visual framing. As our formative studies identified natural gesturing as central to expressive communication, we also analyzed moments when presenters used natural gestures in the logs as an indicator of expressiveness. We additionally analyzed preparation time from the interaction logs as an indicator of system usability and cognitive load, reasoning that more intuitive, lower-effort configuration would lead to shorter setup durations with TalkDirector.

5.1.1 Participants. We recruited 12 participants (7 female, 5 male) from a university sample (M age = 28.42, SD = 2.97, range = 23 – 32). Participants rated their experience with giving online presentations on a 7-point Likert scale (1 = no experience, 7 = very experienced), with M = 4.92, SD = 1.40. They also reported low experience with editing webcam video during online presentations (M = 1.75, SD = 1.60) and in post-processing (M = 2.42, SD = 1.53). Each participant received a $10 e-gift card.

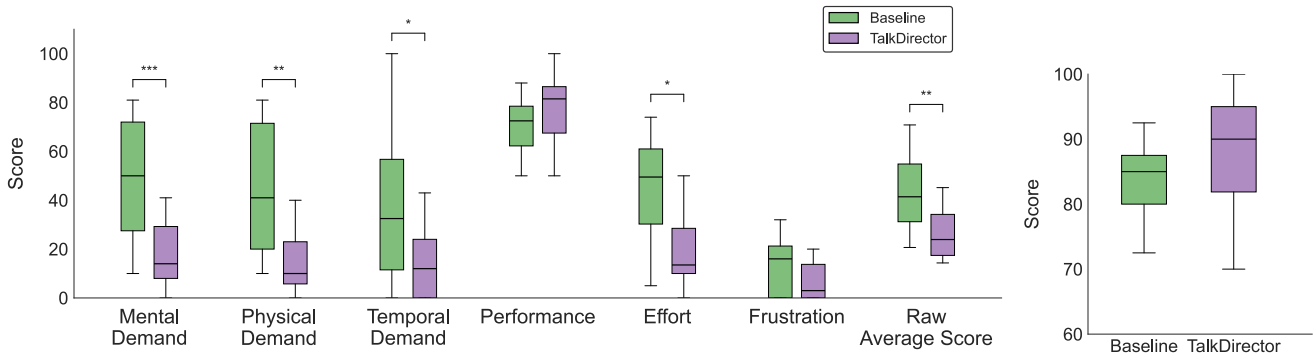
5.1.2 Cognitive Load. The NASA-TLX scores for the two conditions were analyzed using paired-samples t-tests (see Figure 9a).

Mental demand was significantly lower with TalkDirector (M = 17.50, SD = 13.71) compared to the baseline (M = 48.00, SD = 25.36); $t(11) = 3.67$, $p = 0.0014$.
Physical demand was also significantly lower with TalkDirector (M = 14.75, SD = 12.55) than with the baseline (M = 45.25, SD = 26.61); $t(11) = 3.59$, $p = 0.0016$.
Temporal demand showed a significant reduction as well, with TalkDirector at (M = 13.33, SD = 14.49) versus the baseline (M = 37.08, SD = 31.56); $t(11) = 2.37$, $p = 0.0270$.
For **performance**, TalkDirector had a higher mean score (M = 77.50, SD = 15.58) than the baseline (M = 70.00, SD = 12.93), but the difference was not statistically significant; $t(11) = -1.28$, $p = 0.21$.
Effort was significantly lower with TalkDirector (M = 22.67, SD = 23.00) compared to the baseline (M = 44.00, SD = 22.30); $t(11) = 2.31$, $p = 0.031$.
Frustration was also lower with TalkDirector (M = 11.00, SD = 18.62) than with the baseline (M = 18.08, SD = 20.69), though this difference was not statistically significant; $t(11) = 0.88$, $p = 0.39$.
Finally, the **raw total average** NASA-TLX score was significantly lower for TalkDirector (M = 26.13, SD = 10.90) compared to the baseline (M = 43.74, SD = 16.46); $t(11) = 3.09$, $p = 0.0053$.

5.1.3 Procedure. After providing informed consent, participants completed a demographic survey and received an overview of the study. They were then asked to take on the role of a "director" for their online presentation, aiming for an engaging TED Talk-like delivery. Each participant presented two pre-written slide decks, one on their hometown and one on their hobby, while being instructed to present as if the content were their own.

Participants were randomly assigned to start with either TalkDirector or the baseline system, with both system and topic order counterbalanced to control for order effects. Before presenting, they received a demo of each system and were allocated practice time with the interface, including the record, pause, and stop controls.

Both systems featured live transcription subtitles to show when speech was being recognized, ensuring low latency and a consistent user experience. Participants then delivered two presentations, one



(a) Raw-TLX results.

(b) SUS scores.

Fig. 9. Participants' ratings on Raw-TLX and SUS questionnaires in the presenter-side evaluation. The statistic significance is annotated with *, **, or *** (representing p<.05, p<.01, and p<.001, respectively). For SUS scores, No significant difference was found between TalkDirector and the Baseline.

with each system, while we recorded video data to analyze body language as an objective measure of expressiveness.

After each presentation, participants completed a questionnaire rating their experience on several dimensions: engagement, effort, expressiveness, flow, content integration, Systems Usability Scale (SUS) [3] and the NASA Task Load Index (NASA-TLX) [19]. We used a 0-100 scale for the NASA-TLX, and a 7-point Likert scale for all other ratings. To ensure informed comparisons, participants were allowed to revise their ratings for the first system after completing both presentations.

Finally, a post-hoc interview was conducted where participants explained their ratings, indicated their preference for the two systems, and discussed potential application scenarios.

