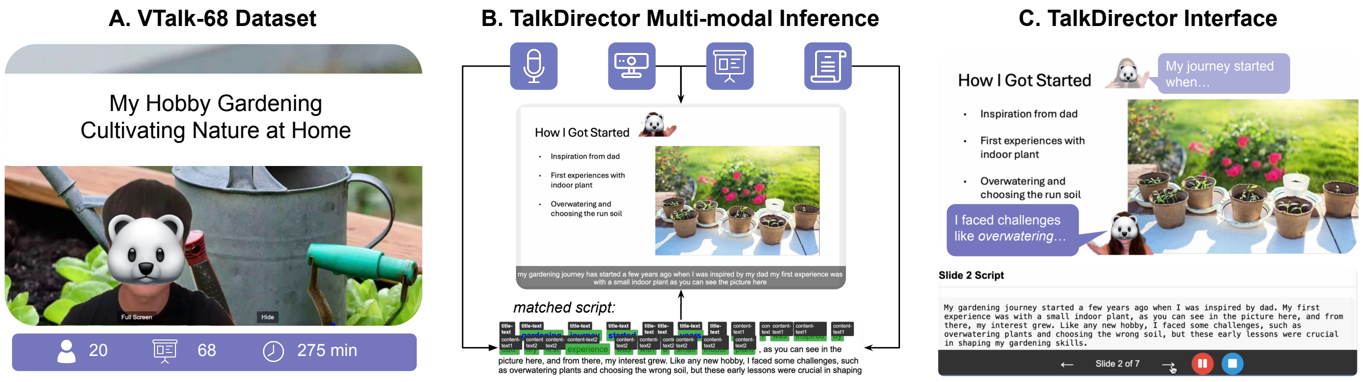
TalkDirector: Interactive Integration of Presenter-Slides for Online Presentations

ANONYMOUS AUTHOR(S)*


A. VTalk-68 Dataset
My Hobby Gardening
Cultivating Nature at Home
Full Screen
Hide
20
68
275 min
B. TalkDirector Multi-modal Inference
How I Got Started
Inspiration from dad
First experiences with indoor plant
Overwatering and choosing the run soil
matched script:
title-text
content-text1
content-text2
C. TalkDirector Interface
How I Got Started
Inspiration from dad
First experiences with indoor plant
Overwatering and choosing the run soil
My journey started when…
I faced challenges like *overwatering*…
Slide 2 Script
←
Slide 2 of 7
→

my gardening journey has started a few years ago when I was inspired by my dad my first experience was with a small indoor plant as you can see the picture here

dad. My first experience was with a small indoor plant, as you can see in the picture here, and from there, my interest grew. Like any new hobby, I faced some challenges, such as overwatering plants and choosing the wrong soil, but these early lessons were crucial in shaping

```
My gardening journey started a few years ago when I was inspired by dad. My first
experience was with a small indoor plant, as you can see in the picture here, and from
there, my interest grew. Like any new hobby, I faced some challenges, such as
overwatering plants and choosing the wrong soil, but these early lessons were crucial
in shaping my gardening skills.
```

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, a dataset of 68 presentations with personalized video integration; (B) TalkDirector's multi-modal pipeline to inference video placement and size based on speech content, gestures, slides, and scripts; and (C) the TalkDirector interface, which runs automatic speech recognition (ASR), background segmentation, gesture detection, and the multi-modal inference pipeline, and interactively renders the presenter's video within the slides.

Effective presentations blend gestures, speech, and slides contents harmoniously. However, current tools for embedding presenters into slides often demand substantial manual effort. We introduce TalkDirector, a novel system that dynamically renders the presenter's video feed within slides, optimizing placement for content relevance, gestures, and layout. Our pipeline leverages automatic speech recognition, real-time segmentation, gesture detection, and large multi-modal models to infer the video feed's placement and size within the slides. We conducted two workshops (n=5) to understand motivations and decision-making for video placements in presentations. Further, we collected VTalk-68, a multimodal dataset of 68 presentations with 20 presenters to gain insights into user preferences for video placements in various presentation contexts. A user evaluation against a baseline prototype (n=12) showed that TalkDirector significantly reduces cognitive load and enhances presentation effectiveness, underscoring the potential of dynamic video integration to improve online presentations.

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

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

ACM Reference Format:
Anonymous Author(s). 20XX. TalkDirector: Interactive Integration of Presenter-Slides for Online Presentations. In *Proceedings of the XX, Nov 26–Dec 1, 20XX, XX, XX*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Online presentations have become ubiquitous in both academic and professional settings. While platforms like Zoom and Microsoft Teams allow presenters to manually overlay their video feeds onto slides, most presentations rely on fixed layouts due to the additional cognitive load of manual adjustments. Inspired by TED-style presentations, where directors dynamically align slides and presenters based on talk content and body language, we asked: can online presentations benefit from automated integration of presenters and slides?

Effective presentations rely on multimodal communication, including slides, verbal communication, and nonverbal cues like gestures, facial expressions, and body movements [3, 25]. These cues emphasize key points, clarify ideas, and keep audience attention [5, 37]. Research shows that combining verbal and nonverbal signals enhances audience engagement and emotional response [4, 8]. Therefore, designing an automatic system to dynamically adjust the video feed can ensure nonverbal cues are highlighted effectively, improving audience comprehension and engagement. Key design criteria would include responsiveness to gestures, presentation contents and seamless integration of presenter video feed and graphical presentation contents. Current works in augmented presentation often addresses some of these criteria, with presenter video feed as background and graphical overlays on top. However, these content over presenter video feed methods don't quite work with predominant slide-based presentations, or presentations that require denser information delivery.

To investigate how might we augment existing slide-based presentation experiences with dynamic multimodal input from presenters, we conducted two workshops with five participants experienced in

XX '2x, Nov 26–Dec 1, XX, XX

presentation video editing to explore their motivations, decision-making processes, and workflow patterns. Additionally, we recruited 20 participants for a data collection study, resulting in VTalk-68 (Figure 1A), a multimodal dataset comprising 68 live talks over various 3 categories comprising of 6 topics, totaling over 275 minutes. Each participant prepared 3-4 talk in a wide range of topics including hometown, hobby, tutorial, lectures, research, and inventions. We compiled the VTalk-68 dataset to include structured data on slide content, preferred video placements , transcripts, and presenter gestures, aiming to inform our design and guide future research.

Drawing from insights gained through data analysis and semi-structured interviews, we developed a presenter-slides layout inference pipeline that computes and adjusts the layout in real-time, dynamically adapting as the presentation unfolds. This pipeline takes multi-modal inputs into account, including the presenter's speech, hand gestures, and the slide's scripts, contents (*e.g.*, text, images, videos), and layouts. Leveraging techniques such as Optical Character Recognition (OCR) and large multimodal generative models, the system continuously calculates video positioning, size, and visibility based on real-time data from the presenter's speech, gestures, and the slide material (Figure 1B).

Finally, we present **TalkDirector**, an online presentation system that interactively integrates the presenter and slides (Figure 1C). TalkDirector runs real-time automatic speech recognition (ASR), background segmentation, and gesture recognition, feeding these intermediate results to the multi-modal layout inference pipeline to dynamically position present's video feed within the slides. We conducted a comparative evaluation of TalkDirector with 12 participants, using a baseline similar to Zoom's "Slide as Virtual Background" feature. Our findings indicate that TalkDirector significantly reduced the time required to prepare for an interactive presentation, and participants reported a notable increase in perceived expressiveness during their presentations. Our contributions include:

- **Design implications** derived from patterns identified in two workshop studies and **VTalk-68**, a multimodal dataset of online presentations, highlighting how users integrate themselves into slides during presentations. These insights informed the development of dynamic presenter-slide integration and transitions.
- **TalkDirector**, a novel interface that enables live-editing of video feeds into presentation slides, enhancing flexibility and engagement in online presentations.
- **A multimodal framework** that dynamically adjusts the presenter's video feed based on speech, gestures, and slide content.
- **An evaluation study** exploring cognitive load and presentation effectiveness, along with a user study to evaluate users' preferences when using TalkDirector.

By open-sourcing our dataset, system, and method[1], we aim to advance research and development in online presentation tools, contributing to more interactive and effective presentations through seamless human-AI collaboration.

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 not only on presentation slides, verbal communication, but also on non-verbal cues such as gestures, facial expressions, eye contact, and body movements [3, 25]. Gestures are integral to speaking [27, 31], and empirical studies have shown that human interpret these multimodal signals as a unified whole in communication [14]. These multimodal cues help emphasize key points, clarify complex ideas, and sustain audience attention by directing focus to critical content areas [5, 37]. Additionally, a great presentation that blends languages with body gestures can shape how an audience interprets meaning and responds emotionally, directly influencing engagement and empathy [4, 8]. For example, gestures synchronized with speech (gestural synchrony) significantly improve content clarity, aiding cognitive processing for both the presenter and the audience [16]. Eye contact and intentional facial expressions further help maintain attention and foster a connection with the audience, both of which are critical for sustained engagement and information retention [26]. Dynamic body movements can evoke emotional responses and create a more immersive experience, leading to deeper audience involvement [1]. Importantly, body interaction within the presentation space can highlight content transitions or draw attention to new visual elements [43]. This alignment of body displacements with contents spatially allows presenters to create a more interactive and engaging environment, adaptable to both in-person and virtual settings [12]. Our system aims to increase presenters' body interaction with contents through dynamically generating the appropriate video feed placement that *spatially aligns with the relevant content*.

2.2 Body-Content Interaction in Augmented Presentations

Current remote presentation technologies strive to bridge the gap between remote and in-person communication, particularly in body-content interactions – allowing user body to more seamlessly interact with the contents. Augmented presentations have long been used in classroom educational settings [11, 17, 40, 51], online explanatory videos [54] and public presentations [44, 45]. Many augmented presentations are created through post-production processes [32], yet with the popularity of livestreaming, video conferencing and augmented reality, HCI researchers are showing increasing interest in creating augmented presentations [2, 10, 17, 19, 28, 33]. Here, we present a unified overview of augmented presentation systems that seamlessly integrate the user's body languages with interactive content.

