An Exploratory Study on Multi-modal Generative AI in AR Storytelling

HYUNGJUN DOH*, Purdue University, USA
JINGYU SHI*, Purdue University, USA
RAHUL JAIN*, Purdue University, USA
HEESOO KIM, Purdue University, USA
KARTHIK RAMANI, Purdue University, USA

arXiv:2505.15973v1 [cs.HC] 21 May 2025

Storytelling in AR has gained attention due to its multi-modality and interactivity. However, generating multi-modal content for AR storytelling requires expertise and efforts for high-quality conveyance of the narrator's intention. Recently, Generative-AI (GenAI) has shown promising applications in multi-modal content generation. Despite the potential benefit, current research calls for validating the effect of AI-generated content (AIGC) in AR Storytelling. Therefore, we conducted an exploratory study to investigate the utilization of GenAI. Analyzing 223 AR videos, we identified a design space for multi-modal AR Storytelling. Based on the design space, we developed a testbed facilitating multi-modal content generation and atomic elements in AR Storytelling. Through two studies with N=30 experienced storytellers and live presenters, we 1. revealed participants' preferences for modalities, 2. evaluated the interactions with AI to generate content, and 3. assessed the quality of the AIGC for AR Storytelling. We further discussed design considerations for future AR Storytelling with GenAI.

CCS Concepts: • **Human-centered computing → Mixed / augmented reality**.

Additional Key Words and Phrases: Storytelling, Augmented Reality, Generative Artificial Intelligence

ACM Reference Format:
Hyungjun Doh, Jingyu Shi, Rahul Jain, Heesoo Kim, and Karthik Ramani. 2018. An Exploratory Study on Multi-modal Generative AI in AR Storytelling. 1, 1 (June 2018), 25 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Storytelling is the art of conveying a narrative or story to an audience through words, images, sounds, or other forms of content, often with the purposes of entertainment [65], education [4], persuasion [104], or preservation of cultural traditions [72]. With the rapid development of Augmented Reality (AR), Storytelling has expanded onto AR platforms, enabling the possibility of delivering multi-modal content to enhance interactive Storytelling [56, 57, 99]. This evolution has opened the discussion on the creation and effects of multi-modality in the research field of AR Storytelling.

*Marked authors contributed equally to this research.

Authors' addresses: Hyungjun Doh, Purdue University, West Lafayette, Indiana, USA, hdoh@purdue.edu; Jingyu Shi, Purdue University, West Lafayette, Indiana, USA, shi537@purdue.edu; Rahul Jain, Purdue University, West Lafayette, Indiana, USA, jain348@purdue.edu; Heesoo Kim, Purdue University, West Lafayette, Indiana, USA, kim2903@purdue.edu; Karthik Ramani, Purdue University, West Lafayette, Indiana, USA, ramani@purdue.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Association for Computing Machinery.
XXXX-XXXX/2018/6-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

Creating multi-modal content in AR Storytelling poses challenges. Thorough preparation is necessary for authors to effectively deliver their narratives to an audience, which encompasses not only organizing the story flow but also creating multi-modal content that goes with the flow, to enhance the audience's comprehension and engagement [56]. However, creating such content often requires mastery of certain software, expertise in visual or auditory design, and time and effort.

To democratize the users or storytellers from the requirement of expertise in content generation, recent work has explored the possibility of deploying Generative Artificial Intelligence (Gen-AI) in AR Storytelling [3, 8, 32]. AI-generated content (AIGC) has introduced a wider range of possible modalities into broader research by enabling multi-modal **functionalities** such as text-to-image [87], text-to-video [47], text-to-audio [19, 58], and text-to-motion [92]. Consequently, the boundaries of AR Storytelling are also to be re-established. The integration of Gen-AI with AR Storytelling is expected to enable the creation of more complex narratives and presentations at a much quicker pace, seamlessly bridging the gap between authors' intention and visual/audio realization [57, 93].

Despite the potential benefit of applying AIGC in AR Storytelling, current HCI research has yet to validate the effect of all possible modalities of AIGC in AR Storytelling. No prior work has provided a systematic validation of the impact of each modality on the core of AR Storytelling, or the design considerations of the necessary interactions for the users to co-create multi-modal content with Gen-AI for AR Storytelling. To explore the field and locate the pivot point for future research and development, critical questions are yet to be addressed. Motivated by the promising landscape of Gen-AI for AR Storytelling, we put forward the essential HCI research questions to shed light on future research and development in this field: **1. How suitable is multi-modal AIGC for augmenting the elements in AR storytelling?** And further, we question **2. What are the key HCI considerations when designing an AR Storytelling system utilizing the multi-modal power of Gen-AI?**.

To reveal the design space of multi-modal content in AR Storytelling in our study, we have investigated 223 videos from the existing video-edited augmented presentations. As a result, we define the design space consisting of the dimensions of **Elements** of AR Storytelling and the **Modalities** of augmentations. We further implemented an authoring platform that leverages Gen-AI models to generate multi-modal content, and an AR interface for live AR Storytelling with the generated augmentations. With our platform as a testbed, we conducted two studies (each $N = 15$) with participants with rich experience in Storytelling and live presentations. We experimented with five stories, with which the participants either were assigned certain elements to augment with AIGC or freely chose the elements and modalities to augment their AR Storytelling experience. We present the analysis of the study results which reveal (1) the participants' preferences for modalities to augment elements, (2) qualitative evaluations of the interactions to co-create the content, (3) qualitative evaluations of AIGC in AR Storytelling in terms of expressiveness, immersion, alignment, and exploration, and (4) empirical comparison of AIGC with human-generated content. Our contributions are four-fold:

- A systematic preliminary analysis of 223 augmented storytelling videos and a resulting design space depicting the elements and modalities in AR Storytelling.
- Implementation of an assembled storytelling content generator with Gen-AI supporting the creation of animations, images, music, and text, along with an AR interface for storytelling presentations, supporting speech-to-text, displaying corresponding generated content, and hand-interactable features.

- An exploratory study evaluating the impact of multi-modality AIGC regarding its quality, interactions to create, and capability of augmenting elements of AR Storytelling.
- An in-depth discussion providing design recommendations for future investigations within the domain of AIGC in AR storytelling.

2 RELATED WORK

In this section, we present the landscape of current research. We divide the related works into two topics: AR Storytelling and Multi-modality in Gen-AI.

2.1 AR Storytelling

Storytelling is the art or practice of a storyteller narrating stories or events, often involving the weaving of characters, settings, and plots, to engage, entertain, educate, or communicate with an audience [61]. This art or practice is a production co-constructed by two parties - tellers and recipients (or authors and audiences) [61, 91].

Recent research in AR Storytelling has demonstrated the suitability of AR as a medium of Storytelling for its presentation and interactivity in situ in multiple modalities [15, 65]. Firstly, AR Storytelling can be deployed on multiple platforms for diverse downstream tasks. For example, Holloway et al. [35] utilize AR Storytelling for educating university students on historical heritage, where the 3D animation of historical or fictional characters of stories is rendered for the users through an AR mobile app. A similar piece of research by Hirsch et al. [33] demonstrated that users' historical understanding can be improved by embedding our interfaces with AR stories. Research has also shown the potential of AR Storytelling in diverse domains such as game design [62, 83, 105], collaborative learning [62, 73, 111], communication [55, 103], and multimedia industry [70, 75].

Secondly, AR Storytelling has shown a promising connection between the authors and the audiences and better understanding and expression of objects that are otherwise hard to describe or share [55]. Research also emphasizes improving the user experience from each end of AR Storytelling. For authors, prior works introduce novel designs of how authors can create content for AR Storytelling [46, 50, 98]. For example, BlocklyXR [42] introduces an interactive visual programming environment for authoring digital storytelling with 3D models. Darzentas et al. [21] combined Internet of Things-inspired approaches and VR/AR experience technologies, with 3D scanning content creation techniques to create mobile and adaptable AR Storytelling. For the audience, research focuses on the understanding of the narratives [20, 30, 37, 79, 95, 96] as well as the immersiveness in the story [55]. Our work focuses on the authoring experiences of the storytellers in AR Storytelling, evaluating the content through the authors' satisfaction and preference.

Thirdly, Storytelling in AR consists of multiple possible modalities. RealityTalk [56] leverages embedded visuals and animation for engaging and expressive storytelling in AR presentations, where users can interactively prompt, move, and manipulate graphical elements through real-time speech and supporting modalities. Visual Captions [57] enables video conferencing augmented with AI-generated visual captions to enrich verbal storytelling. FairyTailor [8] enables the production of a coherent and creative sequence of text and images for visual Storytelling. Besides 2D visual content like images and videos, or 3D content like CAD models or avatar animation, modalities such as text [56, 62, 111] and audio [6, 39, 90, 98, 115] are also utilized in AR Storytelling to enhance the experience.

Despite the variety of modalities available in AR Storytelling serving specific downstream tasks, the authoring of multi-modal content for AR Storytelling can be challenging due to its requirement of mutually exclusive expertise in content creation across diverse modalities. To this end, researchers in this field have begun to utilize Gen-AI to create multi-modal content for AR Storytelling, enabling the authors to create content of modalities outside of their expertise from those they can easily

compose, (e.g. text-to-image [3, 8, 57, 78] and image-to-text [5, 18]). However, prior research has not addressed the impact of multi-modal AIGC on the elements of AR Storytelling. Our research aims to investigate how multi-modal AIGC can affect the experience of the authors in the pipeline of AR Storytelling.

2.2 Multi-modality in Gen-AI

With the significant development of Gen-AI, researchers have proposed various AI models that generate various modality content, including images, videos, audio, and motion.

Recently, Gen-AI has achieved significant progress in image generation and manipulation. Researchers have employed Generative Adversarial Networks (GANs) with CLIP [31, 44, 45]—utilizing image-text representations—for text-driven image manipulation, enabling realistic editing using text [1, 24–26, 29, 66, 74]. Moreover, with the popularity of diffusion models, several large text-to-image models have been proposed, including Imagen [89], DALL-E2 [84], GLIDE [69], Parti [112], CogView2 [23], Stable Diffusion [87], and DreamBooth [88].

In addition to image content, researchers have also investigated video generation using AI. Prior works utilize the text-to-video method [17, 47, 48, 59, 94, 97]. Recently, video generation models have been implemented based on diffusion models [11, 17, 34, 47, 48]. Additionally, researchers have explored using different input types, such as image [68] and motion [47, 59].

