FitVid: Responsive and Flexible Video Content Adaptation

Jeongyeon Kim
School of Computing, KAIST
Daejeon, South Korea
imurs4825@gmail.com

Yubin Choi
School of Computing, KAIST
Daejeon, South Korea
joyda2525@gmail.com

Minsuk Kahng
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, United States
minsuk.kahng@oregonstate.edu

Juho Kim
School of Computing, KAIST
Daejeon, South Korea
juhokim@kaist.ac.kr

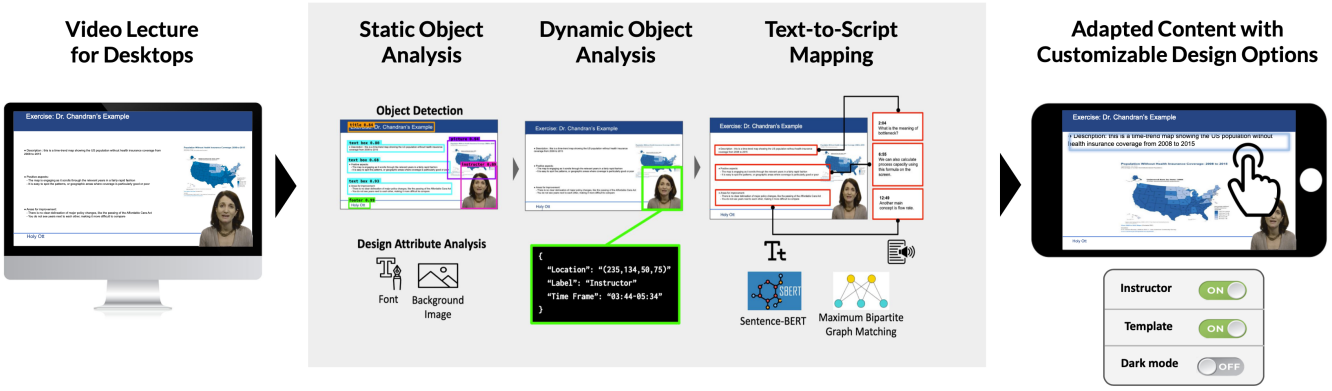

Figure 1: Overview of FitVid, an interactive video interface powered by an automated content adaptation pipeline for mobile video-based learning. The pipeline first retrieves in-video elements (e.g., text, images) from raw pixels in the video by analyzing static and dynamic objects. It then builds mappings between visual elements and the audio narrations. Finally, it generates content adaptations and renders them to mobile screens. Our video player UI supports direct manipulation and design customization with auto-generated adaptations. (Lecture Source and License: Aruna Chandran, Holly Ott)

ABSTRACT

Mobile video-based learning attracts many learners with its mobility and ease of access. However, most lectures are designed for desktops. Our formative study reveals mobile learners' two major needs: more readable content and customizable video design. To support mobile-optimized learning, we present FitVid, a system that provides responsive and customizable video content. Our system consists of (1) 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 (2) 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.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; *User studies*; *Empirical studies in ubiquitous and mobile computing*.

KEYWORDS

Content Adaptation; Responsive Design; Mobile Learning; Video-Based Learning

ACM Reference Format:
Jeongyeon Kim, Yubin Choi, Minsuk Kahng, and Juho Kim. 2022. FitVid: Responsive and Flexible Video Content Adaptation. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3491102.3501948

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501948

1 INTRODUCTION

An increasing number of learners use their mobile devices to watch instructional videos due to their ease of access and mobility. Accordingly, 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 [11, 53, 70].

However, the content adaptation is limited to static content such as websites and ebooks [7, 94], 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 multiple reasons.

First, it is required to decompose video into design elements such as text and images to resize and rearrange them to fit small screen sizes. Nonetheless, 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 non-pixel-based approach to extract metadata from the video content utilizing lecture slides and lecture notes, access to the lecture materials is often limited. Furthermore, video lectures involve dynamic elements such as talking-head instructors and real-time handwriting, which are not included in the lecture slides. Second, a high diversity of lecture designs makes the content adaptation even harder. There have been attempts to adapt the video content by using rule-based methods [43, 107], but creating heuristic rules entails massive manual work and consideration for the combinatorial explosion of possible conditions. This huge cost of the heuristic method limits the generalizability of content adaptation techniques.

To design a system that mitigates the above challenges of video content adaptation, we first investigated mobile learners' needs through a formative study. They wanted to have more readable content and customize the video design depending on their learning contexts. The findings from the formative study lead to the design of a new system for video content adaptation. In this paper, 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 (Fig. 1) includes a computational pipeline that automatically generates a responsive adaptation, which powers an interactive video interface that supports direct manipulation and content customization, including dark mode and instructor/template toggle option. 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 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 applying existing design guidelines for mobile content. To maximally apply the guidelines in limited screen space, we used constrained optimization and a set of heuristics.

However, the automated pipeline 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 by 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.

In summary, the primary contributions of this work are:

- An annotated dataset of 5,527 video frames for design element detection in lecture videos
- An automated pipeline that generates mobile-friendly content adaptation
- A design and implementation of FitVid, a system that provides learners with responsive and customizable video content
- Results of quantitative pipeline evaluation and empirical user study

2 RELATED WORK

Our work is informed by previous work on (1) content adaptation for mobile learning, (2) detection of lecture design elements from pixels, and (3) direct manipulation for video content.

2.1 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 [11, 70]. A responsive design such as excluding menus and images from educational websites increased mobile readability [53]. Other work introduced techniques of responsive content adaptation for informative visualization, which includes repositioning axes, removing labels from charts [35], repositioning legends, and word-wrapping for text overflows [99]. Meanwhile, responsive eBook design techniques allow users to customize font sizes, font styles, and line spacing on mobile devices [79, 94].