2.2.1 Interactive Presentations with Presenter as Foreground. While commercial 2D video conferencing tools like Zoom, Microsoft Teams, and Google Meet can capture gestures, eye contact, and some body language, they fall short in terms of conveying body displacements and spatial dynamics when the presenter is interacting with contents [18, 36, 57]. These platforms mainly emphasize upper body

[1] GitHub link deducted for anonymity.

Project	POC	COP	Adaptive	Input Modality
Tutor In-sight [51]	✓ (avatar)		✓	mouse
Microsoft Teams, Microsoft Cameo, Google Meet [18, 35, 36]				mouse & real-world contents
Zoom 'Slides as Virtual Background' [57]	✓			mouse
OpenMic [23]				mouse
ChatDirector [42]			✓	speech
Charade [2]				hand (glove-tracking)
Bringing physics to the surface [55]	✓		✓	hand
Chalktalk [40]		✓		mouse
RealitySketch [50]		✓	✓	mouse & real-world objects
RealityTalk [32]		✓	✓	speech & gestures
Elastica [7]		✓	✓	pre-defined animations, speech & gestures
Interactive Body-Driven Graphics [46]		✓	✓	gestures & postures
Augmented Chironomia [19]		✓	✓	gestures
ARCADE [49]		✓	✓	hand
ThingShare [24]				mouse
Matulic et al., [33]	✓		✓	gestures & slide elements
Our Work	✓		✓	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.

and facial expressions, which work for some non-verbal cues. However, key aspects like body-content interactions, such as moving around to emphasize points or physically engaging with content are often overlooked, leading to a less natural communication experience [22]. Although body-content interaction in 3D can not be fully replicated in 2D video conferencing, researchers found presentation can be improved through dynamically resizing, repositioning, and seamlessly integrating the presenter's video feed with the content as background. Friedland et al. [15] highlight that separating slides from the presenter's video, which shows facial expressions and gestures, leads to a split-attention effect. To enhance the connection between content and the presenter's gestures, they propose extracting and overlaying the presenter directly onto the presentation feed, improving cognitive association between visual cues and spoken information. Ellis et al. [13] found that the size of the instructor's video feed in online lectures affects students' perceived closeness to the instructor. Larger video sizes create a stronger impression and lead to better learning performance. Zhang et al. discovered instructor's presence and location on screen, in particular on the right side of the screen increases learning performance and satisfaction in students [56]. Recently, commercial tools like Zoom have enabled presenters to integrate and adjust video feed size and position with features like 'Slides as Virtual Background.' [58] However, these adjustments require manual resizing and positioning without context-aware adaptations, or occlusion handling, imposing a high cognitive load [57]. Similarly, Microsoft Cameo allows users to insert their video feed in slide editing software, as a preparation tool for presentations rather than live augmented presentation tool [35]. The 'Dynamic View' feature in Microsoft Teams automatically optimizes the layout of shared content and participant video feeds based on interaction patterns in the. However, it lacks automated adjustments specifically for the presenter's video feed, requiring manual intervention for optimal placement [36]. HCI researchers have also explored manipulating video feeds in multiparty video conferences. Hu et al. proposed dynamically resizing participants' video feeds to represent proxemic metaphors during conversational floor transitions in video conferencing [23]. ChatDirector generates users' video feeds into 3D portrait avatars and proposes a pipeline to render them in a virtual 3D space for space-aware attentional transition [42]. Other works in HCI explored systems such as emulating writing side-by-side on a physical whiteboard with gesture content coupling [21]; using hand gestures to manipulate and simulate physics in presentations [55]; integration of presenter avatars with their presentation content, allowing full gestural manipulation of presentation elements [33]. Our system builds on these insights to enhance communication by adaptively render the presenter's video feed over slide backgrounds, using multimodal inputs ranging from including slide content, presenter gestures, and speech.

2.2.2 Interactive Presentations with Dynamic Graphical Overlays on Presenter. Instead of manipulating presenters video feeds as foreground and slide contents as background, another type of interactive video presentation uses presenter video feeds as background and seamlessly integrate content as graphical overlays. For example, Chalktalk allows users to create sketches and can be later animated via mouse gestures [40]; RealitySketch introduces a way to bind sketches with physical objects in dynamic and responsive ways [50]. Beyond mouse and pens, researchers have explored ways to use presenter body to interact with graphical overlays, work such as Tutor-Insight have been using MR avatars with auto-generated body language to direct student's attention in classroom settings [51]; Interactive Body-Driven Graphics proposes ways for presenters to interact with the graphical elements in real-time with gestures and postures [46]. Hall et al., combined dynamic charts overlaid on presenter's video feed and allowed presenter's bimanual manipulation for expressive movements in communicating numerical data [19]. In ThingShare, users are able to create digital copies of objects in the video feeds and interact with it with gestures [24]. Researchers have also proposed speech-driven system that augments presentation

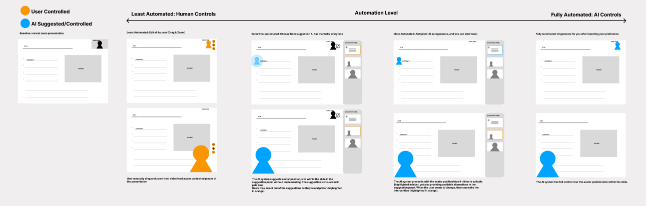

User Controlled
AI Suggested/Controlled
Least Automated: Human Controls
Automation Level
Fully Automated: AI Controls
Baseline: normal zoom presentation
Least Automated: Edit all by user (Drag & Zoom)
Somewhat Automated: Choose from suggestion AI has manually everytime
More Automated: Autopilot (AI autogenerate, and you can intervene)
Fully Automated: AI generate for you after inputting your preference
User manually drag and zoom their video feed avatar on desired places of the presentation
The AI system suggests avatar position/size within the slide in the suggestion panel without implementing. The suggestion is visualized in pale blue
Users may select out of the suggestions as they would prefer (highlighted in orange).
The AI system proceeds with the avatar position/size it thinks is suitable (highlighted in blue), yet also providing available alternatives in the suggestion panel. When the user wants to change, they can make the intervention (highlighted in orange).
The AI system has full control over the avatar position/size within the slide.

Fig. 2. Mock-ups exploring users' decision-making processes and workflows in dynamic presenter video feed generation, based on adaptation level.

with visuals like texts, titles and images that can be interactively prompted, moved, manipulated [32], and with dynamic predefined graphic animations to real-time speech and gestures [7], as well as augmenting presenter integration with 3D holographic objects [49]. Beyond digital presence of content and presenter, dynamic graphical overlays on the presenter has also been explored in the same physical space using holography [28, 29].

While these systems enhance expressive presentation experiences, very few of them adapt with multimodal inputs – such as presenter's speech, hand gestures, slide contents and layouts at the same time. In addition, they lack interaction capabilities specifically with slide content and are more geared towards adding graphical elements (eg. image overlays, animations, text overlays, dynamic data visualisations) on top of presenter video feeds. Since in most existing online presentations and face-to-face presentations, slides play a crucial role in conveying contents, having information only as overlays can be a limitation in information-dense presentations where emphasizing key figures, images, and text is critical.

3 WORKSHOP STUDY

We conducted an expert workshop study to gain insights into user motivations, decision-making processes, workflows, and challenges regarding editing online presentations with the presenter's video feed. The study was organized as two sessions of semi-structured interviews.

3.1 Participants and Procedure

We recruited five participants who have experience in creating online presentation videos where the speaker's video feed is edited in size and/or position alongside their slides. The semi-structured interviews lasted 60-90 minutes and were conducted via Zoom. Participants received a $10 compensation and provided informed consent for audio and video recording.

In the first session (40-50 minutes), we focused on understanding the problem scope and context. We explored participants' motivations for integrating video feeds into their presentations, their preparation processes, preferences for video feed placement and sizing, interactions with slide content, general challenges, and desired features for an ideal tool.

In the second part (20-30 minutes), we gathered feedback on four mock-up designs of automated suggestion tools for presentation video feed editing, each representing a different level of human-AI automation. The least automated version allows users to manually drag and zoom their video avatar. In the semi-automated mock-up, the AI suggests avatar position/size in a suggestion panel without direct placement, allowing users to select their preferred option. The more-automated mock-up depicts user with the top-1 suggested avatar position/size, while allowing users to change the choice. The fully automated version gave the AI full control over the avatar's placement and size. Participants were interviewed about their preferences and comments on these mock-ups as live-editing and post-editing tools, and their desired level of shared control during presentation editing.

3.2 Results and Design Insights

Two researchers organized and analyzed participants' responses with the affinity diagram approach. Our key findings are as follows:

3.2.1 Rehearse, record, and edit presentations. Participants expressed a strong interest in utilizing the proposed mock-ups for rehearsing, recording, and editing presentations. One participant (P4) compared the tool to PowerPoint's designer suggestions, stating that it "gives me some inspiration," highlighting the potential of the mock-ups to spark creative ideas during the preparation process. P4 further suggested an approach where the "1st pass [is] fully automated and then tweak from there," indicating a preference for a system that offers an initial automated draft that can be refined manually. Another participant (P2) emphasized recording video feeds during presentation video editing feels "like a more familiar, more cozy type of presentation" and closely replicates the experience "on the presentation day".

3.2.2 Highlight contents with speech and gestures. Participants emphasized the critical role of highlighting content through both speech and gestures, noting that gestures, in particular, occur naturally during presentations. One participant (P3) remarked that gestures "add some personal aspect" to the presentation, suggesting that using more gestures is especially effective when "the focus is on the speaker." Another participant (P5) expressed interest in incorporating gestures into digital presentations, stating, "if there are ways of deriving how my gestures are in an actual presentation, and if I can get the help of those gestures in my PowerPoint presentation, then I think I would definitely rely on those." They further emphasized the importance of gestures in conveying complex ideas, such as texture, noting that "to show a texture, it's better than to describe a texture." Additionally, P1 reinforced the naturalness of gestures, stating, "Yeah, I think it's more natural. I don't plan. I have the gestures," highlighting that these movements are an instinctive part of their communication style.