5.2 Results for Presenter-Side Evaluation

5.2.1 Video Data Analysis. We analyzed participants' video data to understand their preparation and presentation behavior in both conditions (See Figure 8a and 14). The key findings are as follows:

Preparation Time. Preparation time included reading the script, planning webcam placement, and deciding when to use gestures before recording. TalkDirector significantly reduced preparation time compared to the baseline (M = 58.75, SD = 51.88 vs. M = 137.50, SD = 53.73; $t(11) = 3.65$, $p = 0.0014$).

Gesture Use. To estimate natural gesture use as an indication for expressiveness, we measured the total time participants' hands were visible in the webcam video. Using MediaPipe Hands for detection, we manually excluded non-gesture hand visibility (e.g., holding a microphone). TalkDirector significantly increased hand-visible time (M = 42.58, SD = 21.87) compared to the baseline (M = 16.00, SD = 22.66; $t(11) = -2.92$, $p = 0.0079$), suggesting more frequent gesturing.

Video Repositioning. We counted instances of video repositioning or resizing during the presentation (Figure 14). In the baseline, this required manually pausing to adjust the video, whereas in TalkDirector, repositioning occurred automatically via pointing gestures. We include this metric as an indicator of how actively presenters integrated their presence with content. TalkDirector led to significantly more repositioning events (M = 4.33, SD = 1.67) than the baseline (M = 0.50, SD = 0.67; $t(11) = -7.37$, $p < 0.0001$).

Topic Effects. To verify that observed differences were due to the system rather than presentation content, we analyzed topic effects across the two assigned topics (hometown and hobby). A two-way ANOVA showed no significant main effects of topic on preparation time ($F(1, 20) = 1.01$, $p = 0.33$), hand-visible time ($F(1, 20) = 0.37$, $p = 0.55$), or gesture activation ($F(1, 20) = 0.41$, $p = 0.53$), nor any significant interactions between method and topic.

5.2.2 User Experience. Participants rated their experience on five custom dimensions: engagement, perceived value, expressiveness, flow, and integration (see Figure 8b for full definition). **Engagement** was significantly higher with TalkDirector (M = 6.08, SD = 1.00) compared to the baseline (M = 5.00, SD = 1.21); $t(11) = -2.40$, $p = 0.0253$. **Perceived value** was also rated slightly higher with TalkDirector (M = 5.92, SD = 1.31) than with the baseline (M = 5.42, SD = 1.08), though the difference was not statistically significant; $t(11) = -1.02$, $p = 0.32$. **Expressiveness** showed a significant improvement with TalkDirector (M = 6.25, SD = 0.97) compared to the baseline (M = 4.67, SD = 1.78); $t(11) = -2.71$, $p = 0.0127$. Similarly, **flow** was rated significantly higher for TalkDirector (M = 6.33, SD = 1.15) than for the baseline (M = 4.67, SD = 1.83); $t(11) = -2.67$, $p = 0.0139$. Finally, for **Integration**, TalkDirector again received slightly higher ratings (M = 5.67, SD = 1.30) than the baseline (M = 5.33, SD = 1.44), though the difference was not statistically significant; $t(11) = -0.60$, $p = 0.56$.

5.2.3 System Usability. The **System Usability Scale (SUS)** [3] scores for the two conditions were analyzed using a paired-samples t-test (See Figure 9b). The baseline condition had a mean score of M = 82.71, SD = 7.57, while the TalkDirector condition had a mean score of M = 87.92, SD = 9.10. The t-test revealed no significant difference between the two conditions ($t(11) = -1.52$, $p = 0.14$).

5.2.4 Preference. After using both systems, 10 out of 12 participants (83.3%) preferred TalkDirector over the baseline, indicating a strong overall preference for our system.

Category	Question
Content Comprehension	The presentation interface helped me better understand the presentation content.
Social Connection	The presentation interface helped me feel more socially connected to the presenter.
Willingness to Connect	I would feel comfortable talking with this speaker again in the future.
Distraction	At any point you found the tool distracting from the presenter or the main content.
Attention	The presentation interface helped me pay attention to the primary content (e.g., the presenter's slides or video feed).

Table 3. Subjective questionnaire used in the audience evaluation study.

5.3 Audience-Side Evaluation

Remote presentations often struggle to replicate the sense of connection and engagement found in in-person talks, partly due to the disjointed layout of speaker and slide content. To address this, we explored whether dynamically integrating the presenter's video into slide content enhances audience engagement and perceived connection with the speaker. We hypothesized that this integrated format would lead to greater attention to the speaker, stronger feelings of connection, and improved content retention compared to a conventional side-by-side layout. In a within-subjects study, the audience viewed presentations delivered using either TalkDirector or a standard format typical of platforms like Zoom or Microsoft Teams. We administered post-viewing questionnaires to assess engagement, perceived content comprehension, and connection to the presenter.

5.3.1 Participants. 12 participants (6 female, 6 males, age 25–54) were recruited through an internal mailing list at Anon. All participants reported that they are experienced with online presentations ($M = 6.33$, $SD = 0.49$).

5.3.2 Procedure. Before the study, all participants provided informed consent. Upon arrival, they completed a demographic survey and were seated in front of a laptop.

Each participant viewed two presentation videos: one using a conventional side-by-side layout (baseline) and the other using our TalkDirector system. Topics included either a personal story about the presenter's hobby or an instructional video on time management. Presentation style and topic combinations were counterbalanced to mitigate order effects.