Gen-AI has also been utilized for audio generation, employing a variety of models. Audio generation models have utilized likelihood-based models [43, 64, 71], VAE-based models [76], and GAN-based models [10, 53, 109]. Recently, diffusion models have also been applied to generate audio [52]. Copet et al. introduced MusicGen [19], a text-to-music model comprised of a single-stage Transformer Language Model (LM).

In the case of human motion generation, the text-to-motion task is used to learn a latent space for language and motion. JL2P [2] uses an auto-encoder to learn KIT motion-language dataset [80]. T2M employs a VAE [51] to map a text prompt into a normal distribution in latent space. Additionally, MotionCLIP [101] addresses the problem of data limitation and enables latent space editing by utilizing CLIP [82]. Researchers also suggest diffusion models for motion generation [49, 102, 113].

These prior works utilize easily accessible input, enabling users to generate output without requiring any specific expertise. Text is commonly employed as input to generate content. In addition to text, prior works have explored the use of various modalities as inputs for Gen-AI, such as image [68] and motion [47, 59]. This flexibility allows users to generate more diverse and well-aligned content with their intentions, as they are not limited to a single input modality.

3 DESIGN SPACE OF MULTI-MODAL AR STORYTELLING

To investigate the possible modalities in AR Storytelling as well as the elements of Storytelling created through these modalities, we analyzed 223 videos from existing augmented-reality storytelling videos taken from the YouTube platform.

3.1 Methodology and Ananlysis

3.1.1 Identification of Corpus. To find relevant videos of Storytelling in AR, authors performed a manual search on platforms such as YouTube. Initially, three authors independently used the keywords "Augmented Reality Storytelling" on YouTube. However, this approach yielded few relevant results. This was primarily due to two reasons 1) Augmented reality keywords are less used in the AR Storytelling context. 2) The AR Storytelling videos are often used for different purposes or domains. Therefore, we collected videos based on a manual search with keywords of various other domains such as weather, education, tutorial, sports, and news. We do not claim that

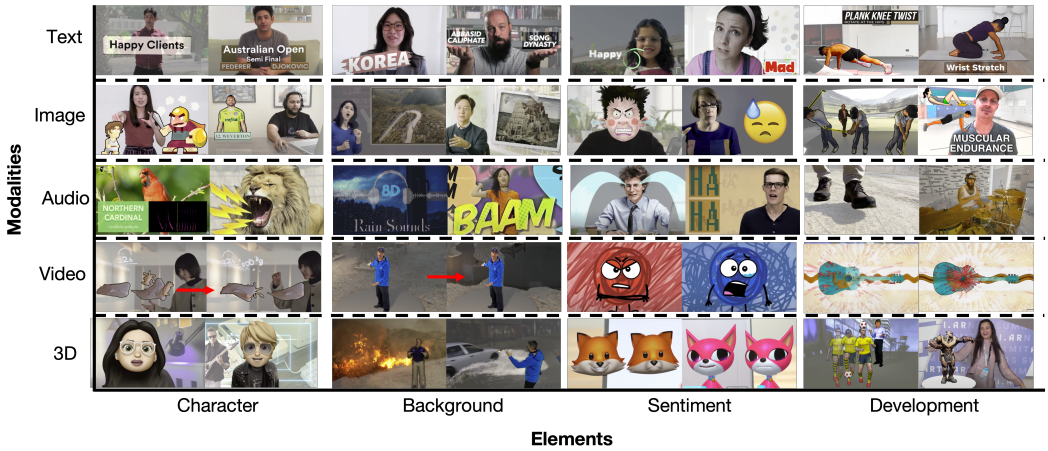


Fig. 1. Our Design Space for Multi-modal AR Storytelling: Modalities and Elements. These examples illustrate how a modality augments each element of AR Storytelling, offering a visual summary of our design space as derived from our analysis of the 223 augmented videos.

our corpus was collected by a systematic search. However, we argue that it is sufficient to serve the main goal of systematic analysis to identify common elements and modalities for AR Storytelling.

3.1.2 Filtering Methodology. We initially collected 732 videos from YouTube for Storytelling across various domains. Then, we conducted a filtering process to exclude the videos that did not possess AR components. Our inclusion criteria for AR components consist of:

- digital information or objects overlaid onto the real-world components in the videos,
- audio or visual effects that are not inherently from the storytellers,
- auxiliary texts displayed in addition to storytellers' narratives,
- animation or dynamic widgets rendered in the videos.

Three authors individually watched all the videos and decided on the inclusion or exclusion of the videos. The inter-rater agreement measure Fleiss' kappa was used during the filtering of the videos. The mean agreement score of $\kappa = 0.76$ indicates substantial agreement. Any conflicts were resolved by discussion among the authors. After the filtering process, we identified 206 videos. Additionally, we added 17 videos based on the author's knowledge from published work in AR Storytelling. In total, the final corpus contains 223 videos that were used for analysis.

3.1.3 Analysis. To obtain the design space of AR components in Storytelling, we conducted a nuanced analysis of the located corpus. Notice that the main focus of our analysis was on the content in AR Storytelling and not on the interactions performed with the content.

Firstly, three of the authors individually conducted open coding to analyze a subset of the final corpus ($N = 50$) and each summarized a preliminary design space, considering the modalities and elements of AR Storytelling in the videos from each subset. Secondly, the three authors discussed and decided on a revised design space. In this step, the purpose of the discussion is to form a clarified definition of each dimension, ensure a clear borderline that differentiates each dimension, and maximize the coverage of the revised dimensions over the three subsets aforementioned. Thirdly, the authors went through the entire final corpus, individually located each video in the revised design space, and merged to discuss the location of the video. This procedure ensured

each video was reviewed according to a set of consistent criteria. Upon disagreement about a video, authors merged and discussed to decide whether existing dimensions were appropriate or add/remove dimensions otherwise. Additionally, our main finding on analyzing the videos suggests that speech modality (speech by the storytellers, which is a necessity) triggers other modalities such as text, visuals, or audio sounds relevant to the narration in AR Storytelling. Therefore, we used the narrated content as our primary modality to design our authoring system which then triggers other modalities in the AR Storytelling.

3.2 Design Space

To this end, we identified the design space of multi-modal AR Storytelling, covering the discussion over possible modalities of the AR components and the elements of AR Storytelling that can be represented and expressed by the modalities (as shown in Figure 1).

3.2.1 Modality. We identified 5 different types of modalities being used in the current field of AR Storytelling:

- ***Text***: In AR Storytelling, the text is overlaid in a real environment to provide background stories characters dialogue, or additional plot details enhancing the overall storytelling experience. Texts are usually used for detailed information delivery such as instructions, context, or any other details, and provide user guidance such as navigating or prompts.
- ***Audio***: Many AR Storytelling scenarios involve background sounds, music, or ambient noises, which contribute to the mood and atmosphere of storytelling. Character voices and dialogues make storytelling experiences more immersive. Realistic sounds are associated with virtual objects making the experience more engaging. Audio cues are used as feedback for user interactions. For example, a sound effect confirms a successful button click.
- ***Image***: Images are used to convey key visual elements of the story, such as the appearance of characters, the vision of scenes, or the composition of objects. Visual elements add depth to the narrative making it more memorable for the users. Images are used to represent characters, and interactive elements and provide visual feedback. Visual images also have a powerful impact on rendering emotions. It enhances the emotional resonance of the narrative. Visual imagery is also used for aesthetics and designs of AR Storytelling.
- ***Video***: Videos can convey dynamic and moving content, allowing for a more fluid and cinematic Storytelling experience. Animation and motion make characters or virtual objects move. Character animations and temporal narrations convey sequences of actions, and changes in the storyline, or illustrate the development of the plot.
- ***3D***: 3D models are used as characters, objects or scenes. 3D models add more dimensions and viewpoints to visual AR content, e.g. 3D Avatars, 3D texts, 3D scenes, and 3D animations.

3.2.2 Elements. We identified four different atomic elements essential to telling a story in AR. These atomic elements, namely Character, Background, Sentiment, and Motion, compose higher-level elements of a complete story such as Conflict, Plot, and Message [28], and eventually serve the highest-level goal of the story - to make the audience understand [91].

- ***Character***: A Character refers to a person, an entity, or a creature that inhabits the story. It can be a human, an animal, an object, or even an abstract concept, depending on the nature of the narrative. A Character can be elicited through its traits, roles, appearance, or other attributes that help shape its figure in the story.
- ***Background***: This element encompasses the setting or environment where the story unfolds. On the one hand, it can be a physical location where the story happens such as a castle, a forest, or a beach. On the other hand, it can be depicted as a virtual context or an abstract

setting, that contributes to the cultural or historical context of the story, e.g. London in the eighteenth century or a world with no elves. The atomic background element solely depicts the "where" and "when" the story happens.

- ***Sentiment***: Sentiment refers to the emotional tone, mood, or atmosphere conveyed through the story experience. In Storytelling, characters (usually humanoid ones) can display their Sentiments that add context and reasoning to the plots, shape the figure of the characters, or stir up sentimental resonance with the audience, e.g. "Jack is sad due to his loss". Furthermore, the narratives can also display Sentiments that render the atmosphere of the story without being derived from the characters, e.g. "The devastating smell resonates in the air".
- ***Development***: Development includes dynamic movements and temporal progress of the Backgrounds, Characters, or Sentiments. In Storytelling, Development depicts the time arc of the story, which emphasizes the progression or change in a significant transition along the timeline, e.g. "Lucy turned twenty-two and graduated from college". The Development element also involves motion, action, or movement that focuses on the behaviors or changes that do not emphasize time, e.g. "The cup keeps spinning" and "Andrew is a lot happier now".

4 EXPLORATORY STUDY

4.1 Study Setup: Testbed Implementation

To investigate our research questions, we conducted an exploratory study to evaluate the effect of multi-modal AIGC in the preparation and presentation stages of AR Storytelling. We thus developed a multi-modality content generator and AR Storytelling platform as the testbed for our study.

