Elastica: Adaptive Live Augmented Presentations with Elastic Mappings Across Modalities

Yining Cao
University of California, San Diego
La Jolla, California, USA
yic069@ucsd.edu

Rubaiat Habib Kazi
Adobe
Seattle, Washington, USA
rhabib@adobe.com

Li-Yi Wei
Adobe
San Jose, California, USA
lwei@adobe.com

Deepali Aneja
Adobe
Seattle, Washington, USA
aneja@adobe.com

Haijun Xia
University of California, San Diego
La Jolla, California, USA
haijunxia@ucsd.edu

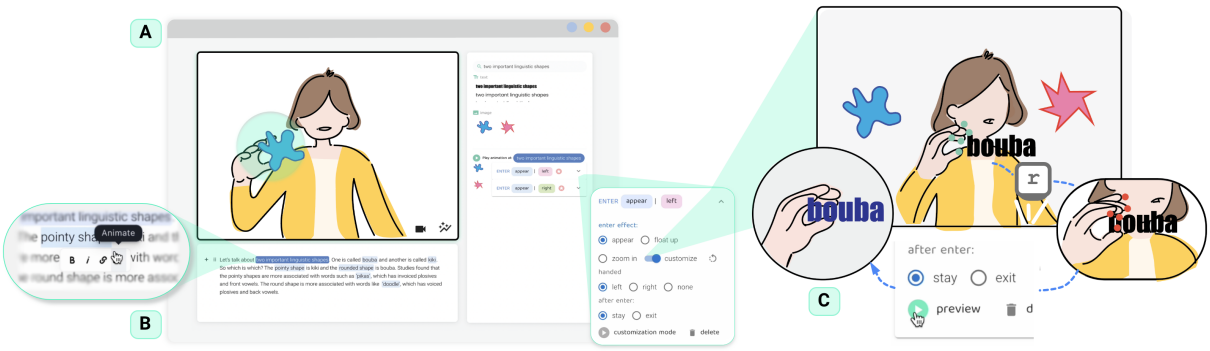


Figure 1: Elastica leverages adaptive animation to (A) allow users to design, rehearse, and present live augmented presentations with (B) script annotations and (C) gestural demonstrations. During a live performance, Elastica adapts the predefined visuals to the presenter's real-time performance to achieve an optimized synchronization between animation, gesture, and speech.

ABSTRACT

Augmented presentations offer compelling storytelling by combining speech content, gestural performance, and animated graphics in a congruent manner. The expressiveness of these presentations stems from the harmonious coordination of spoken words and graphic elements, complemented by smooth animations aligned with the presenter's gestures. However, achieving such desired congruence in a live presentation poses significant challenges due to the unpredictability and imprecision inherent in presenters' real-time actions. Existing methods either leveraged rigid mapping without predefined states or required the presenters to conform to predefined animations. We introduce adaptive presentations that dynamically adjust predefined graphic animations to real-time speech and gestures. Our approach leverages script following and motion warping to establish elastic mappings that generate runtime graphic parameters coordinating speech, gesture, and predefined animation state. Our evaluation demonstrated that the proposed adaptive presentation can effectively mitigate undesired visual artifacts caused by performance deviations and enhance the expressiveness of resulting presentations.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05.
https://doi.org/10.1145/3613904.3642725

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Mixed / augmented reality**.

KEYWORDS

animation, augmented presentation, gestural interaction

ACM Reference Format:
Yining Cao, Rubaiat Habib Kazi, Li-Yi Wei, Deepali Aneja, and Haijun Xia. 2024. Elastica: Adaptive Live Augmented Presentations with Elastic Mappings Across Modalities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3613904.3642725

1 INTRODUCTION

Live augmented presentation is an emerging presentation format that overlays digital content with presenters' real-time speaking and gesturing performance [21, 25, 36]. Its expressiveness stems from the synchronization and interaction between digital content

with the presenters' live performance [42, 43], blending the boundaries between digital and physical realms as well as creating a captivating and engaging experience. As such, it is becoming an increasingly popular communication and demonstration medium for concept explanation [16, 31], education [39], and storytelling [36]. For example, in his powerful demonstration of the relationship between GDP and life expectancy, Rosling masterfully coordinated the animation of the visualizations, his narration, and body poses to deliver engaging and memorable presentations [33–35].

Achieving the congruent coordination between gesture, speech, and graphics during live performances is enormously challenging. For example, a typical approach for authoring a live augmented presentation is to ***pre-define the presentations*** using presentation software that allows video feed integration [2, 29], ***memorize the mappings*** across the modalities through rehearsals, and deliver the final presentation [34]. Because any performance deviation and misalignment of speech, gesture, and visuals can compromise the presentation quality and break the illusion of blending physical and digital worlds, presenters often invest substantial time and effort in content preparation and rehearsal while juggling attention to the graphic content, speech and gesture during the live performance to ensure a high-quality presentation.

Recent work explored an alternative approach to authoring live augmented presentations, which leverages speech and gesture recognition to create one-on-one mappings across the modalities, such as mapping spoken keywords in the narration to trigger animations [25], and animating visuals by attaching them to a body part and following its movement [36]. During the performance, the presenter leveraged the defined mappings to create live-defined presentations. While these methods can deliver robust synchronization between the physical and virtual worlds, the resulting presentations often suffer from inferior visual artifacts, such as uneven graphic layouts and jittery motion. This is because the motion and layout of graphics are ***determined during live performances through rigid mappings***, which are more prone to imperfect physical manipulation and system recognition. For example, the graphics may fail to appear or animate due to incorrect performance or recognition errors, exhibit rigid and jittery motions, or be placed at an undesired final position and scale if the gesture overshoots or undershoots.

We see a gap between ***pre-defined presentations with memorized mappings (PM)***, which requires significant practice and effort for the presenter to fully conform to the presentation, and ***live-defined presentations with rigid mappings (LR)***, which relies on the presenter's on-the-spot manual control but often results in inferior visual quality. We seek to bridge this gap by supporting an approach to authoring live augmented presentations that can strike a balance between maintaining the desired synchronization across modalities and generating consistently high-quality visual artifacts during live performances. Through a comparative analysis of existing approaches for live augmented presentations, we identified two main causes of compromised presentation quality: the reliance on real-time specification of layouts and effects of the visuals, and the rigid mappings between modalities (e.g., keyword–graphic mapping [25], graphic-body joint mapping [36]), make the graphic content sensitive to performance deviations and allow errors of one modality to propagate to another.

We propose *adaptive presentations* for a live augmented presentation that employs ***pre-defined presentations with elastic mappings (PE)***. By pre-defining the layouts and effects of graphics, as one would do in typical presentation applications, along with a set of customizable and algorithmic constraints to allow speech, gesture, and graphics to be *elastically* connected, the prepared presentation can adapt to presenters' real-time gestures and speech for achieving synchronization between presenter's performance and graphical animations, while constraining the adaptation to pre-defined states to ensure visual quality.

We implemented a prototype system, Elastica, which enables users to create and customize expressive adaptive animations for live presentations. When preparing the presentation, Elastica allows users to specify desired animation timing with script annotations; and specify customized animation effects that are synchronized with gestures through demonstration. During the live performance, the pre-defined animations in the presentation are triggered and adapted to users' real-time speech and gesture performance with elastic mappings across modalities, resulting in coherent synchronization that is resilient to imperfect performances and expressive presentation effects.

We evaluated Elastica with a user study with eight participants to author an augmented presentation from planning to presenting using Elastica. Our result showed that adaptive animation can support users' expressive and high-quality augmented presentation while effectively preventing 73.6% of undesired visual results that are caused by live performance deviations. To delve deeper into the potential strengths and limitations of adaptive animation, we conducted a second comparative study with six participants who evaluated their experience with adaptive presentation against existing approaches. Our analysis unveiled that adaptive presentations are the collaborative interplays between human performance and intelligent adaption, yielding high satisfaction for their ability to react flexibly to presenters' performance while ensuring the desired visual layouts and effects. Combined, this paper contributes:

(1) A real-time adaption method that leverages pre-defined states and elastic mappings across modalities to automatically adjust animated graphics to real-time speech and gesture performance to ensure synchronization and visual quality;
(2) A prototype system, Elastica, that enables users to flexibly specify the mappings across the modalities - script (speech), gestures, and graphics, and specify customized gestures for expressive presentation effects;
(3) Evaluations of Elastica and the adaptation mechanism with three studies that demonstrate its effectiveness, and limitations from the perspectives of presenters and viewers.

2 RELATED WORK

As Kang et al. pointed out, effective presentations are a symphony of gestures, language, and props — in our case, graphics [21]. In this section, we first review augmented presentations and then dive into how the synchronizations of gestures, language, and graphics are supported and leveraged in HCI.

2.1 Augmented Presentation

While augmented presentation — blending graphics, speech, and gestures — has become popular recently, the integral of these modalities has long been realized and widely used for explanation and communication in practices, such as in educational settings [11, 31], public presentations [33, 35], and online explanatory videos [45].

Extensive research in psychology has also investigated the integral power of these modalities and the coordination of the different modalities [12, 21, 22, 30]. Kang et al., for example, refer to successful explanatory presentations as the symphony of gesture, language, and props, and found that props such as diagrams served as the backbone of the communicated information, and the gestures and language served to annotate and animate the diagrams [21]. Despite the desired synchrony of these modalities [12, 21], prior work found that they do not follow strict simultaneity in practice. For example, some modalities can be absent [21]; the modalities often only partially overlap, or one modality may proceed with the others [22, 24, 30], suggesting the interplay of these modalities are highly dynamic and context-dependent.

It is perhaps because of the intricate coordination of these signals and technical challenges in creating and capturing interactive presentations, that many augmented presentations are produced via post-production [25]. This allows creators to configure the various elements post hoc, achieving optimal visual and presentation quality. With the growing popularity of augmented reality and live streams, recent HCI research has explored augmenting live videos with interactive graphics to create compelling and engaging presentations [4, 15, 16, 27, 36]

For example, RealityTalk employs a keyword-matching table to link the speech and graphical elements so that pre-defined graphics appear when certain keywords are detected in real-time speech, which can be subsequently manipulated using hand gestures [25]. Saquib et al. explored body-driven graphics, in which prepared graphics are mapped to certain body parts and can follow body movement [36]. Because animation and layout of elements are determined live, the overall presentation style is well-suited for sketchy and improvisational settings. ChalkTalk [31] and Augmented Chironomia [16] employ polished graphics with predefined interactive controls and programmatic animations, which can be interacted live for high-quality visual effects. However, creating flexible interactive experiences requires significant expertise and effort.

Our work takes inspiration from these explorations but addresses an important gap. Instead of requiring users to prepare polished interactive graphics, which requires significant expertise, or requiring users to configure layout and animations live, which can add cognitive load during live presentations with the potential cost of the visual quality, we aim to create presentations that can adapt to users' real-time performance. With user-defined constraints that can be configured using typical GUIs and embedded algorithmic constraints, the adaptive presentation that we propose adapts to users' performance for spontaneity and synchronization while ensuring the visual quality does not deviate significantly from users' prepared content.

2.2 Synchronization of Gestures and Graphics

While significant research has explored using gestures to manipulate graphics on 2D and 3D spaces [46, 47, 50], closely aligned to our work are those where gestures are leveraged for their descriptive and communicative aspects [1, 18]. For example, Holze and Wilson leveraged the descriptive ability of gestures and proposed computationally matching users' descriptive gestures with 3D models of physical objects to enable the retrieval of 3D graphical objects based on users' gesture performance [18]. In the context of animation authoring, MagicalHands presented a set of gestures for complex graphical effects (e.g., particle systems) based on a gesture elicitation study, where participants provide gestures that could describe the intended animation effects [3].