After watching the video (7–8 minutes each), participants filled out a questionnaire assessing perceived understanding, connection with the presenter, distraction, and attention, each on a 7-point Likert scale. Detailed question are listed in Table 3

Finally, semi-structured interview gathered qualitative feedback on participants' experiences and perceptions of each presentation style.

5.4 Results for Audience-Side Evaluation

We evaluated the audience experience through subjective questionnaires and follow-up interviews. We report key findings across comprehension, social connection, attentional experience, and overall preferences.

5.4.1 Content Comprehension. Participants rated their understanding of the presentation content using a 7-point Likert scale. A paired t-test revealed no significant difference between the baseline (M=4.17, SD=2.12) and TalkDirector (M=4.92, SD=1.56) conditions ($t = 0.93$, $p = 0.37$).

5.4.2 Social Connection and Presenter Perception. Participants reported a stronger sense of connection to the presenter when viewing videos created with TalkDirector (M=5.00, SD=1.41) compared to the baseline condition (M=2.67, SD=1.56) ($t = 4.55$, $p < 0.001$). While willingness to talk to the speaker again was not statistically significant, the average rating for TalkDirector (M=5.92, SD=0.90) was higher compared to the baseline (M=5.00, SD=1.54) ($t = 1.65$, $p = 0.13$).

Participants described the baseline videos as feeling "isolated" and reported difficulty connecting with the presenter. One participant noted, *"I wouldn't even be able to remember if the presenter had long hair or not."*

In contrast, several participants reported stronger attentional anchoring with TalkDirector. As P5 noted, *"I looked at the presenter more and relied on them to guide my attention."*

5.4.3 Distraction and Attention. No statistically significant differences were observed in perceived distraction (Baseline: M=3.08, SD=1.56; TalkDirector: M=3.92, SD=1.83; $t = 0.94$, $p = 0.37$) or attention (Baseline: M=4.50, SD=1.83; TalkDirector: M=5.17, SD=1.75; $t = 0.80$, $p = 0.44$). However, 9 participants explicitly mentioned that they paid more attention to what the presenter had to say rather than reading the slide content on their own, indicating a shift in attentional focus toward the speaker.

Notably, 4 participants reported difficulty concentrating during the baseline videos. In contrast, TalkDirector was often described as more engaging. One participant compared it to short-form media, *"This almost felt like watching a TikTok video – in a good way. It kept me engaged and helped me connect the content with the speaker." (P12)*

5.4.4 Presentation Preferences and Qualitative Insights. When asked which presentation format they preferred, 7 out of 12 participants selected TalkDirector. Those who preferred the baseline (5 participants) generally found the dynamic movement of the speaker distracting. Several noted that the effectiveness may depend on learning preferences; for instance, participants who preferred to read first tended to favor the static baseline format.

In contrast, those who favored TalkDirector highlighted reduced fatigue and improved attentional flow. Participants mentioned that

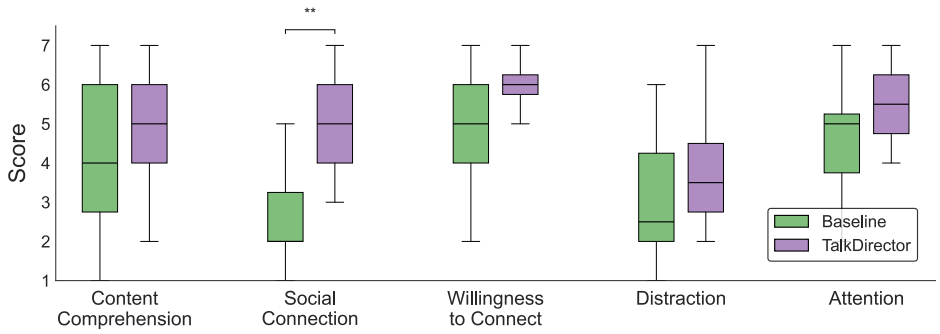


Fig. 10. Participants' Likert-scale ratings in the audience-side evaluation comparing the baseline and TalkDirector interfaces. Statistic significance is annotated with ** ($p < .01$).

the traditional layout required frequent gaze shifts, resulting in cognitive load. Two participants with experience creating online training videos expressed enthusiasm for the system, noting *"This would be especially useful for training videos, where audience can often doze off and not pay attention to the content. I would love to use this."*

Overall, TalkDirector enhanced participants' perception of speaker engagement, helped connect content with gesture and presence, and led to more memorable viewing experiences.

6 Discussion

6.1 Multimodal Integration Enables More Expressive Presentations

TalkDirector significantly increased presenters' expressiveness, both behaviorally and perceptually. Presenters' hand gestures were visible for 2.7× longer (M = 42.6s) compared to the baseline (M = 16.0s; t(11) = −2.92, p = 0.0079), and they made an average of 4.3 dynamic repositionings of the video feed using gestures per presentation, versus only 0.5 manual adjustments in the baseline (t(11) = −7.37, p < 0.0001). These results suggest that TalkDirector's multimodal coordination, using speech-content alignment and gestures to reposition the presenter, frees users from rigid layouts and fosters embodied, expressive delivery.

Subjective expressiveness ratings were also significantly higher for TalkDirector (M = 6.25) than for the baseline (M = 4.67; t(11) = −2.71, p = 0.0127), indicating that participants not only moved more, but also felt more expressive and "part of the slide," as one participant remarked.