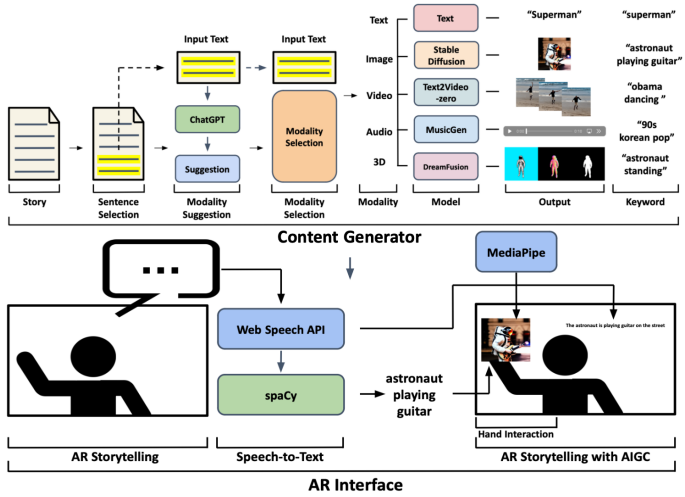


Fig. 2. The testbed workflow. 1) **Content Generator interface:** The user employs the content generator to create AIGC for AR Storytelling, supporting five modalities for the selected sentence. 2) **AR interface:** The user can view the text corresponding to spoken words. Based on the transferred speech text, the user can interact with AIGC using hand.

The testbed contains two separate interfaces. 1) **Content Generator interface:** the author can create the content for the AR Storytelling using AI according to textual narratives as shown in Figure 3-(a). 2) **AR interface:** This AR interface allows users to interact and place the generated content at the desired location and present the entire story augmented with multi-modal content as shown in Figure 3-(b).

The testbed is designed with the following specified considerations: 1) The testbed should generate multi-modal AR components, including text, audio, image, video, and 3D, with a given text narrator. 2) The testbed should be able to generate content with various forms of sentence structure. 3) The testbed should allow the users to decide the modality of the AIGC based on their preference. 4) Through iterations with the users, the testbed should be able to create content aligning with the author's intention. 5) The testbed should allow AR Storytelling with generated content, supporting simple spatial AR interaction built with hand gestures, and speech-based interactions.

4.1.1 Multi-modal Content Generator. Our multi-modal content generator is an interface empowered by multiple Gen-AI models based on textual prompts. For video generation, our interface employs the Motion-Diffusion-Model (MDM) [102]. This model takes text as input and generates detailed character motion (For example, generating a human motion of skipping rope). To extract the verb and noun from the input sentence, we leverage spaCy [36], a Natural Language Processing (NLP) Python library. Then based on the generated motion, our system uses Text2Video-Zero [47] to create a video with the described character and background. For image generation, we utilize Stable Diffusion [87]. The testbed generates the audio with MusicGen [19] which generates audio using text prompts. To generate 3D content, we used text-to-3D Gen-AI model [81].

Figure 3-(a) illustrates the front-end interface of our multi-modality content generator testbed. The front-end communicates with the back-end using FastAPI [27], a web framework for building RESTful APIs [85] in Python. The generator comprises three main sections: (1) **Story input**: The textual input section imports the text file containing the narratives. Users can select the sentence in this section by highlighting the portion, which they want to augment in the storytelling process and (2) **Output**: Based on the users' selection of modality, the output section generates and displays the corresponding AR component. The output is saved, corresponding to the highlighted sentence, which is a trigger during the storytelling process.

4.1.2 AR Interface. Figure 3-(b) shows the interface of our AR Storytelling platform, testing the AIGC in real AR Storytelling scenarios to observe how it impacts the storytelling experience. The AR Storytelling interface incorporates MediaPipe [63] for Hand Landmark detection to support hand interaction to locate the generated content. Additionally, it employs the Web Speech API [67] to showcase the narrator's words on display and trigger the generated content. Users can convey their stories in AR using AIGC with these features.

4.2 Study Design

We conducted two studies to evaluate the effect of multi-modal AIGC in AR Storytelling. Study 1 was performed in a controlled setting to evaluate which modality is best suited for what type of elements of AR Storytelling according to the authors' preference. Study 2 was performed to evaluate users' interactions in generating content for AR Storytelling. Furthermore, we aimed to gain insights into the deployment of multi-modal AIGC in AR Storytelling through subject participant feedback for both studies.

4.2.1 Participants. We invited 30 participants (13 female; and 17 male) via emails, mutual connections, and posters from a university's graduate and undergraduate programs. The user's age ranges between 18 to 31 with a mean of 25 years. All of the users had prior knowledge of Storytelling in the form of presentations, oral speeches, theater performances, etc. All participants had prior experience in live public storytelling such as stand-up comedy shows, presentations to a crowd audience, or being interviewed. 24 out of 30 participants had some experience with AR/VR interfaces. The participants were randomly divided into two groups, each participating in one of the two studies. After the arrival of users, a 5-minute brief introduction was provided about the study. Users

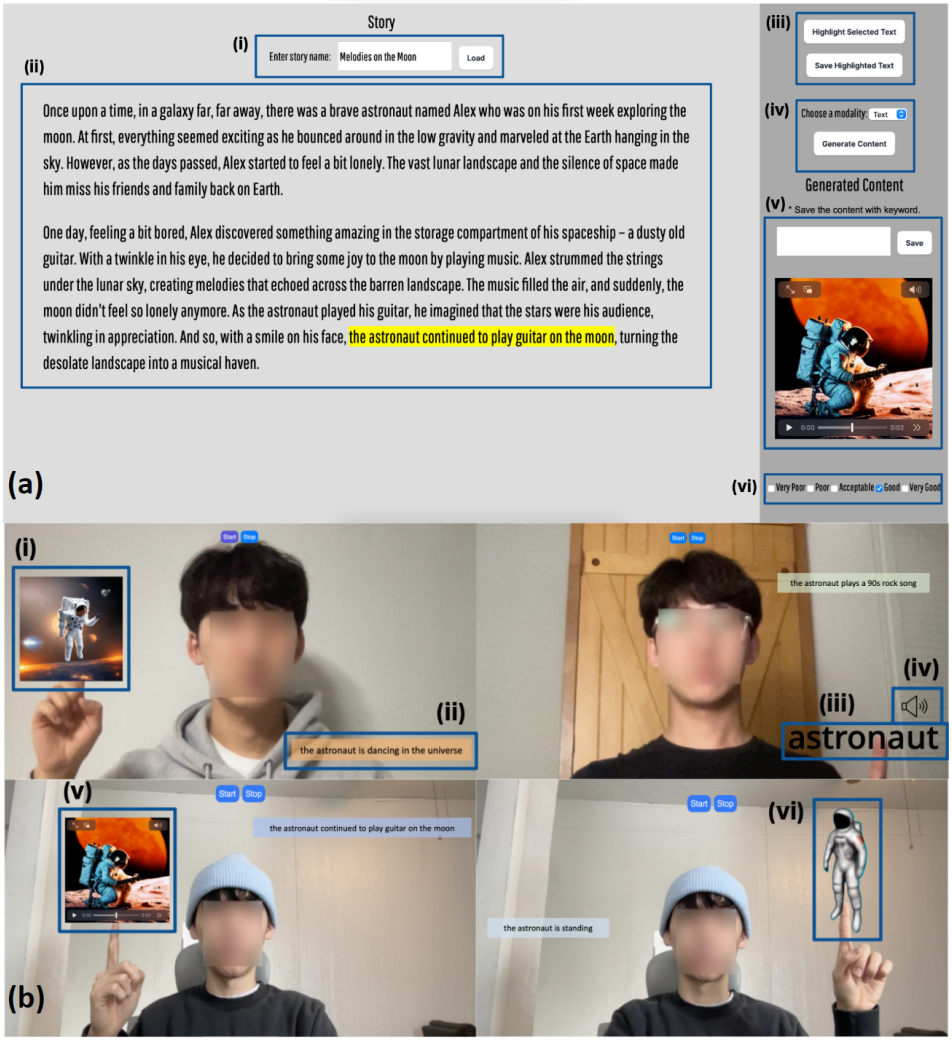


Fig. 3. The testbed. **(a)** The Multi-Modal Content Generator interface. (a-i) The textual input section. The user can import a story text file for storytelling by entering the story title. (a-ii) The loaded story. Where the user can see the loaded story and select a sentence for augmentation by dragging it with the mouse cursor. (a-iii) The highlight (top) and save (bottom) buttons. The top button highlights the selected portion with yellow, indicating to the user that this part will be generated. The bottom button saves the highlighted text to the backend for output generation. (a-iv) The modality selection. The user can choose the desired modality for content generation. (a-v) The output. The testbed displays the generated content, and the user can save it using a unique keyword. This keyword acts as a trigger during the storytelling process. (a-vi) The quality evaluation of the content by the user. **(b)** The AR interface. (b-i) Image. The user can interact with the AIGC using hand landmarks. The content appears on the tip of the user's index finger. (b-ii) The Speech-to-Text box. The interface displays the narrator's words, serving as a trigger for the corresponding content. (b-iii) Text. This modality shows the main keyword and detailed information. (b-iv) Audio. The icon indicates the corresponding audio is playing. (b-v) Video. (b-vi) 3D Content.

agreed to perform the study and signed a consent form. A step-by-step instruction was given to the participants on the usage of the interface and the overall goal of the study Then, the participants were given sufficient time to get familiar with the interface up to 15 minutes. Both studies took 1.5 hours in a closed environment and each participant was compensated with a 15 USD e-gift card.

4.2.2 Study 1 - Designated Augmentation. In this study, we aim to investigate the relationship between AR Storytelling elements and modality. We designed 5 different narratives of stories, each with around 20 atomic elements priorly highlighted by us in the interface. The narratives are shown in Appendix A and the pre-determined elements statistics are shown in Figure 4. For each element previously highlighted in the interface, the participants are asked to select at least one modality of augmentation that they think helps with storytelling the most.

Participants are allowed to preview all modalities of the generated content on the generator interface. They are also allowed to prepare or demonstrate storytelling on the AR interface with the generated content. Each participant was asked to complete the augmentation and practice of AR Storytelling of all 5 stories. Upon completion of selecting the modalities, creating the augmentation, and practicing the AR Storytelling with the AIGC, the participants completed a Likert-scale questionnaire and received a semi-structured interview about the experience.

4.2.3 Study 2 - Freestyle Augmentation. In this study, we aim to investigate participants' preference for modalities to augment their AR Storytelling experience and the interaction between the participants and the AI models in the process of content generation. We utilized the same 5 stories as in Study 1. Differently, participants were not given specific elements from the narratives to augment. Instead, participants were asked to augment their storytelling experience according to their preference, by selecting and highlighting elements from the narratives and augmenting them at will. The content generation process, the preview process, and the practice process were all the same as those in Study 1, with an additional request to give ratings on a scale of 5 for the quality of each augmented element. Participants were also asked to finish augmenting and practicing AR Storytelling with all 5 stories. Upon completion of selecting the modalities, creating the augmentation, and practicing the AR Storytelling with the AIGC, the participants completed a similar Likert-scale questionnaire and received a semi-structured interview about the experience.

