RetAssist: Facilitating Vocabulary Learners with Generative Images in Story Retelling Practices

QIAOYI CHEN, Sun Yat-sen University, China
SIYU LIU, Sun Yat-sen University, China
KAIHUI HUANG, Sun Yat-sen University, China
XINGBO WANG, Cornell University, United States
XIAOJUAN MA, The Hong Kong University of Science and Technology, China
JUNKAI ZHU, Guangdong Polytechnic of Industry and Commerce, China
ZHENHUI PENG*, Sun Yat-sen University, China

Reading and repeatedly retelling a short story is a common and effective approach to learning the meanings and usages of target words. However, learners often struggle with comprehending, recalling, and retelling the story contexts of these target words. Inspired by the Cognitive Theory of Multimedia Learning, we propose a computational workflow to generate relevant images paired with stories. Based on the workflow, we work with learners and teachers to iteratively design an interactive vocabulary learning system named *RetAssist*. It can generate sentence-level images of a story to facilitate the understanding and recall of the target words in the story retelling practices. Our within-subjects study (N=24) shows that compared to a baseline system without generative images, *RetAssist* significantly improves learners' fluency in expressing with target words. Participants also feel that *RetAssist* eases their learning workload and is more useful. We discuss insights into leveraging text-to-image generative models to support learning tasks.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**; *Empirical studies in HCI*.

Additional Key Words and Phrases: Vocabulary learning, story retelling, image generation

ACM Reference Format:
Qiaoyi Chen, Siyu Liu, Kaihui Huang, Xingbo Wang, Xiaojuan Ma, Junkai Zhu, and Zhenhui Peng. 2024. RetAssist: Facilitating Vocabulary Learners with Generative Images in Story Retelling Practices. In *Designing Interactive Systems Conference (DIS '24), July 03–05, 2024, Copenhagen, Denmark.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3643834.3661581

1 INTRODUCTION

Learning vocabulary in meaningful contexts, such as stories and images in language learning textbooks, and video clips from movies, is a common and effective practice as it enables deep and active processing of vocabulary (*e.g.,* word associations, logic) [50]. Human-Computer Interaction (HCI) researchers have explored various technologies to support vocabulary learners with meaningful contexts in various learning activities, *e.g., ViVo* in watching videos [74], *VocabEncounter* in reading online articles [3], and *EnglishBot* in conversing with others [61]. In this paper, we focus on the story retelling activity that encourages English-as-the-Second-Language vocabulary learners to integrate, reconstruct, and demonstrate the contextualized use of the target words in a short story

*Corresponding author.

DIS '24, July, 2024, DK

ACM ISBN 979-8-4007-0583-0/24/07…$15.00
https://doi.org/10.1145/3643834.3661581

[23, 35, 42, 44]. This practice typically involves two stages – story comprehension and repeated retelling [47], *i.e.,* the learner first reads or listens to a short story that contains a set of target words to comprehend its main idea and then verbally retells it for multiple rounds. Several studies on language education have demonstrated the effectiveness of story retelling for vocabulary learning [16, 23, 43], especially in remembering the meanings of target words and using them in verbal expressions [16, 61]. In fact, story retelling has been included in the English test of the College Entrance Examination in China [1].

However, the story retelling practice is often challenging for learners of second language vocabulary. For one thing, in the story comprehension stage, learners need to associate the meanings of target words with the story context and memorize the story flow for the later repeated retelling practice [33, 44, 46]. For another, in the repeated retelling stage, they should repeatedly narrate the read story with requirements on the correct usage of target words and fluency in speaking them out in the story [33, 44, 46]. In other words, it requires learners to understand, memorize, recall, organize, and speak the target words and associated story [33]. This becomes more challenging when there is a time limit for each round of repeated retelling, which could help to develop language fluency under pressure [46]. Images related to the story can help vocabulary learners cope with these two challenges during the story retelling practice. As suggested by the Cognitive Theory of Multimedia Learning (CTML) [53], building mental representations from text and visual elements could facilitate comprehension and recall of words and their contextualized usage [19, 41, 49, 65]. In the context of second language acquisition, individuals tend to subvocally articulate the text associated with visual stimuli in their native language [52]. Thus, compared to learning without visual aids, non-verbal modalities such as images bridge the gap between two different languages, which would enhance the likelihood of recalling the second language [51, 52]. Given these benefits, language educators widely prepare relevant images for the textual stories in course books or online resources, and HCI researchers have proposed vocabulary learning support systems in activities that involve visual elements [1, 29, 74]. For example, *CoSpeak* [1] uses voice recognition techniques to support students to collaboratively and verbally create a story given an image prompt. However, it is time-consuming and often unavailable to prepare relevant images for the story with any set of target words that users wish to learn in the story retelling practices, while irrelevant images would confuse learners and reduce vocabulary learning outcome [26].