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 the following characteristics of video medium; difficulties of editing video content after release and a high diversity of lecture designs

which cannot be easily adapted using a simple set of rules or heuristics. Our work reverse engineers video content at an element level, enabling flexible adaptation even after release. Our deep-learning-based approach also covers a variety of lecture designs overcoming limitations of heuristic methods.

2.2 Detection of Lecture Design Elements from Pixels

Content adaptation requires detecting and identifying design elements for flexible resizing and rearrangements. Existing studies used a non-pixel-based approach to extract design elements [71, 91, 92]. In other words, they utilized lecture slides as a data source, which contains metadata such as locations and types (e.g., text, image) of each design element. However, access to the lecture 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.

On the other hand, pixel-based methods that extract metadata from raw pixels are generalizable to most existing video lectures that have no 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 [5, 43, 100, 103, 104, 106, 107] identify edges in images by looping over the pixel values and classify the design elements based on a set of IF-THEN rules. However, building heuristic rules includes huge manual work and a combinatorial explosion of possible conditions because of the high diversity of lecture designs. On the other hand, deep learning approaches learn such patterns from data. ViZig [102] used Convolutional Neural Networks with their own dataset of images downloaded from image search engines. WiSe [34] and SPaSe [33] contributed an 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 video lecture adaptation. We then trained an 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. Our dataset and trained models enable design element detection tailored to video-based learning content, finally allowing an in-video-level content adaptation.

2.3 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 [38, 84]. 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 [25, 45, 46], and reduced temporal ambiguities by allowing spatial-temporal manipulation [47, 66]. On the other hand, another thread of work introduced zoomable video interfaces to overcome the constraint of small screen sizes [4, 18, 68, 75, 86].

However, the interaction design for mobile video-based learning remains unexplored. Our work investigates users' challenges of the current interaction design of video lectures and suggests a new functionality that enables content customization through directly resizing and repositioning in-video elements.

3 FORMATIVE STUDY

We conducted formative interviews to investigate mobile learners' needs and learning experiences with video lectures.

3.1 Interview Study

3.1.1 Interviewees. We recruited 21 participants (13 male, 8 female) through Amazon Mechanical Turk (AMT) and advertisement posts on online university communities. Their age ranged from 18 to 44 years old. Interviewees were from South Korea (11), Brazil (4), the U.S. (3), India (2), and Canada (1). We ensured that all participants had a mobile video-based learning experience by asking them to upload a mobile screen capture of learning history or certificate from online video-based learning platforms. We provided a $10 Amazon gift card to each participant for a 30-minute long interview.

3.1.2 Protocol. We conducted remote interviews using Zoom and recorded the interviews under consent. We first asked them if they think existing video lectures are suitable for mobile learning. After that, we interviewed them on new features they hope to see to mitigate the challenges, regardless of their technical feasibility. The complete set of questions is included in the Supplementary Materials.

3.1.3 Analysis. We followed an iterative coding process [37]. Two authors independently created a codebook for half of the transcripts, each using an inductive approach, and they merged and refined the codebook through discussions. After reaching a consensus on the codebook, another author reviewed their work to finalize it. The two authors then coded three interview transcripts randomly selected from the entire dataset using the codebook. Finally, we computed Cohen's kappa to access inter-rater reliability. 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 interpretations, address any discrepancies in applying the code set, and then adjust their coded data. Finally, we produced ten subcodes for difficulties derived from inappropriate visual design of lectures.

3.2 Interview Results

To inform the design decisions of our system, we encouraged interviewees to share features they hope to use in their mobile video interfaces. The interviewees submitted 27 ideas total, and we merged overlapping ideas into nine themes that can be largely categorized into two: improving readability and customizing video design.

Improving readability. The following six ideas are related to improving the readability of learning content:

- **Automatic zoom-in feature:** Automatically zooming in small or dense text.
- **Element-wise zoom-in feature:** Allowing users to zoom in on the complete elements of interest (e.g., text boxes, images) on a per-element basis since the current pinch-zoom interaction results in cut-offs of part of the elements.
- **Optimized font sizes for mobile screen sizes:** Providing enlarged font sizes in response to small mobile screens.
- **Optimized amount of text for mobile screen sizes:** Reducing the amount of text on screen to improve readability.
- **Typewriting of hand-written materials:** Replacing the hand-written content with typewriting to improve legibility [21].
- **Highlighting the currently explained spot.** Adding visual cues such as highlights to guide learners' attention to the part of the screen currently being explained by the instructor [59].

Providing customization options. The remaining three ideas are about providing options to users:

- **Text-only or image-only mode:** Allowing users to choose a preferred type for the same content [60] by supporting different types of medium (e.g., text, images) to deliver the same content
- **Dark mode:** Providing dark mode for users in dim places.
- **Toggle a talking head:** Allowing users to turn on and off the talking head view of the instructor to utilize the mobile screen space efficiently [48, 93].

Based on the participants' suggestions, we determined a list of features to implement based on the following criteria: (1) severity of the problem and (2) the idea's novelty considering existing research. More specifically, the following two ideas with a low severity were excluded since only one participant suggested each idea: typewriting of hand-written materials, text-only or image-only mode. Considering the novelty, the idea of highlighting the current explanation was not adopted because they were already implemented by the previous work [43, 71]. This selection process that excluded the three ideas above led to a subset of the originally suggested features, including automatic and element-wise zoom-in, mobile-optimized size and amount of content, dark mode, and toggle talking heads.

4 DESIGN GOALS

The formative study results led us to the following design goals in creating a system that supports responsive and customizable video content.

G1. Automatically generating responsive design for video content

The formative study indicated a need for adapting video content at an element-level in response to small mobile screens (e.g., enlarged text, reduced amount of content). The video content adaptation requires a decomposition of raw video into visual elements (e.g., text, images) to be flexibly resized and rearranged. We aim to create an automated pipeline that extracts in-video elements using deep learning algorithms and generates responsive design without a need for costly human labor or manual work.