6.2 Reducing Cognitive Load and Friction in Preparation

TalkDirector greatly reduced presenters' cognitive effort, as evidenced by both self-reports and behavioral indicators. Participants spent an average of 58.8 seconds preparing with TalkDirector—less than half the 137.5 seconds required for the baseline system (t(11) = 3.65, p = 0.0014). Subjective workload measures via NASA-TLX showed significantly lower mental demand (TalkDirector: M = 17.5 vs. Baseline: M = 48.0; p = 0.0014), physical demand (M = 14.75 vs. 45.25; p = 0.0016), and temporal demand (M = 13.3 vs. 37.1; p = 0.0270).

These findings affirm that automation in TalkDirector—driven by context-aware layout analysis and multimodal input—offloads manual configuration without compromising agency. Importantly, the system maintained a high System Usability Scale (SUS) score (M = 87.9), on par with industry standards and not significantly different from the baseline (M = 82.7; p = 0.14).

Partiocipants relied on the system to draw their attention to the right elements of the content as illustrated by this participant, "I'm just bouncing back and forth (in baseline), and I was catching myself feeling like I was trying to read the content as the speaker went over it. But I wasn't able to get that timing precisely, and it just felt like my eyes were bouncing all over the place. With the second one (TalkDirector), when the speaker was kind of moving around with the content, I think that made it a lot easier to follow along."

6.3 Enhancing Audience Engagement and Connection

TalkDirector also improved the audience experience; participants reported a significantly stronger sense of connection to the speaker when viewing TalkDirector presentations compared to standard video side-by-side formats. Engagement scores rose by nearly a full Likert point (TalkDirector: M = 6.08 vs. Baseline: M = 5.00; t(11) = −2.40, p = 0.0253), and expressiveness was rated significantly higher (M = 6.25 vs. 4.67; p = 0.0127).

Audience comments underscored this perceived connection. One participant shared, *"I paid more attention because [the speaker] moved with the content—it felt human."* Despite being more visually dynamic, TalkDirector was not reported as more distracting, suggesting that alignment between gesture, speech, and layout can enhance social presence without overloading viewers.

6.4 A Novel Format that Feels Personal

Audience participants described TalkDirector as a refreshing departure from conventional formats, emphasizing its immersive and memorable qualities. One viewer shared, "I do love the way that the presenter showing on the screen. It makes me feel more comfortable talking to the presenter after. The presentation feels more connected I would say." highlighting how integrated presenter video supported social connection.

The contrast with the baseline was stark; one participant remarked, *"I wouldn't even be able to remember if the presenter had long hair or not."* Concerns about distraction quickly faded, with another viewer explaining, *"It is not distracting… I can just make a quick gesture to move myself if I want to."*

These reactions suggest that TalkDirector not only improves expressiveness and flow, but also introduces a novel, more human-centered presentation experience.

6.5 Balancing Automation and Agency

While TalkDirector significantly reduced preparation time and cognitive load, our findings highlight a nuanced tension between convenience and control. For many participants, the automation was a welcome relief—particularly in time-constrained or informal contexts. One user remarked, *"I didn't have to think so much about it… I didn't have to spend much effort."* This aligns with the observed drop in preparation time and effort-related NASA-TLX scores, suggesting TalkDirector supports flow by offloading layout decisions.

However, some participants valued the precision and control offered by manual interfaces, especially for high-stakes or stylistically sensitive presentations. One noted, "I want to be sure the presentation is as I intended," while another expressed, *"I don't really trust AI to take the wheel for me."* These perspectives reflect the classic HCI tension between *adaptive* systems, which make decisions for users, and *adaptable* systems, which allow users to tune or override behavior.

Our results suggest that presenter preferences are context-dependent: TalkDirector's adaptive mode suits fast, low-effort delivery, but may require more personalization or override mechanisms for users who prioritize stylistic agency. Supporting gesture-based repositioning helped bridge this gap, but future systems might further benefit from learning user preferences over time, transitioning from adaptive to co-adaptive behavior.

7 Interfaces for Shared Control

TalkDirector enables presenters to balance automation and manual control during slide presentations, accommodating a wide range of delivery styles. From our formative workshop studies, we found some presenters prefer to let the system infer optimal video placement based on gestures, speech, and slide content. Others want to make layout decisions themselves or override system actions when needed. To support this flexibility, TalkDirector includes a set of interface elements that embody shared control. These elements allow users to accept or reject suggestions, make quick layout decisions on the fly, or customize default behavior across a deck.

To illustrate how shared control can be embedded into dynamic presentation systems, we present three example interfaces: a real-time suggestion panel, a quick transition selector, and a preferences-based customization panel. These interface concepts draw directly from presenter needs identified in our study and showcase different ways TalkDirector supports hybrid control.

7.1 Quick Transition Selection

During slide transitions, presenters often have preferences about how their video feed should behave before the next slide appears—whether it should stay visible, become fullscreen for emphasis, or hide to let the slide content speak for itself. To support this, TalkDirector offers a quick-select overlay that briefly appears at the start of each new slide (Figure 11).

The overlay includes options like "fullscreen," "hide," or "default," and disappears after a short timeout. This design enables presenters to make fast, per-slide layout decisions without breaking delivery flow. It is especially useful in high-tempo presentations or when the presenter hasn't pre-scripted video behavior for each slide. While the presenter speaks, they can tap or gesture to indicate the preferred layout mode and continue without interruption.

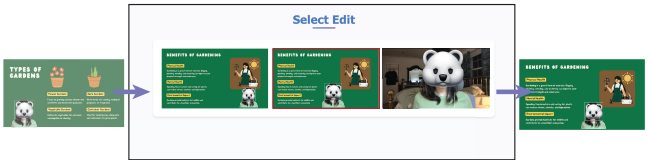


Fig. 11. Quick Transition Selection. Users are able to select their desired appearance at the moment of progressing to the next slide.