In this work, we explore the design and usage of generative images to facilitate the learning of any target word set via reading and repeatedly retelling a short story that contains these words. Our focus is motivated by the benefits of images for vocabulary learning as described above and recent advances in text-to-image generative techniques. For example, the pre-trained Latent Diffusion Model (LDM) [60] is able to generate high-quality and content-relevant images given a text prompt. These generative techniques have been used to support the creations of artworks [21], medical images [20], and game characters [17]. Nevertheless, little work, if any, has explored generative images for supporting vocabulary learning in the story retelling practices where users should master target words' meanings and verbal expressions. Questions arise such as 1) whether and how text-to-image generative techniques can generate relevant images of any story that covers a target word set, 2) if so, what kinds of support that the generative images can offer in the story retelling practices, and 3) how would the support from generative images impact the users' vocabulary learning outcome and experience.

To this end, we seek to provide insights into these questions by designing, developing, and evaluating an intelligent system prototype, *RetAssist*, that can generate relevant images for learning vocabulary in the story retelling practices. Here, we target English-as-the-Second-Language (ESL)

[1]https://gaokao.eol.cn/guang_dong/dongtai/201811/t20181101_1631228.shtml

Chinese learners, *e.g.,* high-school or university students in China. We take an iterative design approach with insights from educational literature and the involvement of ESL learners and English teachers in this process. We first develop a text-to-image computational workflow and validate its capability in generating a series of coherent and relevant sentence-level images given any short textual story that contains IELTS [2] target words. We then conduct an interview study with seven ESL learners to understand their challenges and needs in the story retelling practices and ask for their comments on the generative images. Based on the insights from the interviews and educational literature, we develop a *RetAssist* prototype and seek feedback from another 18 ESL learners and two English teachers to refine it. In the story retelling practice with the refined *RetAssist*, users can first read and listen to the story with generative images aligned to each sentence. Then, during each round of repeated retelling, users can retell the story by viewing the images. After each round, users can review their performance in the expressions of target words and re-read the story with images.

We conduct a within-subjects study with 24 ESL vocabulary learners to evaluate the impact of *RetAssist*'s function on the generative image on the vocabulary learning outcome and experience. The results show that compared with the baseline system without generative images, participants using *RetAssist* significantly outperform in fluently using the target words in verbal expressions. Participants favor the generative images of *RetAssist* for reducing learning workload and aiding recall of the contextual usage of target words in the story. Based on our findings, we highlight the value of text-to-image generative techniques in offering useful learning materials and enjoyable learning experiences. We further discuss design considerations for future vocabulary learning support systems and the impact of our work on generative AIs for education.

Our work makes three contributions. First, we present a vocabulary learning system *RetAssist* that uses generative images to facilitate users to master target words' meanings and expressions via story retelling practices. Second, our design and evaluation of *RetAssist* provide first-hand findings on the feasibility, effectiveness, and user experience of applying text-to-image generative models to vocabulary learning. Third, we propose a story text-to-image generation workflow and offer design considerations of leveraging generative models to support learning tasks.

2 RELATED WORK

We introduce prior studies that motivate, inspire, and support the design of *RetAssist*, including story retelling for vocabulary learning, vocabulary learning systems, and text-to-image generation techniques.

2.1 Story Retelling for Vocabulary Learning

Story retelling is a well-recognized approach that helps students acquire vocabulary and skills like reading, listening, and speaking in language learning and teaching [42, 44, 46, 47]. A story retelling practice normally consists of two stages, *i.e.,* story comprehension in which learners listen or read a given story, and repeated retelling in which they speak it out for several times within a time limit [46, 47]. As suggested by Nation *et al.* [46], it is a practice that properly integrates four typical strands of activities, *i.e.,* meaning-focused input, meaning-focused output, language-focused learning, and fluency development. First, in the story comprehension stage, learners focus on understanding the given story with target words – using language receptively (meaning-focus input) [46]. Next, in the repeated retelling stage, learners are required to correctly and fluently speak the story out – using language productively (meaning-focus output) [46]. Moreover, in the whole

[2]Short for International English Language Testing System, a globally recognized standardized test designed to assess the English language proficiency of individuals.

practice, learners should specifically pay attention to the meanings, pronunciations, and correct usages of the target words – deliberate learning of language features (language-focused learning) [46]. Lastly, the practice requires learners to make the best use of what they already know to perform well in retelling under time pressure [6, 30, 33, 44], which is a typical fluency development learning activity [46]. In all, story retelling encourages learners to integrate, reconstruct, and demonstrate the contextual use of vocabulary [23, 42, 44]. The expected learning outcome is, therefore, not only on memorization of target words' meanings but also on the capacity of using the words correctly and fluently in language expressions [16, 22].

