석사학위논문
Master's Thesis

반응형 및 유동형 비디오 콘텐츠 적응화: 사용자 맞춤형 비디오 인터페이스를 향하여

Responsive and Fluid Video Content Adaptation: Towards Customized Video Interfaces

2022

김정연 (金廷妍 Kim, Jeongyeon)

한국과학기술원

Korea Advanced Institute of Science and Technology

석 사 학 위 논 문

반응형 및 유동형 비디오 콘텐츠 적응화: 사용자 맞춤형 비디오 인터페이스를 향하여

2022

김 정 연

한 국 과 학 기 술 원

전산학부

반응형 및 유동형 비디오 콘텐츠 적응화: 사용자 맞춤형 비디오 인터페이스를 향하여

김 정 연

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2021년 12월 13일

심사위원장 김 주 호 (인)
심 사 위 원 이 기 혁 (인)
심 사 위 원 오 혜 연 (인)

Responsive and Fluid Video Content Adaptation: Towards Customized Video Interfaces

Jeongyeon Kim

Advisor: Juho Kim

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in Computer Science

Daejeon, Korea
December 13, 2021

Approved by

Juho Kim
Professor of Computer Science

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MCS 김정연. 반응형 및 유동형 비디오 콘텐츠 적응화:
사용자 맞춤형 비디오 인터페이스를 향하여. 전산학부 . 2022년. 43+iv 쪽.
지도교수: 김주호. (영문 논문)
Jeongyeon Kim. Responsive and Fluid Video Content Adaptation:
Towards Customized Video Interfaces. School of Computing . 2022.
43+iv pages. Advisor: Juho Kim. (Text in English)

Abstract

Mobile video-based learning attracts many learners with its mobility and ease of access. However, most lectures are designed for desktops. This thesis (1) investigates the gap between mobile learners' challenges and video engineers' considerations using mixed methods, (2) provides design guidelines for creating mobile-friendly MOOC videos, and (3) develops a system that provides responsive and customizable video content. To uncover learners' challenges, we conducted a survey (n=134) and interviews (n=21), and evaluated the mobile adequacy of current MOOCs by analyzing 41,722 video frames from 101 video lectures. Interview results revealed low readability and situationally-induced impairments as major challenges. The content analysis showed a low guideline compliance rate for key design factors. We then interviewed 11 video production engineers to investigate design factors they mainly consider. The engineers mainly focus on the size and amount of content while lacking consideration for color, complex images, and situationally-induced impairments. We then present and validate guidelines for designing mobile-friendly MOOCs, such as providing adaptive and customizable visual design and context-aware accessibility support. Based on the findings from the interviews and surveys, we present FitVid, a system that provides responsive and customizable video content. Our system consists of an adaptation pipeline that reverse-engineers pixels to retrieve design elements (e.g., text, images) from videos, leveraging deep learning with a custom dataset, which powers a UI that enables resizing, repositioning, and toggling in-video elements. The content adaptation improves the guideline compliance rate by 24% and 8% for word count and font size. The content evaluation study (n=198) shows that the adaptation significantly increases readability and user satisfaction. The user study (n=31) indicates that FitVid significantly improves learning experience, interactivity, and concentration. We discuss design implications for responsive and customizable video adaptation.

Keywords Mobile Learning, Video-Based Learning, Content Adaptation, Content Analysis, Responsive Design

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Mobile learning has gained popularity and has been a key driver in enabling ubiquitous learning [7, 8]. In addition, major online learning and Massive Open Online Courses (MOOC) platforms including edX [9], Coursera [10], Khan Academy [11], and Udemy [12] provide mobile apps to support learning on mobile devices. Despite the rise in popularity, mobile video-based learning has physical (e.g., small screen size) as well as environmental (e.g., limitations on sensory channels posed by ambient noise and light) [13] constraints. Existing learning frameworks suggest that tiny font sizes and content-heavy lecture materials on small screens increase learners' cognitive load [14, 15, 16] and lower judgments of learning (JOLs) [17, 18].

Furthermore, most of the existing educational videos have been primarily designed for desktop environments. Although a body of research suggested design guidelines for mobile educational apps and websites [19, 20, 21, 22], few studies have contributed guidelines specific to mobile video-based learning content. Unlike static content, educational videos are temporally dynamic with both audio and visual information, and contain unique design components such as talking-head instructors and real-time handwriting. To mitigate these challenges, the instructional designers, video engineers, and researchers attempted to adapt learning content to small mobile screens. For example, responsive design techniques adapt educational websites to diverse screen sizes by adjusting layouts and amount of content [23, 24, 25].

However, the content adaptation is limited to static content such as websites and ebooks [26, 27], leaving the video content with small fonts and dense text less accessible in mobile environments. This indicates a need for responsive content adaptation of video content, which is, however, challenging for the following reasons.

It is required to decompose video into design elements such as text and images to resize and rearrange them to fit small screen sizes. However, the video becomes a sequence of frames and collection of pixels after encoding, with no access to semantic information of the in-video elements (e.g., text boxes, images). Although the existing research used a pixel-based approach to extract metadata from raw pixels, access to the lecture slides is often limited. Despite the attempts to adapt the video content by using rule-based methods [28, 29], creating heuristic rules entails massive manual work and consideration for the combinatorial explosion of possible conditions. This huge cost and low generalizability of the heuristic method calls for automation of the adaptation process.

To fill this gap, this thesis explores how to provide mobile-friendly video content to learners in two folds: (1) a quantitative and qualitative analysis on the visual design of video content that deteriorates learning experience, (2) a design of a new system for video content adaptation.

1.1 Challenges of Mobile Learners With Visual Design of Video Content

To uncover the challenges learners face in mobile MOOC learning, (1) we surveyed 134 learners and conducted follow-up interviews with 21 learners. The results revealed two main difficulties learners experience with visual content design: readability issues and limitations on sensory channels. We then evaluated whether the current MOOC videos are suitable for mobile learning, thereby quantifying deficiencies of current design and guiding improvement schemes. (2) We analyzed 41,722 video frames sampled from 168,508 frames in 101 courses from MOOC platforms by applying the known design guidelines. The content analysis results showed a low guideline compliance rate for the key readability design factors (2.79% for font size, 74.20% for the amount of text, 0.94% for the font size of the text in the image, 66.22% for color contrast between text and background). (3) Finally, we interviewed 11 video production engineers to investigate how they currently consider the learners' challenges.

The results showed that readability is the biggest concern, and they focus on the font size and amount of content. Meanwhile, little attention has been paid to complex images and images with text, which was the major learner-reported difficulties. In addition, engineers do not pay significant attention to situationally-induced impairments of learners in the design process. The findings imply a gap between learners' challenges and engineers' design considerations.

Based on the noticeable findings from the study of learners and video content, we suggest a set of design guidelines for creating mobile-friendly MOOCs. We then validated the suggested guidelines through an evaluation session with video production engineers. The engineers' ratings on each guideline item demonstrated the guideline's clarity and applicability.

1.2 A System for Video Content Adaptation

The findings from the survey and interviews led us to design a new system for video content adaptation. In the second part of the thesis, we propose FitVid (**Fit** your **Vid**eo), a content adaptation pipeline and video interface that provides responsive and customizable video content for mobile learning. Our system includes a computational pipeline that automatically generates a responsive adaptation, which powers an interactive video interface that supports direct manipulation and design options, including dark mode and instructor/template toggle. The automated pipeline consists of two stages: decomposition and adaptation. First, it is essential to locate and identify in-video elements (e.g., text, images), to resize and rearrange them for adaptation. The decomposition module extracts metadata of in-video elements from raw pixels by leveraging deep learning techniques. We collected and annotated 5,527 lecture video frames and trained a custom object detection model to classify lecture design elements. Second, the adaptation module generates and renders adapted content for mobile devices by following existing design guidelines for mobile content. To maximally apply the guidelines in limited screen space, we apply constrained optimization and a set of heuristics.

However, the pipeline-generated content may sometimes produce incorrect adaptations and may not satisfy every user's needs, for example, not having enough large fonts or proper layout. Thus, FitVid's UI provides users control over the content adaptation instead of automating the whole process. Users can directly manipulate the in-video elements by resizing and repositioning them. Users can further customize the content using design options: applying dark mode to a video and toggling talking-head instructors or slide templates (e.g., university logos) to optimize the mobile screen space.

We demonstrate the effectiveness of FitVid through a quantitative pipeline evaluation and a user study. Our adaptation pipeline increases the guideline compliance rate of design elements by 24%, 8%, and 4% for word count, font size, and text font size in images, respectively. The content evaluation study (n=198) corroborated the quantitative evaluation results, showing that the adaptation significantly increases the perceived readability and design satisfaction. The user study (n=31) indicates that FitVid significantly improves the learning experience with increased interactivity and concentration. We also identified three motivations of direct manipulation usage; to adjust the design, promote concentration, and simply interact with content.

1.3 Thesis Contributions

The primary contributions of this thesis are:

- An identification of mobile learners' challenges and their gap with video production engineers' design considerations

- An automated pipeline that generates mobile-friendly content adaptation
- A design and implementation of a system that provides learners with responsive and customizable video content

Thesis Statement: Video interfaces with automatic content adaptation and customized content design can improve readability and users' satisfaction in mobile environments.

Chapter 2. Related Work

This thesis draws on four bodies of previous research: (1) content design for educational videos, (2) content adaptation for mobile learning, (3) lecture design element detection from pixels, and (4) direct manipulation for video content.

2.1 Content Design for Educational Videos

Studies of content design for educational videos mainly lie in two branches: lecture types and visual design elements. First, studies have explored how different lecture types affect student engagement [30, 31]. For instance, Guo *et al.* [31] analyzed 6.9 million video watching sessions from four courses. They found that talking-head videos and Khan-style tablet drawings are more engaging than slide-based, programming/coding, and recorded lectures.

Other work highlighted the importance of visual design elements in the learning process, such as text and images. Inappropriate font sizes [14] impose cognitive load with reduced readability and lower judgments of learning (JOLs) [17, 18]. Besides, an excessive amount of words increases the cognitive load [14, 16] and information overload [32]. Image elements can also increase the cognitive load by splitting learners' attention when multiple images are presented at once [33, 14] or when the graphics are too complex [34, 35]. Due to the importance of visual designs in a learning context, several studies [36, 37] emphasized readability as a key factor in mobile content design. Also, Cross *et al.* [38] used crowdsourcing to improve the legibility of educational videos.

However, existing literature has only covered content design for instructional videos or mobile learning websites, leaving the intersection area —- mobile video-based learning —- unexplored. Furthermore, previous work mainly focused on learners' perspectives, not involving content creators' viewpoints or interactions between two groups. To fill in this gap, we investigated challenges, perceptions on current content design, and design opportunities from both perspectives of learners and video production engineers.

2.2 Content Adaptation for Mobile Users

In response to the prevalent use of mobile devices, there have been attempts to adapt content to mobile screens. Some work introduced a concept of responsive web design for mobile learning [24, 25]. According to existing work, excluding menus and images increased the mobile readability [23]. Other work proposed techniques of responsive content adaptation for informative visualization [39, 40] and eBook design [27]. However, most approaches for content adaptation are limited to adjusting static content such as text documents or websites. Our research uniquely introduces techniques to adapt dynamic content —-instructional video—-, that has been challenging due to characteristics of video medium; (1) difficulties of editing video content after release and (2) a high diversity of lecture designs which cannot be easily extracted or adapted using a simple set of rules or heuristics.

2.3 Lecture Design Element Detection from Pixels

Content adaptation requires the detection and identification of design elements to be adapted. Several existing studies used a non-pixel-based approach to extract design elements [41, 42, 43]. In other words, they utilized lecture

slides (e.g., PowerPoint files) as a data source, which contains metadata such as locations and types (e.g., text, image) of each design element. However, access to these slides is not always available. Furthermore, video lectures involve dynamic elements such as talking-head instructors and real-time handwriting, which are not included in the lecture slides. From this perspective, pixel-based methods that extract metadata from raw pixels are more generalizable to most existing video lectures that are not provided with accompanying slides. Previous work on pixel-based methods lies in two branches: traditional edge-based techniques and deep learning approaches. First, the traditional edge detection approaches [28, 44, 29, 45, 46, 47, 48] often adopted Canny edge detection and rule-based approaches. However, building heuristic rules includes huge manual work and consideration for the combinatorial explosion of possible conditions due to the high diversity of lecture designs. On the other hand, deep learning approaches learn such patterns from data. ViZig [49] used Convolutional Neural Networks with their own dataset of images downloaded from image search engines. WiSe [50] and SPaSe [51] contributed annotated dataset for the presentation slide segmentation task to train deep learning models. Despite the efforts to utilize deep learning to classify design elements in lecture slides, existing datasets are limited in size and diversity, covering only engineering and science courses and lack semantic groupings considering structural information (e.g., hierarchical bullet points). To fill this gap, we created a new dataset of 5,527 video frames for adapting lecture videos. And we trained a deep-learning-based object detection model by pretraining with a document layout dataset, which is more suitable for our task than general-purpose image datasets and fine-tuning with our new dataset.

2.4 Direct Manipulation for Video Content

Direct manipulation is an interaction style in which users act on displayed objects of interest directly involving rapid, reversible, and incremental actions and feedback [52, 53]. A rich body of work proposed interaction designs that allow users to control the video motions using a video interface that directly reflects their input gestures. For example, some research enabled the in-video object dragging along its motion trajectory [54, 55, 56], and reduced temporal ambiguities by allowing spatial-temporal manipulation [57, 58]. On the other hand, another thread of work introduced zoomable video interfaces to overcome the constraint of small screen sizes [59, 60, 61, 62, 63]. However, the interaction design of direct manipulation and zooming in for video learning content is not fully explored. Our work investigates users' challenges in the current interaction design of video lectures and suggests a new functionality that enables content customization through directly resizing and repositioning in-video elements.

Chapter 3. Mobile-Friendly Content Design for MOOCs

3.1 STUDY1: LEARNER PERSPECTIVES

To investigate the difficulties of mobile learners when consuming video-based lectures, we conducted surveys and follow-up interviews with mobile MOOC learners.

3.1.1 Survey Study

Survey Protocol

The survey questions include demographic questions, mobile learning context (e.g., devices, learning platforms, and subjects, situations in which learners take mobile MOOCs), learning behaviors (e.g., how frequent and how long for mobile MOOC lectures), and difficulties on visual design elements (e.g., text, image, and color). The question on design elements provides ten multiple-choice options (Table 5.1 1st, 2nd column) that are most frequently reported as main visual factors by the existing 25 design guidelines. The 25 guidelines (Table 5.1) are selected from the literature using the following criteria: guidelines on visual design elements of educational content. The survey also contains multiple-choice questions about which lecture type causes the difficulties in mobile learning. We provided eight lecture types based on an existing taxonomy of video lecture styles [2, 3, 4, 5, 6] as multiple-choice options: animation, recorded classroom, programming/coding, hand-drawing, interview/discussion, screencast, slide-based, and talking-head. We ended the survey with a section for respondents to leave their email if they were open to a follow-up interview. We added the detailed survey questions in the supplementary materials.

Respondents and Recruitment

We recruited 134 respondents (43 female, 90 male, one prefer not to specify) with ages ranging from 18 to 74 (18-24: 40, 25-34: 73, 35-44: 16, 45-54: 4, 65-74: 1) from 11 countries (South Korea: 48, Brazil: 25, The United States: 23, India: 22, Others: 16) through Amazon Mechanical Turk (AMT) and advertisement posts on online university communities. We ensured that all respondents had a mobile MOOC learning experience by asking them to upload a mobile screen capture of learning history or certificate from MOOC platforms.

3.1.2 Interview Study

Interviewees

To further understand the learners' challenges and requirements in mobile MOOC learning, we followed up with a subset of respondents who left their email addresses in the survey response. We reached out to 56 surveyors by email, and 21 responded (13 male, eight female). We provided a $10 Amazon gift card as compensation for

each interviewee. Their age ranged from 18 to 44 years old. Interviewees were from South Korea (11), Brazil (4), U.S. (3), India (2), and Canada (1). We refer to these interviewees as P1 through P21.

Interview Protocol

We conducted semi-structured interviews remotely using an online communication tool, Zoom, and audio-recorded the interviews under consent. The interviews took 30-40 minutes with four main sections: (1) whether they think the current visual design of MOOCs is suitable for mobile environments; (2) the main visual design elements that have caused difficulties in mobile MOOC learning; (3) the lecture types they experienced on mobile devices and challenges of each lecture type; and (4) how they currently deal with the difficulties and their wanted features to mitigate the challenges mentioned above, regardless of technical feasibility. The complete set of questions are included in the supplementary materials.

Interview Analysis

To analyze learners' challenges caused by visual design elements and types of lectures, we followed an iterative coding process [64]. The two authors coded three randomly selected interview transcripts from the dataset using the codebook. Finally, to access inter-rater reliability, we computed Cohen's kappa. The average Cohen's kappa score across all codes was 0.85 (SD=0.05, ranging from 0.80 to 0.89) with an average of 92.75% agreement. Each of the two authors then coded the remaining interviews independently. After independent coding, they met to discuss in two 1-hour meetings the interpretations, discrepancies until they reached a consensus on the codebook. They adjusted their coded data accordingly. Finally, we produced subcodes for difficulties on visual design elements and four subcodes for difficulties on the lecture types under structural codes of visual design elements and lecture types.