7.2 Suggestion Panel

When presenting, users may want assistance from the system while still retaining control over what changes occur (Figure 12). The suggestion panel offers a lightweight mechanism for this. It appears contextually, such as after a detected gesture, topic shift, or layout change, and surfaces recommended actions like moving the presenter video, switching to fullscreen, or hiding the video feed temporarily.

These suggestions are not enforced. Instead, the presenter can quickly accept, modify, or ignore them. This model allows the system to assist with awareness of context, while the presenter maintains agency. For example, a presenter discussing a data-heavy slide may be prompted to move the video feed to a less obstructive corner, but can choose to keep it centered if maintaining facial presence is important.

7.3 Preference-driven Control

Not all presenters want to make layout decisions live. For those who prefer to set behavior in advance, TalkDirector includes a preferences panel where users can define default video placement rules, gesture responsiveness, and layout behavior for different types of slides (Figure 13).

These settings can be applied globally or customized per slide. For example, a presenter might want the video feed to be fullscreen during introductory slides, move to the lower-right corner for content-heavy sections, and hide during detailed diagrams. Preferences can also specify how the system should react to specific gestures—e.g., ignoring hand movement unless clearly directed toward the slide.

More advanced configurations may include presets or lightweight scripting to allow presenters to define sequences or behaviors that

Fig. 12. Suggestion Panel. Given the slide layout and user's speech transcription, the system suggests top four likely transitions user will be interested in.

match their presentation narrative. This level of customization supports a wide range of personal styles, from minimalist to expressive, and allows the system to stay aligned with the user's intent.

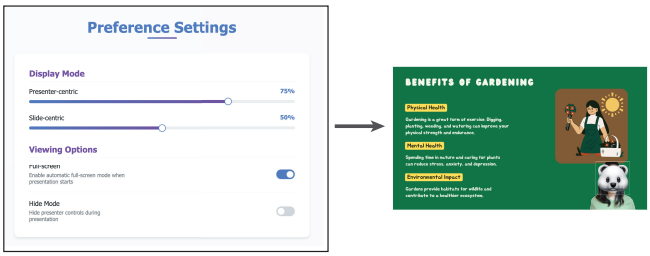


Fig. 13. Preference-driven Control. Users can set their desired features and controls ahead of the pressentation.

8 Limitations and Future Work

While TalkDirector demonstrates robust performance across diverse presentation formats and outperforms traditional systems in both expressiveness and usability, several current boundaries point toward exciting avenues for future development.

First, our evaluation primarily targeted three core types of slide presentations—personal narratives, tutorials, and professional talks—covering a broad and representative range of real-world use. Within these domains, TalkDirector showed strong alignment accuracy, seamless presenter-content integration, and measurable improvements in both engagement and cognitive load. However, our current pipeline has not yet been tested on more dynamic or unconventional formats, such as slides with embedded animations, live coding segments, or densely layered multimedia. These formats may present additional challenges for our layout inference model and real-time multimodal coordination. Expanding TalkDirector's capabilities to handle such high-tempo, nonlinear workflows is a promising next step.

Second, our user studies were conducted in a controlled lab environment using pre-authored scripts and pre-segmented slide decks to ensure systematic comparisons and repeatability. This allowed for a rigorous evaluation of multimodal inference accuracy and user perception. However, in naturalistic presentation settings, speakers often deviate from scripts, improvise gestures, and respond dynamically to live audience input. While TalkDirector already accommodates unscripted speech to some extent through fuzzy semantic matching and gesture-based video positioning, further exploration into unstructured, in-the-wild use cases will be essential for real-world deployment.

Another consideration is TalkDirector's reliance on aligned scripts during preprocessing. Although our system was designed to tolerate spontaneous phrasing variations through fuzzy matching, it still performs best when presenters loosely follow their scripts. In more improvisational or conversational styles—such as fireside chats or panel discussions—alignment strategies based on real-time semantic parsing, discourse tracking, or topic drift detection may be necessary to maintain robust contextual placement.

Moreover, our current gesture recognition pipeline focuses on index-finger pointing, a deliberate design choice grounded in reliability and clarity for targeting slide elements. This approach yielded strong performance in our user studies, enabling precise, low-latency presenter-content alignment. That said, this limits the expressive range of embodied interaction. Broadening the gesture vocabulary to include open-hand sweeps, framing gestures, and posture shifts could further enrich the system's capacity to interpret speaker intent, and support more nuanced communication styles.

Despite these limitations, TalkDirector lays the groundwork for a new class of intelligent, multimodal presentation tools. It introduces a unified real-time pipeline, validated with strong quantitative and qualitative results, and contributes a dataset (VTalk-68) that captures diverse presenter behaviors. Future work will extend these foundations toward even greater adaptability, supporting dynamic content types, broader gesture semantics, and less scripted, more conversational presentation flows—pushing toward more inclusive, expressive, and scalable presentation experiences.

9 Conclusion

Effective presentations depend on the seamless integration of gestures, speech, and visual content. However, current tools for dynamically embedding presenters into slides often require manual interventions, often leading to static video placements that sacrifices the full range of expressiveness and engagement between the presenter and audience. We introduced *TalkDirector*, a system that adaptively integrates the presenter's video feed using multimodal cues such as speech, gestures, slide layout, and content relevance. By combining automatic speech recognition, real-time segmentation, gesture detection, and multimodal models, TalkDirector infers optimal video placement and sizing in real time.