4.2.4 Data Collection. As described in the previous subsubsections, we collected the following data:

(1) 5-point Likert-scale questionnaires indicating the participants' subjective ratings of the creation and practicing of the AR Storytelling experience,
(2) post-study semi-structured interviews,
(3) video recordings during the entire for each participant.

When designing the Likert-scale questionnaires, we took into consideration the Expressiveness, Immersiveness, Exploration, and Alignment in the co-creation process with AI, as suggested in the Mixed-Initiative Creativity Support Index (MICSI) [54]. We only considered a subset of the factors in MICSI to be evaluated, given that our testbed is not an implementation of a systematic application, and thus systematic evaluation metrics such as usability do not contribute to the findings. In addition to the above, we collected the participants' preferences for modalities over the pre-determined elements in Study 1. In Study 2, we collected the participants' ratings of AIGC while creating content. Also, we reviewed the recordings to analyze their failed attempts to augment with some modalities (where participants generated the augmentations but did not apply them later in AR Storytelling practice).

5 RESULTS

In this part, we present our results from the study. Then we summarize and analyze the results according to the interview feedback and our observations.

5.1 Preferences for Modalities to Augment

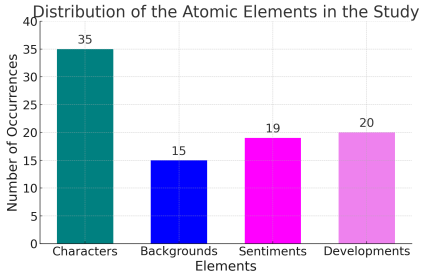


Fig. 4. The distribution of the four elements that emerge in the stories we used for the study

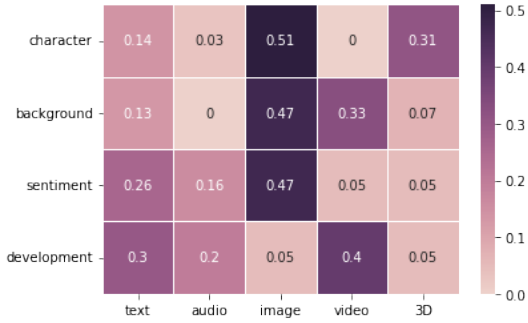


Fig. 5. Results for Preference of modality

We report our findings in Figure 5 which summarizes which modality should be used for what elements of AR Storytelling in Study 1. Characters are highly preferred by images (51%). *"In AR stories, when we see pictures of characters, it feels like they're right there with us, making the story come alive." (P4).* 3D (31%) is also used to represent characters while other modalities are slightly preferred. *"I use 3D characters often in cases for showing the action of the character" (P6).* For background, image (47%) and video (33%) are best suited. *"I prefer 2D visuals in the background" (P9).* The favorite modality to express sentiment is image (47%) followed by text (26%). *"I think emotions and feelings can best be expressed by showing image" (P1).* Text (30%) and video (40%) were preferred by participants for development elements and then audio. *"I feel like the video was easy for me to show change" (P14).*

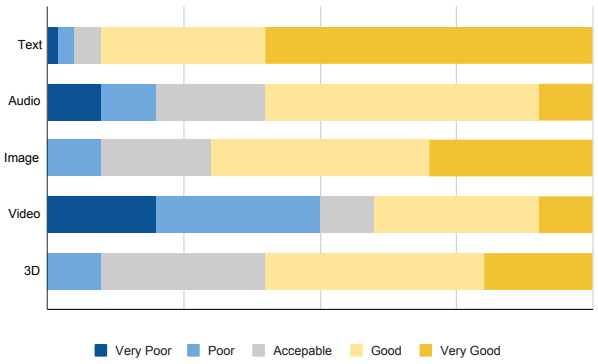


Fig. 6. The participants' ratings of the overall quality of each modality.

5.2 Quality of AIGC in AR Storytelling

We evaluated the quality of the AIGC during the study 2. The results are summarized in Figure 6. Participants have appreciated the quality of the text (AVG=4.43, SD=0.87). *"I think the text is best because it was clear and easy to read during AR presentation and are very much relevant to the story "(P23).* Participants had neutral opinions on the quality of the generated audio from the text (AVG=3.4, SD=1.11). *"Most of the time the sound I was hearing for my content was suitable" (P17)* Participants agree that the images were relevant and good quality (AVG=3.9, SD=0.94)*"The images looked so real as if someone has taken in the real world and were so much similar to the story."* (P19). Participants were not satisfied with the quality and faced difficulty in understanding the content of the video ((AVG=2.8, SD=1.13)). *"I was not able to understand what was the action being performed in the video"(P27).* Many participants acknowledge the 3D content was useful for their story (AVG=3.7, SD=0.90). *"I cannot imagine I can get 3D models of astronaut with a telescope so good"(P25)*

5.3 Interactions with AI to Generate Augmentation

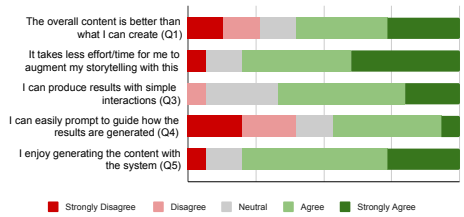


Fig. 7. Results of the Likert-scale questionnaires on the interactions with the testbed to generate augmentations - Study 1

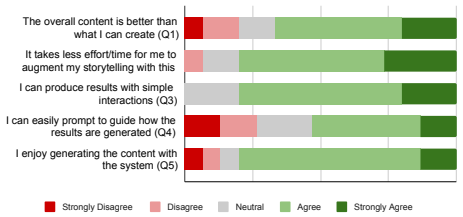


Fig. 8. Results of the Likert-scale questionnaires on the interactions with the testbed to generate augmentations - Study 2

We qualitatively evaluated the interactions of the participants with the testbed in generating the content during both Study 1 and Study 2 as shown in Figure 7 and Figure 8. The participants acknowledge that the final AR Story with the generated content is usable and better than what they can create on their own ((Q1) Study 1: AVG=3.47, SD=1.35; Study 2: AVG=3.60, SD=1.14). *"This is actually good, I would like to use such systems to generate content for my presentations in the future." (P9).* Most of the participants were able to generate relevant content with less time/effort ((Q2) Study 1: AVG= 4.06, SD = 1.06; Study 2: AVG= 4, SD = 0.86). *"In my opinion, this is one of the fastest software I have ever used to make content." (P24)* Participants found that the interaction with the testbed is simple and easy to generate content ((Q3) Study 1: AVG= 3.83, SD=0.82; Study 2:AVG= 4.1, SD=0.63). *"It is definitely not at all a complicated system to use to get some results" (P5).* Participants found it difficult to guide the generation to the desired output using just textual prompts ((Q4) Study 1: AVG=2.93, SD=1.28, Study 2: AVG=3.26, SD=1.23). *"Every time I entered new text, the output looked entirely different and was sometimes not usable." (P8)* Participants enjoyed the generation process of the content using the testbed ((Q5) Study 1: AVG=3.91, SD=0.95; Study 2: AVG=3.73, SD=0.99). *"The system is fun to use, as it gives me some novel options to try for my presentations." (P13)*

Additionally, there is no significant difference in the participants' ratings of the interactions between the two studies. Shapiro-Wilk tests on the ratings showed no normally distributed feedback from either group(Q1:$p_{sw,1} = 0.030$, $p_{sw,2} = 0.036$; Q2:$p_{sw,1} = 0.011$, $p_{sw,2} = 0.001$; Q3:$p_{sw,1} = 0.003$, $p_{sw,2} = 0.050$; Q4:$p_{sw,1} = 0.072$, $p_{sw,2} = 0.037$; Q5:$p_{sw,1} = 0.0005$, $p_{sw,2} = 0.001$.), and Mann-Whitney U tests showed no significant difference between groups(Q1:$p_{mwu} = 0.920$, $Z_{mwu} = 0.103$;

Q2:$p_{mwu} = 0.645, Z_{mwu} = -0.456$; Q3:$p_{mwu} = 0.589, Z_{mwu} = 0.539$; Q4:$p_{mwu} = 0.535, Z_{mwu} = 0.622$; Q5:$p_{mwu} = 0.603, Z_{mwu} = -0.518$.). This suggests that freestyle augmentation and designated augmentation have no significant difference in the interaction experiences of the participants.

5.4 The Suitability of AIGC in AR Storytelling

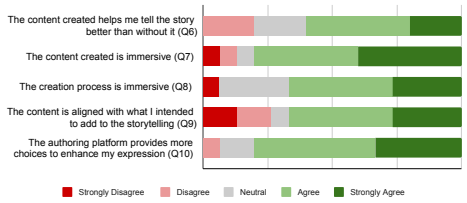


Fig. 9. Results of the Likert-scale questionnaires on Suitability - Study 1

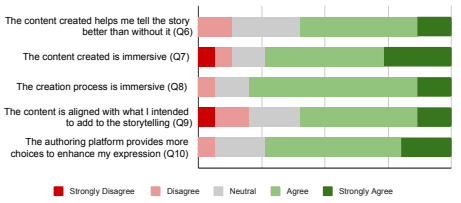


Fig. 10. Results of the Likert-scale questionnaires on Suitability - Study 2

We qualitatively evaluated the usefulness of the content in both Study 1 and Study 2 as shown in Figure 9 and Figure 10. ***Expression***: participants found that the content created is useful and able to tell the story better than without using it ((Q6) Study 1: AVG=3.60, SD= 1.01; Study 2: AVG=3.60, SD=0.87). *"I think additional generated content is good in narrating the story better than without using any content." (P2)* ***Immersion***: participants found the that content created is immersive ((Q7) Study1 : AVG=4.03, SD=1.15; Study 2: AVG=3.80, SD=1.10). *"I just think that the content will increase audience engagement and I feel like the audience will better connect." (P15)* Participants think positively that the content creation process is also immersive ((Q8) Study 1: AVG=3.81, SD= 1.04; Study 2: AVG=3.87, SD=0.71). *"I was so deeply involved with the process that I have to say this is very much authentic." (P17)* ***Alignment***: Participants found that the content is aligned with their intentions and the content effectively conveys the narrative ((Q9) Study 1: AVG=3.53, SD = 1.35; Study 2: AVG=3.46, SD = 1.08). *"I feel like the content was pretty much the same as what I wanted. I was surprised though that I can get what I want easily." (P10)* ***Exploration***: participants reported positively that the authoring platform provides more choices to enhance my expression ((Q10) Study 1: AVG= 4.06, SD= 0.85; Study 2: AVG= 3.86, SD= 0.80). *"I wonder how I am able to get so many variations of the content with just text." (P3)*

Additionally, there is no significant difference in the participants' ratings of the interactions between the two studies. Shapiro-Wilk tests on the ratings showed no normally distributed feedback from either group (Q6:$p_{sw,1} = 0.049$, $p_{sw,2} = 0.041$; Q7:$p_{sw,1} = 0.012$, $p_{sw,2} = 0.002$; Q8:$p_{sw,1} = 0.002$, $p_{sw,2} = 0.017$; Q9:$p_{sw,1} = 0.057$, $p_{sw,2} = 0.013$; Q10:$p_{sw,1} = 0.024$, $p_{sw,2} = 0.012$.), and Mann-Whitney U tests showed no significant difference between groups(Q6:$p_{mwu} = 0.9681, Z_{mwu} = 0.04117$; Q7:$p_{mwu} = 0.53526, Z_{mwu} = -0.62217$; Q8:$p_{mwu} = 0.645, Z_{mwu} = -0.456$; Q9:$p_{mwu} = 0.69654, Z_{mwu} = -0.39404$; Q10:$p_{mwu} = 0.50926, Z_{mwu} = -0.66365$.). This suggests that freestyle augmentation and designated augmentation have no significant difference in the participants' ratings of the suitability of AIGC in their Storytelling experience.