3.1.3 Survey and Interview Results

Results show that the mobile phone and tablet PC were the primary device used for mobile learning. 68% of respondents reported using mobile phones, 31% tablet PC, and 1% laptops. In terms of mobile MOOC learning frequency, 33% of respondents replied 3-5 times a week. The learners did mobile learning during their free time (42%) and while commuting (36%). The respondents' experiences cover a wide range of subjects and lecture types. The complete survey results are in the supplementary material. We analyzed learners' challenges with visual design elements and lecture types in mobile learning environments. The readability issues caused by small font size, dense text, and small text in complex images were primary pain points, followed by challenges related to image elements such as small image sizes and too many images.

Challenges on Visual Design Elements

In general, many interviewees (16 out of 21 in the interview)shared that the visual design of current MOOC content is not suitable for mobile devices (Fig. 3.1). We organize the reasons by visual design elements below.

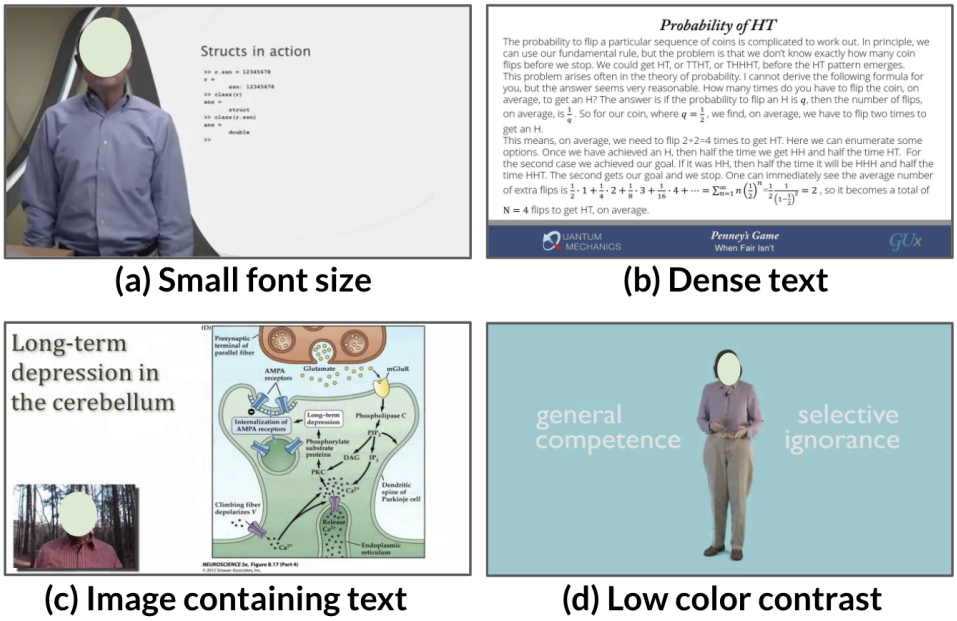

Figure 3.1: Example of learner-reported challenges from sampled video lectures.

Text Element

Small Font Size (survey: 63/134, interview: 19/21). The small font size was the most frequently reported pain point in both the interview and the survey. P19 reported that he "*relies on the audio instead of trying to read the small fonts*" and others (P4, P12, P20) re-watched the content using large-screen devices later. The problem deteriorated further due to the distracting learning environments. For example, P11 stated that "*When I'm on the bus, small fonts cause eye fatigue and motion sickness.*" Meanwhile, the readability issue even causes dropouts. In contrast, interviewees "*could bear with small fonts in stable environments such as home or library*"(P20) even with the small mobile screens. P17 also mentioned that "*I have readability issues when I'm working out, but I don't have any (issue) when I'm home.*" In other words, learners' preferences and requirements differ depending not only on the screen size but their context and environment.

The most commonly used solution to address the small font size was zooming in on the video content. However, all of the interviewees pointed out the inconvenience of the current pinch-zoom interaction of the video interface of MOOC platforms. Some interviewees (P1, P5, P6, P7, P11, P16, P18) found it irritating to keep zooming in content every scene changes. P11 commented on the inconvenience of adjusting the zoomed area, stating that "*It's annoying to move the zoomed area back and forth since zooming always results in part of the content getting cut off.*" P6, on the other hand, encountered technical issues while using zoom interaction, which resulted in unwanted actions such as exiting full-screen mode and scene transitions. Interviewees wanted element-wise and tap-based zoom interaction, which would allow them to enlarge certain design elements such as textboxes or images without cutting off (P10). They suggested an automatic zoom-in feature that enlarges the current spot on the slide being explained.

Dense Text (survey: 36/134, interview: 8/21). Several interviewees (P1, P7, P9, P10, P12, P16, P18) complained that text-heavy content on mobile screens makes it difficult for them to concentrate. Because of the dense text, P12 abandoned mobile learning and used a laptop to rewatch the lecture. P16 suggested that "*chucking long lines of text can mitigate the problem*" and P18 proposed "*reducing the amount of text responsively to fit the mobile display.*" On the other hand, P10 wanted more cues such as highlighting on the current explanation spot since "*it's easy to lose the instructor's explanation spot in the lecture material in the distracting mobile*

environments."

Inappropriate Text Spacing (survey: 18/134, interview: 4/21). P7 and P20 said that the inappropriate text spacing and line spacing make text harder to read on a mobile device.

Inappropriate Font Style (survey: 14/134, interview: 0/21). Inappropriate font style was not mentioned by interviewees, while 14 respondents in the survey reported this issue.

Image Element

Image Containing Text (survey: N/A, interview: 9/21). Since "image containing text" was not included in multiple-choice options in the survey, we have no such data to report as a survey result. However, nine interviewees stated that the main challenge with images was the small text contained in the images. The font sizes of labels in charts and graphs were often too small to read (P2, P10, P20). P10 complained that he "*should rely on the instructor's narration without access to the small labels in the graphs.*" This happened especially when the images were copied and pasted from the sources such as textbooks without adjusting font sizes.

Complex Images (survey: 32/134, interview: 7/21). Complex images include graphs, charts, tables, and infographics. P3 mentioned that "*complex images with subtle details were not recognizable even if they are presented as full-screen images*". They were using pinch-zoom interactions to mitigate the problem. Similar to interviewees' comments on the small font sizes, P10 wanted to use element-wise and tap-based zoom interaction for the complex images.

Small Image Size (survey: 51/134, interview: 5/21). Similar to interviewees' comments on complex images, some interviewees (P1, P11) complained about the text inside the small images, stating "*I had no difficulties with simple and straightforward pictures even their sizes are small, but I could not read the text inside the small-sized images.*" (P1). On the other hand, P3 pointed out that the size of the images was not a problem since most of the complex images are displayed in full-screen mode. However, he needed additional zooming in or cropping for interest in complex images.

Too many images (survey: 23/134, interview: 3/21). One interviewee found it challenging to know which image is currently being explained by an instructor (P10). P1 stated that he is "*easily overwhelmed by many images on small mobile screens compared to desktop environments*".

Color

Low Color Contrast (survey: 19/134, interview: 3/21). Some interviewees (P20, P2) encountered readability issues due to the low color contrast between the fonts and background. According to P2, a fancy background with a low color contrast with text, in particular, decreases the readability of the content: "*It's hard to read text on fancy graphics such as background using the Chroma Key technique. It might make the lecture more engaging in desktop environments, but readability is more important for me on mobile screens.*"

Too Bright Color (survey: 15/134, interview: 2/21). Interviewees explained that bright colors cause eye fatigue when mobile environments have low lighting. P5 and P16 wanted to have a dark mode option for video content.

Challenges per Lecture Type

We now report mobile learners' challenges for each lecture type (Fig. 3.2). The four major pain points were content-heavy lecture material (in slide-based, programmingcoding), low readability & legibility (in slide-based,

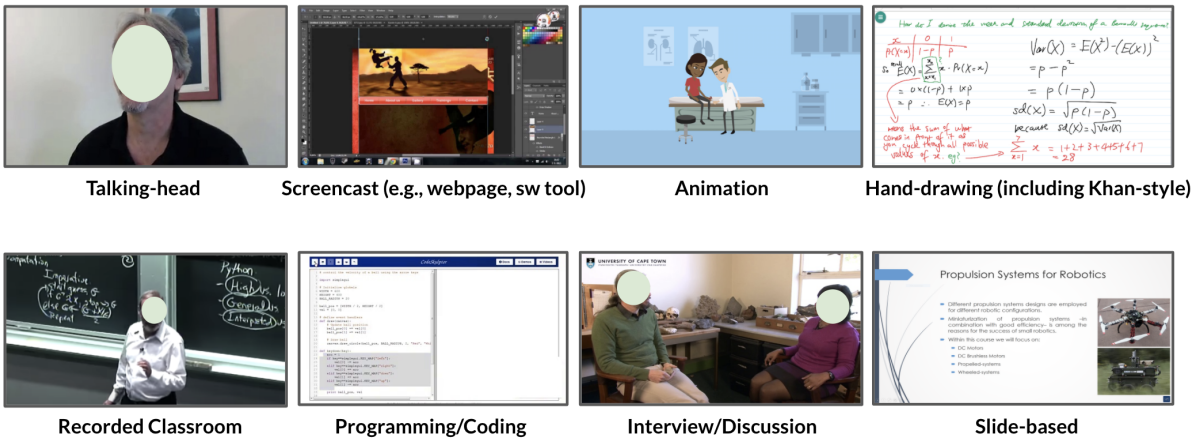

Figure 3.2: Eight lecture types summarized from existing work [2, 3, 4, 5, 6].

screencast, recorded classroom, hand-drawing, programmingcoding), lack of visually organized lecture material (in talking-head, interview/discussion), and the unavailability of following software tutorials or coding practice (in the screencast, programmingcoding).

Talking-head (survey: 89/134, interview: 3/21). The main pain point of talking-head was a lack of visually organized lecture material such as on-screen text and lecture slides. Some interviewees (P7, P9) preferred to have visual lecture material in mobile learning, particularly when their auditory channel becomes unavailable in noisy environments. P7 additionally stated that "*it was hard to get lecture content organized in the head in distracting environments without on-screen lecture material that provides a summary at-a-glance.*" (P7).

Although displaying the instructor's talking-head is engaging, picture-in-picture talking-head along with main content "*decreases readability limiting space for main content*" (P2) and "*split attention in already distracting mobile environments*" (P17, P18). For this reason, P2 wanted to turn on and off the picture-in-picture talking-head as needed, and P19 suggested toggling feature to switch between lecture types.

Slide-based (survey: 84/134, interview: 8/21). While the slide-based lecture is a commonly used lecture type in MOOCs, interviewees found it challenging to consume it in mobile environments due to the content-heavy lecture materials, low readability, and limitations on visual channels. The text and images in lecture slides incurred visual load to interviewees in mobile environments. P12 and P14 pointed out that slide-based lectures usually contain more content than other lecture types, such as taking-head. P12 said that "*I don't take slide-based courses on mobile phones since it's hard to concentrate on text and images on small screens.*" P1 and P5 mentioned the readability issues due to the small text in the slides. Furthermore, the distracting mobile environments, especially when interviewees are on the move, limit access to visual lecture materials (P12).

Hand-drawing (survey: 44/134, interview: 7/21). The main difficulty was the legibility of the handwritten text. Some interviewees (P1, P6) explained that the font size of the handwritten text was small, and others (P2, P4, P14, P15, P21) complained about the poor handwriting of the instructor.

Screencast (survey: 44/134, interview: 4/21). Interviewees found it difficult to follow the screencast lectures for two reasons. First, the shared desktop screens were not readable on mobile devices due to the small UI elements (e.g., mouse pointer, buttons) and tiny text (P7, P18, P21). Second, some interviewees pointed out the unavailability to practice the screencast tutorials using mobile devices. According to P10, "*it is hard to take screencast lectures without actual practice on desktops.*"

Programming/coding (survey: 55/134, interview: 8/21). Similar to screencasts, programming/coding lectures

also had low readability, and interviewees could not practice coding on mobile screens. For example, many interviewees (P2, P7, P14, P15) complained that long code lines were not readable and digestible on small screens. Meanwhile, due to the unavailability of coding practice, some interviewees (P2, P10, P11) did not use mobile devices for the lectures containing coding practices, stating "*I use mobile phones for the theoretical part of programming lectures and move to a laptop for the coding part.*" (P10). P21 mentioned the limited keyboard input as a challenge on mobile coding. Some interviewees suggested providing an interactive code editor embedded in the video content, which enables scrolling on long lines of code (P7) and displaying small chunks of code in a single video frame instead of screencasting the mere code editor (P10).

Recorded Classroom (survey: 60/134, interview: 3/21). Interviewees avoided taking recorded classroom lectures on mobile devices since it is hard to read chalkboard writings: "*the instructor's writing on the chalkboard was not legible*" (P3).

Animation (survey: 63/134, interview: 0/21). Although only two interviewees (P20, P21) have a mobile learning experience with the animation type, they both showed a strong preference for the animation type. P20 found it "*most engaging and interesting*" and P10 said that "*other lecture types such as slide-based type and screencast type should be redesigned to be suitable for mobile devices, but animation type does not need the mobile-specific version.*"

To summarize the survey and interview results, the readability issues were mainly caused by small font size and dense text. In particular, slide-based, hand-drawn, and programming lectures deteriorated readability and legibility problems. On the other hand, learners' situational auditory impairments worsen due to a lack of visually organized learning materials to compensate for the saturated auditory channel, especially in talking-head lectures. Other noticeable findings include that the inappropriate designs can lead mobile learners to dropouts and learners' requirements in mobile environments vary depending not only on the screen size but their context.

3.2 STUDY2: CONTENT ANALYSIS

In this section, we evaluated the mobile adequacy of MOOC videos based on existing visual design guidelines.

3.2.1 Data Set

We evaluated 101 MOOCs selected by Class Central, a search engine and review site for MOOCs. We sampled 51 courses based on their popularity, and 50 courses from top user reviews [65, 66]. Two of our authors selected the video frames considering the diversity of the design elements and lecture material. Our final set comprises courses from 64 institutions in 19 countries across five MOOC platforms, Coursera (54), edX (25), FutureLearn (19), Complexity Explorer (2), and an Independent University (University of Urbino). The complete course list is added in the supplementary materials. The 101 sampled courses contain 3,951 video clips with an average length of 8.1 minutes (min = 12 seconds, max = 54.85 minutes, SD = 6.63 minutes). For each video clip, we detected video frames by calculating edge-based differences to extract unique lecture material [67, 28, 29], ending up with 168,514 frames (M = 1668.5 frames/course, SD = 2599.6). To normalize the number of frames per course, we randomly sampled 500 (1/5 of standard deviation) video frames from each course. For 38 courses containing frames smaller than 500, we used all of the frames without filtering. The average number of frames per course after the normalization was 413, ranging from 31 to 500. In total, we analyzed 41,722 frames.

Design Element	Existing Guidelines	Guideline Compliance Rate
Font size	above 16pt	2.79%
	above 28pt	24.5%
Word count	below 20 words	74.2%
	below 40 words	84.9%
Font size of text in images	above 16pt	0.94%
	above 28pt	5.00%

Table 3.1: The guideline compliance rates of sampled video frames for the three design elements based on existing guidelines.

3.2.2 Evaluated Design Guidelines

Instead of analyzing all the visual design factors, we selected representative ones based on learners' interview results. In the interview, the top four challenges learners faced in mobile MOOC learning for each category were small font size (19 out of 21 learners), images containing text (9 out of 21), dense text (8 out of 21), and low color contrast (3 out of 21 learners). Besides, the interview showed that learners' perception of different visual elements is affected by lecture types. For example, learners suffered from the dense text, mainly in programming/coding lectures. Hence, we analyzed the guideline compliance rate for the four visual design elements—font size, word count, the font size of the text inside an image, color contrast—across different lecture types. For comparison, we estimated the font size in video lectures that are displayed on the most common mobile screen size at the time of analysis: 5.5-inch diagonal size with 1080 x 1920 screen resolution [33, 68].

3.2.3 Results of Guideline Compliance Analysis

The overall compliance rates of each design element are shown in Table 3.1. We present the guideline compliance rate of each design element across different lecture types as follows.

Distribution of Lecture Types

The sampled set covered all eight types of video production. The talking-head (37.8%) was most common followed by slide-based (34.6%), interview-discussion (11.4%), class recording (5.8%), programming/coding (4.7%), screencast (4.3%), hand-drawing (0.8%), and animation(0.5%) among 41,722 frames.