3.2.3 Simple all-in-one software workflow. Participants frequently encounter challenges when switching between multiple software tools during the presentation preparation process, underscoring the need for a simplified, all-in-one editing solution to minimize distractions. They emphasized the importance of an integrated tool that could streamline their workflow and reduce the complexity of post-editing tasks. One participant (P1) described the cumbersome process of constantly adjusting both slides and video, noting that "if I change something in the video, then I change the slides again," highlighting the inefficiency of the current setup. P1 also shared their experience of uploading slides to Canva and then combining them with video, describing it as "a mess...It was really a lot of work," which further illustrates the frustration with the multi-software workflow. Another participant (P2) expressed a desire for a more cohesive experience, suggesting a feature that allows for a "live preview of the overlay," where the software could "sync up...the camera feed, the slide, [and] the audio" in real-time. P2 envisioned a tool that records all elements on separate tracks within a single platform, similar to Adobe Premiere, allowing for easier and more efficient editing later on. This reflects a strong preference for a unified tool that simplifies the entire presentation creation process.

3.2.4 Optimize presenter video transitions based on presentation content. The placement and timing of presenter video feed transitions are influenced by various factors, including the content, layout, and purpose of the slides. Participants highlighted the importance of adapting the video feed to complement the presentation's flow. One participant (P5) described a strategic approach where the presenter's face is prominently displayed during the introduction but shifts to a corner during sections like "Related work" to allow the slide content to take center stage. They suggested that the video feed's position could change "when going from one concept to the other, or introducing a different point," emphasizing the need for flexibility based on the presentation's structure. Similarly, another participant (P3) recommended adjusting the presenter's position depending on the slide content's dynamism. For dynamic content that requires more space, they suggested that "content should be the focus, so move myself to the corner," while the size of the video feed should adhere to the "one-third rule," ensuring that the presenter's video feed does not overshadow the main concept. This approach reflects a nuanced understanding of how to maintain balance between the presenter and the content, enhancing the overall effectiveness of the presentation.

4 INTERACTION PATTERN EXPLORATION STUDY (DATA COLLECTION STUDY)

To investigate preferences for video feed placement relative to presentation content, slide structure, layout, and the presenter's facial expressions, gaze, and gestures, we conducted a data collection study.

4.1 Study Design

We developed three categories of mock-up slide decks to represent diverse presentation types: personal stories (presenter-centric), tutorials (information-centric), and professional presentations (both informational and presenter-centric). Additionally, we created an online presentation software prototype that enables dynamic adjustment of video feed placements.

The study was conducted remotely via Zoom. Participants were asked to imagine preparing the presentations without time constraints, aiming to curate their video feeds and content to achieve the best presentation experience. For each slide category, participants were randomly assigned one of two options. After familiarizing themselves with the slide content, they were instructed to position their video feed optimally for effective presentation. They could place the video feed anywhere on the slides, remove it, or use a full video feed. Once satisfied with the placement, they recorded their presentation using our prototype, allowing for dynamic adjustments to the video feed during the presentation. The position of the video feed, along with the participant's facial expressions, gaze, and gestures, were recorded for each slide. For detailed procedures and instructions, see **??**.

4.2 Mock-up Slide Decks

The three categories of mock-up slides encompass a diverse range of presentation types with varying levels of presenter and information centricity. The decks on topics like hometown, childhood, and gardening are more presenter-centric, focusing on personal stories. In contrast, slides on time management and Fusion360 tutorials are information-centric, emphasizing clear "how-to" guidance. Professional presentations, such as conference talks and product launches, strike a balance between information and presenter-centric content, aiming to highlight both the presented artifacts and the presenter's role in the community.

Each slide deck contains seven slides, starting with an introduction and ending with a conclusion. The decks include one title slide, two text-only slides, one full-figure slide, one full video/GIF slide, and two slides combining text and figures. Slides with the same composition share consistent formatting; for example, all text-and-figure slides use the same layout, as do the text-only slides.

4.3 Prototype

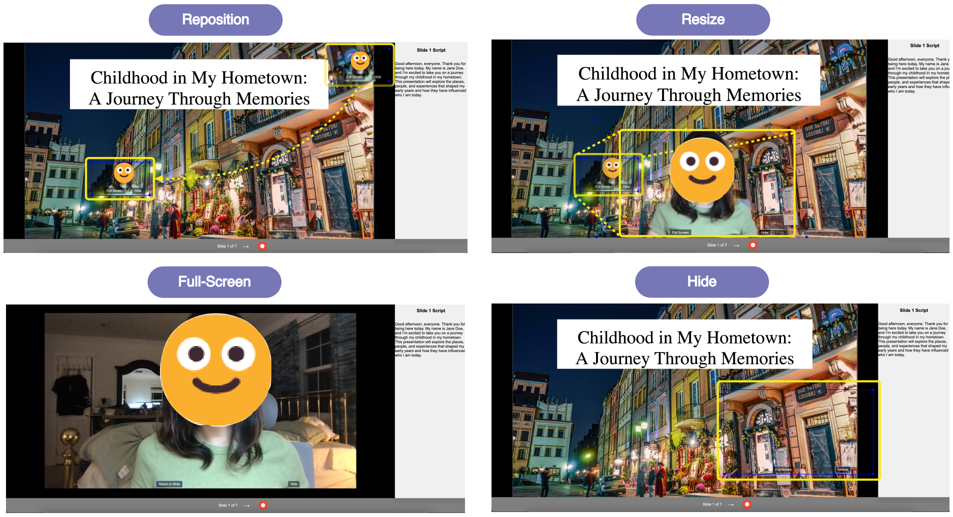
We developed a web application designed to facilitate online PowerPoint presentations, similar to video conferencing platforms such as Zoom's "Slide as Virtual Background" feature, where presenters can manually adjust their video feed sizes and placements to integrate themselves into slides. The application uses the PPTX2HTML library [41] to parse and render PPTX files for the user, enabling seamless navigation through PowerPoint slides while maintaining an integrated webcam video feed. The application accepts two inputs: a slide deck and a script (in .txt format), allowing users to align their presentation content with their spoken narrative.

To enhance the user experience, we integrated the MediaPipe's on-device models [34] to remove the background from the webcam video feed. This background removal allows the presenter's video to be integrated into the slide with minimal obtrusiveness, similar to Zoom's "Slide as Virtual Background" feature, providing a cleaner and more professional appearance.

The application offers several interactive features for managing the webcam video, including (See Figure 3):

- **Repositioning**: Users can click and drag the webcam video to any desired location on the screen.
- **Resizing**: The video can be resized by dragging the corners of the webcam feed.
- **Fullscreen Mode**: Users can expand the webcam video to fullscreen.
- **Hiding Mode**: The webcam video can be hidden or disabled as needed.

A control panel at the bottom of the interface provides options to navigate slides and manage presentation recordings. The application includes record, pause, and stop buttons, enabling users to record their webcam video along with the web application's screen. The



Slide 1 Script

Good afternoon, everyone. Thank you for being here today. My name is Jane Doe, and I'm excited to take you on a journey through my childhood in my hometown. This presentation will explore the places, people, and experiences that shaped my early years and how they have influenced who I am today.

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) hiding, and (d) turning into full-screen mode. The presenter can edit their video and hit the record button on the bottom to record a presentation video.

pause function allows users to reposition and resize their video during the presentation, ensuring that any adjustments are excluded from the final recording.

During recording sessions, we logged various parameters of the webcam video, such as position, size, and mode (*e.g.*, fullscreen, hidden), into a video file for subsequent analysis. This logging enables detailed examination of user interaction patterns with the webcam integrated with presentations.

4.4 Participants

We collected data and conducted semi-structured interviews with 20 participants, 12 self-identified as female, and 8 as male. The mean age is 26.6, SD = 3.23. We collected their self-rated proficiency with PowerPoint/Keynotes (1=not proficient at all, 5=very proficient). The mean self-rated proficiency with PowerPoint/Keynotes is 4.05, SD = 0.60. The mean self-rated presentation skills (1=very amateur, 5=very professional) is 3.73, SD = 0.79. Among them, 16 agreed to release their full video feed as part of the dataset.

4.5 Identified Patterns

Two researchers took field notes during the data collection study and gathered rationales for editing decisions through post-hoc interviews. A codebook was then developed to analyze video recordings of the presentations, enabling the identification of common patterns in how users edited and integrated their video feeds into the slides.

4.5.1 Hidden mode are not used as often. We observed that hidden mode is rarely used, mainly when participants encounter full-screen videos or images they find overwhelming. Occasionally, participants hide their video feeds when the text is dense, preferring the audience to focus on the content instead.

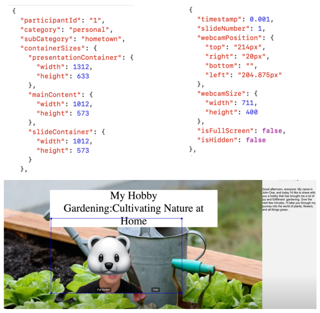
4.5.2 Repositioning primarily happens to avoid occlusion, with frequency dependent on user preferences. Participants usually reposition video feeds to blank spaces to avoid occlusion, ensuring that content remains visible and clear. An exception occurs with title slides featuring picture backgrounds; repositioning is less necessary if the image lacks dense content, often resulting in video feeds being moved to the corners. The size of the repositioned feeds varies based on available blank space, and the frequency of repositioning depends on the participant's desired presentation style. For a more interactive and personal approach, repositioning occurs frequently. In contrast, a formal presentation style involves less repositioning, emphasizing consistency and reducing distractions.