Because human gestures and body postures are notoriously complex due to the rich degrees of freedom our hands and body possess, developing gestural interaction has been challenging. One approach is to reduce the rich input space to a few selected points. For example, RealityTalk [25], and Augmented Chironomia [16] utilize finger point positions of thumbs and index figures for simplicity and precise manipulation. Body-driven Graphics supports attaching graphics to body joints recognized human skeletons [36]. Other approaches leverage programming-by-demonstration [26, 26] or physical simulation [47], by allowing users to flexibly utilize the rich gesture input space for interaction.

Our work builds upon the programming-by-demonstration technique, which enables the flexible update of the transformation of graphical objects based on real-time gestural performance constrained by user-provided sample mappings between gestures and object transformations. Instead of strictly mapping gestures to graphics, which often leads to low-quality visuals due to the noise in gesture performance, we also leverage motion-warping techniques, pioneered in character animation [13, 48], to ensure smooth transformations of graphical objects.

2.3 Synchronization of Language and Graphics

The rich structures embedded in language, including the temporal, linguistic, and narrative structures, have made language-based authoring an increasingly popular content authoring paradigm for a wide range of storytelling content such as videos [19, 41], presentations [23, 49], motion graphics [20], and animation [38]. Instead of manually arranging various elements on a timeline or canvas to achieve the desired temporal and semantic congruence between language and graphics for effective storytelling [49], creators can leverage the high-level structures in the language to efficiently compose the elements that satisfy congruence.

For example, by aligning the timing of motion rhythm detected in videos with music beats, Davis and Agrawala demonstrated a method to automatically create or manipulate the appearance of dance in video for compelling audio-visual effects [10]. Crosscast proposed a set of heuristic-based algorithms to align relevant images to audio travel podcasts based on timing and semantics to automatically create travel videos [51]. With Quickcut, video creators can efficiently identify and place the numerous segments in raw footage into the final video, by aligning the transcripts of verbal annotations of raw footage with the transcript of the voiceover of the final video [41]. DataParticles employs mappings between

natural language and data properties to automate the creation of data articles featuring rich animated data visualizations [7].

In these systems, the timing of animations and transitions are aligned with the timing of the corresponding words and phrases in the script for temporal coherence, and the motion effects are often created to be semantically congruent with the script. Building upon the previous work, our system enables users to create and align corresponding visual elements, effects, and associated gestures with the script during the preparation stage to reduce editing effort. During the presentation, the system matches the presenter's speech with the script, triggering and adapting corresponding animations in real-time to ensure temporal congruence for enhanced presentation delivery.

2.4 Synchronization of Gestures and Language

Gesture in itself is a form of language [6, 8, 44]. Prior research in HCI has explored leveraging gestures as part of the communication language with computers [6, 30, 37]. For example, 'Put-that-there' explored integrating gesture and spoken language to form a mix-modality computer command[6]. Oviatt et al. explored multi-modal interaction by combining speech and pen input [9, 30]. While these works mostly focus on the integration of gestures and spoken language, they have provided insights into how people coordinate gestures and spoken language to express their intentions.

For example, prior work repeatedly reported that gestures often proceed the spoken language [24, 30], suggesting animation may need to adapt to gestures performed before the corresponding words are spoken. This contrasts systems that trigger graphics when keywords are spoken, which can then be manipulated using gestures [25]. Prior work in psychology also found that people often make larger gestures when speaking to others than for themselves [5, 14], indicating that gestures performed live in front of an audience may become larger than prepared or rehearsed, suggesting that an adaptive system should be able to handle gestures that overshoot the intended scale. Similarly, during live performances, users may deviate from their script, skipping the intended words or using alternative words with similar semantic meanings.

In this work, we consider the various cases where gesture and language may deviate from the intended performance and design algorithms to handle these deviations. While the algorithms we propose can not address all possible deviations, they contribute a step toward relaxing the requirement of tightly following the prepared presentation for intended presentation quality, reducing the effort and cognitive load required during the preparation and performance of augmented presentations.

3 COMPARATIVE ANALYSIS AND DESIGN GOALS

The goal of this work is to create engaging and compelling live augmented presentations that can deliver a sense of liveliness by enabling authors to have control of graphics using speech and gestures in a manner that mitigates the side effects of imperfect live performance to ensure high-quality presentation quality.

To achieve this, we conducted a comparative analysis of the two existing approaches of delivering live augmented presentations, the ***pre-defined presentations with memorized mappings*** [2, 29], and ***live-defined presentations with rigid mappings*** [25, 36]. Based on the analysis of their preparation and presentation mechanisms and the challenges in handling speech and manipulation deviations during live performances, we define the design goals of the proposed adaptive presentation, including how the adaptive presentations should be prepared and how they should address deviations to ensure both synchronization and visual quality.

3.1 Comparative Analysis

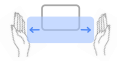
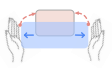
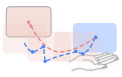
The essence of live augmented presentations is to seamlessly align presenters' performance (i.e., speech and gesture) with visuals (i.e., content and animations). The distinction among various approaches hinges on how mappings of the different modalities are configured during preparation and how these mappings were utilized during presentation to achieve the desired alignment.

*3.1.1 **P**re-defined Presentations with **M**emorized Mappings (PM).* In this approach, visual content and animations are predefined, and their mappings to speech and gestures are memorized and rehearsed. During the presentation, the presenter aligns their performance to the pre-defined sequence, timing, and animations of the visual content to achieve the intended presentation quality. This is often achieved with presentation applications that can add the live video feed from the camera as the background of slides such as Apple Keynote [2] and Microsoft Powerpoint [29]. This serves as real-time visual feedback, allowing them to view alignment on the screen as it would appear for audiences.

*3.1.2 **L**ive-defined Presentations with **R**igid Mappings (LR).* In this approach, the presenter prepares the content and specifies the order or the trigger mechanism (e.g., keywords the presenter needs to say) [25, 36] to make the content appear in the presentation. The control mechanism of the content is also specified, such as the body part that the visual will be attached to, and the hand input the content will be following. Not specified ahead of time are the beginning state (i.e., the beginning set of graphic parameters), the effects (i.e., animation presets), and the end state of the animations of the content, which will be determined during the live performance. During the presentation, when predefined triggers are detected, the content will appear at the position of the specified control input and animate based on the control (e.g., following body movement or hand manipulation). When the control input is detected from the content (e.g., next animation triggered or hand manipulation stopped), the content will stop at the final detected input position.

3.1.3 Comparison Results. Having explained the preparation and presentation mechanisms for both approaches, we then examine how the presentation qualities are impacted by various factors that may occur during the presentation. Figure 2 presents a detailed description and comparison of various conditions.

Lack of pre-defined final states of animations leads to inferior visual quality. From Figure 2, the presentation quality of the *LR* approach suffers from mistakes and imperfections during the real-time performance. Since the position, scale, and other transformations of graphic effects are contingent on the presenter's real-time performance, the uncertainties and imperfections of live performance will inevitably translate to decreased visual quality, which may in turn reduce the overall effectiveness of the concept

		Predefined + Memorized Mapping (PM)	Live-defined + Rigid Mapping (LR)		★ Predefined + Elastic Mapping (PE)
			Speech Driven	Body Driven	
Preparation (mapping configuration)	Trigger	Predefined	Keyword in script / Specific gesture	To be defined live	Gesture + script
	Duration		Predefined		Script
	Begin/End States		To be defined live		Predefined + gesture
	Motion Effects		Specific gesture	Mapped to body part	Customized gesture / Predefined
Presentation (mapping execution)	Trigger	Presenter conform to visual content	Keyword in speech / Specific gesture	Clicker	Speech and gesture
	Duration		Speak conform to predefined timing	Clicker	Speech
	Begin/End States		Hand position / Random position	Body part position	Hand position to predefined position
	Motion Effects		Direct manipulation	Direct manipulation	Realtime gesture
Runtime performance deviations		Affected presentation results			
Runtime speech deviations	Using wrong / alternative word Forgot / Skipped word	Visuals appear earlier / later than the speech content	Fail to trigger animation	*Not applicable due to manual trigger*	Follow the speech content to re-estabilish mapping or using default animation
	Speaking faster / slower		Animation end earlier / later Multiple animations racing each other		Adapt animation duration to runtime speech speed
Runtime gesture deviations	Missing / Wrong gesture Gesture detection error / Target acquisition error	Visuals not aligned with performance	Fail to trigger animation	Unintended animation effect	Adapt animation to runtime gesture
	Gesture started/ended at undesired position		Graphics appear at undesired position 		Align the graphic with gesture and animate to the desired position 
	Gesture span longer/ shorter distance (overshoot/undershoot)		Graphics animate to undesired scale 		Align the graphic with gesture and animate to the desired scale 
	Jittery motion		Jittery motion of the graphics 		Animate with smoothed motion 
Summary HOW and WHY visual quality and synchronization is impacted?	Visual quality How polished is the graphical layout, motion, etc.	Not impacted	Lack of predefined graphic states and rigid mapping causing visual content end at undesired states or have illed motion		Bound graphic to desired state to ensure visual quality
	Synchronization How well coordinated is the animation and performance	Lack of mapping between graphics and performance cause the synchronized effects hard to achieve	Lack of predefined graphic states and rigid mapping cause triggering event vulnerable to performance deviations or detection error		Loose mapping to allow graphics react to realtime performance with constrains

★ Our approach Predefined states Gesture-driven states Adapted states

Figure 2: *Comparative analysis of three difference approaches of authoring an augmented presentation.* we conducted a comparative analysis of the two existing approaches of delivering live augmented presentations, the predefined presentations with memorized mappings (PM), and the live-defined presentations with rigid mappings (LR) and compare their mechanisms with our envisioned adaptive animation presentations, which are featured as predefined presentations with elastic mappings (PE).

delivery. Without the constraint of pre-defined end states, graphics may deviate in position, scale, and alignment.

No mappings or rigid mappings are sensitive to performance errors and deviations. When graphics are unresponsive to real-time performance (*PM* approach), it places the burden on the presenter to synchronize their performance both spatially and temporally, and any misalignment will comprise the presentation quality. Rigid mappings (*LR* approach), on the other hand, synchronize and propagate the errors and deviations that occur in the input modality to others. For example, with rigid mappings, graphics will deviate with jittery gesture motions or fail to appear when the presenter misspeaks the trigger words.

In summary, as both visual quality and synchronization are important in delivering an effective and engaging presentation, we envision a versatile live augmented presentation approach that can overcome the limitations of both approaches.

3.2 Design Goals

The comparative analysis reveals an opportunity to strive for high-quality visual content while fostering spontaneous and expressive live performances. We herein proposed the following design goals for the envisioned adaptive animation:

[DG1] Anchoring the visual content with predefined states, so that graphic elements can latch onto or return back to from deviations or errors that may occur during live performances, thereby guaranteeing the visual quality

[DG2] Establishing elastic and customized mappings between animation and performance to enable graphic elements to adapt to real-time speech and gestures to achieve synchronization and expressive presentation effects.

4 ELASTICA: USER INTERFACE AND WORKFLOW

Elastica is a prototype system that incorporates our design goals (Section 3.2) to allow users to prepare and present augmented presentations with animations that are both bound to predefined states and responsive to the presenter's real-time performance. The interface of Elastica contains three major components: the visual panel, a text editor, and the configuration panel (Figure 3C).