Font size

The suggested body text size in a mobile device is 17pt by the Human Interface Guidelines of Apple [69]] and 16 pt by Google Material Design Guideline [70]]. We also considered guidelines for presentation slides that suggest 28 pt[71, 72] as minimum font size. To evaluate the font size, we unified the different font size units (pt, px, sp) into a point (pt) for ease of comparison. Then we used the Pytesseract OCR engine, which showed reliable accuracy in previous work [73, 47, 29], to detect font sizes. The average font size was 13.9pt (SD=8.57, min=1.14, max=80.50). Adopting Apple and Google's guidelines, 75.5% of the video frames had font sizes smaller than 16pt. Adopting the guidelines for presentation slides, 97.2% of the video frames had font sizes smaller than 28pt. We then investigated the guideline compliance rate across eight lecture types (Fig. 3.3). Overall, more than 50% of

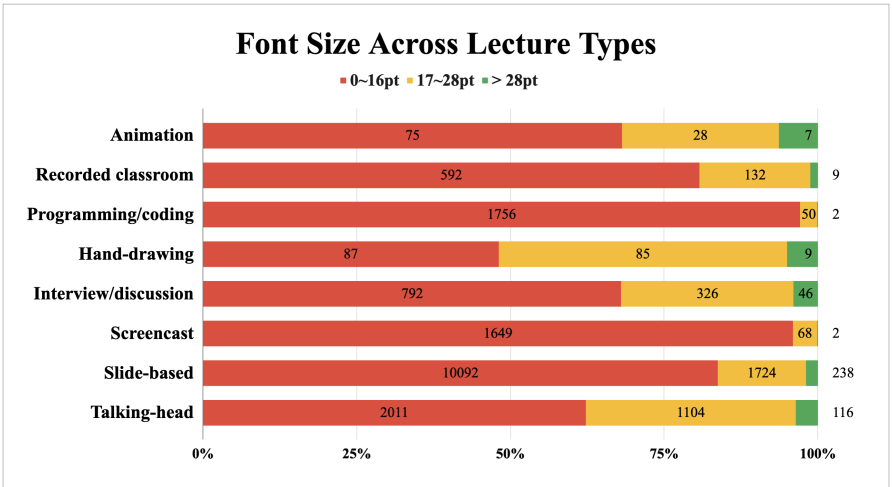

Figure 3.3: The font size compliance rate across different lecture types based upon existing guidelines.

video frames contained font sizes smaller than 16pt except for animation, revealing that font size guidelines are not followed in most lecture types. Especially for programming/coding and screencast lecture type, over 95% of the video frames did not meet the 16pt standard. These results match the learners' survey and interview results: font size was the main pain point across lecture types, particularly for programming/coding and screencast lectures.

Word count

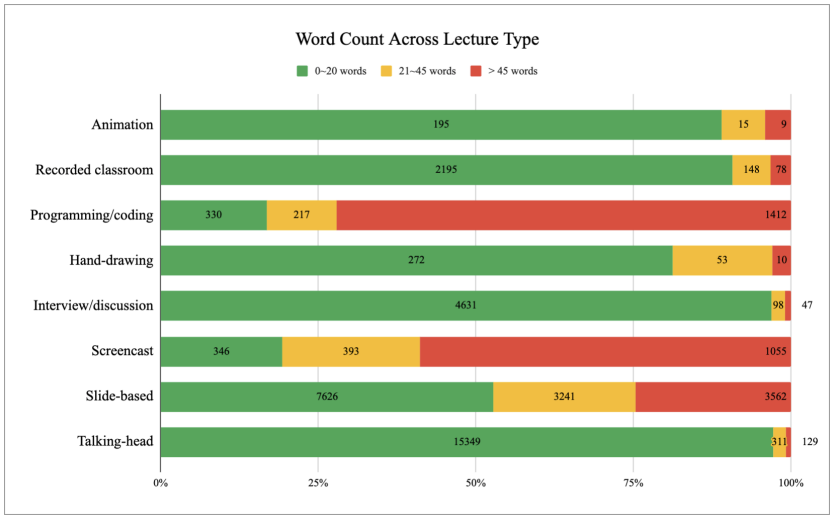

Figure 3.4: The word count compliance rate across different lecture types based upon existing guidelines. Programming/coding, screencast, slide-based lectures were the bottom three.

Using less than 45 words per presentation slide is recommended [74], while stricter guidelines advise using less than 20 words per slide [75, 76]. The average of word counts was 46.2 (SD=73.2, min=1, max=2431). Of the sampled frames, 25.8% had more than 20 words, and 15.1% contained more than 45 words. We used the

Pytesseract OCR engine as in font size to detect the words. The guideline compliance rate across lecture types is shown in Fig. 3.4. Programming/coding, screencast, and slide-based were most text-heavy, with 72.1%, 58.8%, and 24.7% of frames having more than 45 words, respectively. This analysis result was parallel with section 3.1: dense text was the second most frequently reported pain point by learners, especially for programming/coding, screencast, and slide-based lectures.

Font Size of Text Inside Images

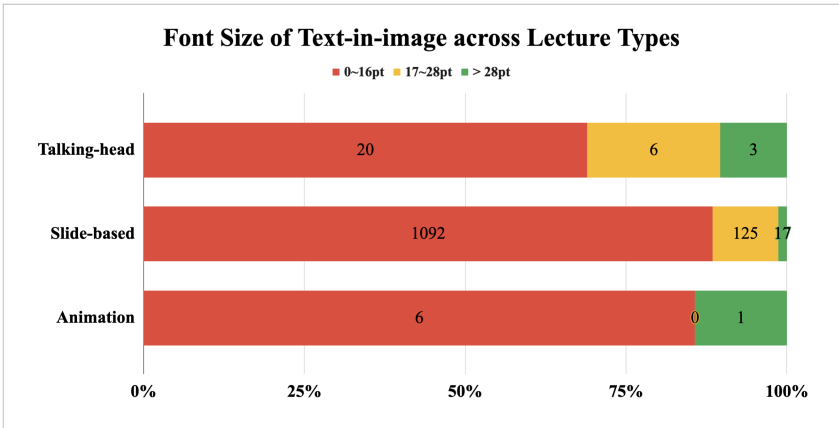

Figure 3.5: The font size compliance rate of text in images across different lecture types based upon existing guidelines. The compliance rates are lower than 15% across all three lecture types.

We evaluated the font size of any text inside images (e.g., graphs, tables, etc.) using the Pytesseract OCR engine. A total of 1,278 video frames had images with text inside them and the mean font size was 11.15pt (SD=6.51pt, min=2.42pt, max=81.65pt). The three lecture types, including talking-head, slide-based, animation, contain images with text from our dataset. Our results show that 95.0% of the frames had font sizes smaller than 16pt, and 99.0% of the frames had font sizes smaller than 28pt. These confirm the findings in learner interviews: among 11 interviewees who claimed difficulties with image elements, 9 of them had problems with the font size of the text inside images. Fig. 3.5 shows the compliance rate across the three lecture types that contain images with text. More than 85% of fonts from all three lecture types violate the guideline. The extremely low guideline compliance rate indicates that engineers' careful consideration of font size is needed when designing images containing text.

Color Contrast

We calculated the color contrast ratio between the background color and font color of 10,420 video frames. Thirty-three frames were excluded for having too low a contrast ratio due to the patterned background or low resolution. We first extracted color palettes from each text box in video frames and compared the color contrast between the most dominant color value and the second frequent color value. Two types of lectures, including slide-based and handwriting types, out of 8 lecture types contain text boxes with which we can estimate the color contrast ratio. Web Content Accessibility Guidelines (WCAG) [77] level AA suggests that the contrast ratio should

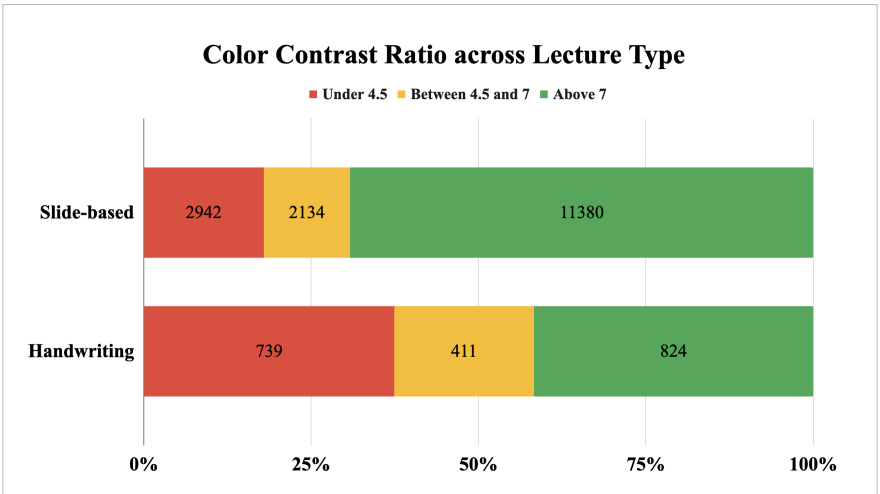

Figure 3.6: The guideline compliance rate of color contrast ratio across different lecture types. About 70% of color contrast from the slide-based lecture type complies with the guideline while 40% from the handwriting type complies with the guideline with a higher ratio than 7.0:1.

be larger than 4.5:1, while level AAA guideline recommends 7.0:1. The analysis results revealed that about 20% of the sampled frames have a contrast ratio lower than 4.5:1 and 34% lower than 7.0:1. The handwritings have a lower guideline compliance rate (42%) compared to the typed text in slide-based lectures (70%) (Fig. 3.6). The color components can be overlooked compared to other design components such as text and images, having less related guidelines, but the current design calls for careful consideration for color contrast for improved readability.

The low guideline compliance rates demonstrated that the current design of video lectures is not suitable to be consumed in mobile learning environments. Several lecture types (programming/coding, screencast, slide-based, and handwriting types) need engineers' attention when considering mobile learning.

3.3 STUDY3: ENGINEER PERSPECTIVE

We interviewed 11 video production engineers to investigate considerations and challenges in designing lecture videos for mobile users.

3.3.1 Participants and Recruitment

We recruited 11 engineers from the U.S. and South Korea via a campus mail list. The participants had 11 years of experience on average and reported their roles as a video editor, video production engineer, and video content designer (Table 3.2). Seven participants were university staff with design experience on MOOC content, and one was an independent engineer with experience in editing and publishing video-based learning content. Three of them were working in video production companies. We used a saturation method [78] to determine the number of participants. We refer to these participants as E1 to E11.

3.3.2 Interview Protocol

We conducted remote semi-structured interviews using ZOOM and audio-recorded the interview under consent. The interviews took 1.5 hours with four main sections: (1) general design process of video lecture content,

ID	Role	Experience (yrs)	Affiliation
E1	Video Editor	1	Freelancer
E2	Video Production Engineer	6	University
E3	Video Editor	10	University
E4	Video Content Designer	10	University
E5	Instructional Designer	6 months	University
E6	Video Content Designer	15	University
E7	Instructional Designer	18	University
E8	Video Editor	25	University
E9	Video Producer/Director	9	Video Production Company
E10	Video Content Designer	17	Video Production Company
E11	Video Content Designer	12	Video Production Company

Table 3.2: Information of interview participants

(2) considerations and challenges in designing content for mobile users, (3) perception on learner survey, interview, and content analysis result, and (4) communication channels with learners. The complete set of questions are in the supplementary materials.

3.3.3 Interview Analysis

Two of our authors and one external researcher who has rich experience in qualitative analysis extracted thematic codes through an open-coding approach [79]. They separately performed open-coding for the interview responses. We used affinity diagramming to cluster the generated codes [80] and iterated until we met a consensus over two 1.5 hour-long meetings. Finally, we identified four themes for challenges in designing mobile-friendly content.

3.3.4 Interview Results

We report the findings from the interviews on considerations and challenges of mobile content design.

General design process of video lecture content

To publish a video lecture, experts from various fields collaborate. The instructor first shared their lecture materials in slides or documents with the video production team. Then they discuss with the video production engineers and instructional designers the specific curriculum, lecture types, and video designs. Based on the discussions, the video production team starts filming. Then video production engineers post-process the recordings: designing subtitles, fonts, images, and video effects. The design process goes through many iterations with instructors. After the instructor's final confirmation, the videos are posted on the platform.

Considerations and challenges in designing content for mobile users

Most engineers know that more and more learners are using mobile to learn. Many already consider them as a target user group. Seven out of eleven engineers keep mobile environments in mind during the design process. But

due to limited resources and time, lectures are designed for desktops first.

Design Element. Much consideration for mobile learners was the readability of content. They simplified the content design to fit the small screen by reducing the amount of content (e.g., segmenting content in one slide into multiple slides). They also enlarge the size of the content. For example, E7 explained, "Font size is our main concern since it determines the readability in mobile screens. We next adjust the amount of dense text that is not digestible in mobile environments". As for the style, they preferred readable fonts such as sans-serif fonts and used font colors that had high color contrast ratio with the background.

Lecture Type. The engineers do not specifically consider mobile environments while deciding the lecture types. They chose the lecture type that is most suitable for the lecture content or preferred by an instructor.

Situationally-Induced Impairments and Disabilities (SIIDs). Most of the engineers have not considered the situationally-induced impairments and disabilities in their design process.

When asked about guidelines for mobile learning environments, they responded that there are no design guidelines specific to mobile environments. They noted that they rely on subjective intuition based on their experiences. In particular, we identified four main challenges for considering mobile devices during the design process. First, their design process in desktop environments makes it challenging to know how the content will look on mobile devices. The diversity of mobile devices deteriorates the problem. For example, E10 said, "In the current design process (on desktops), it's hard to imagine how the content I created will be displayed on very diverse mobile screens including laptops, tablet PCs, and smartphones.". Second, they mentioned the lack of design guidelines for mobile environments. For example, E10 commented, "We currently have guidelines for font size and color contrast. However, they are not mobile-specific and more diverse aspects should be considered (in designing content for mobile learners).". E2 also mentioned that "It is difficult to measure the readability, so for now, we just rely on the subjective intuition of an individual engineer.". Third, they encounter conflicts with instructors. In most cases, the engineers prioritize content readability. A few instructors wanted to use the same offline class learning materials. Some instructors do not prefer changing the original content design since the flow of the lecture can be disrupted. In this case, the engineers and instructors discuss whether to segment or summarize the dense content. Lastly, the engineers pointed out that it's hard to set the design directions due to a lack of understanding of the lecture materials. For example, they find it challenging to reduce word count while preserving the key information with a limited understanding of the lecture.

Perception on learner survey, interview, and content analysis result

Learner Survey and Interviews. All of the engineers replied that it was an expected result that the biggest challenge was the readability issue. All of them responded that it was an expected result. They elaborated that it is difficult to make every content readable on mobile devices due to the challenges in Section (2) considerations and challenges in designing content for mobile users, even though they consider readability a high priority. Meanwhile, the engineers did not expect that the second major difficulty was the SIIDs. They mentioned that they could understand that the learners suffer from SIIDs, but have not considered it in their general design process. When asked about the expected challenges of considering SIIDs in the design process, they mentioned the needs for the design guidelines, elaborating that they cannot think of any possible solutions to alleviate SIIDs.

Content Analysis. The engineers said that both the guideline compliance rate for font sizes and the amount of text are lower than they expected. For example, E2 said, "I expected that 30-50% of the current video lectures follow the guideline. I think it can be because most of the engineers don't have guidelines for mobile content design." In the case of font sizes of text in images, most of them noted that it was an expected result. Some engineers mentioned that the source image provided by an instructor sometimes has poor readability. One participant noted

that ”The instructors provide us with figures with too small text in them since they are not professionals in design. We sometimes redesign the figure based on the provided image, but we should use them as they are in many cases when we have a tough deadline” (E6).

Communication channels with learners

All of the engineers agreed with the need for getting learners' feedback. Some engineers collected user feedback via survey, and others communicated with learners through online bulletin boards on MOOC platforms. However, they noted that most students do not respond to surveys, emphasizing the lack of communication channels. In particular, they have difficulty reflecting feedback on design elements, which requires a complete redesign or reshoot of the video.

In this section, we suggest a set of design guidelines for creating mobile-friendly MOOCs. The guidelines are based on our learner studies (Section 3.1). We validated the guidelines through an evaluation session with the video production engineers. We also discuss design opportunities for mobile-friendly MOOCs. Based on the surveys and interviews with learners, we identify three design opportunities: (1) readability support via adaptive and customizable visual design, (2) context-aware accessibility support, and (3) informed lecture selection. We expand on each design opportunity along with design recommendations derived from noticeable findings and learner suggestions. The summary of findings (F1-9) and design recommendations (G1-9) on visual design elements are shown in Table 3.3.4.

Readability Support via Adaptive and Customizable Visual Design

Existing design techniques such as responsive web design provide customized views across various mobile devices. However, responsive video design is challenging because of the inflexible nature of the video, making it difficult to be edited or deconstructed after release. The lack of support for responsive video design leads to readability issues, in particular, in content-heavy materials (F1, F4). This suggests novel design opportunities for an adaptive and customizable visual design for video content. Specifically, the key is to automatically generate responsive video content and provide customizable design options to fit diverse learners' needs without making separate versions of the content for each device. For instance, Optical Character Recognition (OCR) or edge detection techniques in computer vision can be used to extract in-video elements such as text boxes and images. Once the in-video elements are detected and deconstructed, it becomes easier to adopt them like static content such as websites and ebooks. Then you can enlarge the font size, adjust the layout depending on screen sizes [81, 82] (G1a). Furthermore, the extracted elements can be customized, providing options (e.g., font size, color) for different designs to learners to choose from (G1b, G4a, G4b, G8) by adjusting and redesigning the deconstructed elements based on users' preferences on color or size. On the other hand, learners currently use pinch-zoom interactions to alleviate the readability problem (F2, F3). They complained that zooming sometimes results in missing other important content and that pinch interactions cause unwanted actions such as the exit of fullscreen mode. As an alternative, they mentioned tap or long-press interactions. By deconstructing recorded video into design elements, an improved adaptive video content design technique could enable element-wise zoom interaction by magnifying a complete element (e.g., text box, image) without cut-off parts (G2).