4.5.3 Resizing mainly happens for filling out blank spaces, with a minor category effect. Participants generally resize video feeds for aesthetic and layout purposes, aiming to fill blank spaces in a visually balanced manner. They preferred larger video feeds for presenter-centric approaches and smaller feeds for information-centric presentations.

4.5.4 Full-screen video feeds are most common at the start and end of presentations. Full-screen mode is typically used at the beginning and end of presentations, during introductions or concluding remarks. Presenters often feel that slide content is less critical at these times and prefer to create a lasting impression and engage more directly with the audience.

4.6 Dataset

We are releasing our dataset, ***VTalk-68 Dataset*** Figure 4, 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



```
{
  "participantId": "1",
  "category": "personal",
  "subCategory": "hometown",
  "containerSizes": {
    "presentationContainer": {
      "width": 1312,
      "height": 633
    },
    "mainContent": {
      "width": 1012,
      "height": 573
    },
    "slideContainer": {
      "width": 1012,
      "height": 573
    }
  },
```

```
{
  "timestamp": 0.001,
  "slideNumber": 1,
  "webcamPosition": {
    "top": "214px",
    "right": "20px",
    "bottom": "",
    "left": "204.875px"
  },
  "webcamSize": {
    "width": 711,
    "height": 400
  },
  "isFullScreen": false,
  "isHidden": false
},
```

My Hobby
Gardening:Cultivating Nature at
Home

Good afternoon, everyone. My name is John Doe, and today I'd like to share with you a hobby that has brought me a lot of joy and fulfillment: gardening. Over the next few minutes, I'll take you through my journey into the world of plants, flowers, and all things green.

Fig. 4. VTalk-68 dataset, that collects information on user video feed placements across various slide layouts and contents

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.

Our goal in sharing this dataset is to encourage 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.*

5 TALK DIRECTOR

Based on the insights gained from our data collection study, we developed ***TalkDirector***, an intelligent online presentation interface that dynamically integrates multimodal inputs, including slide content, script, spoken words, and presenter gestures. This system enables real-time adjustments to the position, size, and placement of the presenter's video feed within the slide, enhancing both content relevance and visual engagement.

The workflow of TalkDirector is divided into two stages. The first stage occurs when the user uploads their presentation slides and script to the system, at which point the system processes and analyzes the slide layout and aligns the script with the visual elements. The second stage occurs in real-time during the presentation, where the system uses live transcription and gesture recognition to dynamically adjust the presenter's video feed. Figure 5 provides an overview of the system architecture and its core components.

In the following sections, we will detail each step of the process, from slide analysis and script alignment to live transcription and real-time video positioning.

5.1 Slide Layout Analysis

A key observation from our data collection study revealed that participants frequently repositioned and resized their webcam video to avoid occlusion with important slide content. To address this behavior, we prioritized an accurate understanding of the slide layout as a foundation for dynamically adjusting the presenter's video feed. Our process begins by analyzing the layout of each slide to determine optimal positioning and scaling of the video feed without obstructing key slide information.

We first process the slide deck by converting each slide from a PDF format into individual PNG images for visual analysis. The slide layout is then analyzed in a two-stage approach, utilizing both Optical Character Recognition (OCR) and a large language model with vision capabilities (GPT-4o).

5.1.1 Initial Optical Character Recognition. Our first step involves using Tesseract [48], an open-source OCR engine, to extract text and provide bounding boxes for textual elements. Tesseract is particularly well-suited for handling the wide variety of fonts and text structures found in presentation slides. This stage offers two crucial outputs:

(1) **Text Extraction:** Identifies and extracts textual content, enabling further processing in later stages.
(2) **Bounding Box Creation:** Provides preliminary spatial mapping of text blocks, forming the basis for subsequent refinement.

While Tesseract is effective at extracting text, its bounding box accuracy in complex layouts can be limited, especially when graphical elements such as images or charts are involved. Therefore, a second refinement stage is employed to improve precision.

5.1.2 Deep Learning-Enhanced Layout Analysis. To further refine the bounding box locations and gain a more holistic understanding of the slide layout, we leverage GPT-4o with multi-modal inferencing capabilities [39]: we instruct GPT-4o to refine the text's bounding boxes from Tesseract based on the slide, while simultaneously identifying and categorizing different slide components.

GPT-4o analyzes the slide layout by tagging elements with semantic labels based on their roles within the presentation. Titles are labeled as `<title-text>`, while content text is labeled according to its structure, such as `<content-text1>` for bullet points or sentences. Similarly, images and figures are identified and labeled, for example, as `<image-figure1>`. This dual-stage process of OCR-driven text extraction followed by deep learning-based refinement allows the system to:

- **Enhance Bounding Box Precision:** GPT-4o corrects and refines the bounding boxes based on the Tesseract output, ensuring better spatial accuracy.
- **Label Slide Components:** The model categorizes each slide element as text, image, figure, etc., tagging it accordingly to facilitate video feed positioning.
- **Understand Hierarchical Structure:** GPT-4o detects and preserves the slide's structural hierarchy, such as distinguishing between titles, subtitles, and content blocks.

This detailed slide layout analysis, combining OCR and AI-driven refinement, ensures a robust foundation for dynamically integrating the presenter's video feed without occluding important content. In subsequent sections, we describe how this layout understanding is used to guide video feed placement in real-time presentations.

5.2 Script Labeling with Layout Components

With a comprehensive understanding of the slide layout, our next challenge was to align the presenter's script with the visual elements

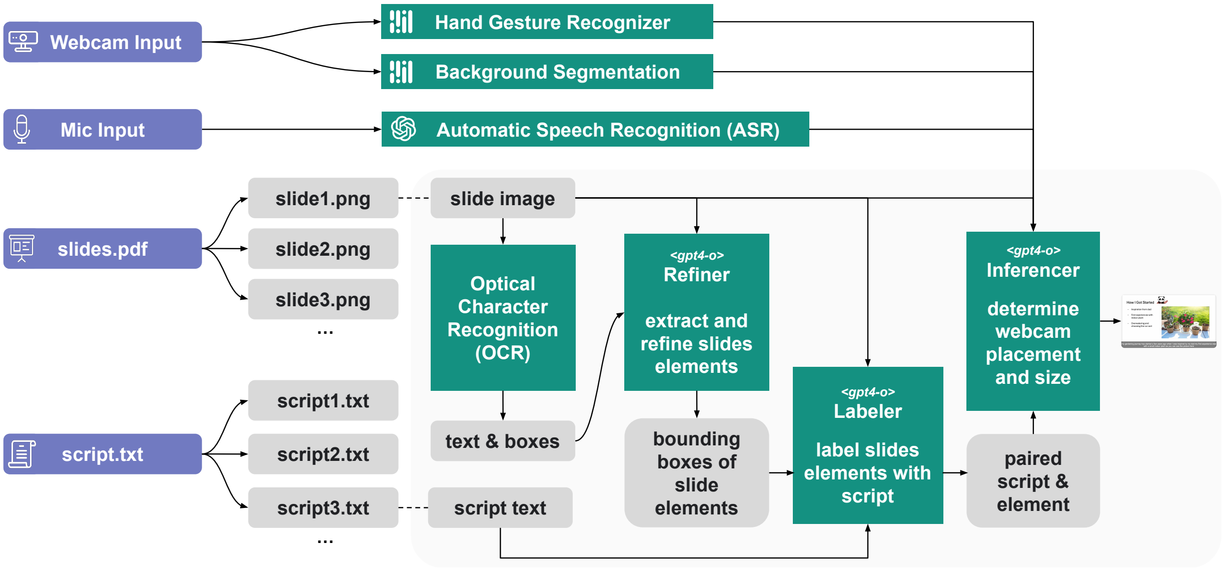


Fig. 5. TalkDirector's multi-modal layout inference pipeline. We first process inputs of slides presentation and scripts with TesseractOCR [48] and Large Generative Multimodal Modal (GPT4-o) with a refiner and a labeler. This pipeline effectively reduces hallucinations of GPT4-o and generate paired script 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 [38], 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.

on each slide. This alignment is critical for enabling features such as real-time content-aware video positioning during the presentation.

5.2.1 Script Parsing and Segmentation. The alignment process begins by parsing the uploaded script to identify its correspondence with individual slides. Our custom parsing algorithm detects both explicit slide references (e.g., "[Slide 1]") and implicit structural cues to segment the script accordingly. This segmentation ensures that each portion of the script is correctly mapped to its respective slide, preserving the logical flow between the speaker's content and the visual elements that accompany it. By pre-processing the script in this way, we maintain a foundation for real-time synchronization during the presentation.

5.2.2 Intelligent Script-Layout Matching. Once the script is segmented, the system matches specific sections of the script to the corresponding visual components identified during the slide layout analysis. This step ensures that textual and visual elements—such as bullet points, titles, and figures—are accurately linked to the script, which forms the foundation for real-time interactions during the presentation.

Using the structured layout information from the previous slide analysis step, we employ GPT-4o [39] to perform context-aware matching between the script and the slide components. The model aligns script segments with relevant layout elements, such as titles (`<title-text>`), bullet points (`<content-text1>`), and figures (`<figure-image1>`). Importantly, when the script refers to a visual element like an image or figure, the system leverages the labels and brief descriptions generated in the layout analysis to ensure precise alignment. For instance, if the script describes a figure, the system can match this reference with the labeled and described `<figure-image1>` element from the layout analysis.

Additionally, during the analysis, we identify whether the slide is the first or last slide in the presentation. If it is the first slide and the presenter is introducing themselves, we add a `<full>` tag to indicate that the presenter's video should be in full-screen mode. Similarly, on the last slide, if the presenter is thanking the audience, the system also applies the `<full>` tag. Moreover, when an image figure covers the entire slide, the system applies a `<hide>` tag, allowing the audience to focus solely on the figure without distraction from the presenter's video. These tags are determined based on patterns identified from our data collection study.