G2. Supporting direct manipulation of in-video elements

The formative study showed that there is a need for element-wise interactions to adjust the design of original content. Furthermore, the automatically generated content may contain incorrect adaptations and may not satisfy every user's needs [1, 80], for example, not having enough large fonts or proper layout. We aim to allow users to have control over the content adaptation instead of automating the whole process by supporting direct manipulation of in-video elements.

G3. Providing options for content customization

The "one-size-fits-all" approach of providing the same video content design to every learner is unlikely to be optimal in mobile learning because of various learning environments (e.g., diverse mobile screen sizes, distracting situations). The formative interviews revealed that users want to customize video design, for example, 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 5 and our user interface that supports direct manipulation (G2) and content customization (G3) in Section 6.

5 COMPUTATIONAL PIPELINE FOR AUTOMATED ADAPTATION

This section describes our computational pipeline that automatically creates an adapted version of video lectures (G1. Automatically generating responsive design for 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. 2. This section describes the main idea of each stage, while we provide additional details in Supplemental Materials for rigor and reproducibility.

5.1 Decomposition Stage

In this stage, video content is decomposed into design elements, and metadata for these elements is extracted to be used for content adaptation. To support adaptation that resizes and rearranges the decomposed elements, which we further describe in Section 5.2, 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 [85]. 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 analyze each lecture slide.
- **Static Object Analysis.** Once the representative frame is identified for each video shot detected, we extract design elements that do not change or move over time (e.g., text, images). These elements are adapted based on the existing

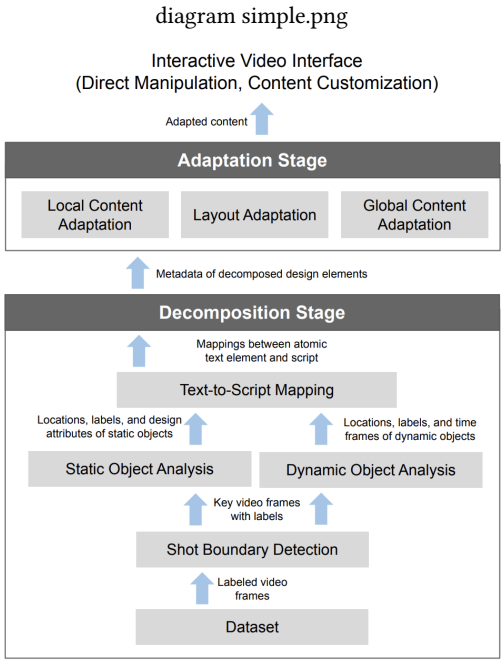

Figure 2: A computational pipeline of FitVid consists of two main stages, decomposition and adaptation. The decomposition stage extracts metadata of in-video elements used for adaptation. The adaptation stage generates and renders adapted content for mobile screens. The detailed version of the diagram is included in Supplementary Materials.

design guideline in the adaptation stage (e.g., enlarged font sizes).

- **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). 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 after adaptation.

5.1.1 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 [6]. It returns a sequence of shots, each consisting of start time, end time, and a single representative video frame. Information about the detailed implementation and threshold is included in the Supplementary Materials.

5.1.2 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 the video frames.

New Labeled Dataset of Lecture Designs. A large and high-quality dataset is crucial to train high-performing object detection models. Previous work released annotated datasets for presentation slides [33, 34], however, these datasets are not applicable to our context primarily for three 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 in learning materials (e.g., hierarchical bullet points). Thus, we annotated dataset of 5,527 video frames sampled from 66 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). In selecting 5,527 frames to be labeled, we first run the shot boundary detection algorithm to extract *keyframes* that naturally select a set of diverse frames filtering out too similar frames across the videos.

We chose 12 class labels for classifying design elements in lecture material [33, 102], which include title, text box, picture, chart, figure, diagram, table, schematic diagram, header, footer, handwriting, and instructor. We labeled design elements based on semantic units, which is considered important in data annotations for design elements [14]. 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 [22] and building microlearning content [54]. Our dataset can also be utilized as a source dataset for ontology construction [72, 76] and as a basic unit in knowledge point recognition [88].

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 [52]. 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 the dataset we collected. 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 [55], SSD based on ResNet [56], EfficientDet [89], and CenterNet [65, 108], the highest *mean Average Precision (mAP)* value (with IoU of 0.5) 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

[1] We used two different ranking measures (i.e., popularity, user reviews) from ClassCentral [40, 41].
[2] https://github.com/imurs34/lecture-design-dataset

detailed information about these experiments and hyperparameters 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 formative study (Section 3) informs 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 the original lecture materials [9], which returns an image without static objects but only the background left.

5.1.3 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 [105]. In order to detect these dynamic objects in a video shot, apart from *Static Object Analysis* (Section 5.1.2), we used OpenCV motion analysis module [26]. We matched the area of the detected motions with the object detection results from Section 5.1.2 to identify dynamic objects. Details can be found in the Supplementary Materials.

5.1.4 Text-to-Script Mapping. Lastly, we determine mappings between the 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 [62, 82, 107], 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 current video frame using Sentence-BERT [77]. After that, we use a *bipartite graph matching* algorithm [28, 44] to find optimal mappings. Details including 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. For example, three bullet points should be displayed in a single frame if an instructor explains them in a non-linear manner, referring to them back and forth. In this case, the three bullet points should not be segmented into three different frames, but need to be grouped as an atomic unit. 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. Second, we merge multiple text elements that an instructor concurrently mentions. If an instructor mentions multiple text elements in a single sentence, then learners should be able to refer to all the elements at once.

5.2 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 intended to refine the adapted results to be consistent and coherent.

5.2.1 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 threshold from the guidelines. 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 5.1.4 even if it violates the guidelines. Since lecture material is a creation of instructors and engineers, we took a preservative approach that maximizes the guideline compliance rate while maintaining the original design as much as possible. The details of the balance between preservation and adaptation, along with exact thresholds are included in the Appendix.