Context-Aware Accessibility Support

Mobile learning is distracted by situational factors, aka Situationally-Induced Impairments and Disabilities (SIIDs)[13, 83]. A design opportunity is to provide context-aware support for addressing SIIDs in mobile video-based learning (G5, G6a). To detect learners' learning contexts, existing detection sensors such as eye trackers or

Challenge	Finding	Guideline	Design Process (Target User Group)
Readability Issues	F1. Learners require different font sizes depending on screen sizes.	G1. Provide options for different font sizes. Let users choose their preferred font size like PowerPoint font size option.	Video Content Design (Platform developers, System researchers)
	F2. Learners' pinch-zoom interaction to mitigate readability issues may cut off parts of elements.	G2. Provide element-wise zoom interaction that magnifies the complete element (e.g., text box, image) instead of parts of elements. Enable zooming in the whole content without cut-offs.	Video Interface Design (Platform developers, System researchers
	F3. Learners' pinch-zoom interaction to mitigate readability issues may cause unwanted actions (e.g., scene transition, exit of fullscreen mode).	G3. Provide alternative zoom methods that can prevent unwanted actions. Provide zoom methods such as tap or press, since current pinch-zoom interactions lead to unintended actions such as the exit of fullscreen mode or scene transitions.	Video Interface Design (Platform developers, System researchers)
	F4. Decorative video designs (e.g., slide transition effect, fancy background using Chroma Key) can decrease readability.	G4a. Avoid using decorative visual effects such as slide animations and fancy backgrounds˜ G4b. Provide a design option to change decorative visual effects. Provide different design modes, for example, basic mode and decorative mode.	Video production engineers, Platform developers, System researchers
Situational Impairments	F5. Ambient light and the mobile state of the learner can cause situational visual impairments.	G5. Provide a context-aware* audio description or extended audio**.	Instructors, Platform developers, System researchers
	F6. Noisy mobile environments can cause situational auditory impairments.	G6a. Provide a context-aware* subtitle. Turn on, for example, subtitle automatically when a learner is in noisy environments. G6b. Provide redundant on-screen text with audio narration. Display, for example, a summary of the audio narration in the form of keywords or a bulleted list.	Instructors, Platform developers, System researchers
	F7. Learners easily lose instructor's current explanation spot in on-screen lecture material.	G7. Display cues or signals on the current explanation spot, both for images and text. Visual cues include underlines, highlighting, and arrows.	Instructors, Video production engineers
	F8. Low ambient light can cause eye fatigue.	G8. Provide dark mode for video content with light color text in a dark background.	Video production engineers
Inaccessible Information	F9. Learners have difficulty knowing if a lecture of interest is mobile-ready, resulting in dropouts due to uninformed lecture selection	G9. Provide information on video content design in the lecture selection stage. Improve information scent about mobile-friendliness. The information scent, for example, includes the design guideline the compliance rate of font sizes or involvement of programming practice.	Video production engineers, Instructional designers

Table 3.3: Summary of notable findings and design recommendations on visual design elements for mobile-friendly MOOCs. *Context-aware learning detects learners' context (e.g., ambient light) and adapts learning materials to match the context [1]). **Audio description is a narration added to the soundtrack to describe important visual details that cannot be understood from the main soundtrack alone. An extended audio description that is added to an audiovisual presentation by pausing the video so that there is time to add additional description.

accelerometers can be used [84]. For example, an audio description for visual lecture material can be provided when learners' eyes are not fixed on the screen. In particular, our findings draw a parallel and contradiction with Mayer's Multimedia Learning Theory [85] at the same time. The learners strongly wanted cues or signals on the currently explained spot due to distraction in mobile environments (F7), corroborating the signaling principle of Multimedia Learning Principles [86]. They even required adding cues or signals on image elements as well as text elements (G7). On the other hand, our findings mildly contradict the redundancy principle, which suggests that the same information presented both on-screen and orally interferes with learning [86]. In mobile environments, learners suffered from the lack of visually organized lecture materials on the screen (F6). For example, when a learner was in noisy environments such as a gym or subway, they wanted redundant on-screen text together with audio narration. In distracting mobile environments, learners' available sensory channels quickly alternate, allowing consistent use of a single channel. Hence, it is recommended to provide redundant information in both auditory and visual modalities to complement each other (G6b). More theoretical and empirical research is required to validate the impact of these mobile learning techniques on learning effects.

Informed Lecture Selection

The illegible and indigestible content may even cause dropouts of mobile learners, as reported in our interview. Learners express the need for information on mobile adequacy of lecture content at the lecture selection stage. This points to an opportunity to support informed decision-making in lecture selection. Our findings revealed that visual design elements and lecture types are critical factors that determine the mobile adequacy of the lecture content. Instructors or platform engineers can provide the information (e.g., font size, word count, the existence of coding) on video content design to improve information scent [87] (G9). Future work can develop a mobile adequacy index or score by investigating the weighted importance of each factor, providing a simple and quantitative measure for mobile adequacy.

3.3.5 Design Guidelines on Mobile-Friendly Lecture Types

The findings on lecture types from section 3.1 indicate that each lecture type has different challenges and needs to be tailored to mobile learning environments. In particular, several findings mildly contradicted the general conclusions from the previous work that assumed desktop environments, calling for the need for mobile-specific design considerations and recommendations. The summary of findings **(F10-14)** and design recommendations **(G10-14)** on lecture types is shown in Table 3.4.

Talking-head. Previous research demonstrated that talking-head videos engaged learners more with personal feelings [31, 89]. On mobile devices, however, learners easily get lost in talking-head due to the lack of visually organized lecture material (F11). The lack of visual material also caused VIIDs and loss of concentration. Hence, we recommend adding visual lecture materials such as text and images beside the instructor's talking head in post-production editing (G11). Another pain point of talking-head was when the picture-in-picture talking-head is displayed along with the main content [88] (F10). First, the picture-in-picture talking-head split learners' attention in distracting mobile environments. Second, the juxtaposition of the main content and talking-head as a split view limits the space for the main content and decreases the readability of the main content. Therefore, we advise providing an option to turn off the picture-in-picture talking-head display (G10).

Programming/coding. Prior work introduced the effectiveness of live-coding in learning programming, including the ease of understanding [90, 91] and a decrease in extraneous cognitive load [92]. However, the programming and screencast were not readable on mobile screens due to small font sizes and an excessive amount

Challenge	Finding	Guideline	Target User Group
Talking-head	F10. Displaying talking-head along with main content decreases r eadability of main content, and splits learner's attention with increased visual clutter in mobile learning environments.	G10. Provide an option to toggle picture-in-picture talking-head window [88].	Video production engineers, Platform developers, System researchers
	F11. Learners easily get lost without visually organized lecture material (e.g., lecture slides) in mobile learning environments.	G11. Add visual lecture material such as overlay text or images to pure talking-head lectures in post-production editing.	Video production engineers, Instructors
Programming/coding	F12. Programming screencasts are not readable in mobile learning environments.	G12a. Encourage instructors to increase the zoom level of code editor screen during live coding. G12b. Zoom in code editor window in post-production editing.	Video production engineers, Production directors
	F13. Despite the small screen size and limited keyboard input, learners want to practice coding on mobile devices.	G13. Provide a lightweight IDE with code snippets for simple coding practice.	Instructors, Platform developers
Hand-writing	F14. Hand-written text is not legible on mobile devices due to cursive fonts and inappropriate text spacing.	G14. Provide a typed version of hand-written text as lecture notes.	Instructors

Table 3.4: Summary of notable findings and design recommendations about lecture types to create mobile-friendly MOOCs

of code (F12). We recommend instructional designers encourage instructors to zoom in on the code editor if instructors insist on the live-coding type (G12a). We recommend zooming in on the code editor at opportune times in post-production editing (G12b). On the other hand, some learners wanted to practice coding on mobile (F13). They complained about the unavailability of mobile IDEs and limited keyboard input. To address this problem, we suggest providing lightweight IDEs to practice coding and providing code snippets so that learners do not have to write long lines of code from scratch (G13).

Hand-drawing. Hand-drawing videos are another recommended lecture type in desktop environments, highly engaging students in desktop environments [93, 31]. However, learners preferred the hand-drawing type the least in mobile learning environments due to its low legibility (F14). We recommend providing corresponding typewriting (print) to the handwritten materials (G14) [38].

Recorded Classroom. Recorded classroom lectures are also not preferred in mobile environments due to their low legibility and small font sizes (F15). Similar to the hand-drawing type, it is recommended to provide lecture notes in digital text to allow the learners to refer to them as needed (G15).

Slide-based. Slide-based videos are prone to contain cluttered materials (F16). As suggested by previous work [19, 14, 94], we recommend segmenting long text into small chunks or summarizing the content into several bullet points (G16).

3.3.6 Expert Evaluation of Design Recommendations

We conducted interviews with 11 video production engineers to investigate the applicability and clarity of our design recommendations.

Procedure

At the end of the interview in section 3.3, we had an expert evaluation session for the design recommendations. The participants were asked to fill out a form that evaluates the items. For each recommendation, we first confirmed the right target group of the item. Since the content design process of video lectures necessarily involves multiple stakeholders such as platform developers, instructors, and instructional designers, the participants might need collaboration or cooperation between multiple teams to implement the recommendations. The form then asked the participants to provide ratings on clarity, understandability, applicability, actionability, and the easiness to work with the recommendations, on a 5-point semantic differential scale (e.g., from 'very confusing' to 'very clear'). We inquired the reasons for the ratings the following question. We also asked anticipated challenges of applying the recommendations. The participants were then requested to improve the recommendations by editing, adding, and deleting the suggested items. Lastly, we asked the participants if there was any wanted system or tool for implementing the recommendations.

Results

In this section, we report the video production engineers' feedback on the clarity and applicability of the recommendations, expected challenges of applying the recommendations, and revisions suggested by the engineers. Fig. 3.7 presents the subjective ratings from video production engineers for each guideline item.

Related Stakeholders. The engineers reported that most of the guideline items are within the scope of their roles. Meanwhile, some of the items require them to collaborate with other teams. For example, they noted that they need to collaborate with learning platform developers to apply some guidelines, including G1-3, G5, G6, G9,

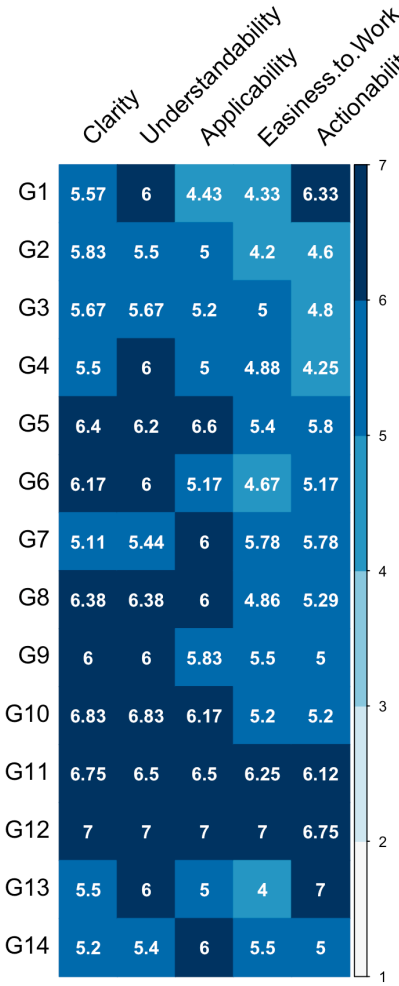

Figure 3.7: Subjective evaluations by video production engineers about the clarity, understandability, applicability, easiness to use, and actionability of the suggested guideline items.

G12, G13, which involve development work for a new video interface or mobile sensor. On the other hand, G14 requires the help of instructors who have a complete understanding of the lecture materials.

Clarity and Understandability. The mean of the ratings for clarity and understandability was 5.99 and 6.07 out of 7, respectively. Some engineers wanted to see the working system or prototype which applies the guidelines for new features in the video interface. For example, E1 commented on G2 that ”It’s hard to imagine how the new video interface (with the element-wise zoom feature) works without the actual system”. They reported that the guidelines are clear and easy to understand overall.

Applicability, Easiness to Use, and Actionability. The mean of the ratings for applicability, easiness to use, and actionability was 5.71, 5.18, and 5.51 out of 7, respectively. For G1 and G2, the engineers rated them less applicable (4.43, 5) and easy to work with (4.33, 4.2) compared to other guidelines. They expressed concerns about the spatial relationships between elements. For example, ”(For G1) a simple adjustment for font sizes can break the layouts of the content.” (E9) and ”(For G2) the element-wise zoom can cover the rest of the content and distort the whole layout.” (E5). For G4 and G9, the engineers mentioned that they need more precise criteria for ”decorative visual effects” and ”mobile-friendly design”. E2 said that ”We will need clear criteria or even experts’ advice to

determine the mobile-friendliness of the video content".

The engineers also indicated that G7 and G8 could pose additional workloads. They explained that visual cues need additional graphic work, and creating a dark mode for the whole content can double the workload. Several engineers (E4, E7, and E10) mentioned the possibility of automating the design process to apply these two guidelines. E4, for example, said, "It would be efficient if there were some programs that extract text and background from the video and convert the color for the dark mode".

Another problem was the difficulty of having a complete understanding of the lecture materials. To apply G11, the engineers "might need close cooperation with instructors, which requires much effort." (E4). Similarly, for G14, they noted that they cannot understand the whole lecture content each time and necessarily need help from instructors to provide correct lecture notes for handwriting.

Usefulness. The engineers agreed on the need to apply the suggested guidelines to improve the mobile learning experience. Many engineers like the idea of giving options to learners customizing, for example, font sizes (G1) and lecture designs (G4, G10). G10 commented that "(For G10) it's a good idea to provide options to users considering their personal preferences because a single design cannot satisfy every user". Some engineers appreciate the guidelines (G1-4) that can alleviate the readability issues in mobile devices. On the other hand, they valued the guideline for providing information for mobile-friendliness (G9), with one expert commenting that "the information scent can benefit many mobile learners." (E5).

Revisions. When asked if there is any guideline they want to add, remove, or edit, the engineers indicated that G2, G5, and G6 have room for improvement. E5 pointed out that G2 overlaps with G3 in that they both suggest a new zoom interaction. E5 also commented that G5 and G6 need to clarify how the context-aware system should be designed, including details such as the minimum level of noise that requires the context-aware subtitle. The rest of the engineers were satisfied with the current guidelines without further need for modifications.

Chapter 4. FitVid: Responsive and Flexible Video Content Adaptation

4.1 Design Goals

The survey and interview results in Chapter 3 led us to the following design goals in creating a system that supports responsive and customizable video content.

G1. Automatically generating responsive design by video decomposition

Video adaptation requires a decomposition of raw video into visual elements (e.g., text, images) so that they can be flexibly resized and rearranged at the element level to improve its readability. We aim to create an automated pipeline that extracts in-video elements by using deep learning algorithms and generates responsive design based on existing design guidelines.

G2. Supporting direct manipulation of in-video elements

The automatically generated content may sometimes produce incorrect adaptations and may not satisfy every user's needs [95, 96], for example, not having enough large fonts or proper layout. We aim to allow users to control the content adaptation instead of automating the whole process by supporting direct manipulation of in-video elements (e.g., resizing elements).

G3. Providing options for content customization

The "one-size-fits-all" approach of providing the same design of video content to every learner is unlikely optimal in mobile learning because of a variety of different learning environments (e.g., small screens, distracting situations). The interviews revealed that users want to customize video design, such as by toggling instructor displays and changing color themes. We aim to provide users with options to adjust video content to mitigate constraints in mobile environments.

We will describe our computational pipeline that automatically generates adapted content (G1) in Section 4.2 and our user interface that supports direct manipulation (G2) and content customization (G3) in Section 4.3.

4.2 Computational Pipeline for Automated Adaptation

This section describes our new computational pipeline that automatically creates an adapted version of video lectures (D1: Automatically adapting video content). Our key idea is to decompose a video into design elements that can potentially be adapted. The pipeline consists of two stages: (1) decomposition and (2) adaptation, as depicted in Fig. 4.1. This section describes the main idea of each stage, while we provide additional details in supplemental materials for rigor and reproducibility.