With Elastica, a user can add graphical objects to the canvas and configure their animations, similar to other presentation applications. To author a live augmented presentation, the user can further map the parameters of animations to speech and gestures to enable graphical objects and animations to adapt to speech and gestures during live performances.

4.1 Create Graphical Objects and Configure Animations (DG1)

With Elastica, the user can add text and images to the canvas and adjust their position and size using direct manipulation. The user can further add and configure animations of the objects using the configuration panel. Elastica supports two types of animations: 'enter' and 'update'. The 'enter' animation allows users to introduce a new graphic object to the canvas, and the 'update' animations transform existing graphic objects. For both 'enter' and 'update' animations, there are three configuration parameters that Elastica supports to enable expressive visual effects.

Effects. Elastica provides simple template effects. There are three types of entering animations, 'zoom in', 'float up', and 'fade in'; and four update animations, 'transform to', 'hand follow', 'seesaw', and 'exit'. Elastica also allows users to author various customized animations with gesture demonstration (Section 4.3).

Handed. This determines which hand gesture the animation will adapt to during the presentation. Users can choose between the 'left', 'right', and 'none' options. Selecting 'none' means that the animation will not be adapted to hand gestures.

After. Elastica allows the user to specify the behavior of the object after the animation is played, including 'stay', 'exit', and 'hand following' options to support different presentation effects.

4.2 Specify Speech-Animation Mapping (DG2)

Instead of mapping an animation to a specific keyword, Elastica enables users to directly specify the intended trigger and duration of animations in relation to the text segments within a script. To define such mappings, the user can select an animation associated with a graphical object from the animation view and use markers in the script to select a text segment. Alternatively, when a text segment is selected in the text editor, the configuration panel allows users to select corresponding graphical elements (i.e., text and images) and the animation view displays all animation parameter options.

In Elastica, the entire script is used as a global timeline, and each mapped text segment serves as a local timeline of the mapped animation. As we will explain later, this mapping is elastic to allow certain deviations.

4.3 Specify Gesture-Animation Mapping (DG2)

In Elastica, the animation and gesture mapping are established by linking graphic states with corresponding hand pose states through demonstrations.

To establish the animation-gesture mapping, users can toggle on the 'customize' mode (Figure 4A). Once enabled, the hand indicators and the border of the visual panel turn green, indicating that the system is in customization mode. The configuration of gesture-graphic mappings in Elastica is achieved by demonstrating the desired graphic states and hand poses simultaneously. As shown in Figure 4A-1, the user demonstrates a pinch gesture and moves the text object to be center-aligned with index and thumb fingers.

Once the user is satisfied with the mapping presented, they can press a button to record the desired mapping between the hand pose and the graphic object state. The hand indicator flashes red to indicate a successful recording (Figure 4A-2). Multiple mappings can be created using the same interaction, allowing users to create a wide range of custom gestures to animate their graphic objects (Figure 4B). If the users are not satisfied with the mapping, they can click the 'reset' button to clear the previous records. Additionally, users can select fallback animations that will be triggered if the system fails to detect any gestures.

Rather than limiting the system to recognizing only specific gestures such as pointing or pinching, the demonstration mechanism allows users to tailor the associations between hand gestures and their preferred animations for the system to computationally respond to in real-time, as detailed in section 4.4.1.

4.4 Adapt Animations Based on Speech and Gestures During Performance (DG2)

In this section, we elaborate on how the speech-animation and gesture-animation mappings are treated algorithmically in Elastica to enable predefined animations to elastically adapt to real-time speech and gestures.

4.4.1 Adaptation Method. While a typical animation can be considered as a function of a graphic's parameters over time, in an

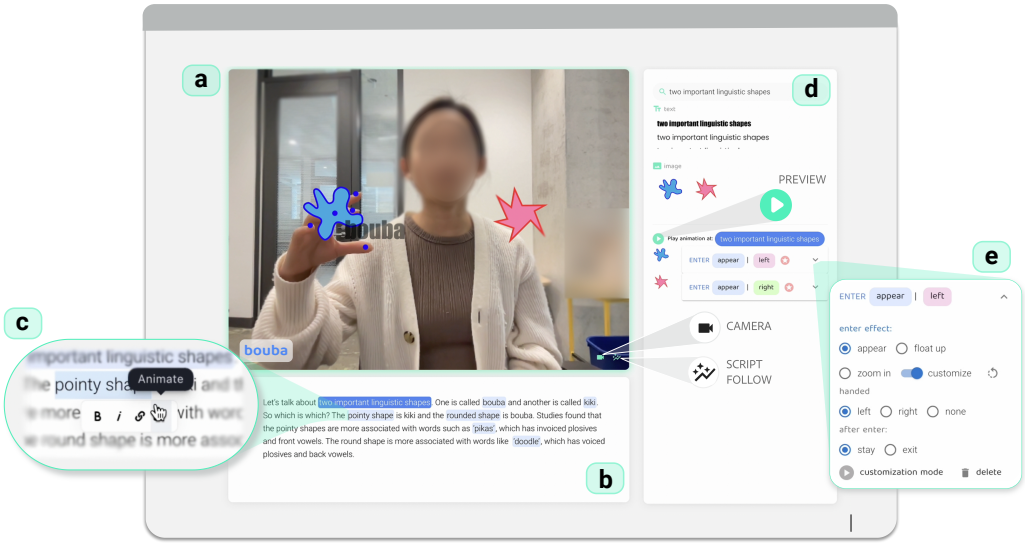


Figure 3: *Elastica Authoring Interface.* (A) the visual panel, (B) the script-based editor, (C) the configuration panel. The user can (d) highlight text segments to add animations linked with the script, (e) specify the corresponding graphical object, (f) configure the animation effects, and (g) perform the presentation.

augmented presentation, animation should additionally adapt these parameters to both speech and gesture inputs. Therefore we define adaptive animation $P_{adp}(t, g_t)$ as a function of graphic parameters that are controlled both by the time of speech (t) and the gesture (g_t) at that time. Specifically, we define adaptive animation as a blend between speech-driven animation $P_{speech}(t)$and gesture-driven animation $P_{gesture}(t)$.

Speech-driven Animation $P_{speech}(t)$ is defined as the animation of a graphic element when only speech is present as an input. This is the same as typical animation which is an interpolation between the predefined start (P_S) and end (P_E) states:

$$P_{speech}(t) = \rho(t)P_S + (1 - \rho(t))\, P_E\,, \tag{1}$$

where $\rho(t)$ is a cubic weight function easing between the start and end states.

Gesture-driven Animation $P_{gesture}(t)$ is the animation of a graphic element, given the current gesture performance. As mentioned in Section 4.3, intended animation effects are mapped to customized gestures by recording a set of n user-created mappings between graphic parameter ($P^{(i)}_{record}$) and gesture performance ($g^{(i)}_{record}$):

$$A_{record} = \left\{\left(P^{(i)}_{record}, g^{(i)}_{record}\right)\right\}_{1\le i\le n}\,, \tag{2}$$

where $g^{(i)}_{record}$ is a hand feature vector constructed using hand landmarks, and P_{record} captures the position, scale, and rotation of a graphic object. We elaborate the details of P_{record}, g_{record} in Appendix A.3.2. Given the limited discrete samples collected during preparation and the continuous space of the gesture vector, we compute the animation during the presentation by a weighted summation of all the recorded animation-gesture mappings based on the similarity between the current gesture (g) and all the recorded gestures:

$$\begin{aligned} P_{gesture}(g_t) &= \sum_{i=1}^{n} \frac{s^{(i)}}{\sum_{i=1}^{n} s^{(i)}} P^{(i)}_{record}\,, \\ s^{(i)} &= e^{-\left(\epsilon_s\left(max\left(0, \|g_t - g^{(i)}_{record}\| - b_s\right)\right)\right)^2}\,, \end{aligned} \tag{3}$$

where (ϵ_s, b_s) are hyper parameters chosen empirically (as detailed in Appendix A.3.3).

Adaptive Animation. $P_{adp}(t, g_t)$ blends the speech-driven animation and gesture-driven animation with a dynamically changing weight w considering the timing t, gesture g_t and the discrepancy between gesture-driven animation $P_{gesture}(g_t)$ and speech-driven animation $P_{speech}(t)$:

$$\begin{aligned} P_{adp}(t, g_t) &= wP_{gesture}(g_t) + (1 - w)P_{speech}(t)\,, \\ w &= F(t, g_t, r_t)\,, \\ r_t &= \|P_{speech}(t) - P_{gesture}(g_t)\|\,. \end{aligned} \tag{4}$$

Because P_{speech} is an interpolation between pre-defined animation states, which serves as the safeguards of animation quality, and $P_{gesture}$ determines the animation states based on live gesture performance, which can lead to drastic deviations, the w that controls the contribution of P_{speech} and $P_{gesture}$ plays a critical role in determining the adaption quality. Therefore, we elaborate on the heuristics and formula for choosing w.

4.4.2 Balancing Speech-driven and Gesture-driven Animations. As formulated in Equation (4), $P_{adp}(t, g_t)$ blends the speech-driven animation and gesture-driven animation with a dynamic changing

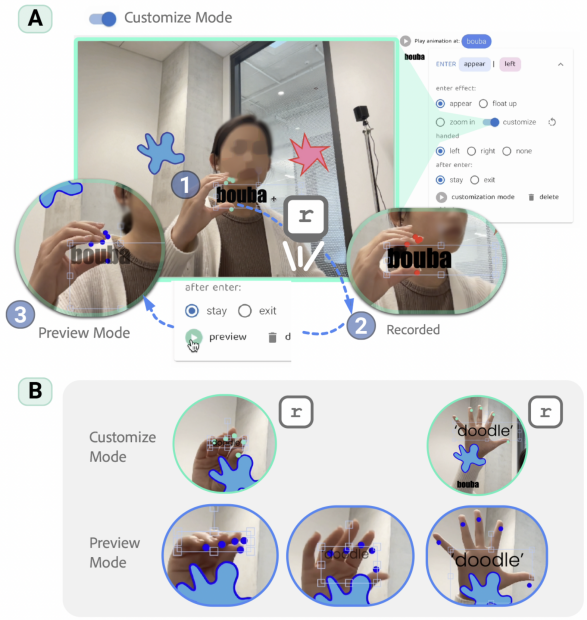


Figure 4: *Customize and Preview Animation Effects via Gesture Demonstrations.* (A) shows the workflow of customizing gesture-driven animation: the user (1) demonstrates the desired mapping states between a hand pose state and a graphic state; then (2) clicks 'r' to record the mapping; finally the user can preview the adapted results by performing the gesture. (B) illustrates how the user authors a 'zoom' effect with gesture: the user first uses the same workflow in (A) to record 2 mappings for start and end states; then previews the adapted effects to the continuously performed gesture.

weight $w = F(t, g_t, r_t)$. A larger w makes the adapted result appear closer to the gesture-driven animation, while a smaller w makes it closer to the speech-driven animation.

Under the assumption of a structured and prepared augmented presentation, these two animations are preferred in different ways: $P_{speech}(t)$ (speech-driven animation) yields grounds visual states of the object to intended timing, preferring congruence between animation and speech content. This is important in a structured presentation where it is important for the animation to be played in the right sequence. On the other hand, the $P_{gesture}(t)$ (gesture-driven animation) enables interactivity between the graphic objects and the presenter. Such congruence between animation and gesture makes the content visual compelling and engaging. While in perfect performance, these two animations are designed to be synchronized, discrepancy is almost inevitable during live presentations. Thus, to generate both visually compelling and grounded animations, the adaptation is designed with the following heuristics:

[H1]: The adapted animation should prioritize the gesture-driven animation at the beginning of the animation to better direct the viewer's attention while it grounds to the defined final state at the end of the animation.