5.2.2 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 [10]. We define the objective function as follows:

$$\min \left(\frac{c-x}{c}\right)^2+\left(\frac{c-y}{c}\right)^2, \text{ where } 0 < x < x_{\max} \text{ and } 0 < y < y_{\max}. \tag{1}$$

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

[3] https://github.com/imurs34/lecture_design_detection

[4] We used Pytesseract OCR [27], DeepFont [97], and color clustering techniques [16] to extract font size, typeface, and font color, respectively.

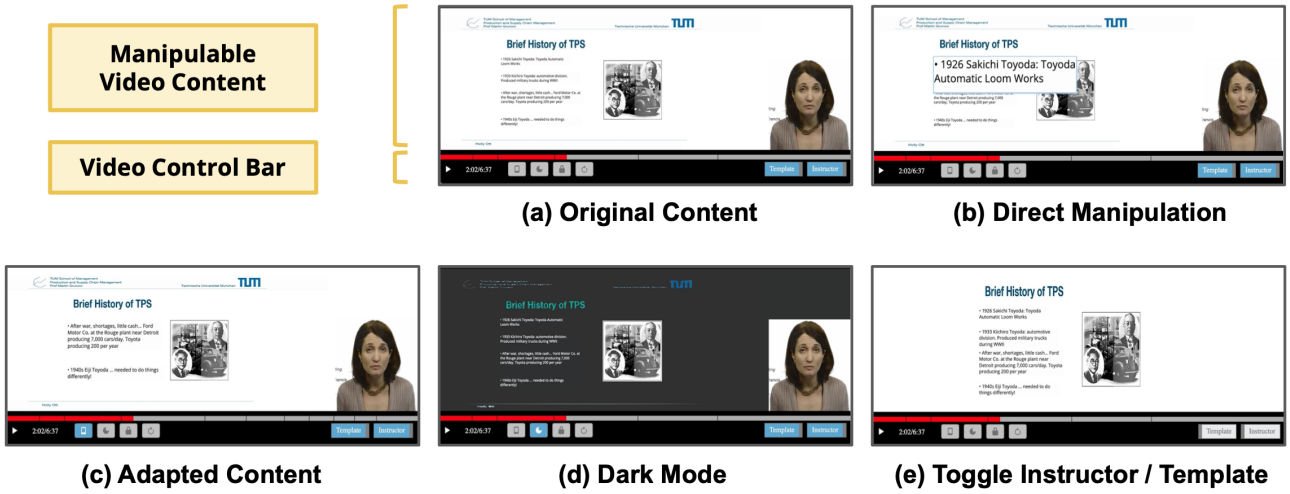

Figure 3: A learner can resize, reposition, and toggle various in-video elements using FitVid's video UI. (a) Original Content: original content without adaptation is displayed to the learner by default, (b) Direct Manipulation: the learner can resize and reposition design elements (e.g., text boxes, 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 (e.g., size and amount of text are adjusted in the figure), (d) Dark Mode: the learner can choose 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 footers). (Lecture Source and License: Holly Ott)

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 [73, 98]. We first extract a column layout and reading orders of learning materials [58]. We then determine a final column layout by adopting the layout that has a higher guideline compliance rate between the original layout and the converted ones with content reflow.

5.2.3 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.

6 VIDEO INTERFACE

We designed and developed an interactive video player of FitVid, depicted in Fig. 3. 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 [1, 83]. FitVid allows users to edit and refine the automated adaptation results through direct manipulation. As shown in Fig. 3 (a) and (b), a learner can directly resize and reposition in-video elements (e.g., text boxes, images, and talking-head instructions) using touch and drag interactions. In particular, a learner can resize elements by dragging their edge. We chose the drag interaction instead of the commonly used pinch-zoom interaction based on the formative study results, in which participants suggested an element-wise zoom feature that does not result in cutting off part of the content. The changes a learner has made are preserved even when they navigate back and forth through a video.

Content Customization. Based on the formative study, we provide customization options for users to determine whether to display talking-head instructors, slide templates, and change background colors. For instance, Fig. 3 (e) shows the turn-off option for the talking-head instructor. Learners can also use the dark theme that provides bright fonts in dark backgrounds (Fig. 3 (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 [1]. 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 web browsers and is implemented using HTML, CSS, and JavaScript.

7 PIPELINE EVALUATION

We evaluated the performance of our computational pipeline through content analysis and content evaluation study. For the evaluation, we sampled 53 video frames from 24 videos, which are included

	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 1: Statistics of design elements from original content and adapted content. The result demonstrates that the adaptation pipeline improves the design guideline compliance rate.

in Supplementary Materials. Specifically, two of the authors examined every keyframe extracted from the dataset for the object detection model training, and selected a subset of them for evaluation to include diverse design factors that are not uniformly distributed across videos. We applied our pipeline from end to end to the sampled videos without correcting propagated errors from the submodules. The detailed evaluation for each submodule without considering dependencies between submodules is included in Supplementary Materials. We also report representative error cases with examples.

7.1 Content Analysis

We conducted the content analysis to compare the design guideline compliance rate before and after the adaptation. The guideline thresholds are in Appendix. In Table 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. The guideline compliance rate increased by 24% and 8%, respectively, for the word count and font sizes. The compliance rate changed from 13% to 17% for the font sizes in images. For typefaces, line spacing, and color contrast, all the target cases of adaptation were successfully adapted to comply with the guidelines. Fig. 4 shows the examples adaptation results. In Fig. 4 (a), the fonts are enlarged, and the text is trimmed down. Fig. 4 (b) shows an example of converting the column design with content reflowing. In Fig. 4 (c), the adaptation algorithm finds the balance between enlarging text and images containing text, compromising the compliance rate of each element. Fig. 4 (d) demonstrates the increased color contrast. Lastly, Fig. 4 (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 (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 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 (g)), the other text elements also fail to meet the guidelines. Third, the number of words could not be reduced below the threshold since the text-to-script mapping module grouped multiple text boxes as atomic units. Meanwhile, the direct manipulation feature can compensate for the system's failure by allowing users, for example, to reposition the overlapping elements or resize text that is not large enough.