Our formative workshops and the VTalk-68 dataset of 68 presentations provided foundational insights into presenter preferences and design considerations for dynamic video integration. A two-part user study showed that TalkDirector reduced presenters' cognitive load, increased expressiveness and usability, and enhanced audience engagement and perceived connection with the speaker. Together, these findings demonstrate the potential of multimodal, context-aware video integration to support more effective and expressive online presentations.

References

[1] Thomas Baudel and Michel Beaudouin-Lafon. 1993. Charade: Remote Control of Objects Using Free-Hand Gestures. *Commun. ACM* 36, 7 (1993), 28–35.

[2] Janet Beavin Bavelas, Nicole Chovil, Linda Coates, and Lori Roe. 1995. Gestures Specialized for Dialogue. *Personality and Social Psychology Bulletin* 21, 4 (1995), 394–405. https://doi.org/10.1177/0146167295214010

[3] J Brooke. 1996. SUS: A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* (1996). https://doi.org/10.1080/10447310802205776

[4] Yining Cao, Rubaiat Habib Kazi, Li-Yi Wei, Deepali Aneja, and Haijun Xia. 2024. Elastica: Adaptive Live Augmented Presentations With Elastic Mappings Across Modalities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 599, 19 pages. https://doi.org/10.1145/3613904.3642725

[5] Josh Urban Davis, Paul Asente, and Xing-Dong Yang. 2023. Multimodal Direct Manipulation in Video Conferencing: Challenges and Opportunities. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1174–1193.

[6] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2024. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. 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 600, 22 pages. https://doi.org/10.1145/3613904.3642579

[7] Neil deGrasse Tyson. 2012. The Inexplicable Universe: Unsolved Mysteries. https://www.thegreatcourses.com/courses/the-inexplicableuniverse-unsolved-mysteries

[8] Nancy Duarte. 2010. *Resonate: Present Visual Stories That Transform Audiences*. John Wiley & Sons.

[9] Michael E Ellis. 1992. Perceived Proxemic Distance and Instructional Videoconferencing: Impact on Student Performance and Attitude. (1992). https://doi.org/10.5688/aj680358

[10] Randi A Engle. 2022. Not Channels But Composite Signals: Speech, Gesture, Diagrams and Object Demonstrations Are Integrated in Multimodal Explanations. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Routledge, Routledge, 321–326. https://doi.org/10.4324/9781315782416-65

[11] Raymond Fok, Nedim Lipka, Tong Sun, and Alexa F Siu. 2024. Marco: Supporting Business Document Workflows via Collection-Centric Information Foraging with Large Language Models. 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 842, 20 pages. https://doi.org/10.1145/3613904.3641969

[12] Gerald Friedland and Raul Rojas. 2007. Anthropocentric Video Segmentation for Lecture Webcasts. *EURASIP Journal on Image and Video Processing* 2008 (2007), 1–10. https://doi.org/10.1155/2008/195743

[13] Susan Goldin-Meadow and Martha Wagner Alibali. 2013. Gesture's Role in Speaking, Learning, and Creating Language. *Annual Review of Psychology* 64, 1 (2013), 257–283. https://doi.org/10.1146/annurev-psych-113011-143802

[14] Jiangtao Gong, Teng Han, Siling Guo, Jiannan Li, Siyu Zha, Liuxin Zhang, Feng Tian, Qianying Wang, and Yong Rui. 2021. Holoboard: A Large-Format Immersive Teaching Board Based on Pseudo Holographics. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 441–456.

[15] Google. 2023. Google Meet. https://meet.google.com/landing. Accessed: 2023.

[16] Philip J. Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: an empirical study of MOOC videos. In *Proceedings of the First ACM Conference on Learning @ Scale Conference* (Atlanta, Georgia, USA) *(L@S '14)*. Association for Computing Machinery, New York, NY, USA, 41–50. https://doi.org/10.1145/2556325.2566239

[17] Brian D Hall, Lyn Bartram, and Matthew Brehmer. 2022. Augmented Chironomia for Presenting Data to Remote Audiences. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–14. https://doi.org/10.1145/3526113.3545614

[18] Yuexing Hao, Zeyu Liu, Robert N. Riter, and Saleh Kalantari. 2024. Advancing Patient-Centered Shared Decision-Making with AI Systems for Older Adult Cancer Patients. 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 437, 20 pages. https://doi.org/10.1145/3613904.3642353

[19] SG Hart. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Human Mental Workload/Elsevier* (1988). https://doi.org/10.1016/s0166-4115

[20] Keita Higuchi, Yinpeng Chen, PhilipA. Chou, Zhengyou Zhang, and Zicheng Liu. 2015. ImmerseBoard: Immersive Telepresence Experience Using a Digital Whiteboard. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/2702123.2702160

[21] Dani Paul Hove and Benjamin Watson. 2022. The Shortcomings of Video Conferencing Technology, Methods for Revealing Them, and Emerging XR Solutions. *PRESENCE: Virtual and Augmented Reality* 31 (2022), 283–305.

[22] Erzhen Hu, Jens Emil Sloth 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*. 1–17.

[23] 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*. 1–22.

[24] Jaeho Jeon, Seongyong Lee, and Hohsung Choe. 2022. Enhancing EFL pre-service teachers' affordance noticing and utilizing with the Synthesis of Qualitative Evidence strategies: An exploratory study of a customizable virtual environment platform. *Computers & Education* 190 (2022), 104620.

[25] Seokmin Kang, Barbara Tversky, and John B Black. 2015. Coordinating Gesture, Word, and Diagram: Explanations for Experts and Novices. *Spatial Cognition & Computation* 15, 1 (2015), 1–26. https://doi.org/10.1080/tand_crossmar_01

[26] Jin K Kim, Michael Chua, Mandy Rickard, and Armando Lorenzo. 2023. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology* 19, 5 (2023), 598–604.