Given these requirements of reading, interpreting, memorizing, and speaking the story with target words [33, 44, 46], story retelling is often challenging for learners. Traditionally, there are additional materials (*e.g.,* images and props) to the textual story and in-situ guidance from teachers (*e.g.,* prompting phrases) to assist learners in the story retelling practices [14, 16]. Images relevant to the story, for example, are beneficial in that they can help learners remember the words' meanings and comprehend the story [2, 19, 49]. Images can also serve as the visual guidance that helps learners recall the story and target words when they get stuck in the repeated retelling stage [68]. According to Cognitive Theory of Multimedia Learning (CTML) [19, 41, 49, 65] and Dual Coding Theory (DCT) [51, 53], building mental representations from text and visual elements could enhance the encoding and retention of information by leveraging dual coding, which taps into both verbal and visual processing systems in the brain. As the extension of DCT, Bilingual Dual Coding Theory (BDCT) [52] suggests that images enhance second language learning since learners covertly pronounce the content of the images in their native language and the content and the images converge on the foreign language responses, increasing the probability of recall relative to the condition without images. Furthermore, Mayer identified the twelve multimedia instructional principles to address the issue of how to structure multimedia instructional practices and employ more effective cognitive strategies to help people learn efficiently [41]. For instance, the spatial contiguity principle [41] indicates that people learn better when corresponding words and images are placed near each other rather than far from each other on the page or screen. In summary, these principles suggest that relevant visuals (*e.g.,* images) can significantly aid in recalling textual information (*e.g.,* story in our case) of vocabulary to facilitate vocabulary learning. Despite the clear benefits, selecting appropriate images that align closely with the textual content remains a considerable challenge [27].

Our work is motivated by the benefits of story retelling practices for enhancing understanding and expression of target words and the helpfulness of images for assisting users in these practices. Instead of requiring human effort to prepare the images, we propose to generate relevant images to any story that covers the target words that users wish to learn.

2.2 Vocabulary Learning Systems

Existing HCI researchers have explored various intelligent systems to support vocabulary learning. Broadly speaking, they are either based on word lists or meaningful contexts. The former type of vocabulary learning system aims to facilitate quick memorization of target words in a list. Previous work has incorporated models of users' memory cycles and individualized learning styles into these systems, such that they can recommend a set of target words with appropriate levels of difficulty and repetition frequency [9, 48, 73]. For example, Chen *et al.* [9] proposed a personalized mobile English vocabulary learning system based on Item Response Theory and the learning memory cycle.

Context-based vocabulary learning systems leverage various forms of materials such as stories [1], videos [74], and online articles [3] to help users learn vocabulary. For example, *VocabEncounter* [3] encapsulates target vocabulary into the context of online articles, while *ARLang* [7] visualizes

bilingual labels on physical objects outdoors in AR environment to support the micro-learning of language within its spatial context. Additionally, *EnglishBot* [61], a language learning chatbot, engages students in interactive conversations on college-related topics to learn English. Learners can click as needed to receive answer prompts provided in their native language, ensuring smooth conversations with *EnglishBot* [61]. In line with *EnglishBot*'s method, *RetAssist* allows users to autonomously click on corresponding images based on their current progress when comprehending or retelling stories. Some studies use images as visual contexts to support vocabulary learning. *AIVAS* [26] uses an image reranking algorithm to select images that prominently contain relevant objects in the middle ground, thus aiding in representing concrete nouns effectively. Furthermore, *FCAI* [27] considers users' personal information, learning time, and location to recommend contextually appropriate images that best represent the target words. Both *AIVAS* and *FCAI* focus on searching for appropriate images for target words, whereas the images in *RetAssist* need to represent the story content, potentially favoring a generative approach. Story retelling also facilitates contextualized vocabulary learning. *CoSpeak* [1] provides an application for learners to practice speaking English by pairing them together to co-create a story with an image prompt based on the ongoing topic in class. Unlike the focus of our study on individual learners using *RetAssist* for vocabulary learning through story retelling, *CoSpeak* concentrates on enhancing English oral expression through dialogues between two individuals in thematic story settings.

Our proposed *RetAssist* falls into the category of context-based vocabulary learning systems. *RetAssist* not only integrates vocabulary into the story to provide textual context, but also generates a set of related images for the story that serve as the visual context to help individual vocabulary learners acquire vocabulary through story retelling.

2.3 Text-to-Image Generation Techniques

As suggested by the Dual Coding Theory [51, 53], textual stories paired with relevant images can facilitate vocabulary learning. Recent advances in text-to-image generative techniques offer great potential for preparing visual aids for any story that covers learners' interested words. Text-to-image generative models normally take a text prompt as input and output one or multiple images that are related to the text content. One of the early representatives is Diffusion Probabilistic Model (DM) [64], which achieved state-of-the-art results in density estimation (*i.e.,* how well the model captures the probability distribution of the dataset) [32] as well as in sample quality (*i.e.,* how well the model generates data samples that closely resemble real data from that distribution) [13]. DMs use deep learning techniques to generate high-quality images from text prompts but have the downside of low inference speed. To address the drawback, recent approaches widely leverage Latent Diffusion Models (LDMs) [60], which work on a compressed latent space of lower dimensionality and speed up inference with almost no reduction in image synthesis quality. In this work, we use a state-of-the-art LDM for generating images from text prompts.