7.2 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 reveals how general users perceive the previous section's adaptation results for the same content.

7.2.1 Participants. We recruited 198 respondents (70 female, 127 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 0.9 USD 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.

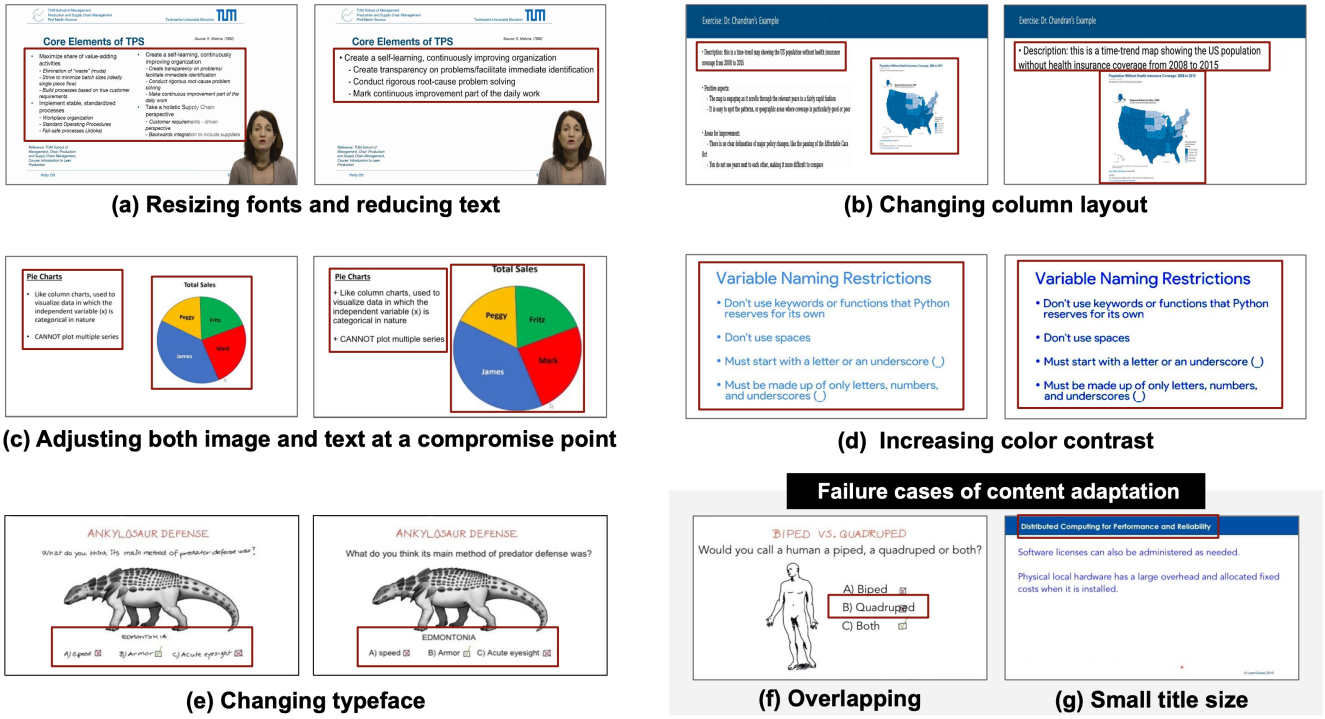

Figure 4: Examples of adapted content. (a) Resized fonts and reduced text, (b) Changed column layout, (c) Adjusted image and text at compromise point, (d) Increased color contrast, (e) Changed typeface. The representative failure cases are: (f) overlappings due to errors from the object detection stage and (g) incomplete text resizing due to its comparative size with titles (Lecture Source and License: Holly Ott, Aruna Chandran, Charlie Nuttelman, Google Career Certificates, Philip John Currie, Jim Sullivan)

7.2.2 Method. We designed a content evaluation survey by utilizing the adaptation results in Section 7.1. 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 [101]. 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 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. A participant could participate in multiple HITs with different image datasets. Our form also included an attention check question [67] to filter invalid responses.

7.2.3 Results. We initially collected 401 responses and removed 152 responses for invalid attention check answers, 44 for invalid screen capture, and 7 as outliers beyond 2 standard deviations from the mean [2]. 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 [20], 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 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 general users' readability and design satisfaction.

8 USER STUDY

This section evaluates 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:

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

- RQ1. How does FitVid's automated content adaptation impact the perceived readability and design satisfaction compared to the baseline video interface?
- RQ2. How do users use and benefit from FitVid's direct manipulation?
- RQ3. How do users use and benefit from FitVid's content customization feature?
- RQ4. 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.

8.1 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 15 USD for up to 70 minutes of participation.

8.2 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 on their mobile phones. 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, perceived learning efficiency, and learning experience for each interface. We used three readability questions from existing work: Design choices made reading harder (fonts, colors, etc.); It was easy for me to lose my place while reading; Overall the content was easy to read [53, 63]. 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.

8.3 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.

8.3.1 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 [20], 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.