5.2.3 Post-Processing and Validation. To maintain accuracy, the system undergoes a post-processing validation step to confirm the integrity of the script-layout alignment. During this phase, we verify that each label applied to the script corresponds correctly to the visual elements identified in the slide layout analysis. This ensures that the alignment remains consistent with the presenter's original content and the intended structure of the presentation. Additionally, this validation process confirms that the hierarchical structure of the slide content is respected, allowing for more effective real-time synchronization in later stages.

5.3 Live Transcription and Script Matching

Following the pre-processing steps of script and layout alignment, we transition to the real-time aspects of our system, where the presenter's speech is transcribed and matched with the previously segmented script and layout components. One of the key behaviors we observed during our formative study was that presenters often wanted to position themselves near the content they were discussing—whether standing next to bullet points to simulate a 'pointer' effect or aligning themselves with images they were referencing. To support this behavior, the system must identify what the presenter is talking about and dynamically adjust the position and size of the video feed accordingly.

To achieve this, we begin by recognizing the presenter's spoken words in real-time. The system transcribes the speech and matches it with the pre-aligned script from the earlier steps, determining which slide component is currently being discussed. This matching process ensures that the presenter's video feed is positioned in the most contextually relevant location on the slide.

The speech recognition process employs a dual approach that balances immediacy with accuracy. For low-latency feedback, we use the Web Speech API, which provides fast, interim transcriptions. This allows the system to immediately begin processing the spoken words and provide visual feedback to both the presenter and the audience. However, to improve accuracy, particularly in acoustically challenging environments, we also integrate OpenAI's Whisper API [38]. While this higher accuracy model processes audio chunks at regular intervals, it ensures that the final transcription is refined and robust against noise or variations in speech.

As the system receives transcriptions, it continuously matches them against the script. We designed a sliding window mechanism that compares recent transcriptions with the labeled script, allowing the system to tolerate variations in speech, rephrasings, or minor recognition errors. By keeping a buffer of the most recent words, the system performs fuzzy matching against the pre-labeled script to locate the closest corresponding segment. Once a match is found, the system updates the script's current position and retrieves the associated layout component label from the earlier script-layout matching phase.

The identification of what content the presenter is referring to—whether it is a bullet point, title, or figure—enables the system to prepare for further interaction with the visual components of the slide.

5.4 Content-Aware Video Positioning

The final component of our system involves dynamically adjusting the presenter's video feed in relation to the slide content. Based on insights from our data collection study, we observed that presenters often wanted to position themselves near specific content they were discussing, such as standing next to a bullet point or an image. However, this does not need to happen every time the presenter speaks. To address this, we enable content-aware video repositioning that is gesture-triggered, allowing presenters to emphasize specific points when necessary. See Figure Figure 6, where the system dynamically adjusts the presenter's video based on both the slide content and the presenter's gestures, ensuring optimal placement for enhanced engagement.

5.4.1 Default Positioning Based on Layout Analysis. Our system provides a default positioning of the presenter's video feed based on the blank spaces identified through the slide layout analysis. Using the bounding boxes from the layout analysis, we determine areas of minimal content where the video feed can be placed without obstructing key information. The positioning process is informed by findings from our data collection study, where participants showed a strong preference for positioning their video in the bottom-right corner of the slide. As a result, the system prioritizes this area for video placement.

If the bottom-right corner lacks sufficient space, the system checks the other three corners of the slide, selecting the one with the largest available space for positioning. This ensures that the presenter's video does not interfere with critical slide content while maintaining a consistent, non-intrusive placement.

Additionally, the system dynamically resizes the video feed based on the available space in the chosen corner, mimicking the behavior observed in the study where users resized their webcams to fit the available space. To prevent extreme size adjustments, we set a minimum and maximum size limit for the video feed.

This default positioning and resizing mechanism serves as the initial setup for the video feed, providing an unobtrusive placement that aligns with user preferences while allowing for further adjustments through gesture-triggered interactions.

5.4.2 Pointing Gesture Recognition. To give presenters control over the positioning of their video feed, we implemented a gesture-triggered interaction. We chose the pointing gesture—where the presenter extends their index finger—based on observations from the formative study. Presenters commonly used this gesture to refer to content while positioning themselves next to it, making it a natural choice for controlling the video feed.

When the system detects a pointing gesture using MediaPipe Hands [], it interprets this action as an intention to reposition the video feed. This gesture-triggered control allows the presenter to adjust their video feed seamlessly without interrupting their speech or using external input devices. By making the interaction intuitive and accessible, the system accommodates a wide range of presenters, including those who may have difficulty with traditional input devices.

5.4.3 Real-time Content-Aware Repositioning and Resizing. When a pointing gesture is recognized, the system doesn't simply move the video feed to the indicated location. Instead, it follows a structured decision-making process to ensure optimal positioning based on the content the presenter is discussing. If the user is reading a part of the script tagged as `<full>` or `<hide>`, the system prioritizes these tags by either turning the presenter's video into full-screen mode or hiding it, respectively.

When the script does not include `<full>` or `<hide>` tags, and the user speaks while a gesture is detected, the system identifies the label of the layout component corresponding to the part of the script being discussed. This label directs the system to reposition and resize the video feed next to the appropriate slide component, such as a bullet point or figure. The system does not consider the direction the user is pointing toward; instead, it directly uses the tagged content in the script to align the video with the related visual element on the slide.

In addition to repositioning, the system adjusts the size of the video feed dynamically based on the available space around the identified slide component. This resizing mechanism mirrors user

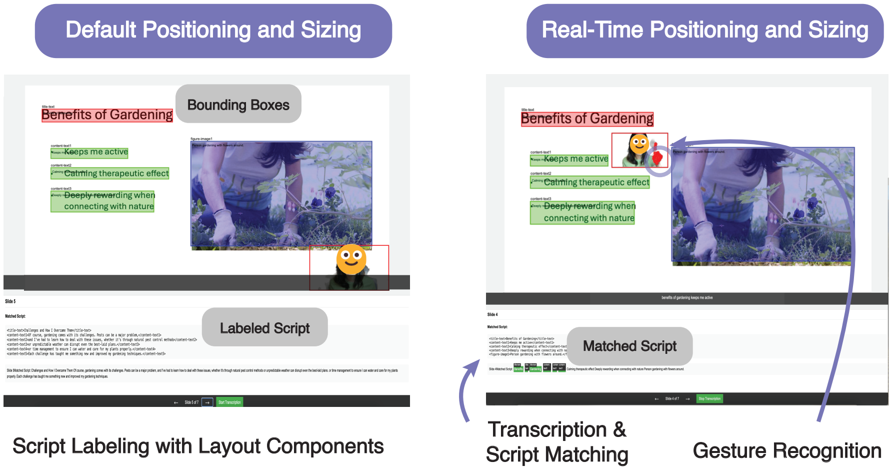

Default Positioning and Sizing
Real-Time Positioning and Sizing
Bounding Boxes
Benefits of Gardening
Keeps me active
Calming therapeutic effect
Deeply rewarding when connecting with nature
Labeled Script
Matched Script
Script Labeling with Layout Components
Transcription & Script Matching
Gesture Recognition

Fig. 6. Visualization of *TalkDirector*'s context-aware video positioning: default positioning and sizing (left) and real-time positioning and sizing (right). 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.

behavior observed during the data collection study, where presenters adjusted their webcam size based on the available space. By following this approach, the presenter's video feed is positioned and sized optimally, while maintaining a smooth visual transition in real-time.

By combining these adaptive repositioning and resizing mechanisms, our system dynamically adjusts the presenter's video feed to align with both the content and the spatial constraints of each slide. This approach ensures a seamless integration of video and slide elements, enhancing the overall presentation experience. As illustrated in Figure 7, the system responds to various slide layouts, presenter gestures, and speech-content matches, demonstrating its flexibility in real-time presentation environments.

6 EVALUATION STUDY

To assess the effectiveness and usability of the TalkDirector system, we conducted a comparison study. We compare our system that dynamically generate presenter video feeds based on slide contents and gestures with a baseline prototype that behaves the same like Zoom's slide as virtual background, with the additional capability to go full screen or hide completely with the presenter video feed. The participants are tasked with preparing and presenting slide decks provided using these two systems. We evaluate both quantitatively and qualitatively on user's subjective experience using these systems. Specifically, participants are asked to fill out a questionnaire on the experience of presenting using these two systems based on self-perceived engagingness of delivery, worthiness of effort, expressiveness, flow, integration with content, Systems Usability Scale (SUS) [6] and the NASA-TLX [20]. We adopted the 0-100 rating scale for the NASA-TLX, and used the 7-point Likert scale for the rest of the ratings. The detailed questionnaire is in the following section. While participants are performing the tasks, we recorded their video data and analyzed their body language as an objective measures of presentation expressiveness. Further, we follow-up with a semi-structured interview aimed to understand participants' experience. The items of the questionnaires used in the evaluation study can be found in Appendix A.

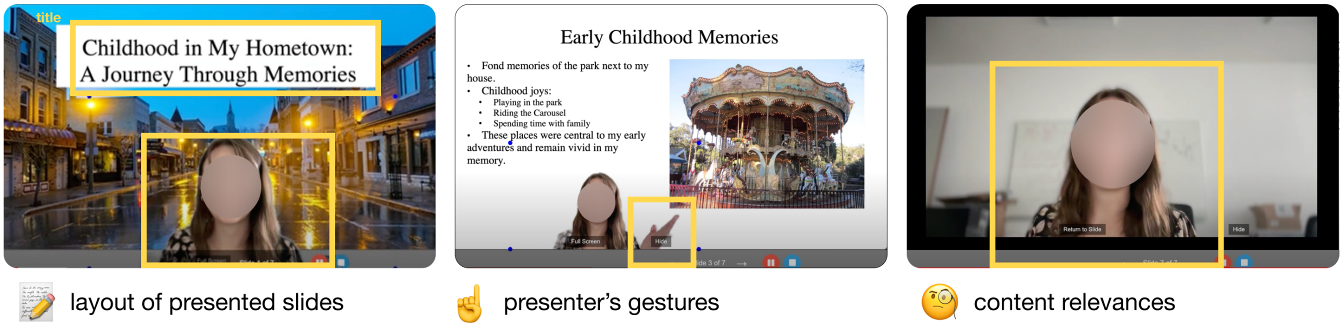
6.1 Participants

A total of 12 participants (7 female, 5 male) were recruited from a university sample, with an average age of 28.42 years (SD = 2.97, range = 23 − 32). Participants were asked to rate their experience with giving an online presentation on a 7-point Likert scale (1 = no experience, 7 = very experienced), with an average score of M = 4.92, (SD = 1.40). They were also asked about their experience editing their webcam video within the slides during an online presentation (M = 1.75, SD = 1.60) and in post-process editing after a presentation (M = 2.42, SD = 1.53), both of which indicated low levels of experience. Participants were compensated with a $10 e-gift card for their time.