Compared to the traditional text-to-image tasks that generate images from a text prompt, image sequence generation for stories is more challenging as it needs to generate a sequence of coherent and consistent images for a story that contains multiple sentences. To enhance the image quality and the consistency of the generated sequences, *StoryGAN* consists of a deep Context Encoder that dynamically tracks the story flow, and two discriminators at the story and image levels [37]. *Neural Storyboard Artist* visualizes the story in the form of a comic strip through the retrieval of multiple related images from the story content and several image rendering steps like segmenting relevant regions and converting the images into cartoon style [10]. Nevertheless, previous story image generation techniques prioritize coherence of the whole story flow, which may overlook the contexts (*e.g.,* sentences) containing target words for learning. To facilitate vocabulary learning based on story retelling practices, we segment the whole story into sentences to provide rich contexts

for target words. The sentences are used as prompts to generate a sequence of relevant images. Specifically, we use a Stable Diffusion model [63] to convert sentences into images and apply a cross-modal model, CLIP [57], to select the most relevant one for each sentence. To improve the visual consistency and coherence of the image sequence, we follow [10] and use a style transfer model [11] to unify the images with cartoon styles. Thereafter, our proposed computational workflow can generate relevant and style-consistent images for story sentences as meaningful contexts to facilitate story retelling for vocabulary learning.

3 DESIGN PROCESS

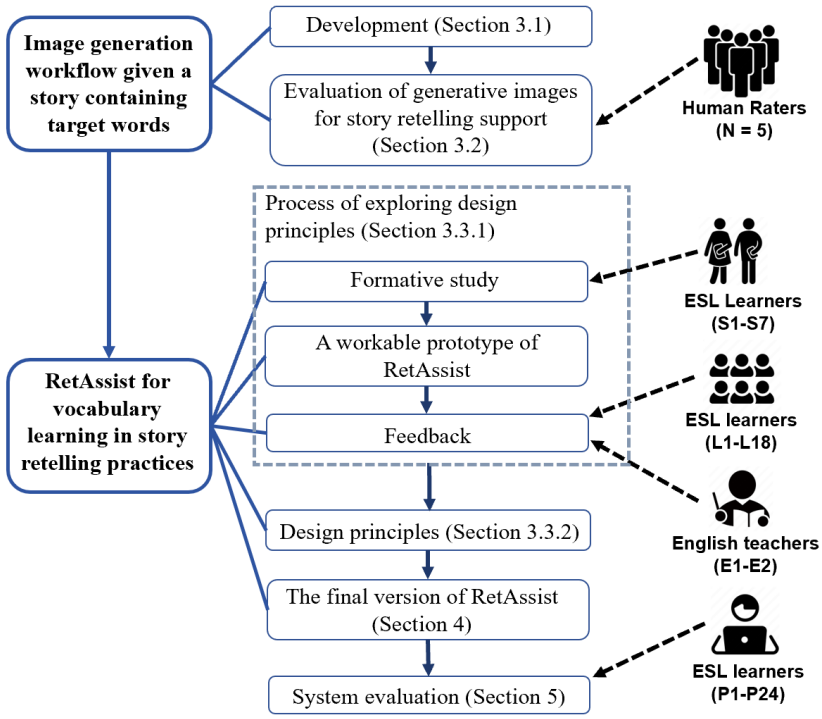

Fig. 1. Our design and development process of *RetAssist* with English teachers and ESL learners.

In this section, we explain how we design and develop *RetAssist* to facilitate vocabulary learners to read and repeatedly retell any story that covers their interested target words (Figure 1). First, we propose a computational workflow for text-to-image generation and validate its feasibility in generating a series of coherent and relevant sentence-level images given any short textual story that contains target words. Then, we work with vocabulary learners and teachers to derive the design principles of *RetAssist*.

3.1 Developing a Computational Workflow for Text-to-Image Generation

To assist users in the story retelling practices, the generative images of a story should satisfy the following two requirements. First, the images should be semantically relevant to the textual story. As suggested by the Dual Coding Theory [51, 53], the brain processes visual and verbal information in distinct regions. The visual channel handles visual data, generating pictorial representations, while

the verbal channel processes verbal information, producing corresponding verbal representations. When the visual and verbal inputs are semantically relevant, people establish mental connections that organize information into cause-and-effect chains [28]. After these connections are formed, there is a significant enhancement in the ability to remember information [54]. Therefore, when serving language learning, the visual information in the image should semantically match with the text content. Second, the images themselves should be coherent in their content and consistent in styles to depict a story [41]. Otherwise, the images could confuse learners in the story retelling practices. These two requirements guide our design choices in the computational workflow, as detailed below (Figure 2).