4.2.1 Decomposition Stage

In this stage, video content is decomposed into several elements, and metadata for these elements is extracted to be used for adapting the content. To support adaptation that resizes and rearranges the decomposed elements,

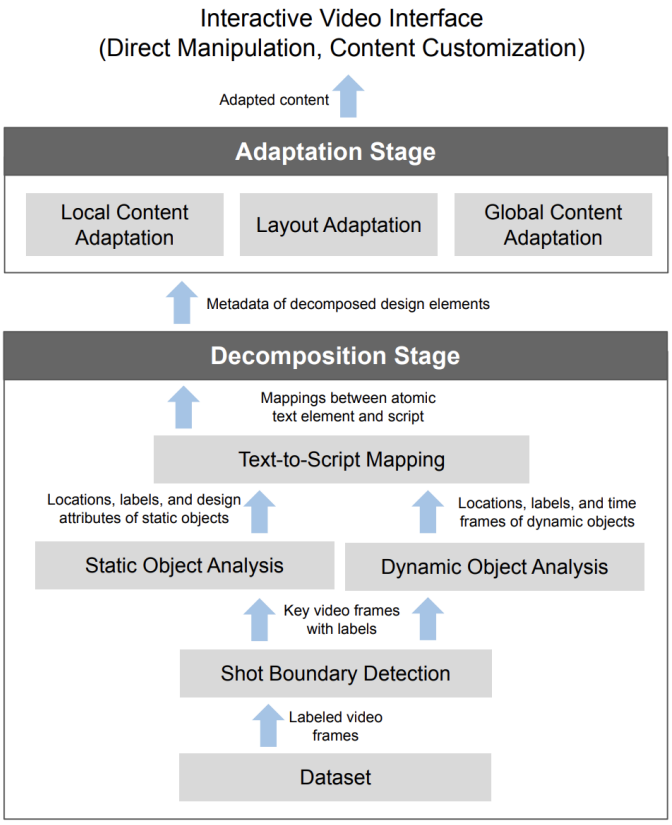

Figure 4.1: A computational pipeline of FitVid consists of two main stages, decomposition and adaptation. The decomposition stage extracts in-video elements and metadata used for adaptation by using several intelligent algorithms and AI techniques, including object detection. The adaptation stage generates and renders adapted content for mobile screens. The detailed version of the diagram is included in Supplementary Materials.

which we further describe in Section 4.2.2, we identify the following information to be extracted from video content:

- **Shot Boundary Detection.** A shot in video analysis refers to a series of frames that runs for an uninterrupted time [97]. For the video lecture domain, we are interested in identifying transitions between lecture slides or scenes in which an instructor narrates continuously so that we can further analyze each lecture slide.
- **Static Object Analysis.** Once the representative frame is identified for each video shot detected, we extract several elements that do not change or move over time (e.g., text, images). These elements are adapted based on the existing design guideline in the adaptation stage (e.g., resized images).
- **Dynamic Object Analysis.** Lecture videos not only contain static elements but dynamic elements that change or move over time within a video shot (e.g., laser pointer). Since the object detection techniques often fail to detect these small objects, we perform a separate analysis to detect these dynamic objects.
- **Text-to-Script Mapping**. For on-screen text elements, we identify the corresponding audio narrations. This is used to determine the right timings to display segmented content units after adaptation.

Shot Boundary Detection

To detect shot boundaries of video lectures (e.g., the transition between lecture slides), we used HSV (Hue, Saturation, and Intensity Value) peak detection and template matching techniques [98]. It returns a sequence of shots, each consisting of start and end time and a single representative image frame. Information about the detailed implementation and threshold is included in the supplementary materials.

Static Object Analysis

Once a video is decomposed into a series of shots with its representative frame, we analyze these frames to identify elements that can potentially be adapted (e.g., text box). We use deep-learning-based object detection models to classify and locate visual design elements in video frames.

New Labeled Dataset of Lecture Designs A large, high-quality dataset is crucial to train high-performing object detection models. Previous work released annotated datasets for presentation slides [50, 51], however, these datasets are not applicable to our context primarily for two reasons: small size (i.e., only consisting of 2,000 slides), limited diversity of subjects (e.g., only engineering and science courses), and the lack of semantic groupings considering structural information (e.g., hierarchical bullet points). Thus, we created a new annotated dataset of 5,527 video frames sampled from 66 popular courses.[1] Our dataset includes lecture videos taken from courses over 44 institutions in 11 countries with the subjects across 14 domains (e.g., computer science, management, and art). Two of our authors selected video frames considering the diversity of the design elements and lecture materials. We chose 12 class labels for classifying design elements in lecture material [51, 49], which include title, text box, diagram, picture, chart, and instructor. We labeled design elements based on semantic units, which is considered important in data annotations for design elements [99]. For example, we grouped multi-level lists with hierarchical relationships as a single text box and complex graphics consisting of separate elements connected with arrows as a single diagram. These semantic groupings enable content adaptation with the semantic relations of the original elements preserved.

We released the dataset to the open dataset repository for further research[2]. The use of this dataset is not limited to the training of object detection models. The semantic annotation of learning materials is an important task in extracting lecture topics [100] and building microlearning content [101]. Our dataset can also be utilized as a source dataset for ontology construction [102, 103] and as a basic unit in knowledge point recognition [104].

Model Training Although one may directly use existing pretrained object detection models, general-purpose detection models trained on natural images (e.g., scenery images) often perform poorly in domain-specific tasks. The detection of design elements in lecture videos is such an example because of the unique characteristics of lecture content, which calls for domain-specific models. Thus we pretrained our model with DocBank, a benchmark dataset of 500K document pages [105]. We chose DocBank because both the DocBank documents and our lecture slides consist of a mix of image and text elements and contain location-sensitive elements such as titles and footers. Once a model is pretrained with DocBank, we fine-tuned the model using our dataset. This transfer learning process improves the performance of object detection models. Without the pretraining step, when we tested with four widely-used deep-learning-based object detection models, namely Faster R-CNN [106], SSD

[1] We used two different ranking measures (i.e., popularity, user reviews) from ClassCentral [65, 66].

[2] https://anonymous.4open.science/r/lecture-design-dataset-F7CF/README.md

based on ResNet [107], EfficientDet [108], and CenterNet [109, 110], the highest *mean Average Precision (mAP)* value (with IoU of 0.5) we obtained was 74% for the CenterNet architecture. The pretraining step with DocBank increased the mAP from 74% to 79%, demonstrating a positive effect of pretraining a model on a document layout dataset for the lecture design detection task. More detailed information about these experiments and hyperparameter information is provided in the supplementary materials. We released both the pretrained model and fine-tuning code as open-source[3]. We expect our model to be used by future researchers as a baseline for lecture design element detection.

Postprocessing for Adaptation We perform two postprocessing analyses for adaptation in the later stage. First, we extract design properties for text elements, including font size, typeface, and font color.[4] The survey and interviews (Section 3) inform us that these properties affect readability. Second, we extract the background of the slides for the reconstruction of slides in the adaptation stage. We used an image in-painting model to remove the detected static objects (e.g., text, figures) from frame images [114], which returns an image without static objects but only the background left.

Dynamic Object Analysis

Lecture videos often contain small objects, such as mouse pointers and handwriting, that dynamically change over time; however, these objects are often not accurately detected by object detection models because of their small sizes [115]. In order to detect these dynamic objects in a video shot, apart from *Static Object Analysis* (Section 4.2.1), we used OpenCV motion analysis module [116]. It matches the area of the detected motions with the object detection results from Section 4.2.1. Details can be found in the supplementary material.

Text-to-Script Mapping

Lastly, we determine mappings between audio and on-screen text so that when adapting content, we can segment a slide that has excessive text into multiple slides while ensuring that we split the slide at the right moment. We developed a rule-based mapping algorithm, which consists of two steps: alignment and grouping.

Alignment Stage We first identify alignments between the on-screen text and the transcript based on two factors: progressive disclosure and semantic similarity. If one new text element appears in a video (e.g., bullet point) and it is currently explained by an instructor [117, 118, 29], we create a mapping between the newly disclosed element and current narration. Otherwise, we calculate the semantic similarity between a narrated sentence and every text element in the same video frame using Sentence-BERT [119]. After that, we use a *bipartite graph matching* algorithm [120, 121] to find optimal mappings. Details and thresholds can be found in the Supplementary Materials.

Grouping Stage Once we find element-level mappings, we combine text elements into units that need to be displayed to learners at once in a single slide. We implement two rules in grouping the element-level mappings. First, we consider the linearity of lecturing. For example, if an instructor does not mention the text elements in a linear order (i.e., from top to bottom, from left to right), we merge all non-linearly mentioned elements into a group.

[3] https://anonymous.4open.science/r/lecture_design_detection-5134/README.md

[4] We used Pytesseract OCR [111], DeepFont [112], and color clustering techniques [113] to extract font size, typeface, and font color, respectively.

Second, we merge multiple text elements that are concurrently mentioned by an instructor. If an instructor mentions multiple text elements in a single sentence, then learners should be able to refer to all the elements at once.

4.2.2 Adaptation Stage

In this stage, the decomposed elements are adapted through multiple strategies. The adaptation stage consists of three modules: (1) local content adaptation, (2) layout adaptation, and (3) global content adaptation. The first and second modules are designed to maximize the guideline compliance rate for mobile learning, while the global adaptation is designed to refine the adapted results to be consistent and coherent.

Local Content Adaptation

We locally adapt content according to design guidelines from the literature (see Appendix). For font size, image size, and line spacing, we enlarge them until they meet the guidelines. For the typeface, we change the handwriting, script, and serif fonts to sans-serif fonts. For color contrast between the fonts and background, we adopt the closest color to the original color of fonts, that exceeds the guidelines threshold. Lastly, the amount of text is adjusted if it exceeds the threshold, being segmented into multiple slides. However, we do not segment a single text box or atomic unit from Section 4.2.1 even if it violates the guidelines. The guideline and exact thresholds adopted from the guidelines are included in Appendix.

Layout Adaptation

When optimizing the structures of the visual elements for mobile devices, a design compromise is required since the guidelines often cannot be fully achieved for every element due to the limited screen space. If there is only one type of design element (e.g., text-only or image-only frames), we can easily enlarge the elements at the same rate to the extent to which they meet the guidelines without overlaps. However, if there are different types of elements in a frame, we have to choose an element to prioritize for enlargement. To determine a point where the overall guideline compliance rate is maximized, we use a constrained optimization technique [122]. We define the objective function as follows:

$$\min \left(\frac{c-x}{c}\right)^2+\left(\frac{c-y}{c}\right)^2, \text{ where } 0<x<x_{\max} \text{ and } 0<y<y_{\max}. \tag{4.1}$$

In this function, x and y represent the average font size of text elements and image elements and c indicates a threshold from the font size guideline. $x_{\max}$ is the largest available font size of text without resizing image elements, and $y_{\max}$ is that in image elements without resizing text elements. Our algorithm aims to minimize deviations between the guideline and content, thereby maximizing the overall guideline compliance rate. We consider the available space on the screen to set constraints for font sizes. To determine x_{max} and y_{max}, we fix the size of one type of element and enlarge the other type of element as long as there is no overlap. The two solutions that minimize the optimization function determine the compromised sizes and locations for all elements. The details of the enlargement and locating methods are in Supplementary Materials.

Lastly, we reconstruct a column layout inspired by the concept of the content reflow in responsive web design which converts multiple columns into a smaller number of columns to fit the width of viewport of devices [123, 124]. We first extract a column layout and reading orders of learning materials [125]. We then determine a

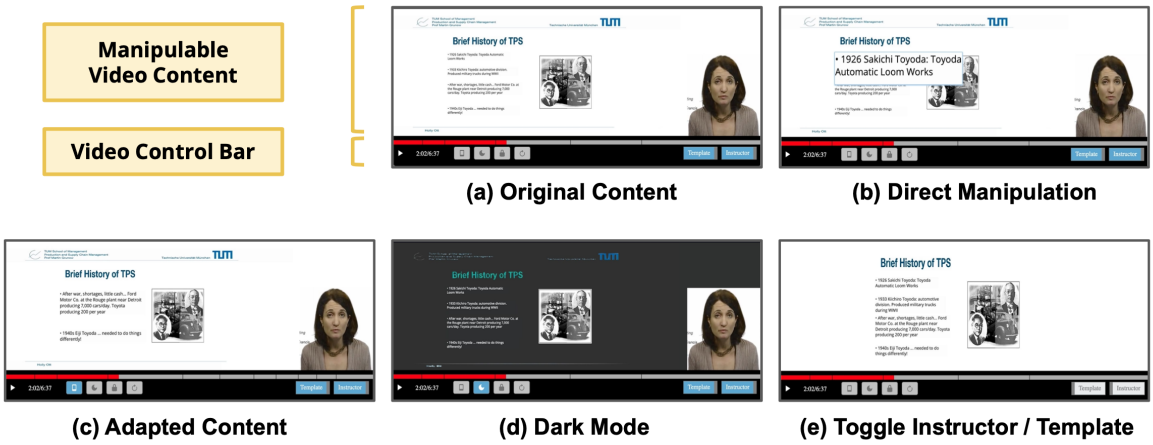

Figure 4.2: A learner can resize, reposition, and toggle various in-video elements using FitVid's video UI. (a) Original Content: original content without adaptation is displayed to the learner by default; (b) Direct Manipulation: the learner can resize and reposition design elements (e.g., text boxes, images, and talking-head instructors) using touch and drag interactions; (c) Adapted Content: the learner can view the adapted content obtained from the automated pipeline; (d) Dark Mode: the learner can toggle between the dark background and bright text of video content; (e) Toggle Instructor and Template: the learner can turn on and off the talking-head instructor view and the slide template (e.g., university logos in headers or headers).

final column layout by adopting the layout that has a higher guideline compliance rate with a larger average font size between the original layout and the converted ones with content reflow.

Global Content Adaptation

The global content adaptation stage refines the local adaptation results in consideration of the consistency of designs. Specifically, we consider font size of title, runt, aspect ratio of images, progressive disclosure, and positional word. The detailed implementation can be found in Supplementary Materials.

4.3 Video Interface

We designed and developed an interactive video player of FitVid, depicted in Fig. 4.2. It renders the automatically adapted results and further supports direct manipulation and content customization.

Direct Manipulation As automatically generated results may produce incorrect adaptations, AI-powered systems are recommended to support the correction of the automated results [95, 126]. FitVid allows users to edit and refine the automated adaptation results through direct manipulation. As shown in Fig. 4.2-**a** and **b**, a learner can directly resize and reposition in-video elements, such as text boxes, images, and talking-head instructions, while watching the videos using touch and drag interactions. In particular, we chose the drag interaction over pinch-zoom interaction for resizing elements since learners suggested an element-wise zoom feature that does not result in cut-offs of part of the content.

Content Customization Based on the survey and interviews, we provide customization options for users to determine whether to display talking-head instructors, slide templates, and background colors. For instance, Fig. 4.2-**e** shows the result of turning off the option for displaying the talking-head instructor. Learners can also use the dark theme that provides bright fonts in dark backgrounds (at Fig. 4.2-**d**).

The control bar of the player shown at the bottom of the interface includes six buttons: mobile-friendly mode, dark theme, lock, refresh, template toggle, and talking-head instructor toggle. The lock button disables direct manipulation in case a learner does not want to manipulate content (e.g., when holding a phone when on the move). Learners can disable the mobile-friendly mode to access the original content before the adaptation. This is to allow users to easily dismiss system-generated results as suggested by the human-AI interaction guideline [95]. The dark gray bars on the video timeline in the play bar indicate the shot transitions, which mostly correspond to lecture slide transitions. The player runs on browsers and is implemented using JavaScript and CSS media queries.

4.4 Pipeline Evaluation

We evaluated the performance of our computational pipeline through content analysis and content evaluation study. We also report representative error cases with examples. For the evaluation, we sampled 53 video frames from 24 videos. Specifically, two of the authors sampled our dataset to include diverse design factors that are not uniformly distributed across videos. The evaluation results of each submodule and the list of sampled videos from the dataset are included in the supplementary material.

4.4.1 Content Analysis

We conducted the content analysis to compare the design guideline compliance rate between before and after the adaptation (Table 4.1). In Table 4.1, the number of target cases indicates how many video frames initially violate the guidelines. The number of adapted cases is the number of video frames successfully adapted to satisfy the guidelines. For the word count and font sizes, the guideline compliance rate increased by 24% and 8%, respectively. For the font sizes in images, the compliance rate changed from 13% to 17%. For typefaces, line spacing, and color contrast, all the target cases of adaptation were successfully adapted to comply with the guidelines. Fig. 4.3 shows the examples adaptation results. In Fig. 4.3 (a), the fonts are enlarged and the text is trimmed down. Fig. 4.3 (b) shows an example of converting the column design with content reflowing. In Fig. 4.3 (c), the adaptation algorithm finds the balance between enlarging text and images containing text, compromising the compliance rate of each element. Fig. 4.3 (d) demonstrates the increased color contrast. Lastly, Fig. 4.3 (e) shows an example of typeface adaptation from handwriting to a sans-serif font. Overall, the results demonstrate that our pipeline is applicable to various types of lecture content designs.

Meanwhile, we identified three representative failure cases. First, an overlap occurs due to the errors of the object detection model. For example, in Fig. 4.3 (f), the checkbox beside the text box 'B) Quadruped' was not detected by the model and caused an overlap. Second, four cases did not satisfy the font size guidelines even after the adaptation (Table 4.1). The issue derived from one of the global adaptation rules, "Font size of the title should be larger than that of the rest of the text elements.". If the size of the title does not comply with the guidelines (Fig. 4.3 (g)), the other text elements also fail to meet the guidelines. Third, the number of words could not be reduced below the threshold since multiple text boxes were grouped as atomic units by the text-to-script mapping module.