[H2]: The adapted animation should be robust to unintentional gestures.

[H3]: The adapted animation should prioritize the predefined states when there are large discrepancies (i.e. deviations) measured by r_t.

The blending weight w captures these heuristics as a function of time t [H1], gesture g_t [H2] and discrepancy r_t [H3]:

$$w = F(t, g_t, r_t) = \Gamma(t) S(g_t) \Phi(r_t, t)\,. \tag{5}$$

$\Gamma(t)$ controls the **timing factor** [H1]. It decreases from 1 to 0 with a cos function:

$$\Gamma(t) = \cos\left(\frac{\pi}{2} t\right)\,. \tag{6}$$

$S(g_t)$ captures the **gesture intentionality** [H2] determined by evaluating the similarity of the performed gesture to a recorded gesture with the same similarity measurement as defined in Equation (3). The gesture intentionality is defined with the largest similarity (i.e. the most similar gesture) by $S(g_t) = max\left(\left\{s^{(i)}\right\}_{1 \le i \le n}\right)$. In the case of an entering animation, the system also takes the extent of hand constancy into consideration, interpreting a static gesture as a cue for intentional staging to reveal the graphic object. To determine hand constancy, the system measures hand center movements within a 0.5-second time window. If the cumulated movements is smaller than a threshold $\kappa = 5(px)$, then $S(g_t) = 1$, otherwise, $S(g_t) = max(s^{(i)})$.

$\Phi(r_t, t)$ adds a **discrepancy penalty** [H3] to the weight when the gesture-driven animation deviates from the speech-driven animation near the end of the intended time period, which usually indicates a timing shift or forgetting certain gestures that were planned. The distance penalty allows a faster convergence to the desired state in such non-ideal cases. $\Phi(r_t, t)$ is given by an inverse quadratic function:

$$\Phi(r_t, t) = \begin{cases} \frac{1}{1+\epsilon_\Phi r_t^2} & \text{if } t \ge t_0 \\ 1 & \text{if } t < t_0 \end{cases} \tag{7}$$

In Elastica, we choose $\epsilon_\Phi = 1, t_0 = 0.8$.

4.4.3 Timing in the Adaptation. The interpolation of the speech-driven animation and the w that controls the mix of speech-driven and gesture animations both depend on where the current time is in the intended period of the animation. Here we describe how time t is determined.

Each animation is linked with a chunk of text as its local timeline. The intended animation period is represented as the start (I_S) and end index (I_E) of the marked text segment in a sentence. For example, in the sentence *'Let's talk about Elastica, a tool for authoring live augmented presentations'*, the highlighted word *'Elastica'* has an intended timing of `[3,4]`. During the presentation, the index of the words is tracked with script following (details in Appendix A.2) and the current index (I^*) of the presenter's speech indicates the current time.

For each local timeline, Elastica determines an **Active Period** where the adaptation is feasible. The active period extends the intended animation period of δ words at the start point: $[I_S - \delta, I_E]$. The decision to have this extended buffer period is due to the fact that gestures often precede lexical items in communication [40].

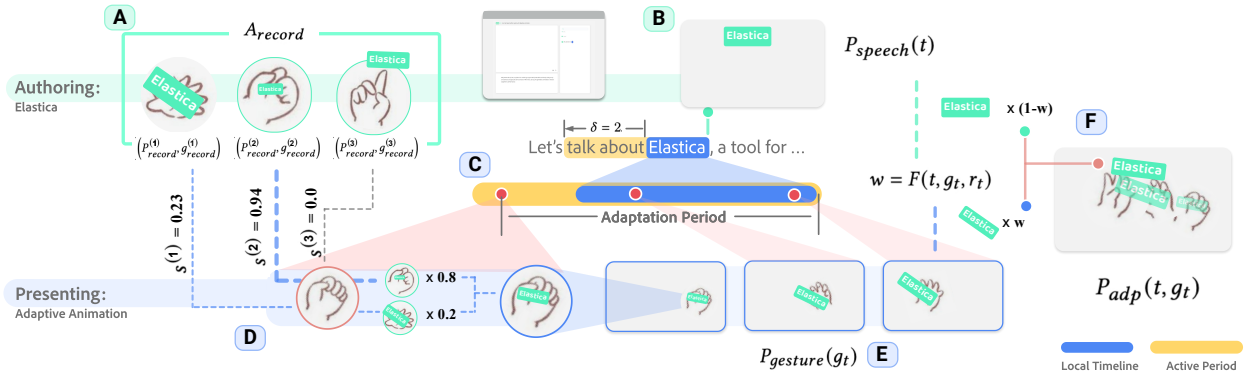


Figure 5: *How Adaptive Animation Works in Elastica.* During the authoring stage, (A) users perform gesture demonstrations in Elastica for defining mappings between gesture and graphic parameters; and then (B) use script annotation in Elastica for specifying intended time segments (local timeline). During the presentation stage, (C) determine the active period based on the local timeline; (D) detect intentional gestures and start the adaptation period; (E) generate gesture-driven animations; (F) make adaptations based on speech and gesture.

By default, we choose $\delta = 2$. For consecutive highlighted words, less buffer is needed: $\delta = 1$; and for the first word in a sentence, no buffer is needed: $\delta = 0$.

During the Active Period, Elastica monitors the gestures of the presenter and starts adaptation when the gesture is considered intentional, which opens an **Adaptation Period** and records the start index of the adaptation period (I_S^*). We estimated a total time duration for the Adaptation period with $T = \omega(I_E - I_S^*)$, where ω is a constant we used for converting word index to timing. We chose $\omega = 400(ms/word)$.

Finally, during the Adaptation Period, the adaptive animation, $P_{adp}(t, g_t)$ (Equation (4)), is calculated based on a the time elapsed since the start of adaptation (Δt) and the estimated total time (T):

$$t = min\left(1, max\left(0, \frac{\Delta t}{T}\right)\right). \tag{8}$$

4.5 Utilizing Elastica for Post Production

While Elastica is primarily designed for live augmented presentations, it also offers valuable capabilities for post-production purposes. These same interaction techniques, integrated with the underlying animation adaptation pipeline, can be effectively utilized to create augmented presentations after the initial recording.

In a post-production context, content creators can upload pre-recorded videos and automatically generate transcripts of the spoken content. By highlighting specific portions of the transcript, users can establish local timelines within the video. This action prompts the system to identify the corresponding video clips that match the highlighted text segments, allowing users to add graphic animations corresponding to the speech content.

Within each local timeline, the configured animations dynamically adapt to the gestures and speech present in the video stream, courtesy of the proposed adaptive animation technology. For more tailored and personalized results, users can take advantage of the gesture customization function, enabling them to fine-tune animations. Additionally, users have the option to import pre-recorded gesture and speech mappings as default settings for convenience and consistency in their post-production work.

5 EVALUATION OF ADAPTIVE PRESENTATION

Having developed the algorithmic methods and user interfaces that can support the preparation and presentation of adaptive presentations, we were motivated to answer the following research questions for a comprehensive understanding of the adaptive presentation approach that we proposed:

> **[RQ1] What is the user experience of creating and delivering adaptive presentations with Elastica**? Specifically, how satisfied are the users with authoring presentations with the adapted animations? How effective is Elastica in supporting preparing and presenting an adaptive presentation? How effective can adaptive animation manage deviations in live performance?
>
> **[RQ2] How does adaptive presentation compare with existing approaches to augmented presentations**? Specifically, what are the potential benefits and drawbacks the adaptation offers in terms of preparation and performance compared with alternative approaches? What are the potential gains and limitations in terms of the viewing experience?

To answer these research questions, we conducted three user studies, investigating the questions from both the presenter's and the viewer's perspectives. In the first study, participants learned and used Elastica to create augmented presentations and reported their evaluations and experiences [RQ1]. In the second study, a new group of participants was recruited to learn and use all three approaches to create augmented presentations and reported their perceptions of the differences among the three approaches [RQ2].

Finally, in the third study, we conducted a survey to gather feedback from viewers on the presentations created in the second study, enabling us to evaluate the three approaches from the viewers' perspectives [RQ2].

6 STUDY 1: CREATION STUDY WITH ELASTICA [RQ1]

We conducted a user study with 8 participants (4 female and 4 male) to evaluate the effectiveness of creating live augmented presentations using Elastica. All participants had experience creating presentations and editing videos. They reported having watched augmented presentations in the past, but none had previously created such content. Participants were recruited through university internal communication channels and mailing lists.

6.1 Study Procedure

The study for each participant lasted for 75-90 minutes. Each participant was paid $40 Amazon gift card. Seven studies were conducted in person within the lab setting. One study (P7) was conducted via Zoom, where the participant accessed Elastica with a web browser. The authoring and presenting processes with Elastica are recorded with screen and audio. The study is conducted in 4 parts:

Introduction and System Walkthrough (~30min). In this part, the experimenter briefly introduces the concept of augmented presentations and a brief motivation behind Elastica. The experimenter then walked through the features of Elastica with a short sample script. During the walk-through, the experimenter described the interaction verbally and demonstrated it with actions. The participant then practiced creating some animation effects and presenting part of the script under the guidance of the experimenter.

Reproduction Task (~25min). Participants were asked to reproduce a video created by Elastica. They were provided with the video and the script used for creating the video. During the authoring, they can use the preview function to rehearse as many times as they want. After they were done with the authoring, they were asked to rehearse and present it.

Creation Task (~20min). Since it can be challenging for participants to prepare scripts within a limited amount of time, to control the study time while allowing for creation freedom, we provided participants with 5 scripts and instructed them to choose one script for authoring and presenting with Elastica.

Questionnaire and Interview (~15min). After finishing all the tasks, participants filled out a questionnaire about their experience with Elastica. The participants were instructed to think aloud while completing the questionnaire.

6.2 Results and Findings

Overall, participants expressed that the whole experience with Elastica was unique and *"a lot of fun"* (P5, P6, P8). Participants on average spent 14.5 minutes on the reproduction task and 7.3 minutes on the creation task. They agreed that Elastica allows them to be expressive and creative while creating an augmented presentation (5 strongly agree; 2 agree; 1 neutral) and would love to use Elastica for future presentation needs (Figure 6A). We further summarize our results and findings in terms of their preparation and presentation experience with adaptive animation.

6.2.1 Planning synchronized content in a 'less rigid timeline'. All the participants felt that our script annotation interface is new, yet easy to learn and use for adaptive animation authoring (Figure 6B). The tight coupling between script and animation allowed participants to *"link (animations) into the real (live) timeframe" (P1)* and *"think how the animation could be aligned with the script"(P7).* P8 saw the script as a *"less rigid timeline"* and considered it useful for an adjustable congruence in the live setting, where the timing is less predictable: *"it doesn't make me think that, oh, I have to pace myself to read this sentence in 10 seconds. [With Elastica] I can just read it."*. Participants also found that being able to advance the presentation content using speech can result in a more natural performance. As p2 puts, *"Now I almost feel it's necessary for this (augmented presentation). You are gesturing and you don't want to perform clicking (to trigger animations)."* However, participants also mentioned that script-based authoring means they always need to map the animations with a particular part of the script, which could be limiting for animations that do not have a clear linkage with the script (P6), or for effects that are only triggered by gestures (P3, P5).