Sub-step 1: Preprocess the Story. We choose to generate one image for each story sentence for two reasons. First, a series of images rather than one image can better reveal the logic of the story [53]. Second, the Segmenting Principal [41] suggests that preparing an image for each story segment can provide natural pauses for learners to absorb the content before proceeding to the next segment [31]. We use the Spacy package in Python to split the story into sentences. To maintain the coherence among the generative images, we further resolve coreferences in the story sentences, *e.g.,* pronouns like "he" and "it" refer to the objects mentioned earlier. Specifically, we adopt a pretrained coreference resolution model named NeuralCoref [71] to select the reference words in the story to replace the pronoun in each split sentence. For example, for the red text in Figure 3, the pronounce "he" in the second and fourth sentences of the example story is replaced by "an old man".

Sub-step 2: Text-to-Image Generation. After preprocessing the story, we proceed to generate multiple images for each story sentence. Specifically, we leverage a state-of-the-art pretrained text-to-image generation model named Stable-Diffusion-v1-5, released by RunwayML and available in the Hugging Face model hub [63], because of its demonstrated capability to generate high-quality images relevant to the text [60]. The model outputs five images given a preprocessed input story sentence.

Sub-step 3: Postprocess generative images. With the candidate generative images, we further select and polish the most relevant image for each story sentence. The selection is based on the semantic similarity between the sentence and its candidate images. Specifically, we use a pretrained cross-modal model named CLIP [57] to encode the sentence and image into vectors and compute the cosine similarity of the image. After selecting the images (*e.g.,* A1-A4 in Figure 3) with the highest similarity scores with the story sentences, we seek to mitigate the potential inconsistencies among the selected images, *e.g.,* the same human character may be visually represented differently

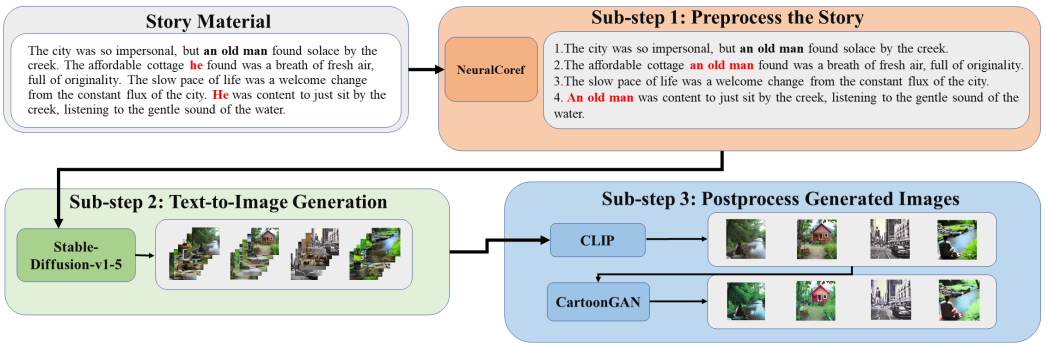

Fig. 2. Our computational workflow of generating relevant images for stories.