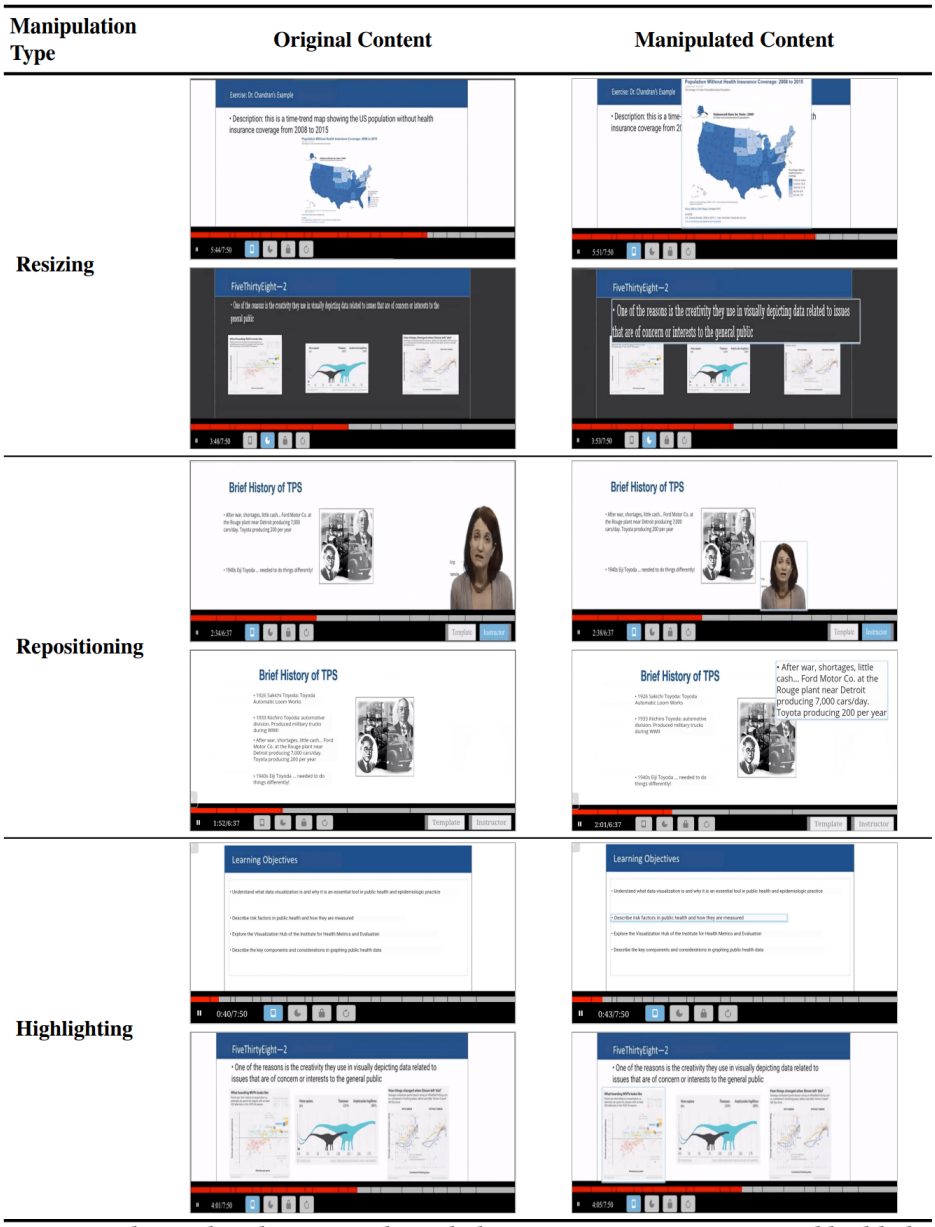

Figure 5: Manipulation types observed in the user study, including resizing, repositioning, and highlighting. (Lecture Source and License: Holly Ott, Aruna Chandran)

8.3.2 RQ2. How do users use and benefit from FitVid's direct manipulation? For a single video, the mean number of direct manipulation interactions (i.e., number of touch interactions) was 21 (min: 0, max: 84). 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. 5 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 3).

To adjust the design. The primary reason for enlarging elements was to improve the readability. The participants noted that direct manipulation allows them to customize the automated adaptation results further. For example, P18 stated, "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 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.

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 or emphasizing to better memorize content	3/31	Resize, Highlight
		2/31	Resize
To interact with content	to enjoy interaction itself	5/31	Resize, Reposition

Table 3: Reasons behind direct manipulation usage. Users used the direct manipulation to adjust the design, promote concentration, and interact with content.

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.".

8.3.3 RQ3. How do users use and benefit from FitVid's content customization feature? Overall, the participants appreciated the content customization option, which allows customization of lecture design. Almost all participants (30/31) want to use the content customization 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, other participants indicated that they did not need it in bright environments and preferred dark fonts in a bright background.

8.3.4 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 perceived 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 ($W = 41$, $p < 0.05$). Nonetheless, 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.

9 DISCUSSION AND FUTURE WORK

We discuss limitations and future research directions for responsive and flexible video content adaptation.

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.

Controlling the degree of adaptation. While the user study results revealed that FitVid improves the overall learning experience with increased concentration and readability of content, the participants did not fully trust the results from automated adaptation. For example, because the adapted content contains less text than the original lecture material, participants checked if the system removed necessary content from the slide by comparing the adapted and original content. They appreciated the feature for accessing original content. The human-AI interaction guidelines [1] are also in line with the participants' responses, which suggest providing users control over automated results and allowing easy correction. Future work may allow learners to choose the degree of adaptation (e.g., only minor adjustments to very aggressive adaptation).

Direct manipulation for video content. We observed a variety of needs for directly adjusting adapted content. Noticeable motivations for direct manipulation include adjusting design, promoting concentration, and interacting with content. One of the unexpected benefits of the manipulation was that learners interacted with the content itself to increase their cognition without pragmatic goals such as resizing or repositioning elements.

Meanwhile, several participants noted that they needed time to get used to the direct manipulation feature while watching the video. While dragging interaction is intuitive to manipulate the sizes and positions, one possible improvement is to provide font size or image size options enabling resizing with fewer touch interactions.

Furthermore, adaptation technique may evolve to be adaptive and personalized, based on usage log or user-specified preferences. For instance, if a user enlarges the font size to 30pt on average using direct manipulation, the system can remember this setting and generate text with the same font size in their future use instead of adopting general design guidelines. The future work can collect users' manipulation log at scale, and the visual design of lectures can even be crowdsourced by using the log data.