6.2.2 Preparing dynamic mappings between gesture and animation. 6 out of 8 participants (P3-P8) mentioned that being able to customize animations by gestures is their most-liked feature, enabling them to be expressive and creative by bringing their own gestures into the performance (P3) and have the agency over the animation effects they wish to achieve(P5). It also allows more natural performance by offloading their stress on remembering pre-defined gestures (such as pinch, and pointing). As P4 and P8 noted:

> *" ...I can DIY a lot of things with the gesture I am familiar with, rather than remember a particular gesture" (P4)*
>
> *"I also liked that the gesture mapping is not rigid. When I designed my gestures for myself, I realized that's all my habits, so I could perform very naturally. " (P8)*

However, compared to the script annotation feature, this feature rated lower in terms of ease of learning and use (Figure 6C). Our interviews with participants revealed two main reasons for this. Firstly, the script annotation is a familiar element from their previous presentation experience, but gesture demonstration is relatively new. Therefore, it required some effort to think about how to incorporate gestures to support their presentations. Secondly, participants reported some challenges in learning the configuration of mappings between static gestures and static states of graphic elements. While this approach resembles animation keyframing, it was not intuitive for some participants. P5 and P8 commented that they typically think of animation as a continuous motion within a period, rather than configuring discrete states. This mismatch between the mental model of users and the system's design might have contributed to the lower rating for ease of use.

6.2.3 Gradually built trust through previewing and presenting. We observed a clear trend that our participants gradually built trust in the adapted result and were satisfied with the adaptive animation generated.

> *"I was a bit confused at the beginning because there were two things happening. So you had this speech, where you had the times prepared, but you also had the gesture mapping, and I think they're defined differently. But at the end, it just worked out fine. So I started not to worry too much about it. " (P5)*

When initially presenting with Elastica, participants usually appeared to be nervous and would always hold their hands to make sure they were visible to the camera and tended to wait for the animations to happen. However, after 1-2 rounds of practice,

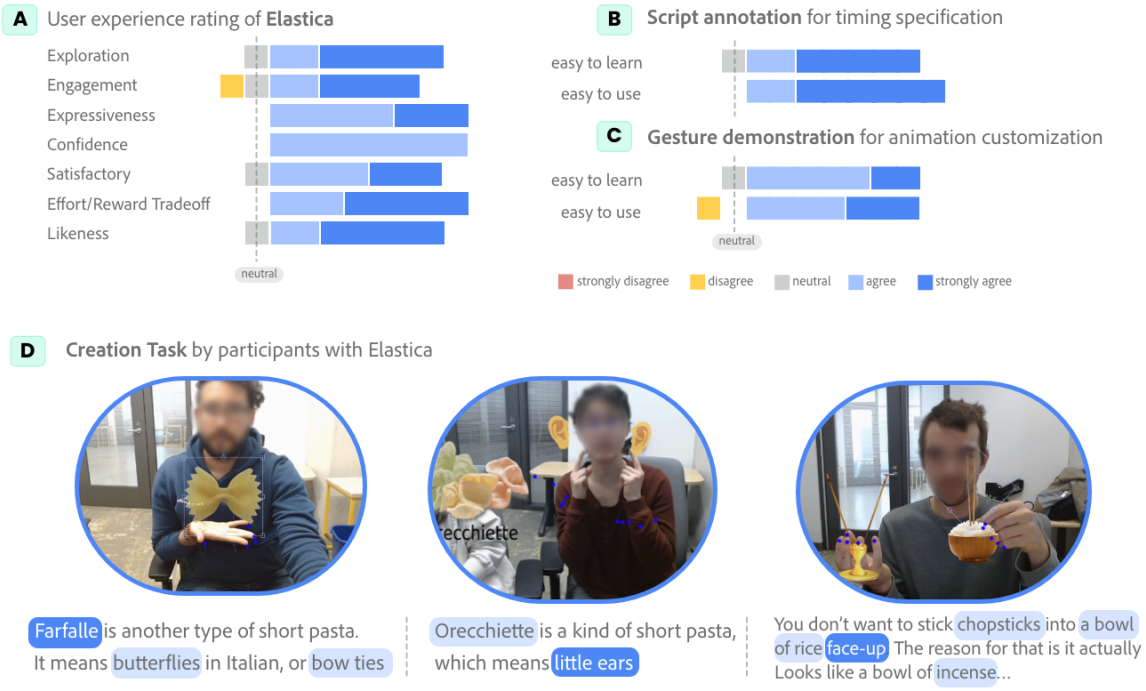


Figure 6: *Study 1 - Creation Study Results.* (A) shows the results related to the overall experience with using Elastica; (B) shows the results for the usability of two core interaction techniques in Elastica; (C) demonstrates some augmented presentation screenshots our participants created with Elastica in task 3.

they became confident and performed more naturally with smooth speech and gestural performance.

6.2.4 Predefined states put minds at ease when presenting. As our adaptation method grounds the visual results to the defined end states, it sets a guard wheel to the animation regarding the deviations that would happen during live augmented presentations. Participants mentioned that being able to perform with *"things will eventually be in the right place"(P2)* in mind can offload many of the concerns and worries while presenting live:

> *"I would've been worried if that's hand following the whole time. But I figured the main thing I wanted would happen anyway." (P3)*
>
> *"Even if I can't remember the specific gesture if I have a hand there, I know it will just pop out with some reasonable effects. So I am not too worried about presenting." (P4)*

6.2.5 Video Coding Results and findings. Several participants (P3, P6, P7, P8) have provided positive feedback on the accuracy of our script tracking and gesture recognition method used for generating adaptive presentations. One participant (P5) referred to it as "satisfying," indicating that it allows them to present the effect they designed ideally.

We also noticed some failure cases during the study. To better understand and evaluate the adaptive presentation's ability to address different types of deviations, we coded all 16 final presentation videos for both reproduction and creation tasks with observed deviations, intended results, and final adapted results.

The video coding results in 91 animations with different deviations, including speech deviations (e.g., missing or mispronouncing words) and gesture deviations (e.g., inaccurate location, overshooting, undershooting, etc.). The adaptation method successfully addressed 73.6% of these deviations, resulting in fairly smooth animations that matched the users' intended effects. We also summarize the 24 failure cases in two categories.

Failure to adapt (5/24) In this case, the system shows the default animation without adapting to the gesture performance. This is due to the failure of detecting an intentional gesture within a given time period. The 5 cases we observed during the study are either linking animation to the first word in a sentence (e.g., *"Farfalle means butterflies"*) or multiple animations are added consecutively with short words (e.g., *"You should not use chopsticks as drumsticks"*). In these cases, we opened a fairly shorter window for detection (Section 4.4.3) making the system less tolerant to the deviation of the gestures.

Flawed adaptations (19/24) In this case, the system adapts to the gestures, but the results are not ideal in terms of timing, location, or effects.

(1) Animation starting early (3/19): This occurred when the performer kept their hand static within the camera view and forgot

about the gesture, causing the system to mistake their current hand pose as an intentional staging gesture.

(2) Animation starting late (5/19): These cases were due to the same reason as the "fail to adapt" cases since the active detection window started late.

(3) Animation ending late (5/19): This happened when there were speech deviations on the highlighted words. As the speech did not follow, the playhead got stuck at certain parts and took more time to pick up the tracking point in the script.

(4) Less smooth animation (2/19): This occurred when the movement was too quick. As the gesture detection cycle was around 80ms, quick movements may result in jumpy animations.

(5) Moving to an undesired position (4/19): This occurred when the presenter forgot about the gesture or when gesture detection was less than ideal, such as poor lighting.

6.2.6 Summary. This study confirmed the effectiveness of Elastica in creating live augmented presentations. Participants find using the script annotation and gesture demonstration can effectively establish flexible mappings between animations and runtime performance. Through the previewing function, participants gradually built trust with the adaptive presentation method, allowing for more natural performances and reducing cognitive workload during presentations. The video coding results showed that, despite some external factors, most failure cases were related to the timing determined by the script, suggesting room for improvement by considering using gesture performance to time the animations.

7 STUDY 2: COMPARISON STUDY OF THREE AUGMENTED PRESENTATION APPROACHES [RQ2]

We were interested in further examining how the proposed approach compares with the other two existing approaches as we contrasted in Section 3. Therefore, we conducted a separate qualitative comparison study encompassing all three approaches to understand participants' perceptions and evaluations of the approaches concerning visual quality, synchronization, learning, and other aspects. To ensure a fair and direct comparison, we implemented the other two mechanisms within the same user interface as Elastica, resulting in three conditions:

[PM] Pre-defined Presentation with Memorized Mappings

[LR] Live-defined Presentation with Rigid Mappings

[PE] Pre-defined Presentation with Elastic Mappings

We recruited another group of six participants (2 female and 4 male) through university internal communication channels and mailing lists. The recruited participants all reported having experience creating presentations and editing videos.

7.1 Study Procedure

The study for each participant lasted for 90-120 minutes with a $40 Amazon gift card for compensation. All six studies were completed in person. The study consisted of three stages:

Introduction (~15min). The experimenter provided a brief overview of augmented presentations and introduced the three approaches participants would use for creating an augmented presentation.

Guided Creation Task (~45min). Participants were asked to author an augmented presentation with a provided script (the same as the script used for the reproduction task in Study 1). Initially, the experimenter guided them through the system on part of the script, and participants then completed the remainder independently. After each creation, the participants were given enough time to practice and rehearse the presentation until they expressed confidence in presenting. Their final presentations using each system were recorded. When introducing the three conditions, the PM condition was always introduced first as it only incorporates speech-following, upon which the other two conditions expand. The other two conditions were randomly assigned (P2, P3, P4, P6 in the order of LR and PE, and P1, P5 in the order of PE and LR).

Questionnaire and Interview (~30min). After finishing all the tasks, participants filled out a questionnaire about their experience with all three different approaches. The participants were instructed to think aloud while completing the questionnaire. The experimenter conducted a follow-up interview.

7.2 Results and Findings

Based on participants' interactions with three different systems, we summarize our findings in terms of implications of trust and control, fluidity in performance, learning curves, and the situational appropriateness of each approach.

7.2.1 Trust and Control. While existing methods provide more control and predictability in animation, they are also more restrictive. Participants expressed higher satisfaction with the performance of adaptive presentations, with 4 out of 6 strongly agreeing that it enhanced their confidence in presenting. They valued the flexibility to use natural gestures, as opposed to limited, predefined ones (P1, P4, P6).

However, participants noted a lack of control with adaptive animations, as these are generated in real-time. As P3 pointed out, even though the final state of the animation was defined, the actual motion path of this approach (PE) was less predictable. We observed that presenters occasionally wanted to intentionally create 'deviated' visual effects. For example, P3 circled their finger when triggering an object to appear. However, this decorated path was smoothed out by the adaptation method. This adaptability also made imagining the exact output challenging for the presenters (P5, P6). This aligns with earlier findings that establishing trust in the system's output requires additional preview and rehearsal, which cost an extra effort compared to existing approaches.

7.2.2 Fluidity of Performance. In Section 3, we compared the three approaches in terms of visual quality of graphics and synchronization. Our study results showed that these mechanisms indeed led to different qualities in terms of the fluidity and authenticity of the performance. The participants recognized the ability to deliver a seamless and genuine performance as a distinctive advantage of the adaptive approach, thanks to its high tolerance to gesture detection errors. As pointed out by P6:

> "You can actually add many expressive and creative details with a simple gesture." (P6)

"... I liked it allows you to perform gestures without worrying too much about whether it makes sense to the system. It will always give me effects that make sense. " (P2)

On the other hand, participants raised concerns about the fluidity of performance for the LR approach. P1 and P3 both commented that using 'pinch' to select and manipulate graphic objects added unwanted pauses in their presentation. P1, P2, P4, and P6 all encountered gesture detection failures that caused unintentional object movement.