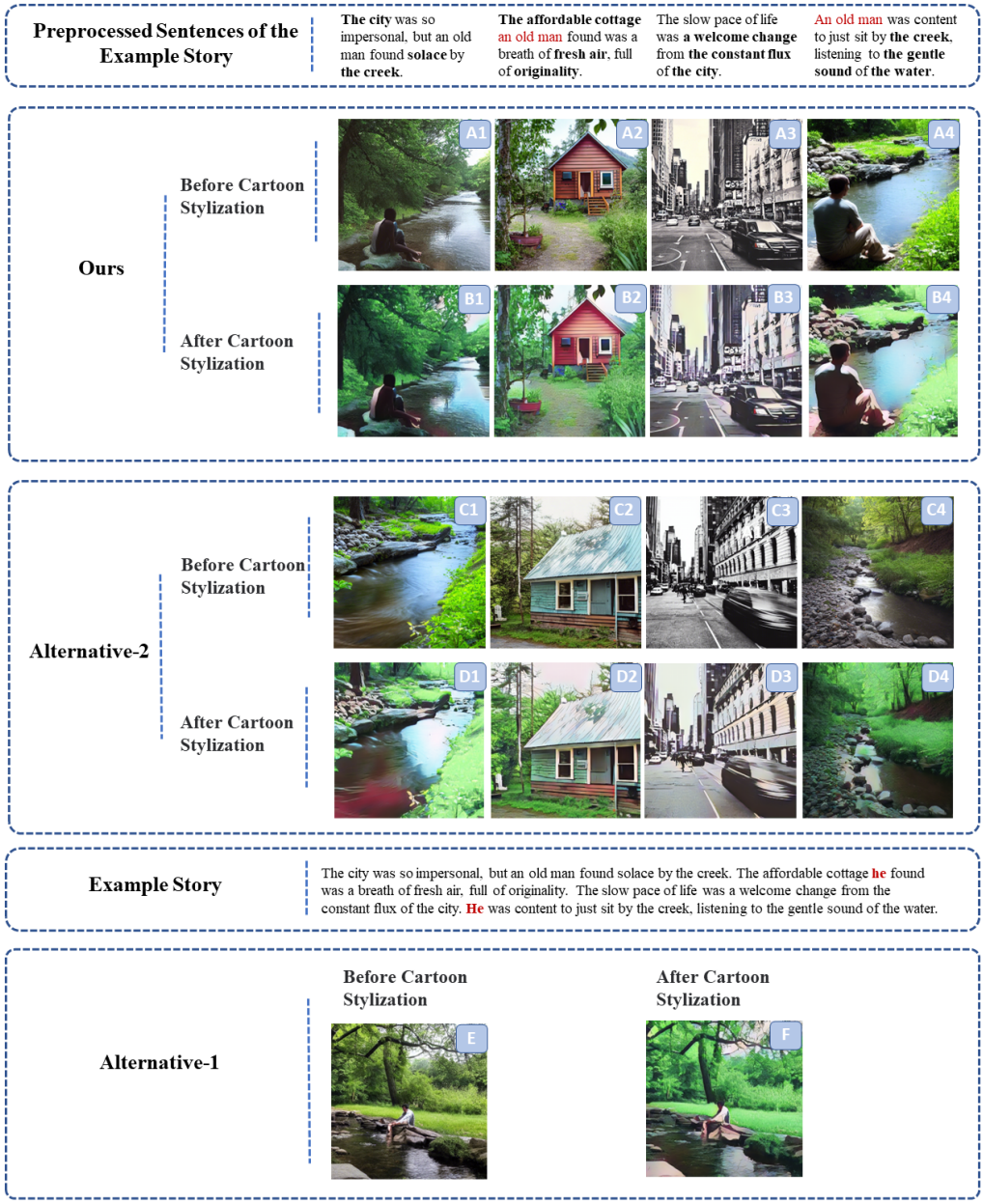

Fig. 3. Given sentences of an example story as input, we compare images generated by our computational workflow with those generated by two alternatives. [**Ours** (sentence-level, sentence-based)] A1-A4: Images generated using the preprocessed sentences as prompts. B1-B4: Cartoon stylization of A1-A4. [**Alternative-2** (sentence-level, keyword-based)] C1-C4: Images generated using the keywords (bold words in the preprocessed sentences of the example story) corresponding to the preprocessed sentences as prompts. D1-D4: Cartoon stylization of C1-C4. [**Alternative-1** (story-level)] E: Images generated using the entire story as a prompt. F: Cartoon stylization of E.

across images, such as variations in hair or facial details. We adopt a cartoon-style transfer model [11] that can convert each image to match a cartoon style while maintaining the original structures, textures, and basic colors of the image (*e.g.,* B1-B4 in Figure 3).

3.2 Evaluating the Feasibility of Generative Images for Story Retelling Support

At this stage, we would like to compare the quality of the images generated by our workflow with those generated by alternative approaches given the same short story. This evaluation aims at validating if the generative images are relevant to the story, have acceptable visual quality, and are perceived as helpful in helping learners comprehend and recall the story. We will assess the effectiveness and user experience of generative images in story retelling in the later experiments with vocabulary learners. Inspired by prior work on text-to-image generation [34], story-related images generation [37], and the usage of images in story retelling practices [16], we derive the following evaluation metrics: **relevance** (*The images are relevant to the story description*), **visual quality** (*The images are close to the real scene*), **perceived effectiveness in aiding comprehension** (*The images are helpful if you are going to do story comprehension*), and **perceived effectiveness in aiding recall** (*The images are helpful if you are going to do repeated retelling*). Each item is rated on a standard five-point Likert Scale (1 for "Strongly disagree" and 5 for "Strongly agree").

3.2.1 Alternative approaches. We compare our computational workflow with two alternative approaches for text-to-image generation. The first one, noted as **Alternative-1**, generates ten images by directly inputting the original story to the Stable-Diffusion-v1-5 model and then selects and stylizes the most relevant one (*e.g.,* F in Figure 3) similar to the sub-step 3 in our workflow. Alternative-1 produces a single image for the entire story refer to *CoSpeak* [1], which provides a single image to assist two English learners to co-create a story through dialogue. The comparison with Alternative-1 (*i.e.,* sentence-level vs. story-level) aims at checking if generating a series of sentence-level images could be more helpful than generating one story-level image. The second approach, noted as **Alternative-2**, uses TextRank to extract keywords (*e.g.,* the bold ones in Figure 3) as prompts to the Stable-Diffusion-v1-5 model to generate five images [38]. It then selects and polishes the most relevant image (D1-5 in Figure 3) for each sentence using the same postprocess methods in our proposed workflow. By comparing our workflow to Alternative-2 (*i.e.,* sentence-based vs. keyword-based), we aim to examine if the sentence-based prompt would be better than the keyword-based prompt, as a related work suggests that these two prompts were comparable in text-to-image generation tasks [38].