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 [50, 90], older adults [19], and dyslexia [23, 78] populations, based on the design guidelines for these populations. For 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 [42, 95]. Our work can be an initial step toward enhancing the accessibility of visual video content for various user groups with different ability profiles.

Generalizing to other video domains. While this work mainly focused on video lectures, FitVid can also be applied to other types of informational videos, such as tutorials, news, and educational talks. Depending on the characteristics of the video contents, people may adjust the parameters in our computational pipeline. For example, TED talk videos often involve more dynamic elements such as moving speakers compared to slide-based lectures, then a lower threshold for the shot boundary detection algorithm can be adopted.

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.

10 CONCLUSION

This paper introduces FitVid, a system that enables automated content adaptation and design customization for mobile learning environments. FitVid consists of an adaptation pipeline that reverse-engineers pixels to retrieve design elements (e.g., text, images) from videos, using deep learning with a custom dataset, which powers a UI that enables resizing, repositioning, and toggling in-video elements. The content adaptation results improve the design guideline compliance rate by 24% for word count and 8% for font sizes compared to the original content. To demonstrate the effectiveness and usefulness of our system, we conducted a user study and content evaluation study. We find that FitVid provides an improved learning experience, significantly increasing perceived readability and the level of concentration. We expect to apply the proposed techniques to enhance the accessibility of video content, lowering the barrier not only for mobile users but for diverse user groups under different contexts, abilities, and preferences.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1C1C1007587). This work was also supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government (MSIT) (No.2021-0-01347,Video Interaction Technologies Using Object-Oriented Video Modeling).

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen,

et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

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

[3] Aries Arditi and Jianna Cho. 2005. Serifs and font legibility. *Vision research* 45, 23 (2005), 2926–2933.

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

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

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

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

[8] Sofie Beier. 2009. *Typeface legibility: towards defining familiarity*. Royal College of Art (United Kingdom).

[9] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. 2001. 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*, Vol. 1. IEEE, I–I.

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

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

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

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

[14] Sara Bunian, Kai Li, Chaima Jemmali, Casper Harteveld, Yun Fu, and Magy Seif Seif El-Nasr. 2021. VINS: Visual Search for Mobile User Interface Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[15] Kirsten R Butcher. 2014. The multimedia principle. (2014).

[16] (c) Python Software Foundation. 2021 (accessed April 10, 2021). *Python extcolors*. https://pypi.org/project/extcolors/

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

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

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

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

[21] Andrew Cross, Mydhili Bayyapunedi, Dilip Ravindran, Edward Cutrell, and William Thies. 2014. 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*. 1167–1175.

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

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

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

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

[26] Open Source Computer Vision enhanced by Google. 2021 (accessed April 8, 2021). *OpenCV Motion Analysis and Object Tracking*. https://docs.opencv.org/4.5.2/d7/df3/group__imgproc__motion.html

[27] Python Software Foundation. 2020 (accessed September 10, 2020). *pytesseract 0.3.6*. https://pypi.org/project/pytesseract/

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

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

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

[31] Jay A Harolds. 2012. Tips for giving a memorable presentation, Part IV: Using and composing PowerPoint slides. *Clinical nuclear medicine* 37, 10 (2012), 977–980.

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

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

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

[35] Jane Hoffswell, Wilmot Li, and Zhicheng Liu. 2020. Techniques for Flexible Responsive Visualization Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

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

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

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

[39] Apple Inc. 2021. *Typography (Human Interface Guidelines)*. https://developer.apple.com/design/human-interface-guidelines/ios/visual-design/typography/

[40] ClassCentral Inc. 2021. *The Best Online Courses of All Time*. https://www.classcentral.com/collection/top-free-online-courses

[41] ClassCentral Inc. 2021. *The Top 100 Most Popular Free Online Course (2021 Edition)*. https://www.classcentral.com/report/100-most-popular-online-courses-2021/

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

[43] Hyeungshik Jung, Hijung Valentina Shin, and Juho Kim. 2018. 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*. 33–41.

[44] Richard M Karp. 1980. An algorithm to solve the m× n assignment problem in expected time O (mn log n). *Networks* 10, 2 (1980), 143–152.

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

[46] Thorsten Karrer, Moritz Wittenhagen, and Jan Borchers. 2009. 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*. 1–3.

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

[48] René F Kizilcec, Jeremy N Bailenson, and Charles J Gomez. 2015. The instructor's face in video instruction: Evidence from two large-scale field studies. *Journal of Educational Psychology* 107, 3 (2015), 724.

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

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

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

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

[53] Qisheng Li, Meredith Ringel Morris, Adam Fourney, Kevin Larson, and Katharina Reinecke. 2019. 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*. 1–12.

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

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

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

[57] Google LLC. 2021. *Thetypesystem(MaterialDesign).* https://material.io/design/typography/the-type-system.html#type-scale

[58] Github lolipopshock. 2021 (accessed April 10, 2021). *Deep Layout Parsing.* https://github.com/Layout-Parser/layout-parser/blob/master/docs/example/deep_layout_parsing/index.rst

[59] Patricia D Mautone and Richard E Mayer. 2001. Signaling as a cognitive guide in multimedia learning. *Journal of educational Psychology* 93, 2 (2001), 377.

[60] Martin Merkt, Sonja Weigand, Anke Heier, and Stephan Schwan. 2011. Learning with videos vs. learning with print: The role of interactive features. *Learning and Instruction* 21, 6 (2011), 687–704.

[61] Aliaksei Miniukovich, Antonella De Angeli, Simone Sulpizio, and Paola Venuti. 2017. Design guidelines for web readability. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 285–296.

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

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

[64] Derek A Muller, Kester J Lee, and Manjula D Sharma. 2008. Coherence or interest: Which is most important in online multimedia learning? *Australasian Journal of Educational Technology* 24, 2 (2008).

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

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

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

[68] Derek Pang, Sherif Halawa, Ngai-Man Cheung, and Bernd Girod. 2011. 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*. 7–12.