	Original Content	Adapted Content	# of Target Cases	# of Adapted Cases
Word Count	0-20 words: 24% 20-45 words: 40% Above 45 words: 36% Average: 38.58 words	0-20 words: 39% 20-45 words: 49% Above 45 words: 12% Average: 26.14 words	20	20
Font Size	1-16 pt: 86% 16-28 pt: 14% Above 28 pt: 0% Average: 11.19 pt	1-16 pt: 78% 16-28 pt: 22% Above 28 pt: 0% Average: 12.04 pt	35	31
Typeface	Serif, Script, Handwritten	Sans-serif	15	15
Line Spacing	Average: 127.5%	Average: 150%	2	2
Font Size in Images	1-16pt: 88% 16-28pt: 13% Above 28pt: 0% Average: 9.72 pt	1-16pt: 83% 16-28pt: 17% Above 28pt: 0% Average: 11.43 pt	19	19
Color Contrast	5.03	7.0	6	6

Table 4.1: Statistics of design elements from original content and adapted content. The result demonstrates that the adaptation pipeline improves the design guidelines.

4.4.2 Content Evaluation Study

The content evaluation study evaluated perceived readability and ratings of the adaptation results by comparing them with the original content. This large-scale evaluation can reveal how general users perceive the adaptation results for the same content in the previous section.

Participants

We recruited 198 respondents (127 female, 70 male, 1 prefer not to specify) with ages ranging from 20 to 69 (20-29: 86, 30-39: 93, 40-49: 12, 50-59: 5, 60-69: 2) through Amazon Mechanical Turk (AMT). They were provided with USD 0.9 for a 10-minute long survey for a Human Intelligence Task (HIT) on AMT. Participants' highest level of education was as follows: less than high school (1.0%), high school degree (11.7%), some college (9.2%), bachelor's degree (62.6%), Master's degree (0.5%), and graduate degree (15.0%). We ensured that all respondents evaluated the content adaptation results on their mobile device by asking them to upload a mobile screen capture of their survey response screen.

Method

We designed a content evaluation survey by utilizing the adaptation results in Section 4.4.1. The study began with an introduction of the experiment with instructions, and a demographic questionnaire followed. We then presented five pairs of the original and adapted content in sequence using a paired comparison method [127]. The participants were required to rate the subjective readability and design satisfaction for six design elements: font sizes, amount of text, typeface, line spacing, image sizes, color contrast between fonts and background. The

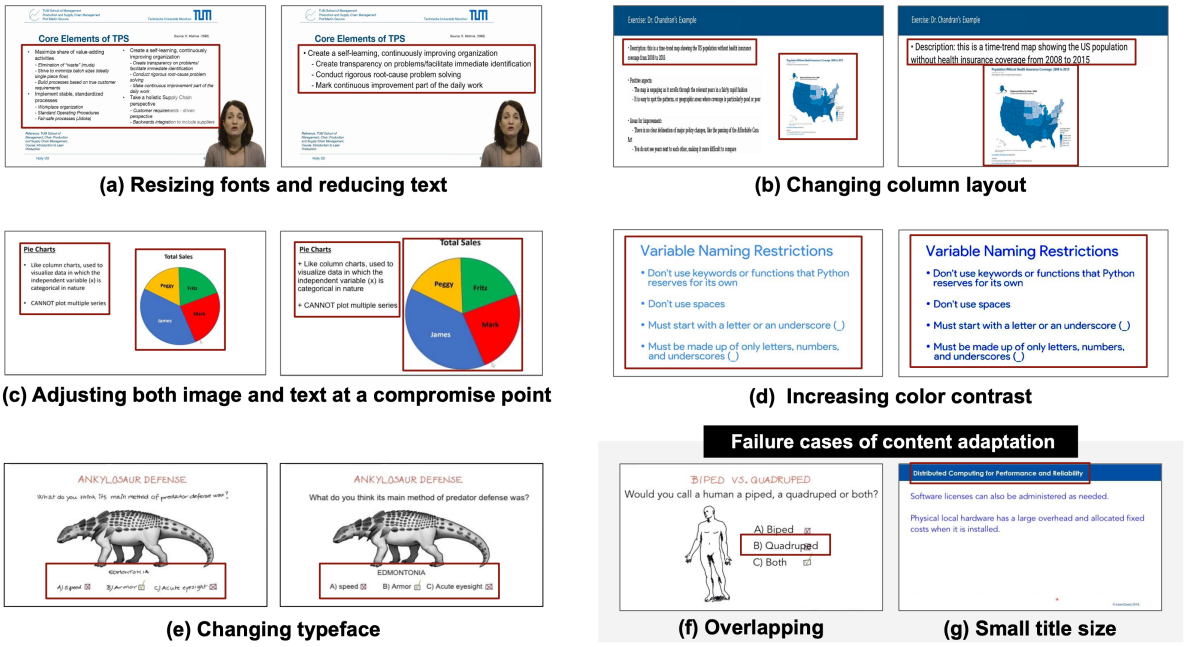

Figure 4.3: Examples of adapted content. (a) Resized fonts and reduced text, (b) Changed column layout, (c) Adjusted both image and text at compromise point, (d) Increased color contrast, (e) Changed typeface. The representative failure cases are: (f) overlappings due to errors from the object detection stage and (g) incomplete text resizing due to its comparative size with titles

question was on a 7-point ordinal scale and included the 'Not Applicable' option. The order of the presented content and condition was randomized. We published a total of 10 HITs on AMT, three with six pairs and seven with five pairs of comparisons. Our form also included an attention check question [128] to filter invalid responses.

Results

We initially collected 401 responses and removed 151 responses for invalid attention check answers, 44 for invalid screen capture, and 7 as outliers beyond 2 standard deviations from the mean [129]. Finally, we had valid responses from 198 participants with at least 16 ratings for each pair of the original and the adapted content.

We tested the internal consistency of the responses using Cronbach's alpha [130], and it showed high reliability with 0.78 on average (min: 0.63, max: 0.90). Thus, we averaged participants' ratings for each item. We then conducted a Wilcoxon signed-rank test for analysis due to the ordinal nature of scales. On a 7-point scale question (1: very poor, 7: very good), the participants rated that the adapted content is significantly more satisfactory than the original content on all seven design elements: font size ($p < 0.0001$), amount of text ($p < 0.0001$), typeface ($p < 0.0001$), line spacing ($p < 0.0001$), image size ($p < 0.0001$), and color contrast ($p < 0.0001$) (Table 4.2). They also rated the readability of the adapted content significantly higher than the original content ($p < 0.0001$). The result showed that content adaptation improves the readability and design satisfaction for general users.

	Original Content		Adapted Content		
	M	**SD**	**M**	**SD**	**p-value**
Font Size	5.10	1.19	5.83	0.75	**<.0001**
Amount of Text	5.38	1.04	5.76	0.80	**<.0001**
Typeface	5.26	1.02	5.65	0.82	**<.0001**
Line Spacing	5.41	1.04	5.76	0.80	**<.0001**
Image Size	5.46	0.91	5.78	0.81	**<.0001**
Color Contrast	5.60	0.95	5.88	0.79	**<.0001**
Readability	5.36	1.01	5.91	0.73	**<.0001**

Table 4.2: Subjective content rating results demonstrate that FitVid significantly increases the users' design satisfaction. Significant p-values are in bold.

4.5 User Study

In this section, we evaluate the users' learning experience and perceptions using our system. In addition to the quantitative study in the previous section, we investigate the usage cases in a learning situation. We conducted a controlled user study that compares FitVid's interface with the baseline interface without content adaptation and customization. We designed our study to answer the following research questions:

- RQ1. How does FitVid's automated content adaptation impact perceive readability and design satisfaction compared to the baseline video interface?
- RQ2. How do users use and benefit from FitVid's direct manipulation?
- RQ3. How does FitVid affect learning experience, concentration, and cognitive demand compared to the baseline video interface?

The study was a within-subjects design, where each participant used two different video players: (1) baseline interface and (2) FitVid with adapted content and UI that provides direct manipulation and content customization. To maintain the uniformity in the look and feel of both interfaces, the baseline used the same interface design as our system. We selected two videos each from two courses considering the diversity of lecture designs (C1: Lean Production (edX), C2: Essential Epidemiologic Tools for Public Health Practice (Coursera)). Each video has a similar length (C1: 6:37, 10:22, C2: 7:50, 7:44) in slide-based lecture type.

4.5.1 Participants

We recruited 31 participants [P1-P31] (15 male and 16 female) through social media posting. They were college students, graduate students, and office workers who had a prior mobile learning experience. They received USD 15 for up to 70 minutes of participation.

4.5.2 Procedure

The study was conducted remotely using Zoom, and the informed consent was collected via email. We first introduced the interface of our system. The participants then familiarized themselves with our system for as long as they wanted. After the exploration, the participants were required to watch two lectures using two different

video players in counterbalanced order. They were randomly assigned to watch two videos from one of the two courses. After the watching session, we interviewed their perception of each video player and the reasons behind their manipulations. They then completed a questionnaire on difficulty, cognitive load, concentration, easiness to use, readability, learning efficiency, and learning experience for each interface. We used three readability questions from existing work (e.g., design choices made reading harder (fonts, colors, etc.) [131, 23]. The questionnaire also includes scoring four design elements of content: the size of content, amount of content, line spacing, and typeface. The complete questionnaire is included in Supplementary Materials.

4.5.3 Results

We summarize the results and describe the main findings with a focus on the research questions, system usage patterns, and our system's usefulness.

RQ1. How does FitVid's automated content adaptation impact the perceived readability and design satisfaction compared to the baseline video interface?

Most of the participants (28/31) watched the video primarily with the 'mobile mode' on, which provides the adapted content. Except for one participant (P21), all participants expressed willingness to use the mobile mode for their daily mobile learning. P21 commented that he would not use mobile devices for video learning due to their limited screen sizes, even with mobile mode.

To analyze the survey results, we used a Wilcoxon signed-rank test due to the ordinal nature of Likert-type scales. For three readability questions, we tested internal consistency using Cronbach's alpha [130], which was all higher than 0.65. Thus we used the average of the three responses. On a 7-point Likert scale question (1: strongly disagree, 7: strongly agree), the participants reported that the adapted content is significantly more readable compared to the original content ($W = 9$, $p < 0.0001$). They rated that the adapted content is significantly more appropriate for mobile devices in size of content ($W = 6.5$, $p < 0.0001$), amount of content ($W = 8$, $p < 0.0001$), line spacing ($W = 3.5$, $p < 0.0001$), and typeface ($W = 19$, $p < 0.01$). The interview responses confirmed the survey results. P24 emphasized the increased readability, "I would have quit watching the lecture without the mobile mode with good readability.". On the other hand, P2 mentioned that he has astigmatism and stated that "The original content was almost painful to watch, while the adapted content was readable enough.". Meanwhile, some learners had concerns about missing necessary content since adaptation sometimes reduced the amount of content. However, they noted that they could check there is no missing content after referring to the original content by turning off the mobile mode.

RQ2. How do users use and benefit from FitVid's direct manipulation?

The mean number of direct manipulation interactions (i.e., number of touch interactions) was 21 (min: 0, max: 84) for a single video. Almost all participants (30/31) were willing to use direct manipulation in their daily mobile learning, while one participant said that he does not need direct manipulation if he can have the adapted content with sufficient readability. Based on the post-interview, we identified three high-level reasons for using direct manipulation. Fig. 4.4 shows the behaviors of manipulation, including resizing, repositioning, and highlighting. We report the reasons behind the manipulations below, which include adjusting the design, promoting concentration, and interacting with content (Table 4.3).

To adjust the design. The primary reason for enlarging elements was to improve the readability. The participants noted that the direct manipulation allows them further customize the automated adaptation results. For example, P18 stated that "Although the adaptation optimizes the content sizes, I sometimes wanted to see them bigger. In that case, I utilized the direct manipulation." When asked to compare the direct manipulation with the

Category of Reasons	Reasons for Using Direct Manipulation	Participants	Direct Manipulation
To adjust design	To enlarge content for improved readability and legibility	20/31	resize
	To arrange content into preferred layout	3/31	reposition
To promote concentration	To put away unnecessary content that learner finishes watching and focus on current important content	7/31	resize, reposition
	To highlight where instructor is explaining	3/31	resize, highlight
	or emphasizing to better memorize content	2/31	resize
To interact with content	to enjoy interaction itself	5/31	resize, reposition

Table 4.3: Reasons behind direct manipulation usage. Users used the direct manipulation to adjust the design, promote concentration, and interact with content.

pinch-zoom interaction that is supported by most video interfaces, the participants noted that "direct manipulation does not cut off the content, allowing to see the whole part of the zoomed content." (P15). They also mentioned the convenience of zooming in on the individual content selectively.

To promote concentration. The participants manipulated the content to aid their concentration. They used direct manipulation to put away the content they finished watching, highlight important parts, and better memorize the materials. Some participants put away elements one by one as they finished watching them. Others merely touched an element with its edges highlighted to mark where they are reading or focusing. P4 elaborated that "I could focus on the items on screen more easily when I am manipulating them.".

To interact with content. Several participants enjoyed the feeling of interaction itself. They expressed excitement about the interactivity, for example, "I have never seen this feature in existing video players, so it was interesting to use it [direct manipulation]." (P27). P22 stated that "[Using the baseline player,] I found the lecture boring. The increased interactivity by moving and touching the element made it more engaging and less tedious.".

RQ3. How do users use and benefit from FitVid's design options feature?

Overall, the participants appreciated the design options that allow customization of lecture design. Almost all participants (30/31) want to use the design options feature for their daily mobile learning. For the toggle instructor feature, 58% of the participants watched the video with the instructor not displayed on the screen. They turned off the instructor display for two reasons. First, they tried to make the most of the mobile screen space since they could have larger text and images without the instructor. Second, they found it easier to focus on the main content without the instructor. They explained that talking head split their attention. Other participants had different opinions, noting that the presence of an instructor gives a feeling of engagement and two-way communication.

For the toggle template, 93% of the participants watched the video with the template hidden. They wanted to remove the irrelevant content to focus on the main content and increase the readability of the main content by utilizing the space taken by the template.

For the dark mode, 58% of participants watched the video in dark mode. They highlighted the reduced eye fatigue using it. For example, P24 said that "The video content basically does not support the dark theme, so I did not watch a video before going to bed in dim light. But now I can watch them in comfort.". Other participants switched original and dark themes for a change, refreshing themselves with a different color theme. Meanwhile,

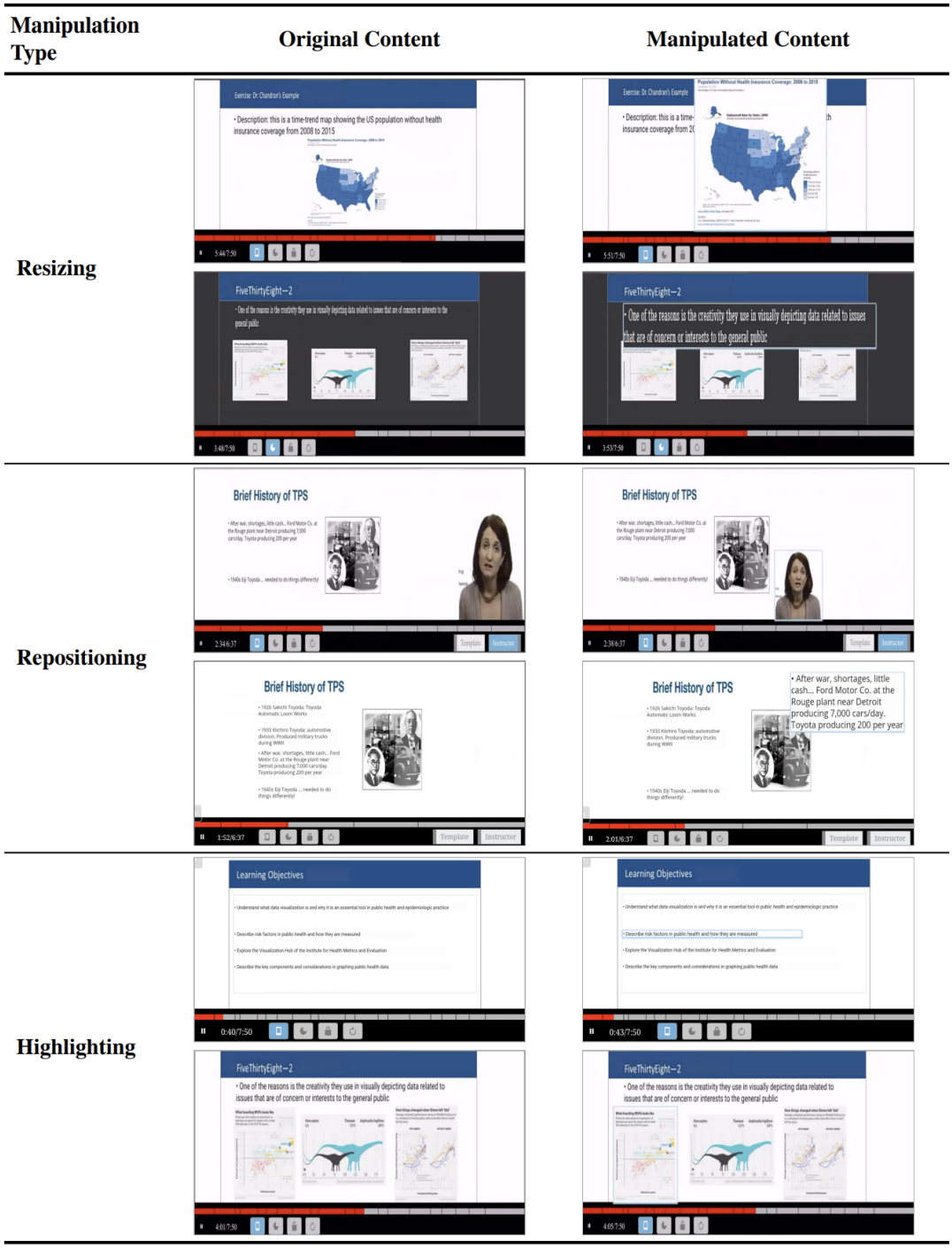

Figure 4.4: Manipulation types observed in the user study, including resizing, repositioning, and highlighting.

other participants indicated that they did not need it in bright environments and preferred dark fonts in a bright background.