5.5 Result Summary and Analysis

5.5.1 Suitability of modality. It was evident that images were highly preferred for character, background, and sentiment. This is because of progress in algorithmic advancement in image generation as compared to that in other modalities. 3D content was mostly used for depicting characters. This is mostly because 3D models are enough to introduce the rough figure of the

characters while the background, scenes, or minute details such as the expression, development, or appearance of the character are not easily and accurately rendered in 3D models. Though some participants highly preferred the text regarding its quality, others felt that the text did not add details. *"I mean I have used text for almost everything but the text is not that engaging as compared to visuals or audio". (P13)* The quality of the video was not good enough so participants only used video at a big scale in the development element where details are safe to ignore, while for small details in development, participants preferred text and audio (P4, P2, P9, P13). *"We used video for development mostly for expressing bigger ideas because of its poor quality. I feel like small details are not good" (P9).* Furthermore, the quality of the videos was reported as the major reason for the misalignment between the content and the intention of the participants. As the participants' speech narration made up the majority of the audio content in the storytelling, generated audio was less used for all the elements and mostly functions as a supplement to describe some of the elements. Audio is also used for showing emotions, and sometimes the temporal progress of the scenes (P1, P6, P15).

5.5.2 Empirical comparison to human-generated content. Many participants appreciated that the overall AR Storytelling with AIGC was useful. Some of the participants felt the content was as good as human-generated content while some felt that the quality of the content limits their freedom to use some modality for some element. The most negative feedback we received from the participants was on the quality of the videos. While participants reported that the generated videos were far from being as good as human-recorded or generated ones, we argue that this is due to the limitations in the algorithmic development in content generation rather than the suitability of the video (or others) as a modality for augmenting storytelling.

Interestingly, participants reported that they could tell the difference between the human-generated content from the AIGC according to the interview results (P1-4, P6, P9-P12, P16-23, P25). However, most of these participants commented that it does not affect the quality of their storytelling, since it is not the story itself but just the augmentation. *"I can tell that the image is definitely not drawn by human[s] but I kind of think that it is still related to my story." (P5)* Yet, some commented that the unrealistic or artificial content may bring confusion to the audience. We further discuss the future research induced by this finding in section 6

5.5.3 Interacting with Gen-AI is easy; Guiding it is hard. When asked about the interactions with the testbed, many appreciated that this testbed required less effort and content could be generated faster than the creation by humans. This provides a unique edge over existing platforms (P5, P23, P24, P29). *"I feel like even though I have I have to do multiple interactions but still the content we can get quickly. I have only seen searching as this quick."* Also, co-creation with Gen-AI is fun, making participants more indulged. However, some reported that the generation of the outputs was not controllable, there were unexpected or unintended outputs, and some were ethically incorrect because of incorrect prompts. Many participants find it difficult to generate the desired content because of incorrect and incomplete prompts. *"I had to perform many iterations to get the desired output. It would have been much better if there was a way my input could give the desired output in one or two iterations." (P5)*

The workflow allows participants to explore various outputs in a short period. This testbed generates more creative content (P23, P27). *I can't believe that I was able to get an extraordinary image of foggy nights, I feel like I found content generated far better than my imagination" (P23).* Participants found that the generated content was easily immersed in the AR environment and was supplementing it. Most of the participants found that the generated content was aligned with their intentions and also with the story.

6 DISCUSSION

In this section, we discuss the primary results of the study and contrast them with prior works. Furthermore, we provide more insights into the future landscape of AIGC in AR Storytelling, as well as design considerations for future AR Storytelling Systems and Applications.

6.1 Modalities of AIGC in AR Storytelling

Our first research question was whether multi-modal AIGC is suitable for AR Storytelling. In section 5, we show results indicating users' preferences for modalities to augment their storytelling. We showed how participants are positively satisfied with the overall quality of each modality in terms of expressiveness, alignment with their intention, immersion into telling and listening, and exploration of options. From the results, we further identified open challenges to address in order to situate AIGC in AR Storytelling. We discuss the challenges in the following three aspects.

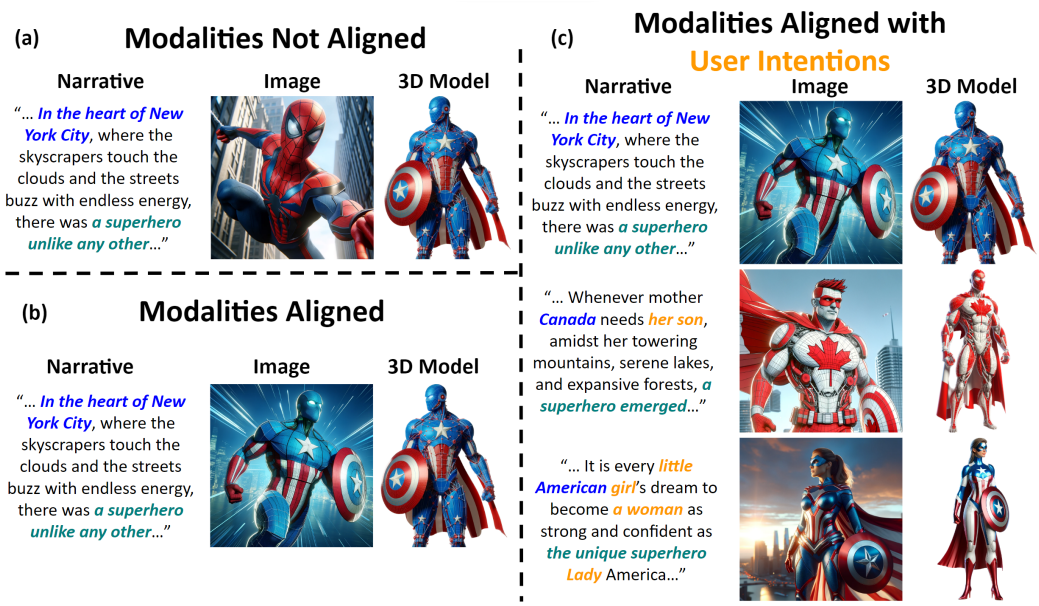


Fig. 11. An illustration of the Alignment among modalities in AR Storytelling. All images are generated by ChatGPT [1] with proper prompts. In (a), the image shows a different character from that shown by the 3D model, even though they are generated from the same textual narrative. In (b), we prompt the image generation with a screenshot of the 3D model, then the generated image is aligned with consistent character information from the 3D model. In (c), we prompted the image and 3D generation with user-intended information, such as nationality and gender, in order to obtain AIGC aligned among modalities as well as the users.

6.1.1 Alignment Among the Modalities and the Users' Intention. Although single augmentation for a single element meets the satisfaction of the participants, it has been reported that the multi-modal augmentations for a single element appeared unaligned, i.e. the information of the element conveyed in one modality is different from that conveyed in another. "*I used both the image and the 3D model to introduce the character of a painter. I thought it would be cool to have the static image on my left hand and the rotating model of it on my right hand, but the model showed a different*

[1] https://chat.openai.com/

character from the image. (P21)" This finding echos with one heated field of Gen-AI research, where prior works [12, 14, 22, 82] propose to train models on a large scale of data to be adapted into various downstream tasks. By sharing the latent features and knowledge of the elements, the model can generate consistent multi-modal augmentations of the same elements from the story. Yet, to bridge between the Gen-AI research and the realm of AR Storytelling, we anticipate the future one step further, where the human intention is also aligned within the co-creation process, by ***evaluating not only the alignment among the modalities but also the modalities and the users' intention*** [3, 54]. The alignment with the users is rather important in addressing the identity (e.g., race, gender, youth, or nationality) of both the tellers and the audience [7, 13, 86].

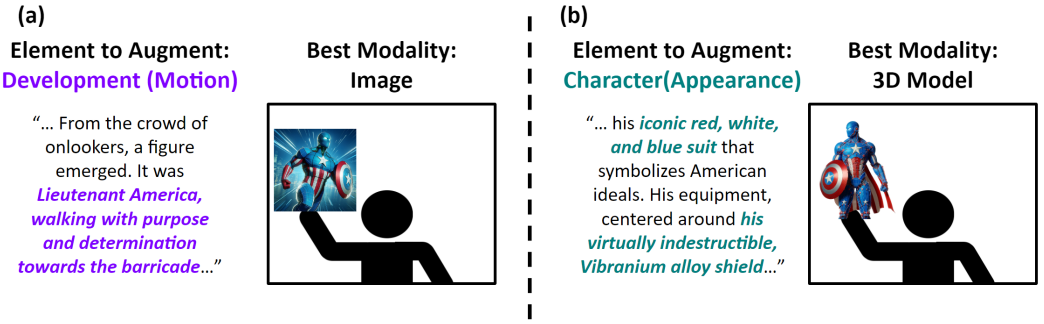


Fig. 12. An illustration of the concept of Selective Augmentation. In (a), the narrative depicts the motion of a character entering the scene. In (b), the narrative introduces the appearance of a character. Selective Augmentation suggests that the system should infer the best modality to augment this narrative, without excessively augmenting it with too many modalities or poorly augmenting it with a less supportive modality.