7.2.3 Preparation Cost and Outcome Expectation. The adaptive presentation approach was seen as fun, engaging, and creative. However, participants felt that mastering it required more time and practice compared to the existing approaches.

" It takes some time to be more familiar with like how to configure the gesture and the animation. But I can see like over time when I'm more familiar with it, it's not too hard." (P1)

In contrast, the other two conditions employ a more direct mapping and require fewer configurations. This simplicity made them easier to grasp, allowing users to quickly understand and confidently utilize them while recognizing the boundaries of the mechanism. As noted by P2:

"It (LR) is super easy because there's absolutely no setup work that I have to do. Yeah. All I have to know is what element is going to appear. That's the only thing I have to remember...And my practice is also pretty straightforward."

Interestingly, the original goal of the adaptive presentation approach was to let users perform gestures without worrying about precision, this is compromised due to the heightened cost of customizing the mappings between animations and gestures. Due to the extra required effort, users began to aim for the exact gestures they specified to achieve optimal visual outcomes. This was evident during the rehearsal stage, where 4 out of 6 presenters posed a similar question: "What did I customize for this animation?" While continued practice could boost users' proficiency with the adaptive animation method, our results suggested that a steeper learning and preparation curve can set higher outcome expectations, potentially placing additional stress on presenters.

7.2.4 Well-suited Settings. Participants noted that our adaptive presentation approach is best suited for scenarios where content needs to be delivered with both high visual quality and engagement (e.g., pitching to investors). Participants found allowing the content to initially follow gestures and then settle into the intended position is particularly valuable for storytelling, where the presenter relies on animations to guide the viewer's attention. The animation created considering both gesture and defined state can *"add a layer of importance in presentation"* (P2) While the LR approach required the least amount of configuration, as graphic states were defined in real-time, participants expressed reservations about using this method for formal and structured presentations. However, they believed it is better suited for informal discussions and brainstorming sessions (P1, P2, P4). Surprisingly, the PM approach garnered substantial praise for its smooth motion and complete predictability. However, content created with the PM approach was perceived as more mundane compared to the other two methods (P1, P6). P1 mentioned their preference for using the PM approach when presentations demand precise control over visual elements.

Criteria	Rating (5-likert scale)	
	1 (lowest)	5 (highest)
expressiveness the animation's ability to convey concepts in a rich and engaging manner	dull and lifeless	highly expressive and engaging
synchronization how well the animation aligns with the presenter's physical movements	completely deviate from each other	perfectly synchronized
smoothness the fluidity and seamless transition of the animation	choppy and inconsistent	extremely smooth and seamless
non-distracting the animation's effectiveness in complementing rather than detracting from the content	highly distracting	enhances focus on content
overall visual quality general assessment of the visual appeal and impact of the animation	poor quality: awkward to watch	exceptional quality: pleasant to watch

Figure 7: *Viewer's Evaluation Rubrics*

7.3 Summary

Our comparison study results revealed that the adaptive presentation, as a collaborative interplay between the presenter and the algorithm, results in high satisfaction for its flexibility, natural gesture support, and fluidity of the presenter's performance compared to existing approaches. However, its unpredictability on the motion effect can sometimes make the presenters feel a lack of control and trust over the system. Its relatively more complex configuration will likely set higher outcome expectations.

8 STUDY 3: VIEWING EXPERIENCE OF THREE AUGMENTED PRESENTATION APPROACHES [RQ2]

While the above two studies focused on evaluating our approach from the presenters' perspectives, we wanted to further investigate the impact of adaptive presentation on *viewers' perception* and how it differs from the previous approaches.

8.1 Evaluation Material Preparation

We sought to use videos recorded from Study 2 as comparison materials. However, these videos contain participants' performance errors, which can arbitrarily affect our evaluation. Our major interest was to investigate the effects of the inherent characteristics of the three approaches on viewers' experiences. This comparison required materials where all presentations were *well performed* across three conditions. To this end, we curated clips from the presentations created in Study 2 with the following criteria: (a) *Consistent Quality.* We selected clips that exhibited smooth, error-free performances across all three conditions, to minimize the influence of technical glitches or performance mishaps. (b) *Diverse Representation.* We aimed to include a broad range of participants, to avoid a single individual's performance disproportionately influencing the

results. (c) *Comprehensive Coverage.* We curated clips that collectively encompass the full narrative of the presentation, ensuring viewers could fully grasp the content being presented.

Since our adaption approach emphasized adapting the animation to gestural movement, we were also interested in a focused assessment of the perceived animation quality by viewers. Therefore, we segmented the clips from each condition into smaller segments, each featuring a singular animation event (e.g., the entrance animation of the word 'Elastica' as depicted in Figure 5). We removed the speech component from these smaller segments, as the corresponding speech was often incomplete, with words cut into syllables. The resulting clips were saved as GIFs. The compiled videos and individual clips are included in the supplemental materials.

8.2 Procedure

The evaluation was conducted with online survey method. We distributed our online survey through multiple listservs for a wide variety of participants.

The survey comprised four sections. The first three sections presented participants with three sets of GIFs selected randomly from the aforementioned animation segments. Each set was drawn from specific segments of a speaker's presentation. In the fourth section, participants were presented with the compiled videos of all three conditions sequentially. Following the viewing of the animation segments/videos, participants were asked to rate the clips based on 5 criteria described in Figure 7, using a 5-point scale. To reduce bias, the order of viewing sequences for the three conditions was randomized. At the end of the survey, we collected additional comments on each approach to gather more nuanced feedback.

8.3 Results and Findings

We received 42 valid responses from our survey, including ratings for a set of compiled video clips and three sets of individual animated GIFs. Ratings for three individual animated GIFs are aggregated as one score during the analysis. We calculated the average viewer's ratings on each metric across three conditions. For each metric in our rubric, we performed a pairwise t-test to identify any significant differences among the three conditions. The results of these tests are presented in Figure 8.

Overall, our results indicated that the output created by our approach (PE) was able to provide a distinctively engaging and comprehensible viewing experience. As shown in Figure 8 (a), it significantly outperforms the PM approach in terms of expressiveness (p = .0005) and synchronization (p = .0191) and surpasses the LR approach in smoothness (p = .0015), unobtrusiveness (p = .0459), and overall visual quality (p = .0005). Notably, viewers' perception of visual quality varies across individual animations and within the broader context of a presentation. The LR approach, in particular, received higher ratings for individual animation segments (Figure 8 (b)). We elaborate on these findings in the following sections.

Note that, since the three approaches ultimately lead to three different types of performance, and the way we select video clips relies on subjective assessment, we recognize the potential limitations of our selection process and its impact on the generalizability of our findings. Hence, the statistical observations are mainly used to drive our observations and elicit discussions around the understanding of the viewer's perception of these augmented presentations. To better inform the findings, we supplemented these statistical results with qualitative feedback gleaned from participant responses.

8.3.1 Balance between Expressiveness and Unobtrusiveness in Animation. The study results highlight the need for a balance between expressiveness and unobtrusiveness in animations used within presentations.

Our method effectively delivered engaging visuals that were both expressive and non-distracting, avoiding overshadowing the content. As one viewer noted, *"I liked that performance 3 was more dynamic than 1, but less distracting than 2. I feel like it was a good balance between the two."* (P17) This aligns with our goal of providing grounded while reactive visuals that complement the presenter's performance. Despite viewers acknowledging the animations' synchronicity and smoothness, the added effects were sometimes seen as overwhelming for content consumption. The easing effect in the algorithm occasionally made animations feel slower and more lagging compared to other methods, leading to moments where they seemed to *"overly linger on the presenter's movements"* (P32).

8.3.2 Perception of visual quality varies under different contexts. The results reveal that the perception of visual quality varies between individual animations and within the context of a full presentation.

The LR approach was noted for its expressiveness ($p_{12} < .0001$, $p_{23} = .0039$) and synchronicity ($p_{12} < .0001$, $p_{23} = .0002$) in individual animations. However, within a presentation setting, this condition introduced noticeable misalignments between speech and gesture, along with graphical quality issues, as also noted in Study 2. This led to significant distractions ($p_{12} < .0001, p_{23} = .0459$) for viewers and diminished the overall perceived quality ($p_{12} = .0355, p_{23} = .0005$): *"Very distracting that your eyes are drawn to the shaky and constantly moving graphics, and make you wonder, would the presenter successfully have them stop at the desired position."* (P2). Conversely, the PM approach's emphasis on control and simplicity garnered notably higher perceived quality ratings within the presentation context, showing viewers' preference for unobtrusiveness over expressiveness and synchronization. As exemplified by one comment, *"The synchronization is slightly off, but it's more acceptable than animations being distracting as in other conditions."* (P39)

8.4 Summary

This study highlights that viewers' perceived visual quality varies in different contexts, emphasizing the need for a balance between expressiveness and unobtrusiveness in animations used in presentation settings. Our approach makes a step towards this objective, by delivering engaging visuals without overshadowing content, and was appreciated for its collaborative aspect. However, it occasionally felt overwhelming due to additional animation effects.

9 DISCUSSION AND FUTURE WORK

9.1 Adaptive Animation: Can it facilitate?

Our comparative analysis in section 3 explored three approaches based on mechanisms and theoretical capabilities. The analysis revealed an opportunity to integrate grounded visual content [DG1]

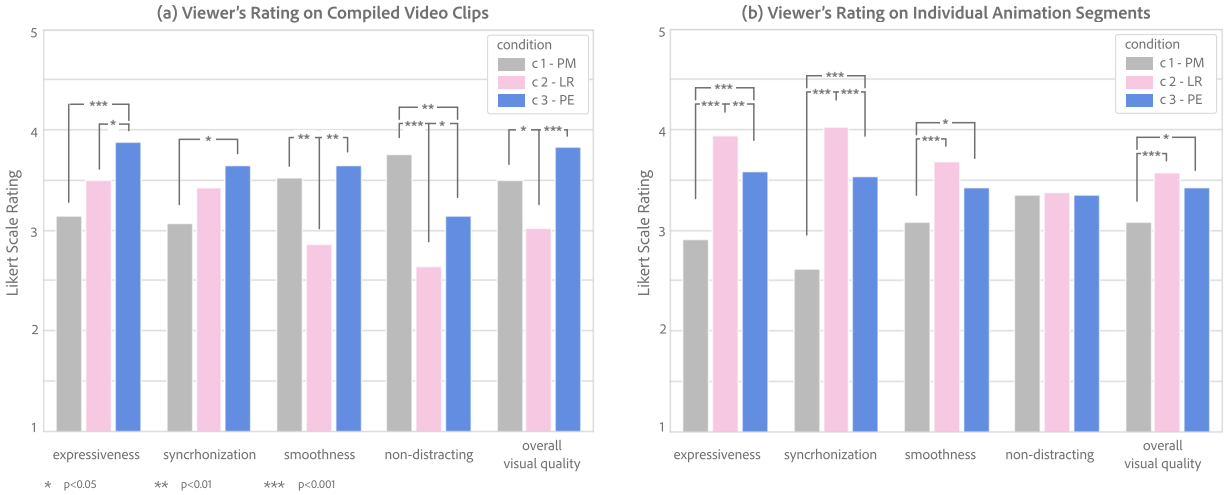


Figure 8: *Study 3 - Comparative Viewer Rating Analysis Results.* (a) shows the average viewer ratings for the compiled video clips, with the significance of differences across three distinct conditions; (b) shows the aggregated average ratings for individual animation segments, with the significance of differences across three distinct conditions.

and foster expressive animations [DG2] in the presentation. We proposed adaptive presentation with predefined visual states and elastic mappings to bridge the gap. Our evaluations with both presenters and viewers highlighted our adaptation method provides benefits in effectively balancing the expressiveness and unobtrusiveness of animations within an augmented presentation while also revealing certain challenges in terms of animation controllability. Moreover, these empirical insights shed light on the nuances of the other two approaches in terms of their strength, weaknesses, and suitable scenarios. We comparatively summarized our findings as shown in Figure 9.