6.2 Procedure

After obtaining informed consent, participants were asked to fill out a demographic survey and were then provided with an overview of the experiment and the procedure. Participants were introduced to the task of becoming presenters for online talks, where their goal was to act as the "director" of their presentation, aiming for an engaging, TED Talk-like experience. Each participant was required to present two pre-prepared slide decks, one on the topic of their hometown and the other on their hobby. Although the slide decks and scripts were pre-prepared, participants were asked to embody the role as if they were actually presenting their own hometown and hobby to the audience.

Participants were randomly assigned to one of the two systems first (TalkDirector or baseline), and the order of the system and topic was counter-balanced across participants to control for any order effects. Before beginning the presentations, participants were



Childhood in My Hometown:
A Journey Through Memories

Early Childhood Memories

- Fond memories of the park next to my house.
- Childhood joys:
  - Playing in the park
  - Riding the Carousel
  - Spending time with family
- These places were central to my early adventures and remain vivid in my memory.

Fig. 7. Example adaptive presenter video placement using generation using our TalkDirector system across various slide layouts and contents, user gestures, and speech-content matches. In the leftmost image, TalkDirector generates video feeds by recognizing the layout and content of the slide. It places the presenter's video in the blank space, in a medium to large size, since it identifies this slide as the title and introduction slide. In the middle image, TalkDirector generates the presenter video feed based on gestures and speech-content matches, as the presenter is talking about the carousel and using pointing gestures. In the rightmost image, TalkDirector goes full screen with the presenter video feed as it understands that the presenter is wrapping up the presentation and there is less content in speech matching the slides.

given a demo of each system and allowed training time to familiarize themselves with the interfaces, including the record, pause, and stop buttons, which enabled them to control the flow of their presentations.

Both systems included live transcription subtitles to help participants track when the system was processing their spoken words, ensuring minimal latency and consistency between the two interfaces. After the demo and training, participants were asked to deliver two presentations—one using the baseline system and one using the TalkDirector system. After each presentation, participants filled out a questionnaire rating their experience. Participants were allowed to adjust their ratings for the first system after completing both presentations to ensure their final ratings reflected the full experience.

Following the presentations, a post-hoc interview was conducted where participants were asked to explain the rationale behind their ratings, indicate their preference between the two systems, and discuss potential application scenarios for the systems.

6.3 Results

6.3.1 Video Data Analysis. We analyzed participants' video data to understand their preparation and presentation behavior in both conditions (See Figure 8a and 8b). The following results were obtained:

First, we measured the preparation time, which included tasks such as reading the script in advance, planning webcam placement, and determining when to use pointing gestures before the presentation recording started. TalkDirector significantly reduced preparation time compared to the baseline condition, with a mean preparation time of $M = 58.75$, $SD = 51.88$ for TalkDirector and $M = 137.50$, $SD = 53.73$ for the baseline ($t(11) = 3.65$, $p = 0.0014$).

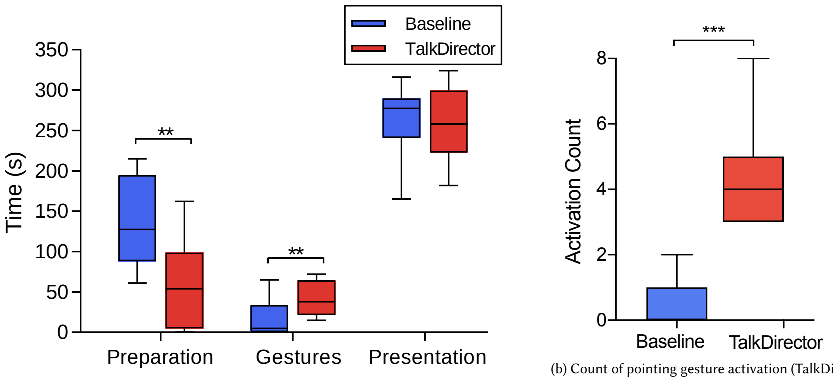
We also analyzed the actual presentation time, meaning the recorded duration of the presentation video. There was no significant difference between the two conditions, with TalkDirector having a mean presentation time of $M = 258.67$, $SD = 45.29$, and the baseline having a mean of $M = 261.92$, $SD = 47.28$ ($t(11) = 0.17$, $p = 0.87$).

Next, we measured the time that participants' hands were visible in the video to estimate the duration during which gestures were made. To do this, we post-processed the participants' webcam videos using MediaPipe Hands to detect when hands were visible in the webcam video. We then manually excluded times when hands were visible but no gestures were made, such as when participants were holding a microphone. TalkDirector resulted in significantly more hand-visible time compared to the baseline, with a mean time of $M = 42.58$, $SD = 21.87$ for TalkDirector and $M = 16.00$, $SD = 22.66$ for the baseline ($t(11) = -2.92$, $p = 0.0079$), implying participants made more gestures with TalkDirector.

We further analyzed how often the video moved within the slide, such as repositioning or resizing during the presentation. For the baseline condition, this happened when participants paused the recording to manually reposition or resize the webcam video. In contrast, for TalkDirector, this occurred when participants made pointing gestures, automatically repositioning the video. TalkDirector led to significantly more instances of video repositioning, with a mean of $M = 4.33$, $SD = 1.67$, compared to the baseline, which had a mean of $M = 0.50$, $SD = 0.67$ ($t(11) = -7.37$, $p < 0.0001$).

To ensure that the results were not affected by the topic (hobby or hometown) participants were asked to present, we conducted a two-way ANOVA. The results confirmed that the significant effects observed in preparation time, hand-visible time, and pointing gesture activation were driven by the method (TalkDirector vs. baseline) and not by the topic. Specifically, there was no significant main effect 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 pointing gesture activation ($F(1, 20) = 0.41$, $p = 0.53$). Additionally, there were no significant interactions between method and topic in any of these measures.

6.3.2 User Experience. Participants were asked custom questions regarding their experience with engagement, value, expressiveness, flow, and integration (See Figure 9). The results are as follows:



(a) Time spent on talk preparation, presentation, and using gestures.

(b) Count of pointing gesture activation (TalkDirector) and movement within slide(Baseline)

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.

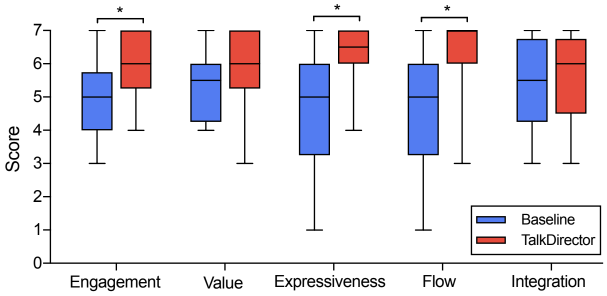


Fig. 9. User preferences between Baseline and TalkDirector conditions along five attributes: engagement, value. The statistic significance is annotated with * (representing p<.05).

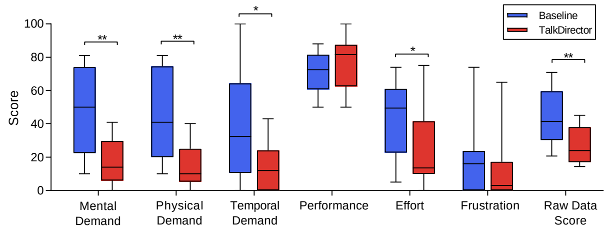


Fig. 10. Raw-TLX results. The statistic significance is annotated with *, **, or *** (representing p<.05, p<.01, and p<.001, respectively).

For engagement, TalkDirector resulted in significantly higher engagement compared to the baseline condition, with a mean score of $M = 6.08$, $SD = 1.00$ for TalkDirector and $M = 5.00$, $SD = 1.21$ for the baseline ($t(11) = -2.40$, $p = 0.0253$).

In terms of perceived value, TalkDirector had a slightly higher mean score ($M = 5.92$, $SD = 1.31$) compared to the baseline ($M = 5.42$, $SD = 1.08$), but no statistically significant difference was detected ($t(11) = -1.02$, $p = 0.32$).

For expressiveness, participants rated TalkDirector significantly higher ($M = 6.25$, $SD = 0.97$) than the baseline condition ($M = 4.67$, $SD = 1.78$) ($t(11) = -2.71$, $p = 0.0127$).

Similarly, for flow, TalkDirector had significantly higher ratings ($M = 6.33$, $SD = 1.15$) compared to the baseline ($M = 4.67$, $SD = 1.83$) ($t(11) = -2.67$, $p = 0.0139$).

In terms of integration, TalkDirector had a slightly higher mean score ($M = 5.67$, $SD = 1.30$) compared to the baseline ($M = 5.33$, $SD = 1.44$), but no statistically significant difference was detected ($t(11) = -0.60$, $p = 0.56$).