3.2.2 Preparing target word sets and short stories. We prepare 20 short stories, each containing a given target word set, to compare the images generated by our proposed workflow with those generated by alternative approaches. The target words are from the vocabulary pool (3,672 words in total) suggested by the International English Language Testing System (IELTS) [12]. Three authors of this paper randomly select non-easy IELTS words (*e.g.,* not the words like "easy" and "general") that they did not know before, which are randomly assigned to 20 sets, each with six or seven words. This manipulation simulates the case in which learners would like to learn any interested target word set via story retelling. To prepare a story for each target word set, we first query ChatGPT [5] with "generate a short story that has no more than 60 words and must contain the words '[word 1]', '[word 2]', ..., and '[word n]' ". This approach leverages the capability of the recent large language models to generate a short story that contains any target word set [55]. Compared to using existing short stories validated by English teachers, stories generated by ChatGPT can be flexibly adapted to learners' needs and interests on mastering any target words. The first author then refines the generated stories to improve their readability. Finally, we get 20 short stories (average word length:

60, average number of sentences: 5) that cover topics like funny animals, disasters, everyday life, and travel.

3.2.3 Procedure and Results.

Vs. Alternative-1. We prepare a document that lists the 20 stories; following each, there is a series of images generated by our workflow, an image generated by Alternative-1, and spaces for raters to input their scores for each metric. We distribute this document to five human raters (3 males, 2 females, age: $Mean = 20.6, SD = 0.49$) recruited from a local university. For each metric of the generative image(s) for a story, we average the scores of five raters as the final score. Next, we use paired-sample Wilcoxon signed rank tests to analyze the differences between our workflow and Alternative-1 on each metric. As depicted in Figure 4, our workflow performs significantly better in generating relevant image(s) to the story than the Alternative-1 ($p < 0.05, z = 2.023$, Cohen's d = 1.208). Raters also perceive that our generative images are significantly more effective in aiding comprehension ($p < 0.05, z = 2.023$, Cohen's d = 1.417) and recall of the story ($p < 0.05, z = 2.023$, Cohen's d = 1.537). These results indicate that generating a series of sentence-level images about a story could be more helpful in story retelling than generating one story-level image.

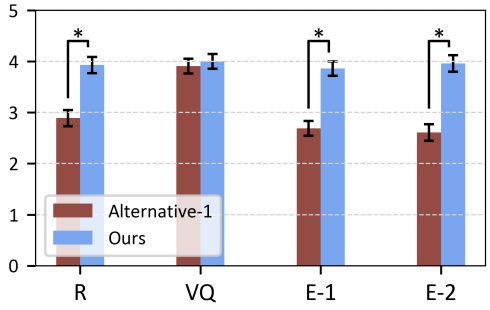

Fig. 4. Means and Standard Errors of human ratings on the quality of generative images; 1/5 - strongly disagree/agree; *: $p < .05$ using paired samples Wilcoxon signed rank tests. We compare **Alternative-1** (story-level) with **Ours** (sentence-level) on the images' relevance (R) to the story, visual quality (VQ), and effectiveness in aiding story comprehension (E-1) and recall (E-2).

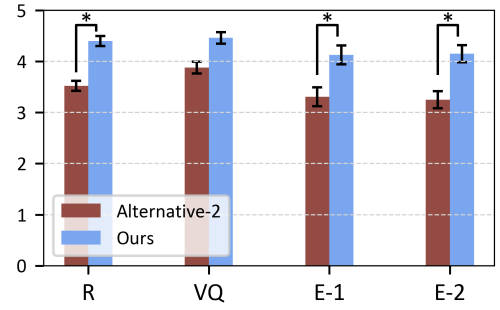

Fig. 5. Means and Standard Errors of human ratings on the quality of generative images; 1/5 - strongly disagree/agree; *: $p < .05$ using paired samples Wilcoxon signed rank tests. We compare **Alternative-2** (keyword-based) with **Ours** (sentence-based) on the images' relevance (R) to the story, visual quality (VQ), and effectiveness in aiding story comprehension (E-1) and recall (E-2).

Vs. Alternative-2. Similar to the procedure in comparing with Alternative-1, we recruit another five human raters (3 males, 2 females, age: $Mean = 20.4, SD = 0.27$) from the local university to score the images generated by our workflow and Alternative-2 on each metric and conduct paired-sample Wilcoxon signed rank tests. The order of encountering images of each story in the rating document is randomized and blind to the raters. As shown in Figure 5, compared with the Alternative-2, images generated by our workflow are significantly more relevant to the story ($p < 0.05, z = 2.023$, Cohen's d = 1.809) and are perceived significantly more effective in aiding comprehension ($p < 0.05, z = 2.032$, Cohen's d = 0.872) and recall ($p < 0.05, z = 2.032$, Cohen's d = 1.06) of the story. These results indicate that generating a series of sentence-level images about a

story using the sentence-based prompt could be more helpful in story retelling than generating the image series using the keyword-based prompt.

To sum up, the results of the evaluation study support our choices to generate sentence-level images using sentence-based prompts. The means of the four metrics on the images generated by our computational workflow are all larger than or equal to 4 out of 5 points, indicating its feasibility for generating images that are relevant to the story, of high visual quality, and potentially helpful to support story retelling. We then proceed to explore how our generative images can be used to support vocabulary learners in their story retelling practices.