9.2 Adaption, to What Extent?

With the algorithmic method we proposed, we wished to conduct a comprehensive technical evaluation of its performance. However, traditional methods like error rate calculation were impractical given the complexity of the adaption algorithm, which involves multiple factors and weights calculated as numerical similarity rather than categorization. In addition, it has proved challenging for us to collect reliable testing cases from users' real performances, as it was difficult to control how much a user should deviate from an intended performance. We have considered controlled technical evaluations using simulated avatars, but the vast array of possible gestural deviations in 3D space made this approach infeasible.

Therefore, we opted for assessing participants' real-world performance in our user study, identifying failure cases, and understanding major failure reasons, as we reported in Section 6.2.5. However, defining a standard metric for "successful adaptation" remained elusive. In our current evaluation, we considered an adaptation successful if the visual object appeared at the intended end state with smooth animation. However, this criterion did not consistently correspond with the intended outcomes by the presenters or viewers. For example, Study 2 highlighted instances where presenters intentionally deviated animations for emotional impact; Study 3 shows viewers occasionally prefer more constrained and minimal animations to avoid distractions. These findings emphasized the importance of considering users' perception in evaluating animation effectiveness, beyond just the technical aspects of the animation.

THe adjustable hyperparameters in our approach enable a nuanced balance between more pronounced and restrained animations. However, the absence of quantitative metrics to assess perceived visual quality complicates both the evaluation and optimization processes. This challenge highlights a broader issue on human-center evaluation within computational generative methods: the critical need to develop metrics that prioritize human perception, rather than solely relying on analytical standards.

9.3 Understanding the Semantic Associations of Gesture and Speech

Although gesture and speech are considered semantically parallel as they refer to the same underlying cognitive representation [17], our research focuses on their discrepancies during presentations. These discrepancies, although often imperceptible in real-life communication, can add an extra challenge in generating automatic solutions to facilitate information consumption by leveraging information from both channels simultaneously.

To develop the algorithm for adaptive presentation, we relied on prior research that found a relationship between speech and gesture, where gestures usually precede lexical items and end before a full meaning is communicated through speech. We incorporated these empirical findings into fixed parameters. However, during evaluation, we encountered failure cases that highlighted the variability in the timing of speech and gesture under different contexts. This could be compensated by understanding the semantic association between gestures and speech to further improve the algorithm's performance. These semantic associations can also be leveraged to

		Predefined + Memorized Mapping (PM)	Live-defined + Rigid Mapping (LR)	★ Predefined + Elastic Mapping (PE)
Strength	perceived by presenters	full predictability with highly grounded visual content; animation has clean motion	require minimal setup; has high controllability; allow flexible improvisation	support flexible natural gestures; has grounded visual content; require low rehearsal cost with high animation quality
	perceived by viewers	clean, simple and smooth animation with minimal distraction to the content	snappy, lively animation; well-synchronized with gesture performance providing the direct interactive feeling	well-synchronized with both speech and gesture; expressive yet unobtrusive, enhancing focus on the content
Weakness	perceived by presenters	require high rehearsal cost for desired synchronization; lack of gesture flexibility	disruptive to presentation flow; high cognitive load and insecurity when presenting; less polished animations	difficult to predict/control the resulting animation; higher cost on learning and content preparation
	perceived by viewers	less synchronized and spontaneous; visual content less engaging	highly distracting due to the interrupted presentation flow caused by speaker's stutters; less formal visual output	contain lagging animations due to gesture smoothing/tuning, introducing unnecessary motions
Suitable Scenarios		**Content-Focused Presentations** scenarios prioritizing content over visual appeal, such as report presentations, and academic lectures. Emphasis is on clear, direct delivery of information.	**Informal and Dynamic Presentations** less formal environments that require live improvisation, such as informal discussions and brainstorming sessions. These settings benefit from a relaxed, interactive approach.	**Highly Engaging Visual Presentations** situations where both high visual quality and audience engagement are crucial. This includes commercial teasers, educational materials for children, where vivid visuals and captivating presentation styles are key.

★ Our approach

Figure 9: ***Extended comparison of three augmented presentation authoring approaches from our evaluation results***

suggest suitable gestures that are congruent with speech or desired visual effects. Although gestures are ubiquitous in communication, designing gestures can be challenging for end users. Our future work will explore methods to suggest appropriate gestures, given the speech and visual content.

9.4 On a Human-Machine Collaborative Perspective

Authoring an adaptive presentation can be seen as a collaboration between human performance and intelligent adaption. From that perspective, the PM approach delivers a presentation with completely machine-generated animation, which is perceived as 'precise' and 'clean', but 'less spontaneous'; while the LR approach relies on fully human control, which results in 'snappy', 'lively' viewing experience, but could be 'messy' and 'distracting' in terms of content delivery. The elastic mapping we proposed aims to establish a collaborative setting where predefined parameters provide a common ground for human-machine interaction.

Nonetheless, identifying additional signals for the machine to utilize could potentially enhance its adaptation. For instance, to improve the visual quality of impromptu presentations, we can consider inferring desired graphical layouts and effects by drawing from existing graphics. This could involve harnessing Gestalt Theory to structure graphics effectively (as demonstrated in [32]), aligning with the semantics embedded in the speech content to infer graphic structures (as demonstrated in [49]); to determine the nuanced timing and effects of an animation, we can incorporate the semantics of gestures and better adjust animations based on the nature of continuous gestural motion.

In the pursuit of creating an ideal collaborative output, we believe it is important to establish flexible and context-aware mechanisms that adjust the division of control between humans and machines, which would benefit from combining contextual signals with established theoretical principles.

10 CONCLUSION

Through a comparative analysis of the existing approaches in supporting live augmented presentations, we identify a need for supporting the authoring of live augmented presentations that are adaptive to live performances to achieve synchronization while ensuring visual quality. We fulfill this need by proposing pre-defined presentations with elastic mappings. We integrate the adaptive presentation concept into a prototype system, Elastica, featured with script annotation and gesture demonstration for configuring the elastic mappings. Our evaluation demonstrates the usefulness of Elastica in supporting the creative authoring of augmented presentations and the effectiveness of our adaptation method in achieving the desired expressiveness, synchronization, and unobtrusiveness in an augmented presentation.

ACKNOWLEDGMENTS

We extend our gratitude to Fuling Sun and Peiling Jiang for their invaluable assistance with the user study; to Ana Maria Cardenas Gasca for her contributions to the creation of the figures and her insightful feedback during the early stages of this work; and to Jane E for visual designs and proofreading the manuscript.

REFERENCES

[1] Roland Aigner, Daniel Wigdor, Hrvoje Benko, Michael Haller, David Lindbauer, Alexandra Ion, Shengdong Zhao, and JTKV Koh. 2012. Understanding mid-air hand gestures: A study of human preferences in usage of gesture types for hci. *Microsoft Research TechReport MSR-TR-2012-111* 2 (2012), 30.

[2] Apple. 2024. Features available with Keynote. https://www.apple.com/keynote/features/

[3] Rahul Arora, Rubaiat Habib Kazi, Danny M. Kaufman, Wilmot Li, and Karan Singh. 2019. MagicalHands: Mid-Air Hand Gestures for Animating in VR. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 463–477. https://doi.org/10.1145/3332165.3347942

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

[5] Janet Beavin Bavelas, Nicole Chovil, Linda Coates, and Lori Roe. 1995. Gestures specialized for dialogue. *Personality and social psychology bulletin* 21, 4 (1995), 394–405.

[6] Richard A. Bolt. 1980. "Put-That-There": Voice and Gesture at the Graphics Interface. *SIGGRAPH Comput. Graph.* 14, 3 (jul 1980), 262–270. https://doi.org/10.1145/965105.807503

[7] Yining Cao, Jane L E, Zhutian Chen, and Haijun Xia. 2023. DataParticles: Block-based and language-oriented authoring of animated unit visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15.

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

[9] Philip R Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow. 1997. Quickset: Multimodal interaction for distributed applications. In *Proceedings of the fifth ACM international conference on Multimedia*. ACM, New York, NY, USA, 31–40.

[10] Abe Davis and Maneesh Agrawala. 2018. Visual Rhythm and Beat. *ACM Trans. Graph.* 37, 4, Article 122 (jul 2018), 11 pages. https://doi.org/10.1145/3197517.3201371

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

[12] Randi A Engle. 2022. Not channels but composite signals: Speech, gesture, diagrams and object demonstrations are integrated in multimodal explanations. In *Proceedings of the twentieth annual conference of the cognitive science society*. Routledge, London, United Kingdom, 321–326.

[13] Michael Gleicher. 1998. Retargetting Motion to New Characters. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '98)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/280814.280820

[14] Susan Goldin-Meadow. 2005. *Hearing gesture: How our hands help us think*. Harvard University Press, Cambridge, MA, USA.

[15] Jiangtao Gong, Teng Han, Siling Guo, Jiannan Li, Siyu Zha, Liuxin Zhang, Feng Tian, Qianying Wang, and Yong Rui. 2021. HoloBoard: a Large-format Immersive Teaching Board based on pseudo HoloGraphics. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 441–456.

[16] Brian D Hall, Lyn Bartram, and Matthew Brehmer. 2022. Augmented Chironomia for Presenting Data to Remote Audiences. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 1–14.

[17] Judith Holler and Geoffrey Beattie. 2003. How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica* 141-1, 4 (2003), 81–116.

[18] Christian Holz and Andrew Wilson. 2011. Data Miming: Inferring Spatial Object Descriptions from Human Gesture. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 811–820. https://doi.org/10.1145/1978942.1979060

[19] Bernd Huber, Hijung Valentina Shin, Bryan Russell, Oliver Wang, and Gautham J Mysore. 2019. B-script: Transcript-based B-roll video editing with recommendations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–11.

[20] Amir Jahanlou and Parmit K Chilana. 2022. Katika: An End-to-End System for Authoring Amateur Explainer Motion Graphics Videos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14.

[21] Seokmin Kang, Barbara Tversky, and John B Black. 2015. Coordinating gesture, word, and diagram: explanations for experts and novices. *Spatial Cognition & Computation* 15, 1 (2015), 1–26.

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

[23] Mackenzie Leake, Hijung Valentina Shin, Joy O Kim, and Maneesh Agrawala. 2020. Generating Audio-Visual Slideshows from Text Articles Using Word Concreteness.. In *CHI*, Vol. 20. ACM, New York, NY, USA, 25–30.

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

[25] Jian Liao, Adnan Karim, Shivesh Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. 2022. RealityTalk: Real-Time Speech-Driven Augmented Presentation for AR Live Storytelling, In Proceedings of the 35rd Annual ACM Symposium on User Interface Software and Technology. *arXiv preprint arXiv:2208.06350*, 1–12.