6.3.3 Cognitive Load. The NASA-TLX scores for the two conditions were analyzed using paired-samples t-tests (See Figure 10). For mental demand, TalkDirector required significantly less mental effort compared to the baseline condition, with a mean score of $M = 17.50$, $SD = 13.71$ for TalkDirector and $M = 48.00$, $SD = 25.36$ for the baseline ($t(11) = 3.67$, $p = 0.0014$).

Similarly, for physical demand, TalkDirector required significantly less physical effort, with a mean score of $M = 14.75$, $SD = 12.55$ for TalkDirector and $M = 45.25$, $SD = 26.61$ for the baseline ($t(11) = 3.59$, $p = 0.0016$).

For temporal demand, TalkDirector also showed a significant reduction, with mean scores of $M = 13.33$, $SD = 14.49$ for TalkDirector and $M = 37.08$, $SD = 31.56$ for the baseline ($t(11) = 2.37$, $p = 0.0270$).

In terms of performance, TalkDirector had a higher mean score ($M = 77.50$, $SD = 15.58$) than the baseline ($M = 70.00$, $SD = 12.93$), but no statistically significant difference was observed ($t(11) = -1.28$, $p = 0.21$).

For effort, participants reported significantly lower effort 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 levels were lower in TalkDirector ($M = 11.00$, $SD = 18.62$) compared to the baseline ($M = 18.08$, $SD = 20.69$), but the difference was not statistically significant ($t(11) = 0.88$, $p = 0.39$).

Finally, for the raw total average, TalkDirector resulted in a significantly lower overall score ($M = 26.13$, $SD = 10.90$) compared to the baseline ($M = 43.74$, $SD = 16.46$) ($t(11) = 3.09$, $p = 0.0053$).

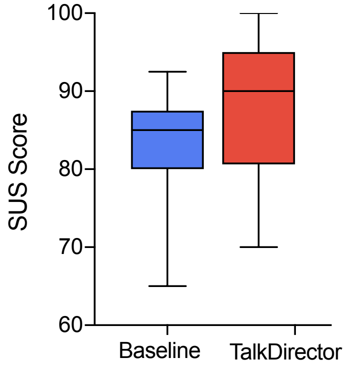


Fig. 11. SUS scores between Baseline and TalkDirector. No significant difference was found.

6.3.4 System Usability. The System Usability Scale (SUS) scores for the two conditions were analyzed using a paired-samples t-test (See Figure 11). 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 ($t(11) = -1.52$, $p = 0.14$).

6.3.5 Preference. After completing the trials with both systems, participants were asked to indicate their overall preference. A total of 10 out of 12 participants (83.3%) preferred the TalkDirector system, while 2 participants (16.7%) preferred the baseline system. This suggests a strong preference for the TalkDirector system over the baseline.

7 DISCUSSION

7.1 Sense of Agency and Convenience Tradeoff

Our findings reveal a nuanced tradeoff between the convenience provided by TalkDirector and the sense of agency users experience when controlling their presented content. On one hand, participants appreciated the convenience and reduction in cognitive load, particularly in situations where they had less time to prepare. This is evident in interviews and data collected on cognitive load and preparation time. For example, Participant P12 noted, *"I liked the second interface because I didn't have to think so much about it, especially the second time... I didn't have to spend much effort."* This ease of use was particularly valued in less formal or impromptu presentations, where users were more focused on delivering content than on precise control over presentation elements.

Conversely, for some participants, the sense of agency and control offered by manual interfaces was a priority. Participants who favored the baseline method expressed a preference for the freedom to adjust every aspect of their presentation. As P5 described, *"I preferred the manual interface because I like to have more sense of control. I am very particular about the way I present things, and I want to be sure the presentation is as I intended."* This sentiment was echoed by others who valued the ability to fine-tune their presentations to match their personal style, particularly for more important, high-stakes presentations. P4 captured this sentiment, saying, *"I have my own way of doing my presentations, and I don't really trust AI to take over and take the wheel for me."*

This dynamic highlights the broader debate between "adaptive" versus "adaptable" interfaces. While adaptive systems, like TalkDirector, aim to reduce cognitive load by making decisions for the user, the challenge lies in aligning AI predictions and suggestions with user preferences. As P3 noted, *"I like that it's less preparation, but I could also see that if you're a person who has to be in control of everything, you might prefer the first method."* This raises important questions about how well AI-driven interfaces can predict user preferences and the potential for systems to become more adaptable over time, learning from user behavior and adjusting to individual styles.

Finally, participants also recognized the value of maintaining flow during their presentations, particularly when they had limited preparation time. Custom flow questions in our evaluation captured a strong preference for uninterrupted presentations. P2 and P10 both remarked that *"this will be more appropriate for presentations where I don't have much time to prepare."* This suggests that the convenience of reduced cognitive load may outweigh the need for granular control in situations where preparation time is constrained, but further investigation is needed to fully understand how different presentation contexts influence this tradeoff.

7.2 Integrating Speaker Video in Slides

For many participants, the experience of integrating their video directly into the slides was a novel and eye-opening one. Most had not considered this approach before, and it encouraged them to rethink how they engage with their audience in virtual settings. This feature, which allowed presenters to embed themselves into the slide content, opened new possibilities for expressiveness and engagement.

Several participants noted that they had not previously thought about how incorporating their video into slides could influence audience attention and interaction. P3 observed, *"I think one of the reasons students don't really pay attention during online lectures is because there's no human interaction... if you have someone on the screen actually moving and talking, it would help them focus a lot more."* This perspective highlights a key challenge of virtual presentations—maintaining audience focus. By having the presenter dynamically integrated into the slide content, participants saw the potential to combat disengagement, particularly in educational or remote settings where attention can wane.

Others,(P4, P10), reflected on their initial concerns about distraction, only to find that the system provided a smooth and non-intrusive experience. P10 stated*"It is not distracting, or taking away the attention as I thought it would... I can just make a quick gesture to move myself if I want to."* This realization highlighted the system's ability to balance presence and content, allowing users to control their video feed without disrupting the flow of the presentation. For P10 and others, the ability to reposition the video provided a sense of control and adaptability, fostering a more dynamic presentation environment.

Participants also highlighted the impact on expressiveness, with P11 sharing how the integration of video helped them convey excitement and engagement, particularly in competitive or formal settings. *"It would help to show them that I'm actually excited about what I'm doing."* This sentiment points to the system's capacity to amplify not just the content being delivered, but also the emotions and energy behind it—an essential aspect in presentations aimed at persuasion or conveying enthusiasm.

While the focus was often on how the integration affected the presenter, some participants reflected on how it would likely impact their audience as well. P12 commented on the increased immersion, explaining that, *"it's so good to be immersed in the presentation... it will allow people to engage more because when somebody is presenting, I also tend to look at their faces."* This insight connects the experience of both presenter and audience, suggesting that the integration of video can foster a more engaging and immersive experience on both sides of the screen.

Quantitative results from our engagement and expressiveness scores reinforced these qualitative insights, with participants consistently rating the experience as more engaging than traditional methods. By dynamically integrating the presenter's video into the slide content, the system helped bridge the gap between static presentations and the human-centered interaction that is often missing in virtual environments.

These findings suggest that integrating speaker video into slides has the potential to transform the virtual presentation experience. It provides presenters with the tools to be more expressive and engaged, while also enhancing the audience's ability to focus and connect with the speaker. As virtual and remote presentations become more prevalent, this approach offers a promising avenue for creating more interactive and human-centered experiences.

7.3 Future Direction: Shared Controls and Personalized Presentation Styles

As we explored throughout our study, the dynamic integration of presenter video and slides has significant potential, but it also invites further opportunities to enhance control and personalization. Our findings suggest that users desire a balance between automation and manual control, as well as the ability to adapt the system to their personal presentation styles. These observations reveal important design implications for creating more flexible and user-centered presentation tools.

7.3.1 Shared Control with AI and Human. The concept of semi-automated control, or more aptly, shared control between humans and AI, has gained significant traction in recent years, particularly within the HCI community. With the rise of LLMs and VLMs, the potential for collaboration between human intelligence and AI systems has evolved considerably. Prior research [9, 47] demonstrates that combining human and artificial intelligence can lead to better outcomes than either agent working alone. This synergy between human adaptability and AI precision can enhance decision-making processes, improve task performance, and increase system robustness.

For TalkDirector, this concept of shared control can be applied by introducing semi-automated features where the system suggests transitions, video placements, or gesture-based interactions, while the presenter retains the final decision-making power. Participants in our study expressed interest in a hybrid approach, where they could benefit from AI-driven suggestions without relinquishing control over their presentations entirely. As P10 remarked, *"I might have liked a blend of both worlds... give me choices or something maybe."* This aligns with the idea of shared control, where the system assists the user but leaves room for human discretion and customization.

In practical terms, shared control can be implemented through suggestion panels, offering users potential actions based on real-time context. This approach not only reduces cognitive load but also ensures that the system adapts to user preferences and presentation styles. The design implication here is that presentation tools must provide a balance between automation and manual control, empowering users to make the final choices while benefiting from AI assistance.

By leveraging the insights from prior work [9, 47], TalkDirector can move toward a more collaborative framework, where human and AI agents work together to create more dynamic and responsive presentations. This balance of control can enhance the presenter's ability to engage their audience, while also adapting to real-time changes during the presentation.

7.3.2 Personalized Presentation Experiences. Beyond shared control, personalization emerged as another key area of interest for participants in our study. During the workshop, several participants expressed that their personality traits and presentation styles would influence how they prefer to interact with the system. For example, some presenters leaned toward a more dynamic, hands-on approach, while others preferred a more automated, static style. This highlights the need for future iterations of TalkDirector to account for these variations, tailoring the system to each presenter's unique style.

To achieve deeper personalization, future versions of TalkDirector could leverage additional machine learning models beyond those currently employed. Personalized experiences could be achieved by analyzing user behavior, learning patterns over time, or incorporating user-defined settings. For example, few-shot learning approaches could allow the system to adapt based on minimal input, or users could input text-based instructions on how they would like the system to behave during their presentations. This is similar to Tilekbay et al.'s work on video editing [52], where user input guides automatic edits. Applying such methods would enable presenters to customize their interaction with the system, giving them greater control over how their content is dynamically integrated.