RQ4. How does FitVid affect learning experience, concentration, and cognitive demand compared to the baseline video interface?

The videos used in the study had a similar reported level of difficulty ($W = 56$, $p > 0.05$). For a 7-point scale question, FitVid significantly improved learning experience ($W = 12.5$, $p < 0.0001$) and learning efficiency ($W =$

20.5, $p < 0.0001$). We also found a significant difference on the levels of concentration ($W = 35$, $p = 0.0002$). The participants reported greater willingness to use FitVid in their daily mobile learning ($W = 14.5$, $p < 0.0001$) and greater easiness to use ($W = 39.5$, $p < 0.001$) compared to baseline. However, FitVid turns out to be significantly more confusing to use compared to baseline ($W = 41$, $p < 0.05$). However, there was no significant difference in cognitive load ($W = 133$, $p = 0.0680$) and FitVid significantly decreased the frustration level while watching the video ($W = 45.5$, $p = 0.0001$).

Most participants expressed that our system was beneficial for their learning. Some participants reported that the increased readability made it easy for them to follow the lecture. Other participants noted the benefits of manipulating talking-head instructor, "I felt more engaged and focused after I moved the position of the instructor to the center of the screen." (P11). P12 said that "I felt like I have more control over my learning because I could adjust the content in my own way.".

Concerning the cognitive demand, some participants noted that reading text with FitVid was less demanding with more legible fonts compared to the original content. They appreciated that they could remove unwanted design elements, including the instructor, template, and cursive typefaces, which can pose an unnecessary cognitive load. On the other hand, some participants stated that they needed time to get used to manipulating content in real-time while watching the video.

Chapter 5. Discussion

We investigated mobile video-based learning using a mixed-method analysis on learners, MOOC content, and video production engineers. Based on the results of the mixed-method analysis, we built a system for responsive and customizable video content adaptation. In this section, we discuss extended design implications and future work.

5.1 Gap between Learners and Engineers

First, several design factors need to be paid more attention to in the design process. As shown in Table 5.1, engineers may be paying relatively little attention to complex images, images containing text, and too bright color. The learners complained about complex images with intricate details and small unreadable text usually accompanied by complex images. We recommend adjusting text design when bringing images containing text from other sources such as textbooks instead of copying and pasting them directly. We also encourage engineers to crop or zoom in on the area of interest in the complex image. Meanwhile, the video production engineers try to make lectures more engaging by adding video effects. However, readability was more critical than engagement for mobile learners. On the other hand, engineers were unaware or had not considered SIIDs in the mobile learning environments, implying the need to extend existing guidelines, education, or design tools for engineers. Second, some design factors did not meet learners' needs despite the engineers' considerations. For example, the small font size was learners' primary pain point, although they were also engineers' main consideration, with a guideline compliance rate of 2.79%. The engineers' challenges resulting in low compliance rates include the difficulty in considering diverse mobile screen sizes (e.g., smartphones, tablet PCs, and laptops), not having design guidelines for mobile environments, coordinating with instructors, and deciding design directions with limited understanding of the lecture content.

5.2 Responsive design for various display settings

Although our work mainly focused on smartphone devices, FitVid can be extended to support various display settings, such as tablets, smartwatches, and very large screens. For example, content can be optimized for smartwatches by compressing information into keywords or short summaries. FitVid's pipeline can also support large screens or multiple screens. For example, for a video lecture to be displayed on a large screen in a conference hall, the system can adjust the font size to be visible to audiences or combine content across multiple video shots into a single slide for effectively presenting the information. On the other hand, a learner with a dual monitor can have one screen dedicated to lecture slides and the other to talking-head instructors, with a customized layout.

5.3 Advanced accessibility with content customization

The idea of content customization can be used to increase the accessibility of videos in various contexts. FitVid can readily generate adapted content for specific populations, such as for low vision [141, 142], older adults [143], and dyslexia [144, 145] populations, based on the design guidelines for these populations. For

Visual Design Element	Pain Point	Learners		Engineers	Guideline	Prior Work
		Surveys	Interviews	Interviews	Compliance Rate	
Text Element	Small Font Size	63/134	19/21	O	2.79%-24.5%	CL [14], JOLs [17, 18], DG [132, 133, 72, 69, 134] [14, 70, 135]
	Dense Text	36/134	8/21	O	74.2%-84.9%	CL [14, 15, 16], IO [32], DG [74, 133, 72, 69, 136]
	Inappropriate Text Spacing	18/134	4/21			DG [135]
	Inappropriate Font Style	14/134	0/21			DG [137, 133, 72, 69, 14] [70, 135, 136]
Image Element	Images Containing Text		9/21		0.94%-5.00%	DG [132, 133, 72, 69, 134], [14, 70, 135]
	Complex Images	32/134	7/21			CL [34, 35, 33], DG [138, 14, 139]
	Small Image Size	51/134	6/21	O		DG [133, 139]
	Too Many Images	23/134	3/21	O		CL [14], IO [32], DG [134, 14]
Color	Low Color Contrast	19/134	3/21			DG [133, 140, 72, 69], [14, 135, 136]
	Too Bright Color	15/134	2/21			DG [140, 72, 69, 135, 136]

Table 5.1: Triangulation of learner-reported difficulties, design considerations of video production engineers, and guideline compliance rate of current MOOC content, including ties to existing learning framework. *CL: Cognitive Load, JOLs: Judgements Of Learning, IO: Information Overload, DG: Design Guidelines.

example, we provide color settings for color blindness by providing them with an inclusive color palette, instead of the dark theme we provided for mobile learners [146, 147]. Our work can be an initial step toward enhancing the accessibility of visual video content for various user groups with different ability profiles.

5.4 Ecosystem of mobile-friendly video content

While this work targets learners, FitVid can potentially benefit other stakeholders, including instructors and learning platform engineers. For example, our automated adaptation technique can be helpful for instructors to create mobile-friendly videos without receiving help from people with professional video editing skills. FitVid can also benefit learning platform engineers by automatically rendering content to fit mobile devices and providing design options to learners without additional development work.

Chapter 6. Conclusion

6.1 Conclusion

This thesis argues that the current video-based learning content is not accessible and digestible for users in diverse contexts. In response, this thesis investigates the challenges of learners in mobile environments and presents an interactive system that supports responsive and fluid adaptation of video content.

This thesis makes contributions in 3 areas: (1) an identification of mobile learners' challenges and their gap with video production engineers' design considerations, (2) an automated pipeline that generates mobile-friendly content adaptation, and (3) a design and implementation of a system that provides learners with responsive and customizable video content.

Curriculum Vitae

Name : Jeongyeon Kim

E-mail : imurs4825@gmail.com

Educations

2015. 3. – 2019. 8. Sogang University (B.S. in CSE)

2019. 9. – 2022. 2. Korea Advanced Institute of Science and Technology (M.S. in CS)

Publications

1. **Jeongyeon Kim**, Yubin Choi, Minsuk Kahng, Juho Kim, "FitVid: Responsive and Flexible Video Content Adaptation", *Accepted with minor revisions for CHI '22*
2. **Jeongyeon Kim**, Yubin Choi, Meng Xia, Juho Kim, "Mobile-Friendly Content Design for MOOCs: Challenges, Requirements, and Design Opportunities", *Accepted with minor revisions for CHI '22*
3. Seoyoung Kim, Donghoon Shin, **Jeongyeon Kim**, Soonwoo Kwon, Juho Kim, "How Older Adults Use Online Videos for Learning", *Under R&R CHI '22*
4. Tae Soo Kim, Nitesh Goyal, **Jeongyeon Kim**, Juho Kim, Sungsoo Ray Hong, "Supporting Collaborative Sequencing of Small Groups through Visual Awareness", *CSCW '21*
5. **Jeongyeon Kim** Juho Kim, "Guideline-Based Evaluation and Design Opportunities for Mobile Video-Based Learning", *CHI '21 LBW*
6. **Jeongyeon Kim**, Junyong Park, I-Hao Lu, "HyperButton: In-video Question Answering via Interactive Buttons and Hyperlinks", *CHI '21 Asian CHI Symposium*
7. **Jeongyeon Kim** Juho Kim, "FitVid: Towards Development of Responsive and Fluid Video Content Adaptation", *AAAI '21 Workshop on Imagining Post-COVID Education with AI*
8. Tae Soo Kim, Sungsoo Ray Hong, Nitesh Goyal, **Jeongyeon Kim**, Juho Kim, "Consensus Building in Collaborative Sequencing with Visual Awareness", *CHI '20 LBW*
9. **Jeongyeon Kim**, Yoonah Lee, Inho Seo, "Math Graphs for the Visually Impaired: Audio Presentation of Elements of Mathematical Graphs", *CHI '19 SRC*

Bibliography

[1] Aziz Hasanov, Teemu H Laine, and Tae-Sun Chung. A survey of adaptive context-aware learning environments. *Journal of Ambient Intelligence and Smart Environments*, 11(5):403–428, 2019.

[2] Hung-Tao M Chen and Megan Thomas. Effects of lecture video styles on engagement and learning. *Educational Technology Research and Development*, pages 1–18, 2020.

[3] Konstantinos Chorianopoulos. A taxonomy of asynchronous instructional video styles. *International Review of Research in Open and Distributed Learning*, 19(1), 2018.

[4] Jared Danielson, Vanessa Preast, Holly Bender, and Lesya Hassall. Is the effectiveness of lecture capture related to teaching approach or content type? *Computers & Education*, 72:121–131, 2014.

[5] Christina Ilioudi, Michail N Giannakos, and Konstantinos Chorianopoulos. Investigating differences among the commonly used video lecture styles. 2013.

[6] Ozlem Ozan and Yasin Ozarslan. Video lecture watching behaviors of learners in online courses. *Educational Media International*, 53(1):27–41, 2016.

[7] Aziz Naciri, Mohamed Amine Baba, Abderrahmane Achbani, and Ahmed Kharbach. Mobile learning in higher education: Unavoidable alternative during covid-19. *Aquademia*, 4(1):ep20016, 2020.

[8] José-María Romero-Rodríguez, Inmaculada Aznar-Díaz, Francisco-Javier Hinojo-Lucena, and Gerardo Gómez-García. Mobile learning in higher education: Structural equation model for good teaching practices. *IEEE Access*, 8:91761–91769, 2020.

[9] edX Inc. Take edx on the go, 2021.

[10] Coursera Inc. Coursera: Download on ios and android, 2021.

[11] Khan Academy Inc. Downloads: Khan academy, 2021.

[12] Udemy Inc. Learn anywhere with udemy for ios and android, 2021.

[13] Zhanna Sarsenbayeva, Niels van Berkel, Chu Luo, Vassilis Kostakos, and Jorge Goncalves. Challenges of situational impairments during interaction with mobile devices. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, pages 477–481, 2017.

[14] Petra J Lewis. Brain friendly teaching—reducing learner's cognitive load. *Academic radiology*, 23(7):877–880, 2016.

[15] John Sweller. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, 4(4):295–312, 1994.

[16] John Sweller, Jeroen JG Van Merrienboer, and Fred GWC Paas. Cognitive architecture and instructional design. *Educational psychology review*, 10(3):251–296, 1998.

[17] Vered Halamish. Can very small font size enhance memory? *Memory & cognition*, 46(6):979–993, 2018.

[18] Matthew G Rhodes and Alan D Castel. Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of experimental psychology: General*, 137(4):615, 2008.

[19] Tanya Elias. 71. universal instructional design principles for mobile learning. *International Review of Research in Open and Distributed Learning*, 12(2):143–156, 2011.

[20] Jiyou Jia and Bilan Zhang. Design guidelines for mobile mooc learning—an empirical study. In *International Conference on Blended Learning*, pages 347–356. Springer, 2018.

[21] Charles Miller and Aaron Doering. *The new landscape of mobile learning: Redesigning education in an app-based world*. Routledge, 2014.

[22] Natalia Spyropoulou, Christos Pierrakeas, and Achilles Kameas. Creating mooc guidelines based on best practices. *Edulearn14 Proceedings*, 20:6981–6990, 2014.

[23] Qisheng Li, Meredith Ringel Morris, Adam Fourney, Kevin Larson, and Katharina Reinecke. The impact of web browser reader views on reading speed and user experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[24] Vivek Bhuttoo, Kamlesh Soman, and Roopesh Kevin Sungkur. Responsive design and content adaptation for e-learning on mobile devices. In *2017 1st International Conference on Next Generation Computing Applications (NextComp)*, pages 163–168. IEEE, 2017.

[25] Wenhui Peng and Yaling Zhou. The design and research of responsive web supporting mobile learning devices. In *2015 International Symposium on Educational Technology (ISET)*, pages 163–167. IEEE, 2015.

[26] Meltem Huri Baturay and Murat Birtane. Responsive web design: a new type of design for web-based instructional content. *Procedia-Social and Behavioral Sciences*, 106:2275–2279, 2013.

[27] Nicholas Vanderschantz, Claire Timpany, and Annika Hinze. Design exploration of ebook interfaces for personal digital libraries on tablet devices. In *Proceedings of the 15th New Zealand Conference on Human-Computer Interaction*, pages 21–30, 2015.

[28] Hyeungshik Jung, Hijung Valentina Shin, and Juho Kim. Dynamicslide: Exploring the design space of reference-based interaction techniques for slide-based lecture videos. In *Proceedings of the 2018 Workshop on Multimedia for Accessible Human Computer Interface*, pages 33–41, 2018.

[29] Baoquan Zhao, Songhua Xu, Shujin Lin, Ruomei Wang, and Xiaonan Luo. A new visual interface for searching and navigating slide-based lecture videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 928–933. IEEE, 2019.

[30] May Kristine Jonson Carlon, Nopphon Keerativoranan, and Jeffrey S Cross. Content type distribution and readability of moocs. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*, pages 401–404, 2020.

[31] Philip J Guo, Juho Kim, and Rob Rubin. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 41–50, 2014.

[32] Mohamed Ally. Using learning theories to design instruction for mobile learning devices. *Mobile learning anytime everywhere*, pages 5–8, 2005.

[33] Hyunjeong Lee, Jan L Plass, and Bruce D Homer. Optimizing cognitive load for learning from computer-based science simulations. *Journal of educational psychology*, 98(4):902, 2006.

[34] Simon Harper, Eleni Michailidou, and Robert Stevens. Toward a definition of visual complexity as an implicit measure of cognitive load. *ACM Transactions on Applied Perception (TAP)*, 6(2):1–18, 2009.

[35] Qiuzhen Wang, Sa Yang, Manlu Liu, Zike Cao, and Qingguo Ma. An eye-tracking study of website complexity from cognitive load perspective. *Decision support systems*, 62:1–10, 2014.

[36] Fatt Cheong Choy. Digital library services: towards mobile learning. 2010.

[37] Rabail Tahir and Fahim Arif. A measurement model based on usability metrics for mobile learning user interface for children. *The International Journal of E-Learning and Educational Technologies in the Digital Media*, 1(1):16–31, 2015.

[38] Andrew Cross, Mydhili Bayyapunedi, Dilip Ravindran, Edward Cutrell, and William Thies. Vidwiki: Enabling the crowd to improve the legibility of online educational videos. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1167–1175, 2014.

[39] Jane Hoffswell, Wilmot Li, and Zhicheng Liu. Techniques for flexible responsive visualization design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[40] Aoyu Wu, Wai Tong, Tim Dwyer, Bongshin Lee, Petra Isenberg, and Huamin Qu. Mobilevisfixer: Tailoring web visualizations for mobile phones leveraging an explainable reinforcement learning framework. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):464–474, 2020.

[41] Yi-Hao Peng, JiWoong Jang, Jeffrey P Bigham, and Amy Pavel. Say it all: Feedback for improving non-visual presentation accessibility. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2021.

[42] Miles Thorogood. Slidedeck. js: A platform for generating accessible and interactive web-based course content. In *Proceedings of the 21st Western Canadian Conference on Computing Education*, pages 1–5, 2016.