6.1.2 Selective Augmentation - What to Augment. Despite the consistently positive ratings of AIGC augmentations, we have witnessed differences between the choices of elements in Study 2 (freestyle) and the elements we have assigned in Study 1. In Study 2 where the choices of elements to augment were free, participants tended to augment fewer elements than we assigned in Study 1. "*Not all elements are necessarily needed to be augmented. Also, I do not tend to use multiple modalities to augment one simple element. For example for the background, just an image will do. I do not want too many visuals to show up in my presentation and confuse my audience. (P27)*" This finding aligns with the fact that (1) storytellers possess individual preferences and strategies for choosing what to emphasize in their stories [107], that (2) excessive augmentation leads to convoluted presentation of the story [56], and that (3) redundant information adds up to the mental cost of the audience to understand the story [77]. Prior works [3, 8] have suggested tackling the problem by adapting to individual co-creation in visual story generation, such that the system interprets what the authors would like to augment with visuals.

In addition, this adaptation should also consider the modality of the augmentation. Given a system adaptive to and aware of what to augment, the expression is then constrained by the attributes of the element. For example, in our study, we discovered that videos are preferred for the element of development, since "*...it (video) tells what the action is. No other option there could properly describe "spent days and nights". (P4)*". We also conclude from our research that the mapping of the best modalities to augment an element is not bijective. Although video can show dynamic elements in the story, an image can sometimes be precise and vivid enough to depict a monotonous motion and is thus preferred over videos. It is significantly important to understand the purpose of

the elements in the story, in order to select the best modality or modalities. These above findings inspire us to design authoring systems for AR Storytelling that adaptively infer ***what to augment*** (in what modalities) based on user input and the narratives of the story.

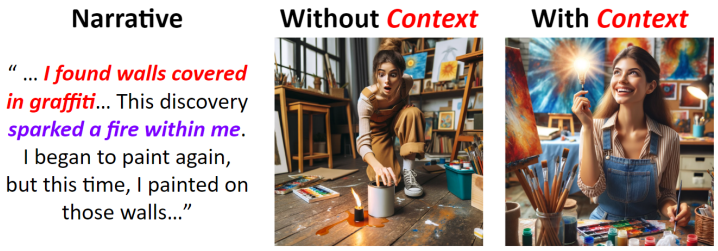


Fig. 13. An illustration of context-awareness AIGC in AR Storytelling. Given a narrative, Gen-AI can generate illustrative images for storytelling. Prompting without the context of "sparked a fire" results in an image falsely illustrating the literal meaning, while prompting with the context of discovery artwork "sparked a fire" gives a correct illustration of the metaphor of an Eureka moment.

6.1.3 Context-aware AIGC in AR Storytelling. Among the four attributes we evaluated (expression, alignment, exploration, and immersion), we identified the lowest rating of the alignment between the AIGC and the participants' intention, which echos with the lowest Likert-scale score when the participants were asked about the easiness of prompting to guide the result generation. These defects reflect that, on the one hand, the Gen-AI does not understand the context as the participants, "*... I would like to augment "sparked a fire within me" with an image depicting a moment of Eureka or something, but the image was generated with literally a fire being sparked inside... I don't think it understood metaphors. (P18)*" Implicit expressions and metaphorical elements are proven to be engaging and leave room for audience imagination [100]. Although prior works [106] have demonstrated the ability of Gen-AI to comprehend metaphors, we argue that ***context-awareness of the Gen-AI is needed for augmenting storytelling***, enabling the Gen-AI to keep track of the elements of the story and understand what happens before and after the prompt.

6.2 Interactions to Author AIGC in AR Storytelling

Our second research question is to explore the key aspect when designing the interactions to create AIGC for AR Storytelling. We showed in section 5 that co-creating with Gen-AI to augment Storytelling is easy in some modalities but faced with difficulty with some other modalities individually, and participants were overall satisfied with the ease and richness in variation compared to human-created content empirically. Yet, our results showed that participants were not overwhelmingly positive about interactions during co-creation with Gen-AI, especially about how to guide the Gen-AI to generate their intended results. From our analysis, we further discuss the following factors to be taken into consideration when designing the interactions.

6.2.1 Interacting in AR to Prompt. According to our results, participants gave the lowest rating when evaluating the easiness of prompting to guide the generation of the system. In our experiment setup, we only enabled text-to-other content generation, where participants prompted with textual narratives. For modalities that require more details, prompting with texts is not intuitive in guiding the generation. "*I had trouble generating the motion of someone wandering and then stumbling. I think the video is the best way to express this. I tried different text inputs of the same meaning, but none of the generated videos did it the way I pictured in my mind... I would rather have demonstrated it myself.*

(P2)" Prior works have demonstrated the potential applications of utilizing spatial interactions in AR to interact with general AI, e.g. authoring through hand gesture [108], sketch-to-image generation [9, 25, 40], motion-to-animation [92, 110]. AR is a platform enabling multi-modalities not only for outputs but also inputs. Inspired by the results, we anticipate future research on AIGC in AR to reveal more possibilities for ***AR interactions prompting GenAI*** to enable intuitive content creation in AR, more than just text.

6.2.2 Live Authoring During Storytelling - When to Augment. In our study setup, we recruited experienced storytellers and presenters to augment their storytelling experience and examine it with our testbed. We have witnessed some negative ratings of the experience's expressiveness and immersion. As some participants pointed out, the audience is an essential part of the storytelling and their reaction is a deciding factor for the storytellers on when they should emphasize or elaborate part of the story. "*I think a better way to use this system is that when I actually tell the story and I figure out, from the audience's faces or my own improvisation, that I need something to show them so that they know what I am saying. You know, it is more of a feeling than some program. (P29)*" Storytelling is a two-way communication [77], even though the audience is silent most of the time. It is shown by the prior works such as Visual Captions [57] that when the content creation to augment the conversation is automated, communicators prefer live augmentations of their expression. With Gen-AI taking the burden of content creation, we suggest that the storytellers have more freedom to decide ***When to Augment*** their story during live presentations. Further, we anticipate future systems to have the ability to actively infer and augment AR Storytelling when the presenters need it during live presentations.

7 STUDY LIMITATIONS

For our study, we have developed a testbed consisting of a multi-modality content generator and an interactive AR Storytelling platform. The collected study results are heavily based on users' experiences within our specific testbed environment. Although the testbed serves the general purpose of exploring the effect of multi-modal AIGC in AR Storytelling, we did not design it in consideration of all specific or dedicated types of Storytelling tasks in diverse AR platforms. We acknowledge that for a target task in AR Storytelling, there can be specific design considerations or requirements for the generated content, and thus dedicated metrics for evaluating task-related output.

Additionally, Our testbed is built upon a desktop application with a webcam. The advances in AR headsets have enabled many interactions and experiences that are different from a desktop AR experience. This difference may limit the ability to fully immerse users in a three-dimensional, interactive environment. AR headsets can offer a hands-free experience. In contrast, using a laptop camera might require users to stay in a specific position, limiting their mobility and potentially affecting the overall engagement during Storytelling in AR. Most of the available multi-modal generative AI for AR Storytelling remains on desktop applications (e.g. [3, 18, 57]). In this paper, we aim to answer ***whether and why*** we should deploy multi-modal AIGC in AR Storytelling while acknowledging ***how to do it*** on the most popular platforms as another open research question, which is ongoing in the field (e.g. [38, 41, 114]) and also discussed in this paper. Yet, the contribution of our setup is a testbed for the study rather than a comprehensive system. Although the choice of platform may change the spatial experience, we anticipate that the impact of AIGC on AR Storytelling can also be revealed via experiments on desktop-based AR[16, 60].

Lastly, we focus on the modalities of the content in AR Storytelling, while the interactivity of AR Storytelling is also a key and promising aspect that we do not comprehensively cover. Such interactions involve (1) the necessary interactions the storytellers need to generate the content

(during generation) as well as (2) the interactions between the tellers and the authored content to understand the story (during presentation). With a focus on the modalities of AIGC, we mainly investigated (1) the ***interactions to create the content***. It is essential to recognize that Storytelling can manifest in diverse forms, and the limitations associated with hand interaction may influence the observed limitation. Thus, we suggest interpreting the results as a reference for utilizing Gen-AI in the creation of multi-modality content specifically for AR Storytelling with hand interaction. We anticipate future research to explore the applicability of these findings to other interaction modalities in AR Storytelling or to broader contexts beyond the scope of our developed testbed.

8 CONCLUSION

In this paper, we have presented an exploratory study of Mutli-modal Gen-AI in AR Storytelling. We created an AR-based test bed that is capable of creating multi-modal content as well as using the content in AR for storytelling. We initially reviewed 223 videos to find the design space for AR storytelling. We found two dimensions: 1) Modalities - text, audio, image, video, and 3D and 2) Elements - Character, Background, Sentiment, and Development. We conducted two studies to evaluate the effect of multi-modal generated content in AR Storytelling. We invited 30 participants, who were then divided into 2 groups of 15 each. From the analysis of both qualitative and quantitative results of the study, we provide insights into the future directions of AIGC in AR storytelling. We further discuss design considerations for future AR storytelling systems.

REFERENCES

[1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. 2022. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 conference proceedings*. 1–9.

[2] Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*. IEEE, 719–728.

[3] Victor Nikhil Antony and Chien-Ming Huang. 2023. ID. 8: Co-Creating Visual Stories with Generative AI. *arXiv preprint arXiv:2309.14228* (2023).

[4] National Storytelling Association et al. 2012. What Storytelling is. An attempt at defining the art form.

[5] Paulo Bala, Stuart James, Alessio Del Bue, and Valentina Nisi. 2022. Writing with (Digital) Scissors: Designing a Text Editing Tool for Assisted Storytelling Using Crowd-Generated Content. In *International Conference on Interactive Digital Storytelling*. Springer, 139–158.

[6] Valentin Bauer, Anna Nagele, Chris Baume, Tim Cowlishaw, Henry Cooke, Chris Pike, and Patrick GT Healey. 2019. Designing an interactive and collaborative experience in audio augmented reality. In *Virtual Reality and Augmented Reality: 16th EuroVR International Conference, EuroVR 2019, Tallinn, Estonia, October 23–25, 2019, Proceedings 16*. Springer, 305–311.

[7] Lee Anne Bell. 2010. *Storytelling for social justice: Connecting narrative and the arts in antiracist teaching*. Routledge.