[26] Hao Lü and Yang Li. 2013. Gesture studio: authoring multi-touch interactions through demonstration and declaration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 257–266.

[27] Fabrice Matulic, Lars Engeln, Christoph Träger, and Raimund Dachselt. 2016. Embodied interactions for novel immersive presentational experiences. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1713–1720.

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

[29] Microsoft. 2024. Presenting with cameo. https://support.microsoft.com/en-gb/office/presenting-with-cameo-83abdb2e-948a-47d0-932d-86815ae1317a

[30] Sharon Oviatt. 1999. Ten Myths of Multimodal Interaction. *Commun. ACM* 42, 11 (nov 1999), 74–81. https://doi.org/10.1145/319382.319398

[31] Ken Perlin, Zhenyi He, and Karl Rosenberg. 2018. Chalktalk: A Visualization and Communication Language–As a Tool in the Domain of Computer Science Education. *arXiv preprint arXiv:1809.07166* (2018). https://doi.org/10.48550/arXiv.1809.07166 arXiv:arXiv:1809.07166

[32] Florian Perteneder, Martin Bresler, Eva-Maria Grossauer, Joanne Leong, and Michael Haller. 2015. CLuster: Smart Clustering of Free-Hand Sketches on Large Interactive Surfaces. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 37–46. https://doi.org/10.1145/2807442.2807455

[33] Hans Rosling. 2006. *Debunking third-world myths with the best stats you've ever seen*. TED.

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

[35] Hans Rosling. 2013. The River of Myths.

[36] Nazmus Saquib, Rubaiat Habib Kazi, Li-Yi Wei, and Wilmot Li. 2019. Interactive body-driven graphics for augmented video performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM New York, NY, New York, NY, USA, 1–12.

[37] Arjun Srinivasan, Bongshin Lee, Nathalie Henry Riche, Steven M. Drucker, and Ken Hinckley. 2020. InChorus: Designing Consistent Multimodal Interactions for Data Visualization on Tablet Devices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376782

[38] Hariharan Subramonyam, Wilmot Li, Eytan Adar, and Mira Dontcheva. 2018. Taketoons: Script-driven performance animation. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 663–674.

[39] Ryo Suzuki, Rubaiat Habib Kazi, Li-yi Wei, Stephen DiVerdi, Wilmot Li, and Daniel Leithinger. 2020. RealitySketch: Embedding Responsive Graphics and Visualizations in AR through Dynamic Sketching. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*. Association for Computing Machinery, New York, NY, USA, 166–181. https://doi.org/10.1145/3379337.3415892

[40] Marlijn Ter Bekke, Linda Drijvers, and Judith Holler. 2020. The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech. (2020). https://doi.org/10.31234/osf.io/b5zq7

[41] Anh Truong, Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2016. Quickcut: An interactive tool for editing narrated video. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 497–507.

[42] Barbara Tversky and Azadeh Jamalian. 2021. Thinking Tools: Gestures Change Thought About Time. *Topics in Cognitive Science* 13, 4 (2021), 750–776. https://doi.org/10.1111/tops.12566 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12566

[43] Barbara Tversky, Julie Bauer Morrison, and Mireille Betrancourt. 2002. Animation: can it facilitate? *International journal of human-computer studies* 57, 4 (2002), 247–262.

[44] Clayton Valli and Ceil Lucas. 2000. *Linguistics of American sign language: An introduction*. Gallaudet University Press.

[45] Vox. 2015. Obama on what most Americans get wrong about foreign aid. Retrieved 2023 from https://youtu.be/nzL_avUlEE

[46] Andrew D Wilson and Hrvoje Benko. 2010. Combining multiple depth cameras and projectors for interactions on, above and between surfaces. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, New York, NY, USA, 273–282.

[47] Andrew D Wilson, Shahram Izadi, Otmar Hilliges, Armando Garcia-Mendoza, and David Kirk. 2008. Bringing physics to the surface. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*. ACM, New York, NY, USA, 67–76.

[48] Andrew Witkin and Zoran Popovic. 1995. Motion warping. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. Association for Computing Machinery, New York, NY, USA, 105–108.
[49] Haijun Xia. 2020. Crosspower: Bridging graphics and linguistics. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 722–734.
[50] Haijun Xia, Michael Glueck, Michelle Annett, Michael Wang, and Daniel Wigdor. 2022. Iteratively Designing Gesture Vocabularies: A Survey and Analysis of Best Practices in the HCI Literature. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–54.
[51] Haijun Xia, Jennifer Jacobs, and Maneesh Agrawala. 2020. Crosscast: Adding Visuals to Audio Travel Podcasts. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 735–746. https://doi.org/10.1145/3379337.3415882

A APPENDIX

A.1 Elastica Source Code

https://github.com/Rrrima/Elastica.git

A.2 Script Following Details

The goal of the script following tech is to enable graphical effects to be anchored to a text script and triggered in real-time via the presenter's speech. The input to the system is the text script provided by the user and the streaming audio. The system components used in this approach are described as follows:

A.2.1 Speech-to-Text Transcription. This initial component employs the Microsoft Azure Speech-to-Text service to perform real-time transcription of spoken language from an audio source into written textual form. It transforms continuous audio input into a textual script for further processing.

A.2.2 ScriptLocationPredictor. The ScriptLocationPredictor is a pivotal model within our framework that enables the synchronization of graphical effects with the presenter's speech. It takes the text script and streaming audio as the input. Its objective is to probabilistically identify key positions within the script where specific trigger words or phrases are expected to occur. We leverage standard NLP techniques and libraries for sentence tokenization, indexing, and fuzzy word matching to predict the locations in the script where trigger words are likely to occur.

A.2.3 ScriptAdvancer. The ScriptAdvancer serves as the orchestrating entity responsible for dynamically managing the progression of the presentation script in real-time. It utilizes the output of the ScriptLocationPredictor to smoothly advance the script position as the speaker delivers their presentation. When a trigger word or phrase is detected in the audio stream, the ScriptAdvancer identifies the corresponding location within the script and instructs the system to initiate the associated animation effect. This orchestration ensures a synchronized and engaging presentation experience.

A.2.4 Trigger Animation Effect. The Trigger Animation Effect module is an integral part of our system, activated in response to the identification of specific trigger words or positions within the presentation script by the ScriptAdvancer. Upon activation, this module engages a library of predefined animation effects, associating each effect with particular trigger words or positions. These effects are seamlessly integrated into the presentation, enhancing the visual and interactive aspects of the overall presentation experience.

One of the advantages here is that we do not get false positive triggers as compared to when we have single keyword-based speech triggers because we have more context while matching the triggered word. We can handle spoken errors and mistakes, improvisations and transcription errors in the speech while following the input script.

A.3 Adaptive Animation Method Details

We report all the defaults including hyperparameters, functions, and feature vectors that we used in constructing the adaptive animation method described in Section 4.4. All the parameters are tuned based on their performance while developing. Note that these parameters can also be exposed to users for further customization of animation, even though this is not yet exposed in our current prototype interface for user studies.

A.3.1 $\rho(t)$ in equation 1. By default, animations in Elastica are tweening with cubic in/out ease function. With the local $t \in [0, 1]$ specified in equation 8, we define the weight function in equation 1 as:

$$\rho(t) = \begin{cases} 1 - 4t^3 & t \leq \frac{1}{2} \\ 4(1-t)^3 & t > \frac{1}{2} \end{cases} \tag{9}$$

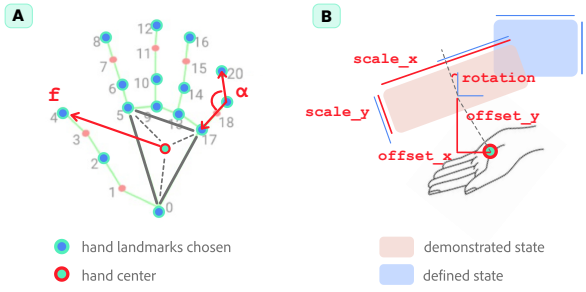


Figure 10: Visualizing g_{record} (A) and P_{record} (B).

A.3.2 g_{record}, P_{record}. In Elastica, a hand gesture is characterized by a hand feature vector, which is constructed using the hand landmarks detected through Mediapipe [28], which captures the direction of each finger pointing as a 3-dimensional vector f and the bend angle of each finger as α (Figure 10-A). The 3-dimensional vector f is calculated with the start point as the hand center, which is the geometric center of landmarks (0, 5, 17).

Given the influence of different scales of f and α, we applied normalization with:

$$\alpha^* = \frac{\alpha}{90} - 1, f^* = f * 0.1 \tag{10}$$

The feature vector for each finger v_j, is a concatenate of f^* and α^*. The hand feature vector $g^{(i)}_{record}$ in each record i is hence a 20 dimension concatenation of vectors v_j, $j \in$ {thumb, index, middle, ring, pinky}.

The graphic vector $g^{(i)}_{record}$ in each record i is a 5-dimensional vector constructed with scales, offsets from the hand center and

the rotation angle (Figure 10-B). Note that while scales and rotation angle parameters are relative to the defined state of the object (i.e., the end state defined with script annotations), the positional parameters are relative to the hand center. Thus, when computing the adapted value of an object, the absolute position of the object on the screen is determined by the computed offsets and the current position of the hand center.

The goal of choosing (ϵ_s, b_s) is to ensure that we can: (1) detect the intentional gestures with a certain tolerance to the deviations caused by irrelevant factors (e.g., detection error, camera angle, etc.); (2) give reasonable weights to similar gestures to allow mapping discrete gesture states to continuous gesture states; (3) filter out unintentional gestures.

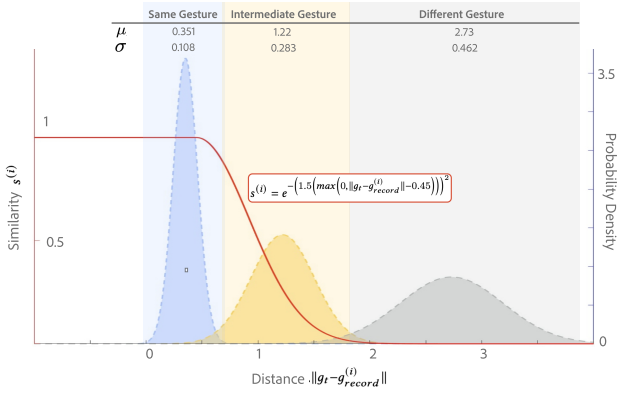


Figure 11: Distributions of measured gesture distances and the fitted hyper-parameters for $s^{(i)}$ in Equation (3).

A.3.3 (ϵ_s, b_s). We derived these hyper-parameters from experimental measurements. We recorded 3 different gestures: pinch, pointing, and palm upstaging. For each of them, three types of gestures were collected: (1) intentionally performed same gestures; (2) intermediate gestures (e.g., for pinch gesture, change the distance between the index finger and thumb finger); (3) completely different gestures. We collected the output distances calculated with recorded and performed hand feature vectors $\|g_t - g^{(i)}_{record}\|$.

We collected 637 data points in total and the distributions of these 3 types of distances are shown in Figure 11. With the result we chose $\epsilon_s = 1.5, b_s = 0.45$ to serve the above mentioned goals: (1) most of the same gestures (blue area) have $s^{(i)} = 1$; (2) intermediate gesture (yellow area) covers a wide range of similarity values; (3) most completely different gestures have $s^{(i)} = 0$.