[43] Shoko Tsujimura, Kazumasa Yamamoto, and Seiichi Nakagawa. Automatic explanation spot estimation method targeted at text and figures in lecture slides. In *INTERSPEECH*, pages 2764–2768, 2017.

[44] Baoquan Zhao, Shujin Lin, Xiaonan Luo, Songhua Xu, and Ruomei Wang. A novel system for visual navigation of educational videos using multimodal cues. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1680–1688, 2017.

[45] Chengpei Xu, Ruomei Wang, Shujin Lin, Xiaonan Luo, Baoquan Zhao, Lijie Shao, and Mengqiu Hu. Lecture2note: Automatic generation of lecture notes from slide-based educational videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 898–903. IEEE, 2019.

[46] Haojin Yang and Christoph Meinel. Content based lecture video retrieval using speech and video text information. *IEEE transactions on learning technologies*, 7(2):142–154, 2014.

[47] Haojin Yang, Maria Siebert, Patrick Luhne, Harald Sack, and Christoph Meinel. Automatic lecture video indexing using video ocr technology. In *2011 IEEE International Symposium on Multimedia*, pages 111–116. IEEE, 2011.

[48] Esha Baidya and Sanjay Goel. Lecturekhoj: automatic tagging and semantic segmentation of online lecture videos. In *2014 Seventh international conference on contemporary computing (IC3)*, pages 37–43. IEEE, 2014.

[49] Kuldeep Yadav, Ankit Gandhi, Arijit Biswas, Kundan Shrivastava, Saurabh Srivastava, and Om Deshmukh. Vizig: Anchor points based non-linear navigation and summarization in educational videos. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 407–418, 2016.

[50] Monica Haurilet, Alina Roitberg, Manuel Martinez, and Rainer Stiefelhagen. Wise—slide segmentation in the wild. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 343–348. IEEE, 2019.

[51] Monica Haurilet, Ziad Al-Halah, and Rainer Stiefelhagen. Spase-multi-label page segmentation for presentation slides. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 726–734. IEEE, 2019.

[52] Ben Shneiderman. Direct manipulation for comprehensible, predictable and controllable user interfaces. In *Proceedings of the 2nd international conference on Intelligent user interfaces*, pages 33–39, 1997.

[53] Edwin L Hutchins, James D Hollan, and Donald A Norman. Direct manipulation interfaces. *Human–computer interaction*, 1(4):311–338, 1985.

[54] Pierre Dragicevic, Gonzalo Ramos, Jacobo Bibliowitcz, Derek Nowrouzezahrai, Ravin Balakrishnan, and Karan Singh. Video browsing by direct manipulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246, 2008.

[55] Thorsten Karrer, Malte Weiss, Eric Lee, and Jan Borchers. Dragon: a direct manipulation interface for frame-accurate in-scene video navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 247–250, 2008.

[56] Thorsten Karrer, Moritz Wittenhagen, and Jan Borchers. Pocketdragon: a direct manipulation video navigation interface for mobile devices. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–3, 2009.

[57] Thorsten Karrer, Moritz Wittenhagen, and Jan Borchers. Draglocks: handling temporal ambiguities in direct manipulation video navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 623–626, 2012.

[58] Cuong Nguyen, Yuzhen Niu, and Feng Liu. Direct manipulation video navigation in 3d. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1169–1172, 2013.

[59] Derek Pang, Sherif Halawa, Ngai-Man Cheung, and Bernd Girod. Mobile interactive region-of-interest video streaming with crowd-driven prefetching. In *Proceedings of the 2011 international ACM workshop on Interactive multimedia on mobile and portable devices*, pages 7–12, 2011.

[60] Wei Song, Dian W Tjondronegoro, Shu-Hsien Wang, and Michael J Docherty. Impact of zooming and enhancing region of interests for optimizing user experience on mobile sports video. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 321–330, 2010.

[61] Ngo Quang Minh Khiem, Guntur Ravindra, Axel Carlier, and Wei Tsang Ooi. Supporting zoomable video streams with dynamic region-of-interest cropping. In *Proceedings of the first annual ACM SIGMM conference on Multimedia systems*, pages 259–270, 2010.

[62] Carlier Axel, Guntur Ravindra, and Ooi Wei Tsang. Towards characterizing users' interaction with zoomable video. In *Proceedings of the 2010 ACM workshop on Social, adaptive and personalized multimedia interaction and access*, pages 21–24, 2010.

[63] Axel Carlier, Guntur Ravindra, Vincent Charvillat, and Wei Tsang Ooi. Combining content-based analysis and crowdsourcing to improve user interaction with zoomable video. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 43–52, 2011.

[64] Daniel J Hruschka, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard A Jenkins, and James W Carey. Reliability in coding open-ended data: Lessons learned from hiv behavioral research. *Field methods*, 16(3):307–331, 2004.

[65] ClassCentral Inc. The best online courses of all time, 2021.

[66] ClassCentral Inc. The top 100 most popular free online course (2021 edition), 2021.

[67] Leon Barnard, Ji Soo Yi, Julie A Jacko, and Andrew Sears. Capturing the effects of context on human performance in mobile computing systems. *Personal and Ubiquitous Computing*, 11(2):81–96, 2007.

[68] Afilias Technologies Limited. Viewport, resolution, diagonal screen size and dpi for the most popular smartphones, 2019.

[69] Apple Inc. Typography (human interface guidelines), 2021.

[70] Google LLC. Thetypesystem(materialdesign), 2021.

[71] H David Brecht. Learning from online video lectures. *Journal of Information Technology Education*, 11(1):227–250, 2012.

[72] J Holzl. Twelve tips for effective powerpoint presentations for the technologically challenged. *Medical Teacher*, 19(3):175–179, 1997.

[73] Moula Husain, SM Meena, Akash K Sabarad, Harish Hebballi, Shiddu M Nagaralli, and Sonal Shetty. Counting occurrences of textual words in lecture video frames using apache hadoop framework. In *2015 IEEE International Advance Computing Conference (IACC)*, pages 1144–1147. IEEE, 2015.

[74] Sabra Brock, Yogini Joglekar, and Eli Cohen. Empowering powerpoint: Slides and teaching effectiveness. *Interdisciplinary Journal of Information, Knowledge, and Management*, 6(1):85–94, 2011.

[75] Kirsten R Butcher. The multimedia principle. 2014.

[76] Karen Stein. The dos and don'ts of powerpoint presentations. *Journal of the American Dietetic Association*, 106(11):1745–1748, 2006.

[77] Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, Wendy Chisholm, John Slatin, and Jason White. Web content accessibility guidelines (wcag) 2.0. *WWW Consortium (W3C)*, 290, 2008.

[78] H Russell Bernard and Harvey Russell Bernard. *Social research methods: Qualitative and quantitative approaches*. Sage, 2013.

[79] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. sage, 2006.

[80] Judith S Olson and Wendy A Kellogg. *Ways of Knowing in HCI*, volume 2. Springer, 2014.

[81] Erica Sadun. *iOS Auto Layout Demystified*. Addison-Wesley Professional, 2013.

[82] Estelle Weyl. *Flexbox in CSS*. " O'Reilly Media, Inc.", 2017.

[83] Jacob O Wobbrock. Situationally-induced impairments and disabilities. In *Web Accessibility*, pages 59–92. Springer, 2019.

[84] Mayank Goel, Leah Findlater, and Jacob Wobbrock. Walktype: using accelerometer data to accomodate situational impairments in mobile touch screen text entry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2687–2696, 2012.

[85] Richard E Mayer. Multimedia learning. In *Psychology of learning and motivation*, volume 41, pages 85–139. Elsevier, 2002.

[86] Richard E Mayer. Introduction to multimedia learning. *The Cambridge handbook of multimedia learning*, 2:1–24, 2005.

[87] Ed H Chi, Peter Pirolli, and James Pitkow. The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 161–168, 2000.

[88] René F Kizilcec, Kathryn Papadopoulos, and Lalida Sritanyaratana. Showing face in video instruction: effects on information retention, visual attention, and affect. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2095–2102, 2014.

[89] Khe Foon Hew and Chung Kwan Lo. Comparing video styles and study strategies during video-recorded lectures: Effects on secondary school mathematics students' preference and learning. *Interactive Learning Environments*, 28(7):847–864, 2020.

[90] Charles H Chen and Philip J Guo. Improv: Teaching programming at scale via live coding. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, pages 1–10, 2019.

[91] Adalbert Gerald Soosai Raj, Jignesh M Patel, Richard Halverson, and Erica Rosenfeld Halverson. Role of live-coding in learning introductory programming. In *Proceedings of the 18th Koli Calling International Conference on Computing Education Research*, pages 1–8, 2018.

[92] Adalbert Gerald Soosai Raj, Pan Gu, Eda Zhang, Jim Williams, Richard Halverson, and Jignesh M Patel. Live-coding vs static code examples: Which is better with respect to student learning and cognitive load? In *Proceedings of the Twenty-Second Australasian Computing Education Conference*, pages 152–159, 2020.

[93] Sarah Dart. Khan-style video engagement in undergraduate engineering: Influence of video duration, content type and course. In *Proceedings of the 31st Annual Conference of the Australasian Association for Engineering Education (AAEE 2020)*. Engineers Australia, 2020.

[94] Genevieve Stanton and Jacques Ophoff. Towards a method for mobile learning design. In *Proceedings of the informing science and information Technology education conference*, pages 501–523. Informing Science Institute, 2013.

[95] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[96] Quentin Roy, Futian Zhang, and Daniel Vogel. Automation accuracy is good, but high controllability may be better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2019.

[97] Robert Sklar. *Film: An international history of the medium*. Prentice Hall, 1993.

[98] Deepika Bajaj and Shanu Sharma. Comparative analysis of shot boundary detection algorithms for video summarization. *CSI transactions on ICT*, 4(2-4):265–269, 2016.

[99] Sara Bunian, Kai Li, Chaima Jemmali, Casper Harteveld, Yun Fu, and Magy Seif Seif El-Nasr. Vins: Visual search for mobile user interface design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[100] Ananda Das and Partha Pratim Das. Automatic semantic segmentation and annotation of mooc lecture videos. In *International Conference on Asian Digital Libraries*, pages 181–188. Springer, 2019.

[101] Jiayin Lin, Geng Sun, Tingru Cui, Jun Shen, Dongming Xu, Ghassan Beydoun, Ping Yu, David Pritchard, Li Li, and Shiping Chen. From ideal to reality: segmentation, annotation, and recommendation, the vital trajectory of intelligent micro learning. *World Wide Web*, 23(3):1747–1767, 2020.

[102] S Krishnamurthy Ramnandan, Amol Mittal, Craig A Knoblock, and Pedro Szekely. Assigning semantic labels to data sources. In *European Semantic Web Conference*, pages 403–417. Springer, 2015.

[103] Minh Pham, Suresh Alse, Craig A Knoblock, and Pedro Szekely. Semantic labeling: a domain-independent approach. In *International Semantic Web Conference*, pages 446–462. Springer, 2016.

[104] Ying Su and Yong Zhang. Automatic construction of subject knowledge graph based on educational big data. In *Proceedings of the 2020 The 3rd International Conference on Big Data and Education*, pages 30–36, 2020.

[105] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*, 2020.

[106] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[107] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[108] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[109] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[110] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[111] Python Software Foundation. *pytesseract 0.3.6*, 2020 (accessed September 10, 2020).

[112] Zhangyang Wang, Jianchao Yang, Hailin Jin, Eli Shechtman, Aseem Agarwala, Jonathan Brandt, and Thomas S Huang. Deepfont: Identify your font from an image. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 451–459, 2015.

[113] (c) Python Software Foundation. *Python extcolors*, 2021 (accessed April 10, 2021).

[114] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[115] Xuehui Yu, Zhenjun Han, Yuqi Gong, Nan Jan, Jian Zhao, Qixiang Ye, Jie Chen, Yuan Feng, Bin Zhang, Xiaodi Wang, et al. The 1st tiny object detection challenge: Methods and results. In *European Conference on Computer Vision*, pages 315–323. Springer, 2020.

[116] Open Source Computer Vision enhanced by Google. *OpenCV Motion Analysis and Object Tracking*, 2021 (accessed April 8, 2021).

[117] Toni-Jan Keith Palma Monserrat, Shengdong Zhao, Kevin McGee, and Anshul Vikram Pandey. Notevideo: facilitating navigation of blackboard-style lecture videos. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1139–1148, 2013.

[118] Hijung Valentina Shin, Floraine Berthouzoz, Wilmot Li, and Frédo Durand. Visual transcripts: lecture notes from blackboard-style lecture videos. *ACM Transactions on Graphics (TOG)*, 34(6):1–10, 2015.

[119] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[120] Zvi Galil. Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys (CSUR)*, 18(1):23–38, 1986.

[121] Richard M Karp. An algorithm to solve the m$\times$ n assignment problem in expected time o (mn log n). *Networks*, 10(2):143–152, 1980.

[122] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[123] Aryo Pinandito, Hanifah Muslimah Az-zahra, Lutfi Fanani, and Anggi Valeria Putri. Analysis of web content delivery effectiveness and efficiency in responsive web design using material design guidelines and user centered design. In *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, pages 435–441. IEEE, 2017.

[124] Suroyya Wongsalam and Twittie Senivongse. Visual design and code generation of user interface based on responsive web design approach. In *Proceedings of the 2019 3rd International Conference on Software and e-Business*, pages 51–59, 2019.

[125] Github lolipopshock. *Deep Layout Parsing*, 2021 (accessed April 10, 2021).

[126] Ben Shneiderman. Eight golden rules of interface design. *Disponible en*, 172, 1986.

[127] Qianqian Xu, Ming Yan, Chendi Huang, Jiechao Xiong, Qingming Huang, and Yuan Yao. Exploring outliers in crowdsourced ranking for qoe. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1540–1548, 2017.

[128] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, 45(4):867–872, 2009.

[129] Brett G Amidan, Thomas A Ferryman, and Scott K Cooley. Data outlier detection using the chebyshev theorem. In *2005 IEEE Aerospace Conference*, pages 3814–3819. IEEE, 2005.

[130] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951.

[131] Meredith Ringel Morris, Adam Fourney, Abdullah Ali, and Laura Vonessen. Understanding the needs of searchers with dyslexia. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[132] Jun Gong, Peter Tarasewich, et al. Guidelines for handheld mobile device interface design. In *Proceedings of DSI 2004 Annual Meeting*, pages 3751–3756. Citeseer, 2004.

[133] Jay A Harolds. Tips for giving a memorable presentation, part iv: Using and composing powerpoint slides. *Clinical nuclear medicine*, 37(10):977–980, 2012.

[134] Natasha Larocque, Stephanie Kenny, and Matthew DF McInnes. Medical school radiology lectures: what are determinants of lecture satisfaction? *American Journal of Roentgenology*, 204(5):913–918, 2015.

[135] Lesley Pugsley. How to... design an effective power point presentation. *Education for Primary Care*, 21(1):51–53, 2010.

[136] Minjuan Wang and Ruimin Shen. Message design for mobile learning: Learning theories, human cognition and design principles. *British Journal of Educational Technology*, 43(4):561–575, 2012.

[137] Cindy Ann Dell, Thomas F Dell, and Terry L Blackwell. Applying universal design for learning in online courses: Pedagogical and practical considerations. *Journal of Educators Online*, 12(2):166–192, 2015.

[138] Antonella Grasso and Teresa Roselli. Guidelines for designing and developing contents for mobile learning. In *IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE'05)*, pages 123–127. IEEE, 2005.

[139] David Parsons, Hokyoung Ryu, and Mark Cranshaw. A design requirements framework for mobile learning environments. *JCP*, 2(4):1–8, 2007.

[140] Taralynn Hartsell and Steve Chi-Yin Yuen. Video streaming in online learning. *AACE Journal*, 14(1):31–43, 2006.

[141] Mary Frances Theofanos and Janice Ginny Redish. Helping low-vision and other users with web sites that meet their needs: Is one site for all feasible? *Technical communication*, 52(1):9–20, 2005.

[142] Gordon E Legge. Reading digital with low vision. *Visible language*, 50(2):102, 2016.

[143] Neil Charness, Elizabeth A Bosman, et al. Human factors and design for older adults. *Handbook of the psychology of aging*, 3:446–464, 1990.

[144] Luz Rello, Gaurang Kanvinde, and Ricardo Baeza-Yates. Layout guidelines for web text and a web service to improve accessibility for dyslexics. In *Proceedings of the international cross-disciplinary conference on web accessibility*, pages 1–9, 2012.

[145] Vagner Figueredo de Santana, Rosimeire de Oliveira, Leonelo Dell Anhol Almeida, and Maria Cecília Calani Baranauskas. Web accessibility and people with dyslexia: a survey on techniques and guidelines. In *Proceedings of the international cross-disciplinary conference on web accessibility*, pages 1–9, 2012.

[146] Luke Jefferson and Richard Harvey. Accommodating color blind computer users. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 40–47, 2006.

[147] Jan Walraven and Johan W Alferdinck. Color displays for the color blind. In *Color and Imaging Conference*, volume 1997, pages 17–22. Society for Imaging Science and Technology, 1997.