[27] Minju Kim, Jungjin Lee, Wolfgang Stuerzlinger, and Kwangyun Wohn. 2016. HoloStation: Augmented Visualization and Presentation. In *SIGGRAPH Asia 2016 Symposium on Visualization*. 1–9.

[28] Minju Kim and Kwangyun Wohn. 2018. HoloBox: Augmented Visualization and Presentation With Spatially Integrated Presenter. *Interacting With Computers* 30, 3 (2018), 224–242. https://doi.org/10.1093/iwc/iwy007

[29] Charlotte Kobiella, Yarhy Said Flores López, Franz Waltenberger, Fiona Draxler, and Albrecht Schmidt. 2024. "If the Machine Is As Good As Me, Then What Use Am I?" – How the Use of ChatGPT Changes Young Professionals' Perception of Productivity and Accomplishment. 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 1018, 16 pages. https://doi.org/10.1145/3613904.3641964

[30] Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM. 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 766, 28 pages. https://doi.org/10.1145/3613904.3642830

[31] Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. PaperWeaver: Enriching Topical Paper Alerts by Contextualizing Recommended Papers with User-collected Papers. 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 19, 19 pages. https://doi.org/10.1145/3613904.3642196

[32] Jian Liao, Adnan Karim, Shivesh Singh Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. 2022. Realitytalk: Real-Time Speech-Driven Augmented Presentation for Ar Live Storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–12. https://doi.org/10.1145/3526113.3545702

[33] Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, and Ruofei Du. 2023. Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals. 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 108, 20 pages. https://doi.org/10.1145/3544548.3581566

[34] Fabrice Matulic, Lars Engeln, Christoph Träger, and Raimund Dachselt. 2016. Embodied Interactions for Novel Immersive Presentational Experiences. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1713–1720. https://doi.org/10.1145/2851581.2892501

[35] Richard E Mayer. 2005. *The Cambridge handbook of multimedia learning*. Cambridge university press.

[36] MediaPipe. 2023. MediaPipe. https://developers.google.com/mediapipe. Accessed: 2023.

[37] Microsoft Support. 2023. Presenting With Cameo. https://support.microsoft.com/en-gb/office/presenting-with-cameo-83abdb2e-948a-47d0-932d-86815ae1317a Accessed: 2024-09-08.

[38] Microsoft Tech Community. 2021. Now in Public Preview: Dynamic View. https://techcommunity.microsoft.com/t5/microsoft-teams-public-preview/now-in-public-preview-dynamic-view/m-p/2264831. Accessed: 2023.

[39] Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill, and Sedinha Teßendorf. 2013. Body-Language-Communication. *An International Handbook on Multimodality in Human Interaction* 1, 1 (2013), 131–232. https://doi.org/10.1007/978-3-031-70064-4

[40] OpenAI. 2022. Whisper: OpenAI's Speech Recognition Model. https://openai.com/research/whisper Accessed: 2024-09-09.