[8] Eden Bensaid, Mauro Martino, Benjamin Hoover, and Hendrik Strobelt. 2021. Fairytailor: A multimodal generative framework for storytelling. *arXiv preprint arXiv:2108.04324* (2021).

[9] Sukanya Bhattacharjee and Parag Chaudhuri. 2020. A survey on sketch based content creation: from the desktop to virtual and augmented reality. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 757–780.

[10] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan. 2019. High fidelity speech synthesis with adversarial networks. *arXiv preprint arXiv:1909.11646* (2019).

[11] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22563–22575.

[12] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[13] Anjuli Joshi Brekke, Ralina Joseph, and Naheed Gina Aaftaab. 2021. "I address race because race addresses me": women of color show receipts through digital storytelling. *Review of Communication* 21, 1 (2021), 44–57.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural*

information processing systems 33 (2020), 1877–1901.
[15] Licia Calvi. 2020. What do we know about AR Storytelling?. In *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*. 278–280.
[16] Jorge Camba, Manuel Contero, and Gustavo Salvador-Herranz. 2014. Desktop vs. mobile: A comparative study of augmented reality systems for engineering visualizations in education. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*. IEEE, 1–8.
[17] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. 2023. Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models. *arXiv preprint arXiv:2305.13840* (2023).
[18] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: visual sketching of story generation with pretrained language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–4.
[19] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and Controllable Music Generation. *arXiv preprint arXiv:2306.05284* (2023).
[20] Delneshin Danaei, Hamid R Jamali, Yazdan Mansourian, and Hassan Rastegarpour. 2020. Comparing reading comprehension between children reading augmented reality and print storybooks. *Computers & Education* 153 (2020), 103900.
[21] Dimitrios Darzentas, Martin Flintham, and Steve Benford. 2018. Object-focused mixed reality storytelling: technology-driven content creation and dissemination for engaging user experiences. In *Proceedings of the 22nd Pan-Hellenic Conference on Informatics*. 278–281.
[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[23] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems* 35 (2022), 16890–16902.
[24] Runlin Duan, Shao-Kang Hsia, Yuzhao Chen, Yichen Hu, Ming Yin, and Karthik Ramani. 2025. Investigating Creativity in Humans and Generative AI Through Circles Exercises. *arXiv preprint arXiv:2502.07292* (2025).
[25] Runlin Duan, Nachiketh Karthik, Jingyu Shi, Rahul Jain, Maria C Yang, and Karthik Ramani. 2024. ConceptVis: Generating and Exploring Design Concepts for Early-Stage Ideation Using Large Language Model. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 88377. American Society of Mechanical Engineers, V03BT03A042.
[26] Runlin Duan, Chenfei Zhu, Yuzhao Chen, Yichen Hu, Jingyu Shi, and Karthik Ramani. 2025. DesignFromX: Empowering Consumer-Driven Design Space Exploration through Feature Composition of Referenced Products. *arXiv preprint arXiv:2505.11666* (2025).
[27] FastAPI. [n. d.]. FastAPI. https://fastapi.tiangolo.com.
[28] Klaus Fog, Christian Budtz, and Baris Yakaboylu. 2005. *Storytelling*. Springer.
[29] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.
[30] Ze Gao, Anqi Wang, Pan Hui, and Tristan Braud. 2022. Bridging curatorial intent and visiting experience: Using ar guidance as a storytelling tool. In *Proceedings of the 18th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*. 1–10.
[31] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).
[32] Ariel Han and Zhenyao Cai. 2023. Design implications of generative AI systems for visual storytelling for young learners. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 470–474.
[33] Linda Hirsch, Robin Welsch, Beat Rossmy, and Andreas Butz. 2022. Embedded AR Storytelling Supports Active Indexing at Historical Places. In *Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction*. 1–12.
[34] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).
[35] Lissa Holloway-Attaway and Lars Vipsjö. 2020. Using augmented reality, gaming technologies, and transmedial storytelling to develop and co-design local cultural heritage experiences. *Visual Computing for Cultural Heritage* (2020), 177–204.
[36] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spaCy: Industrial-strength natural language processing in python. (2020).
[37] Xiyun Hu, Dizhi Ma, Fengming He, Zhengzhe Zhu, Shao-Kang Hsia, Chenfei Zhu, Ziyi Liu, and Karthik Ramani. 2025. GesPrompt: Leveraging Co-Speech Gestures to Augment LLM-Based Interaction in Virtual Reality. *arXiv preprint arXiv:2505.05441* (2025).

[38] Yongquan Hu, Mingyue Yuan, Kaiqi Xian, Don Samitha Elvitigala, and Aaron Quigley. 2023. Exploring the Design Space of Employing AI-Generated Content for Augmented Reality Display. *arXiv preprint arXiv:2303.16593* (2023).
[39] Reinis Indans, Eva Hauthal, and Dirk Burghardt. 2019. Towards an audio-locative mobile application for immersive storytelling. *KN-Journal of Cartography and Geographic Information* 69 (2019), 41–50.
[40] Myunggeun Ji and Junchul Chun. 2020. A Sketch-based 3D Object Retrieval Approach for Augmented Reality Models Using Deep Learning. *Journal of Korean Society for Internet Information* 21, 1 (2020).
[41] BRENNAN JONES, YAN XU, MARY ANNE HOOD, MOHAMMAD SHAHIDUL KADER, and HAMID EGHBALZADEH. [n. d.]. Using Generative AI to Produce Situated Action Recommendations in Augmented Reality for High-Level Goals. ([n. d.]).
[42] Kwanghee Jung, Vinh T Nguyen, and Jaehoon Lee. 2021. Blocklyxr: An interactive extended reality toolkit for digital storytelling. *Applied Sciences* 11, 3 (2021), 1073.
[43] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient neural audio synthesis. In *International Conference on Machine Learning*. PMLR, 2410–2419.
[44] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* 34 (2021), 852–863.
[45] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.
[46] Sarah Ketchell, Winyu Chinthammit, and Ulrich Engelke. 2019. Situated storytelling with SLAM enabled augmented reality. In *Proceedings of the 17th International Conference on Virtual-Reality Continuum and Its Applications in Industry*. 1–9.
[47] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439* (2023).
[48] Doyeon Kim, Donggyu Joo, and Junmo Kim. 2020. Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access* 8 (2020), 153113–153122.
[49] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. 2023. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 8255–8263.
[50] Kiyoung Kim, Noh-young Park, and Woontack Woo. 2014. Vision-based all-in-one solution for augmented reality and its storytelling applications. *The Visual Computer* 30 (2014), 417–429.
[51] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
[52] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761* (2020).
[53] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems* 32 (2019).
[54] Tomas Lawton, Francisco J Ibarrola, Dan Ventura, and Kazjon Grace. 2023. Drawing with Reframer: Emergence and Control in Co-Creative AI. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 264–277.
[55] Kyungjun Lee, Hong Li, Muhammad Rizky Wellyanto, Yu Jiang Tham, Andrés Monroy-Hernández, Fannie Liu, Brian A Smith, and Rajan Vaish. 2023. Exploring Immersive Interpersonal Communication via AR. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–25.
[56] Jian Liao, Adnan Karim, Shivesh Singh Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. 2022. RealityTalk: Real-Time Speech-Driven Augmented Presentation for AR Live Storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–12.
[57] Xingyu" Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang" Anthony" Chen, and Ruofei Du. 2023. Visual Captions: Augmenting Verbal Communication With On-the-Fly Visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
[58] Ziyi Liu, Zhengzhe Zhu, Lijun Zhu, Enze Jiang, Xiyun Hu, Kylie A Peppler, and Karthik Ramani. 2024. Classmeta: Designing interactive virtual classmate to promote VR classroom participation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
[59] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. 2023. Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos. *arXiv preprint arXiv:2304.01186* (2023).
[60] Tiago Madeira, Bernardo Marques, Pedro Neves, Paulo Dias, and Beatriz Sousa Santos. 2022. Comparing desktop vs. Mobile interaction for the creation of pervasive augmented reality experiences. *Journal of Imaging* 8, 3 (2022), 79.
[61] Jenny Mandelbaum. 2012. Storytelling in conversation. *The handbook of conversation analysis* (2012), 492–507.
[62] Dimitrios Markouzis and Georgios Fessakis. 2015. Interactive storytelling and mobile augmented reality applications for learning and entertainment—a rapid prototyping perspective. In *2015 International Conference on Interactive Mobile*

Communication Technologies and Learning (IMCL). IEEE, 4–8.
[63] MediaPipe. [n. d.]. MediaPipe. https://mediapipe.dev/.
[64] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. 2016. SampleRNN: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837* (2016).
[65] Carolyn Handler Miller. 2019. *Digital Storytelling 4e: A creator's guide to interactive entertainment.* CRC Press.
[66] Ron Mokady, Omer Tov, Michal Yarom, Oran Lang, Inbar Mosseri, Tali Dekel, Daniel Cohen-Or, and Michal Irani. 2022. Self-distilled stylegan: Towards generation from internet photos. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.
[67] mozilla. [n. d.]. Web Speech API. https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API.
[68] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. 2023. Conditional Image-to-Video Generation with Latent Flow Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18444–18455.
[69] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).
[70] Jennifer O'Meara and Kata Szita. 2021. AR cinema: Visual storytelling and embodied experiences with augmented reality filters and backgrounds. *PRESENCE: Virtual and Augmented Reality* 30 (2021), 99–123.
[71] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).
[72] Augusto Palombini. 2017. Storytelling and telling history. Towards a grammar of narratives for Cultural Heritage dissemination in the Digital Era. *Journal of cultural heritage* 24 (2017), 134–139.
[73] Seung-Bo Park, Jason J Jung, and EunSoon You. 2015. Storytelling of collaborative learning system on augmented reality. *New Trends in Computational Collective Intelligence* (2015), 139–147.
[74] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2085–2094.
[75] John V Pavlik and Frank Bridges. 2013. The emergence of augmented reality (AR) as a storytelling medium in journalism. *Journalism & Communication Monographs* 15, 1 (2013), 4–59.
[76] Kainan Peng, Wei Ping, Zhao Song, and Kexin Zhao. 2020. Non-autoregressive neural text-to-speech. In *International conference on machine learning*. PMLR, 7586–7598.
[77] Eric E Peterson and Kristin M Langellier. 2006. Communication as storytelling. *Communication as... Perspectives on Theory* (2006), 123–131.
[78] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
[79] Judith Pintar. 2023. Invisible, aesthetic, and enrolled listeners across storytelling modalities: Immersive preference as situated player type. *Convergence* (2023), 13548565231206505.
[80] Matthias Plappert, Christian Mandery, and Tamim Asfour. 2016. The KIT motion-language dataset. *Big data* 4, 4 (2016), 236–252.
[81] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).
[82] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
[83] Gideon Raeburn, Martin Welton, and Laurissa Tokarchuk. 2022. Developing a play-anywhere handheld AR storytelling app using remote data collection. *Frontiers in Computer Science* 4 (2022), 927177.
[84] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
[85] resfulapi. [n. d.]. RESTful API. https://restfulapi.net/.
[86] Rosalie Rolón-Dow. 2011. Race (ing) stories: Digital storytelling as a tool for critical race scholarship. *Race Ethnicity and Education* 14, 2 (2011), 159–173.
[87] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[88] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

[89] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

[90] Kari Salo, Diana Giova, and Tommi Mikkonen. 2016. Backend infrastructure supporting audio augmented reality and storytelling. In *Human Interface and the Management of Information: Applications and Services: 18th International Conference, HCI International 2016 Toronto, Canada, July 17-22, 2016. Proceedings, Part II 18*. Springer, 325–335.