Personalization would allow presenters to not only interact with TalkDirector in ways that suit their individual styles but also adapt the system to their preferences, resulting in more engaging and tailored presentation experiences. This approach could further enhance user satisfaction by making the system feel more intuitive and responsive to personal needs.

7.3.3 Design Implications for Presentation Tools. These observations reveal broader design implications for future presentation systems. The tradeoff between manual control and automation suggests that future systems need to be flexible and adaptable, catering to diverse user preferences. A system that offers semi-automated suggestions, while allowing the user to retain final decision-making power, could strike the right balance between efficiency and control. At the same time, personalization features that adapt to individual styles based on behavior or input could provide a more tailored experience, improving user satisfaction and engagement.

As virtual presentations continue to evolve, there is a growing need for systems that not only reduce the cognitive load on presenters but also enable more expressive and engaging presentations. By integrating semi-automated controls and personalized features, systems like TalkDirector can better support presenters, whether they are preparing for high-stakes professional events or more informal educational settings.

8 LIMITATIONS AND FUTURE WORK

While TalkDirector introduces a dynamic and interactive presentation interface, several limitations need to be addressed in future iterations.

8.1 System Latency and User Experience

One key limitation of our system is the latency in real-time interactions, especially during transcription and script matching. While no significant difference in overall presentation time was observed between TalkDirector and the baseline, P6 noted, *"I talked slightly slower with that interface because I wanted the features to work properly, and I noticed it takes a split second for the subtitles to show up."* This subtle delay may affect the presentation fluidity, as users may adjust their speech to accommodate the system.

To mitigate this, we could explore optimizing the transcription pipeline, perhaps by leveraging on-device processing for certain tasks to reduce network latency. Additionally, a preview feature that provides visual feedback on upcoming transitions could enhance usability by allowing users to anticipate changes, rather than waiting for the system to respond.

8.2 Handling Complex Slide Designs

Our system also faces limitations when processing complex slide designs. While GPT-4o helps identify layout components, its accuracy in generating bounding boxes for these components is not always reliable, particularly in slides with intricate or non-standard layouts. Although we mitigated this by using Tesseract to obtain text bounding boxes, recognizing images with text (*e.g.*, road signs) and complex figures remains challenging.

Future work could involve integrating advanced computer vision techniques, such as saliency detection, to enhance the identification of key visual elements on slides. Saliency detection could prioritize prominent areas, aiding in the detection of figures and important content. Additionally, leveraging object detection models like YOLO [53], or SAM [30] may further improve the recognition of figures and images in complex slides.

8.3 Script-to-Speech Matching and Flexibility

Our script-to-speech matching mechanism effectively synchronizes the presenter's speech with prepared content but limits spontaneity, requiring presenters to adhere closely to their scripts and reducing system flexibility. P1 suggested that it would be beneficial to allow the presenter to "go off script and talk about a slightly different matter", and in such cases, the system could automatically adapt by switching the presenter's video to full-screen mode when detecting off-script speech.

To enhance flexibility, future iterations of the system could incorporate a more dynamic script-to-speech matching mechanism. This approach would enable the system to recognize related content even if the presenter's wording deviates from the script. Integrating a user feedback loop, where the system visualizes its matching process and allows presenters to confirm or adjust its understanding, could further enhance adaptability.

Such enhancements would provide a balance between structure and spontaneity, making the system more responsive to the presenter's flow of thought during live presentations.

8.4 Assessing Audience Engagement

While our evaluation study results indicated that speakers felt more engaged when using TalkDirector, it remains unclear if the audience perceives a similar level of engagement. Future research could explore this by comparing presentation videos created with our method against a baseline. By analyzing audience eye gaze, we can assess whether viewers' attention aligns with the content emphasized by the speaker, as shown in prior studies on engagement in educational videos [56].

Additionally, incorporating information recall tests or subjective engagement assessments after viewing could quantify our system's impact on audience engagement. This would provide deeper insights into how dynamic video integration into slides influences both presenter and audience experiences, and whether it improves audience attention and comprehension compared to traditional presentation interfaces.

9 CONCLUSION

In summary, we present TalkDirector, a novel system for dynamically integrating presenter video feeds into presentation slides, to significantly enhance online presentation effectiveness. Through multimodal inputs and shared control mechanisms, TalkDirector optimizes presenter's video placement and size in real time, improving both the presenter's experience and audience engagement.

Our evaluations against existing tools, such as Zoom's "Slides as Virtual Background", reveal significant improvements in editing efficiency and viewer interaction. By blending automation with user control, our approach addresses key pain points in current online presentation workflows, making it easier for presenters to create engaging and contextually relevant presentations.

We believe that our open-source dataset, system, and methods will spur further research and development in this area, paving the way for more interactive and adaptive online presentation tools. Future work will explore extending our framework to incorporate additional multimodal inputs and investigate broader applications in various presentation contexts.

REFERENCES

[1] John Maxwell Atkinson and Max Atkinson. 2004. *Lend me your ears: All you need to know about making speeches and presentations.* Random House.

[2] Thomas Baudel and Michel Beaudouin-Lafon. 1993. Charade: remote control of objects using free-hand gestures. *Commun. ACM* 36, 7 (1993), 28–35.

[3] 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.

[4] Jens F Binder. 2023. Establishing conversational engagement and being effective: The role of body movement in mediated communication. *Acta Psychologica* 233 (2023), 103840.

[5] Mark Bowden. 2015. Winning body language: Control the conversation, command attention, and convey the right message without saying a word.

[6] J Brooke. 1996. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry* (1996).

[7] 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

[8] Erica A Cartmill, Sian Beilock, and Susan Goldin-Meadow. 2012. A word in the hand: action, gesture and mental representation in humans and non-human primates. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367, 1585 (2012), 129–143.

[9] Gabriele Cimolino and T.C. Nicholas Graham. 2022. Two Heads Are Better Than One: A Dimension Space for Unifying Human and Artificial Intelligence in Shared Control. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 8, 21 pages. https://doi.org/10.1145/3491102.3517610

[10] 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.

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

[12] Nancy Duarte. 2010. *Resonate: Present visual stories that transform audiences*. John Wiley & Sons.

[13] Michael E Ellis. 1992. Perceived Proxemic Distance and Instructional Videoconferencing: Impact on Student Performance and Attitude. (1992).

[14] 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, 321–326.

[15] Gerald Friedland and Raul Rojas. 2007. Anthropocentric video segmentation for lecture webcasts. *EURASIP Journal on Image and Video Processing* 2008 (2007), 1–10.

[16] 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.

[17] 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.

[18] Google. n.d.. Google Meet. https://meet.google.com/landing. Accessed: 2023.

[19] 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.

[20] SG Hart. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload/Elsevier* (1988).

[21] Keita Higuchi, Yinpeng Chen, Philip A 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*. 2383–2392.

[22] 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.

[23] 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.

[24] 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.

[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.

[26] Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

[27] Adam Kendon et al. 1980. Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication* 25, 1980 (1980), 207–227.

[28] Minju Kim, Jungjin Lee, Wolfgang Stuerzlinger, and Kwangyun Wohn. 2016. HoloStation: augmented visualization and presentation. In *SIGGRAPH Asia 2016 Symposium on Visualization*. 1–9.

[29] Minju Kim and Kwangyun Wohn. 2018. HoloBox: Augmented visualization and presentation with spatially integrated presenter. *Interacting with Computers* 30, 3 (2018), 224–242.

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Doll'ar, and Ross B. Girshick. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4015–4026.

[31] Willem JM Levelt, Graham Richardson, and Wido La Heij. 1985. Pointing and voicing in deictic expressions. *Journal of memory and language* 24, 2 (1985), 133–164.

[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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

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

[39] OpenAI. 2023. GPT-4: OpenAI's Multimodal Large Language Model. https://openai.com/research/gpt-4 Accessed: 2024-09-09.

[40] 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).

[41] PPTX2HTML. n.d.. PPTX2HTML: Convert PowerPoint to HTML. https://g21589.github.io/PPTX2HTML/. Accessed: 2023.

[42] 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.

[43] Garr Reynolds. 2011. *Presentation Zen: Simple ideas on presentation design and delivery*. New Riders.

[44] Hans Rosling. 2010. 200 Countries, 200 Years, 4 Minutes. BBC.

[45] Hans Rosling. 2013. The River of Myths. Self-published or other source, if applicable.

[46] 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.

[47] Yang Shi, Tian Gao, Xiaohan Jiao, and Nan Cao. 2023. Understanding design collaboration between designers and artificial intelligence: a systematic literature review. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–35.

[48] Ray Smith and Google. 2005. Tesseract OCR: An Open-Source Optical Character Recognition Engine. https://github.com/tesseract-ocr/tesseract Accessed: 2024-09-09.

[49] Murphy Stein. 2012. ARCADE: a system for augmenting gesture-based computer graphic presentations. In *ACM SIGGRAPH 2012 Computer Animation Festival*. 77–77.

[50] 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.

[51] 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.

[52] Bekzat Tilekbay, Saelyne Yang, Michal Adam Lewkowicz, Alex Suryapranata, and Juho Kim. 2024. ExpressEdit: Video Editing with Natural Language and Sketching. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 515–536.

[53] Ultralytics. 2022. YOLOv5. https://github.com/ultralytics/yolov5. Accessed: 11-Sep-2024.

[54] Vox. 2015. Obama on what most Americans get wrong about foreign aid. YouTube video. Retrieved 2023 from https://youtu.be/nzL_avUIlEE.

[55] 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.

[56] 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.

[57] Zoom Support. n.d.. Sharing Slides as a Virtual Background. https://support.zoom.com/hc/en/article?id=zm_kb&sysparm_article=KB0067697. Accessed: 2023.

[58] 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.

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?