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

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

[71] Yi-Hao Peng, JiWoong Jang, Jeffrey P Bigham, and Amy Pavel. 2021. Say It All: Feedback for Improving Non-Visual Presentation Accessibility. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

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

[73] Aryo Pinandito, Hanifah Muslimah Az-zahra, Lutfi Fanani, and Anggi Valeria Putri. 2017. 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)*. IEEE, 435–441.

[74] Lesley Pugsley. 2010. How To... Design an effective power point presentation. *Education for Primary Care* 21, 1 (2010), 51–53.

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

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

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

[78] Luz Rello, Gaurang Kanvinde, and Ricardo Baeza-Yates. 2012. 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*. 1–9.

[79] Luz Rello, Gaurang Kanvinde, and Ricardo Baeza-Yates. 2012. A mobile application for displaying more accessible eBooks for people with Dyslexia. *Procedia Computer Science* 14 (2012), 226–233.

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

[81] Audrey Dawn Shaikh. 2007. *Psychology of onscreen type: Investigations regarding typeface personality, appropriateness, and impact on document perception.* Ph. D. Dissertation.

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

[83] Ben Shneiderman. 1986. Eight golden rules of interface design. *Disponible en* 172 (1986).

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

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

[86] Wei Song, Dian W Tjondronegoro, Shu-Hsien Wang, and Michael J Docherty. 2010. 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*. 321–330.

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

[88] Ying Su and Yong Zhang. 2020. 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*. 30–36.

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

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

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

[92] Shoko Tsujimura, Kazumasa Yamamoto, and Seiichi Nakagawa. 2017. Automatic Explanation Spot Estimation Method Targeted at Text and Figures in Lecture Slides.. In *INTERSPEECH*. 2764–2768.

[93] Margot van Wermeskerken and Tamara van Gog. 2017. Seeing the instructor's face and gaze in demonstration video examples affects attention allocation but not learning. *Computers & Education* 113 (2017), 98–107.

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

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

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

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

[98] Suroyya Wongsalam and Twittie Senivongse. 2019. 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*. 51–59.

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

[100] Chengpei Xu, Ruomei Wang, Shujin Lin, Xiaonan Luo, Baoquan Zhao, Lijie Shao, and Mengqiu Hu. 2019. Lecture2Note: Automatic Generation of Lecture Notes from Slide-Based Educational Videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 898–903.

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

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

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

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

[105] Xuehui Yu, Zhenjun Han, Yuqi Gong, Nan Jan, Jian Zhao, Qixiang Ye, Jie Chen, Yuan Feng, Bin Zhang, Xiaodi Wang, et al. 2020. The 1st tiny object detection challenge: Methods and results. In *European Conference on Computer Vision*. Springer, 315–323.
[106] Baoquan Zhao, Shujin Lin, Xiaonan Luo, Songhua Xu, and Ruomei Wang. 2017. A novel system for visual navigation of educational videos using multimodal cues. In *Proceedings of the 25th ACM international conference on Multimedia*. 1680–1688.
[107] Baoquan Zhao, Songhua Xu, Shujin Lin, Ruomei Wang, and Xiaonan Luo. 2019. A New Visual Interface for Searching and Navigating Slide-Based Lecture Videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 928–933.
[108] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. 2019. Objects as points. *arXiv preprint arXiv:1904.07850* (2019).

A APPENDIX

A.1 Design Guidelines from the literature

Category	Design Element	Design Guideline	Prior Work
Text Element	Font Size	Above 16 - 28 pt	[29, 31, 36, 39, 49, 51, 57, 74]
	Number of Words	Below 20-45 words	[13, 31, 36, 39, 96]
	Typeface	Avoid handwriting and script fonts	[3, 8, 24, 31, 36, 39, 51, 57, 64, 74, 81, 96]
	Line Spacing	1.5 * font size	[74]
	Letter Spacing	0.12 * font size	[17]
Image Element	Small and Complex Images	As large as possible	[30, 31, 51, 69]
Color	Color Contrast	above 4.5 - 7.0	[17, 31, 32, 36, 39, 51, 74, 96]

Table 4: Design guidelines from the literature

- **Font Size.** Apple's Human Interface Guidelines suggest using 17 pt as body text size [39] while Google Material Design Guidelines recommend 16 pt [57]. We also considered guidelines for presentation slides that suggest 28 pt [12, 36] as minimum font size.
- **Number of Words.** Using less than 45 words per presentation slide is recommended for readability [13], while stricter guidelines advise using less than 20 words per slide [15, 87].
- **Small and Complex Images.** It is recommended to enlarge the images in slides as large as possible [31]. Our formative study results indicated that learners' main pain point was a complex image that contains text, such as charts, tables, and diagrams. Based on users' feedback, we applied font size guidelines in adaptation for complex images with text, resizing the image until the fonts within the image meet the guidelines.
- **Line Spacing.** Several guidelines including WCAG [17, 61] suggest the line height to be at least 1.5 times the font size.
- **Typeface.** The existing guidelines for typeface design recommend avoiding handwriting and script fonts [3, 8, 64, 81].
- **Color Contrast.** With regard to the color contrast between the font and background, WCAG suggests the color contrast ratio of 4.5:1 (Level AA) and 7.0:1 (Level AAA).

To set up basic design principles for content adaptation, we used the average values from the guidelines since there is no rule-of-thumb guideline with hard numbers. We do not claim that this value is a strict standard, but it can vary depending on the context and purpose of the adaptation. We suggest regarding it as an initial step to establish a metric for learning content design in mobile environments. The averaged design principles we used for content adaptation are: (1) average font size of a frame should be above 21.4 pt, (2) the number of words per frame should be below 30 words, (3) decorative fonts such as handwriting and script typeface should be adapted to sans-serif, (4) line spacing should be above 1.5 * font size, (5) complex images containing text should be enlarged for font size to be above 21.4pt, (6) color contrast ratio between text and background should be above 5.75:1.