[91] Emanuel A Schegloff. 1997. " Narrative analysis" thirty years later. *Journal of narrative and life history* 7, 1-4 (1997), 97–106.

[92] Jingyu Shi, Rahul Jain, Seunggeun Chi, Hyungjun Doh, Hyung-gun Chi, Alexander J Quinn, and Karthik Ramani. 2025. CARING-AI: Towards Authoring Context-aware Augmented Reality INstruction through Generative Artificial Intelligence. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[93] Jingyu Shi, Rahul Jain, Hyungjun Doh, Ryo Suzuki, and Karthik Ramani. 2023. An HCI-centric survey and taxonomy of human-generative-AI interactions. *arXiv preprint arXiv:2310.07127* (2023).

[94] Jingyu Shi, Rahul Jain, Runlin Duan, and Karthik Ramani. 2023. Understanding Generative AI in Art: An Interview Study with Artists on G-AI from an HCI Perspective. *arXiv preprint arXiv:2310.13149* (2023).

[95] Jae-eun Shin, Boram Yoon, Dooyoung Kim, and Woontack Woo. 2022. The Effects of Spatial Complexity on Narrative Experience in Space-Adaptive AR Storytelling. *IEEE Transactions on Visualization and Computer Graphics* (2022).

[96] Bilal Şimşek and Bekir Direkçi. 2023. The effects of augmented reality storybooks on student's reading comprehension. *British Journal of Educational Technology* 54, 3 (2023), 754–772.

[97] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).

[98] Abbey Singh, Ramanpreet Kaur, Peter Haltner, Matthew Peachey, Mar Gonzalez-Franco, Joseph Malloch, and Derek Reilly. 2021. Story creatar: a toolkit for spatially-adaptive augmented reality storytelling. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 713–722.

[99] Nayia Stylianidou, Angelos Sofianidis, Elpiniki Manoli, and Maria Meletiou-Mavrotheris. 2020. "Helping Nemo!"—Using augmented reality and alternate reality games in the context of universal design for learning. *Education Sciences* 10, 4 (2020), 95.

[100] Murray Taylor, Mauricio Marrone, Mark Tayar, and Beate Mueller. 2018. Digital storytelling and visual metaphor in lectures: a study of student engagement. *Accounting Education* 27, 6 (2018), 552–569.

[101] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*. Springer, 358–374.

[102] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

[103] Anastasia Tyurina. 2023. Leveraging AR-Driven Visual Storytelling to Enhance Communication of Complex Social Issues: Principles, Strategies, and Multifaceted Roles of Visuals. In *SIGGRAPH Asia 2023 Educator's Forum*. 1–2.

[104] Tom Van Laer, Stephanie Feiereisen, and Luca M Visconti. 2019. Storytelling in the digital era: A meta-analysis of relevant moderators of the narrative transportation effect. *Journal of Business Research* 96 (2019), 135–146.

[105] Fernando Vera and J Alfredo Sánchez. 2016. A model for in-situ augmented reality content creation based on storytelling and gamification. In *Proceedings of the 6th Mexican Conference on Human-Computer Interaction*. 39–42.

[106] Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 Grasp Metaphors? Identifying Metaphor Mappings with Generative Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1018–1032.

[107] Josephine Walwema. 2015. The Art of Storytelling. *Writing & Pedagogy* 7, 1 (2015).

[108] Tianyi Wang, Xun Qian, Fengming He, Xiyun Hu, Yuanzhi Cao, and Karthik Ramani. 2021. Gesturar: An authoring system for creating freehand interactive augmented reality applications. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 552–567.

[109] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6199–6203.

[110] Hui Ye, Kin Chung Kwan, Wanchao Su, and Hongbo Fu. 2020. ARAnimator: In-situ character animation in mobile AR with user-defined motion gestures. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 83–1.

[111] Rabia Meryem Yilmaz and Yuksel Goktas. 2017. Using augmented reality technology in storytelling activities: examining elementary students' narrative skill and creativity. *Virtual Reality* 21 (2017), 75–89.

[112] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* 2, 3 (2022), 5.

[113] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001* (2022).

[114] Zhenjie Zhao and Xiaojuan Ma. 2018. A compensation method of two-stage image generation for human-ai collaborated in-situ fashion design in augmented reality environment. In *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 76–83.

[115] ZhiYing Zhou, Adrian David Cheok, JiunHorng Pan, and Yu Li. 2004. An interactive 3D exploration narrative interface for storytelling. In *Proceedings of the 2004 conference on Interaction design and children: building a community*. 155–156.

A STORIES

In this section, we manifest the five stories we used in our study. We marked four atomic elements to augment in color, namely Characters, Backgrounds, Sentiments, and Development.

A.1 Story - 1

Once in a small, mountain-ringed village, lived Emily, a young painter whose heart brimmed with passion but also frustration. She loved the natural world, yet felt her paintings never truly captured its essence.

One day, Emily stumbled upon a hidden waterfall in the forest, a scene so breathtaking it seemed otherworldly. Inspired yet challenged, she set out to paint it. Her usual methods felt inadequate against the waterfall's majesty. She spent days and nights experimenting with new styles and techniques, striving to capture not just the sight, but the emotion it stirred within her.

Through this tireless effort, Emily realized her art shouldn't just replicate a scene; it needed to convey the feelings it evoked. Finally, she created a masterpiece painting, vastly different from her previous works. It was vibrant, alive, a visual echo of the waterfall's spirit.

When exhibited, her painting moved the villagers, evoking the sensation of being right there at the waterfall. Emily's journey transformed her artistic style and understanding of art itself. She learned that true art was about conveying emotion and connecting deeply with others. Her new approach brought not just satisfaction, but recognition, and she continued exploring and capturing the world's beauty, infusing her emotions into her canvases.

A.2 Story - 2

I've always been drawn to the stars. Ever since I was a child, lying on the grass, gazing up at the night sky, I felt a connection to the cosmos. My name is Alex, and I am an astronomer.

My life revolves around telescopes and celestial charts, seeking to unlock the mysteries of the universe. But it was one particular night that changed everything for me. As I peered through my telescope, I discovered a celestial phenomenon - a star that seemed to pulsate with an unusual light. This discovery consumed me, and I dedicated months to studying it, often forgetting the world around me. The more I studied, the more I realized that this star's behavior defied all our known laws of astrophysics.

This revelation was not just a scientific breakthrough; it made me question everything we knew about the universe. It humbled me, reminding me of the infinite mysteries of space. Sharing my findings with the world brought new excitement to the astronomical community, and together, we embarked on a journey of redefining our understanding of the cosmos.

A.3 Story - 3

On a remote coastline, battered by relentless waves, stood an old lighthouse, it lights a solitary beacon in the foggy nights. The lighthouse keeper, an old seafarer named Captain Holloway,

spent his days in , tending to the lighthouse and reminiscing about his long-lost love, a woman as enigmatic and wild as the sea itself. Their love was a tempest, intense and fleeting, as she disappeared one stormy night, leaving Holloway with only memories. Each evening, he would play a melancholic melody on his violin, the notes carried by the wind, hoping they would reach her wherever she might be.

One stormy night, as Holloway played his violin, a figure appeared in the rain - it was her, drawn back by the sorrowful melody. They shared one last night together, their love rekindled amidst the storm. By morning, she was gone again, but this time, Holloway felt a sense of closure. He continued to tend the lighthouse, his heart no longer heavy with loss, but filled with gratitude for the brief reunion granted by the sea's mysterious ways.

A.4 Story - 4

Deep in the heart of the dense jungle, hidden under layers of time and nature, lay the ruins of an ancient city, once a bustling center of an advanced civilization, now forgotten by the world. Its sole guardian was an ageless being known as Arion, a creature both human and spirit, bound to the city by an ancient curse. Arion's existence was a solitary one, his only purpose to protect the secrets and artifacts of the lost city. For centuries, he kept intruders at bay, using his powers to create illusions and manipulate the jungle's treacherous paths.

His resolve was unyielding until the arrival of a team of explorers, led by a determined archaeologist named Dr. Elena Mendoza. Unlike others before, they sought not treasure, but knowledge. Moved by their genuine quest for understanding, Arion revealed himself and the city's history. He guided them through the ruins, each structure and artifact a testament to the city's once-great civilization. With Arion's help, Dr. Mendoza and her team were able to unravel the mysteries of the city, bringing its forgotten history to light. Arion, having fulfilled his purpose, was finally released from his curse, allowing the spirit of the lost city to live on through the pages of history.

A.5 Story - 5

My name is Isabella, and I am an artist, though the world has long forgotten me. I used to paint vibrant landscapes and portraits full of emotion, but as the years passed, my art fell into obscurity.

Living in a small, quiet town, my days were spent walking through the streets, my heart heavy with unspoken stories and unseen art. One day, while wandering, I stumbled upon an abandoned building. Inside, I found walls covered in graffiti - a tapestry of untold stories. This discovery sparked a fire within me. I began to paint again, but this time, I painted on those walls. My art became a fusion of my traditional style and the raw, expressive nature of graffiti. As I transformed the walls into a gallery of emotions, people started to visit, drawn by the curious blend of art styles. My forgotten art found a new audience, and more importantly, I rediscovered my passion.

Through this unexpected journey, I learned that art is not just about recognition, but about the joy of creation and the impact it has on others, even in the most unexpected places.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009