[41] OpenAI. 2023. GPT-4: OpenAI's Multimodal Large Language Model. https://openai.com/research/gpt-4 Accessed: 2024-09-09.
[42] Ken Perlin, Zhenyi He, and Karl Rosenberg. 2018. Chalktalk: A Visualization and Communication Language-As a Tool in the Domain of Computer Science Education. *ArXiv Preprint ArXiv:1809.07166* (2018).
[43] PPTX2HTML. 2023. PPTX2HTML: Convert PowerPoint to HTML. https://g21589.github.io/PPTX2HTML/. Accessed: 2023.
[44] Xun Qian, Feitong Tan, Yinda Zhang, Brian Moreno Collins, David Kim, Alex Olwal, Karthik Ramani, and Ruofei Du. 2024. ChatDirector: Enhancing Video Conferencing With Space-Aware Scene Rendering and Speech-Driven Layout Transition. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.
[45] Garr Reynolds. 2011. *Presentation Zen: Simple Ideas on Presentation Design and Delivery*. New Riders.
[46] Hans Rosling. 2010. 200 Countries, 200 Years, 4 Minutes. BBC.
[47] Hans Rosling. 2013. The River of Myths. Self-published or other source, if applicable.
[48] Nazmus Saquib, Rubaiat Habib Kazi, Li-Yi Wei, and Wilmot Li. 2019. Interactive Body-Driven Graphics for Augmented Video Performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3290605.3300852
[49] Ray Smith and Google. 2005. Tesseract OCR: An Open-Source Optical Character Recognition Engine. https://github.com/tesseract-ocr/tesseract Accessed: 2024-09-09.
[50] Murphy Stein. 2012. ARCADE: A System for Augmenting Gesture-Based Computer Graphic Presentations. In *ACM SIGGRAPH 2012 Computer Animation Festival*. 77–77.
[51] Ryo Suzuki, Rubaiat Habib Kazi, Li-Yi Wei, Stephen DiVerdi, Wilmot Li, and Daniel Leithinger. 2020. Realitysketch: Embedding Responsive Graphics and Visualizations in AR Through Dynamic Sketching. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 166–181. https://doi.org/10.1145/3379337.3415892
[52] Yilin Tang, Liuqing Chen, Ziyu Chen, Wenkai Chen, Yu Cai, Yao Du, Fan Yang, and Lingyun Sun. 2024. EmoEden: Applying Generative Artificial Intelligence to Emotional Learning for Children with High-Function Autism. 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 1001, 20 pages. https://doi.org/10.1145/3613904.3642899
[53] Santawat Thanyadit, Matthias Heintz, and Effie LC Law. 2023. Tutor In-Sight: Guiding and Visualizing Students' Attention With Mixed Reality Avatar Presentation Tools. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
[54] Vox. 2015. Obama on What Most Americans Get Wrong About Foreign Aid. YouTube video. Retrieved 2023 from https://youtu.be/nzL_avUllEE.
[55] Jiyao Wang, Haolong Hu, Zuyuan Wang, Song Yan, Youyu Sheng, and Dengbo He. 2024. Evaluating Large Language Models on Academic Literature Understanding and Review: An Empirical Study among Early-stage Scholars. 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 12, 18 pages. https://doi.org/10.1145/3613904.3641917
[56] Zijie J. Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. 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 976, 40 pages. https://doi.org/10.1145/3613904.3642335
[57] Andrew D Wilson, Shahram Izadi, Otmar Hilliges, Armando Garcia-Mendoza, and David Kirk. 2008. Bringing Physics to the Surface. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology*. 67–76.
[58] Anna Xygkou, Chee Siang Ang, Panote Siriaraya, Jonasz Piotr Kopecki, Alexandra Covaci, Eiman Kanjo, and Wan-Jou She. 2024. MindTalker: Navigating the Complexities of AI-Enhanced Social Engagement for People with Early-Stage Dementia. 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 96, 15 pages. https://doi.org/10.1145/3613904.3642538
[59] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing Biomedical Literature to Calibrate Clinicians' Trust in AI Decision Support Systems. 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 14, 14 pages. https://doi.org/10.1145/3544548.3581393
[60] Yi Zhang, Ke Xu, Zhongling Pi, and Jiumin Yang. 2022. Instructor's Position Affects Learning From Video Lectures in Chinese Context: An Eye-Tracking Study. *Behaviour & Information Technology* 41, 9 (2022), 1988–1997. https://doi.org/10.1080/tand_crossmar_01
[61] Zhongyi Zhou, Jing Jin, Vrushank Phadnis, Xiuxiu Yuan, Jun Jiang, Xun Qian, Jingtao Zhou, Yiyi Huang, Zheng Xu, Yinda Zhang, Kristen Wright, Jason Mayes, Mark Sherwood, Johnny Lee, Alex Olwal, David Kim, Ram Iyengar, Na Li, and Ruofei Du. 2025. InstructPipe: Building Visual Programming Pipelines in Visual Blocks With Human Instructions Using LLMs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1–30. https://doi.org/10.1145/3706598.3713905
[62] Zoom Support. 2023. Sharing Slides As a Virtual Background. https://support.zoom.com/hc/en/article?id=zm_kb&sysparm_article=KB0067697. Accessed: 2023.
[63] Zoom Video Communications, Inc. 2022. Sharing Slides As a Virtual Background. https://support.zoom.us/hc/en-us/articles/360046912351-Sharing-slides-as-a-Virtual-Background. Feature introduced in Zoom version 5.2.0.
[64] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation. 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 450, 18 pages. https://doi.org/10.1145/3613904.3642450

A Evaulation Study Questionnaires

A.1 Custom 7-Point Likert-Scale Questionnaire

(1) **Engagement**: I feel that this system helps me give engaging presentations.
(2) **Value**: The presentation outcome is worth the effort that I put into preparing, and delivering using this system.
(3) **Expressiveness**: I feel that I am able to convey ideas, emotions, and messages effectively through both verbal and non-verbal communication.
(4) **Flow**: I feel that I am immersed and uninterrupted during the presentation.
(5) **Integration**: I feel like my video feed is well-integrated with the content.

A.2 System Usability Scale (7-Point Likert-Scale)

We adapted from the SUS.

(1) I think that I would like to use this system frequently.
(2) I found the system unnecessarily complex.
(3) I thought this product was easy to use.
(4) I think that I would need the support of a technical person to be able to use this product.
(5) I found the various functions in the system were well integrated.
(6) I thought there was too much inconsistency in this system.
(7) I imagine that most people would learn to use this system very quickly.
(8) I found the system very awkward to use.
(9) I felt very confident using the system.
(10) I needed to learn a lot of things before I could get going with this system.

A.3 NASA-TLX

(1) **Mental Demand**: How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc)? Was the task easy or demanding, simple or complex, exacting or forgiving?
(2) **Physical Demand**: How much physical activity was required (e.g., dragging-and-dropping, clicking, typing, pushing, pulling, turning, controlling, activating, etc)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
(3) **Temporal Demand**: How much time pressure did you feel due to the rate of pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
(4) **Performance**: How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
(5) **Effort**: How hard did you have to work (mentally and physically) to accomplish your level of performance?
(6) **Frustration**: How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

A.4 Post-hoc Semi-Structured Interview

(1) Can you explain your rationale for the ratings you give for the items in the questionnaire?
(2) What is your preference with the two systems? Can you explain?
(3) Can you imagine specific use of the system in your life?
(4) What features would you like to add to the system?
(5) What will your dream online presentation system look like in the future regardless of technical constraints?

A.5 Statistical Results from User Studies

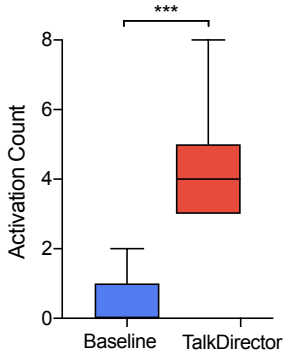


Fig. 14. Number of Presenter Video Feed Adjustment within Slides