3.3 Exploring Design Principles of *RetAssist*

With our computational workflow for text-to-image generation as the backbone of *RetAssist*, we work with vocabulary learners and teachers to derive design principles of *RetAssist*.

3.3.1 Process of exploring design principles. To put forward design principles on how to build a vocabulary learning system that uses generative images in story retelling practices, we first conduct a formative study with seven ESL (English-as-Second-Language) learners. Then, we develop a workable prototype of *RetAssist*. Next, we evaluate the *RetAssist* prototype through a within-subjects study with 18 ESL learners. According to user feedback on the prototype, we prepare a revision plan on *RetAssist* and solicit feedback from two English teachers.

Formative study with seven ESL learners. To understand user needs and requirements for a system that provides generative images in the story retelling practices, we conduct a formative study with seven ESL college students (S1-S7, 1 male, 6 females, age: $Mean = 20.57, SD = 0.82$) in China. Focusing on gathering the feedback and suggestions of ESL learners, we do not specifically balance the order of retelling with and without generative images in this instance. We first invite them to conduct one story retelling practice with generative images [3] and the other without images. Then, we ask questions about their perceptions of the practices and their expectations for a system using generative images to support story retelling. The findings here underpin DP1, DP2, DP4 and DP5 in Section 3.3.2.

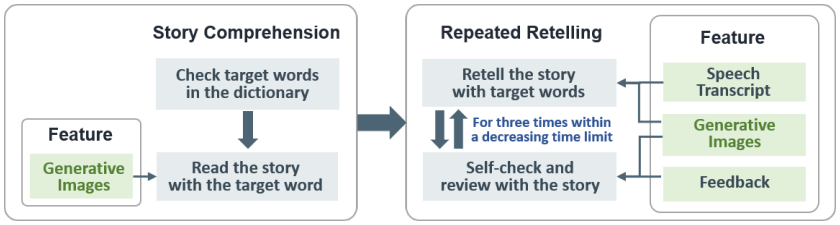

Fig. 6. The structured story retelling practice flow with the story comprehension and repeated retelling stages in *RetAssist* and baseline systems. In the evaluation of *RetAssist* prototype, the baseline system does not have features of speech transcript, generative images, and feedback. In the user study of the final version of *RetAssist*, the baseline system does not have the generative images but has other features like *RetAssist*.

Prototype of RetAssist. Based on the results of the formative study, we develop a workable prototype of *RetAssist*. This prototype structures the procedure of story retelling practice as used in the final version of *RetAssist* (Figure 6, detailed in Section 4) but has several features different from the final version of *RetAssist*. For example, the generative images are sequentially fixed in the interface and are not interactive. Inspired by the study of Gu *et al.* [24], this prototype will

[3]The stories and images are listed in a Word file and come from the materials used in the evaluation of our workflow in Section 3.2.2.

prompt the next sentence that masks the keyword when users get stuck for five seconds during the repeated retelling stage. Besides, after each round of repeated retelling, this prototype provides feedback about the incorrect use of target words and the associated sentence but does not provide the story and generative images for review before the next round of retelling. These features are discarded or refined in the final version of *RetAssist* based on the feedback from ESL learners and English teachers, as discussed in Section 3.3.2.

Evaluation of the RetAssist prototype. To probe user experience of the *RetAssist* prototype and feedback to improve it, we conduct a within-subjects study with 18 ESL learners (L1-L18, 14 females, 4 males, age: $Mean = 20.56, SD = 1.17$). The task and procedure are similar to the later user study of the final *RetAssist* (detailed in Section 5.3), except that we do not have the pretest and the two posttests in this study. During the within-subjects study, we get their qualitative feedback on how the features of the *RetAssist* prototype affect their learning process. We compare our *RetAssist* prototype and a baseline system without generative images, speech transcription, adaptive prompts, and feedback to explore the necessity of these system features. Consistent with the user study of final *RetAssist*, we counterbalance the order of the used systems and encountered word sets using Latin Square. After the learning sessions, we ask about their experience in the story retelling practices, their perception towards the two systems, and suggestions for improvement. The findings here underpin DP1 - DP5 in Section 3.3.2.

Feedback from two English teachers. Based on the user feedback, we prepare a revision plan in a PowerPoint file that draws possible designs for features about the interaction with generative images, prompts in the retelling stage, and feedback on user performance. We bring this plan and our *RetAssist* prototype to two English teachers (E1, female, age:27; E2, male, age: 27) and ask for their critiques and suggestions. The findings here underpin DP1 - DP5 in Section 3.3.2.

3.3.2 Design principles. We finalize five design principles (**DP**s) based on the results from the design process.

DP1: In the story comprehension stage, the system should provide generative images to facilitate users in understanding and remembering the storyline. Previous educational literature suggests that images depicting the story could help learners to understand stories efficiently [19, 49]. Our participants in the formative study favor the condition with images in their story retelling practices since the generative images can help them quickly and correctly understand the story. "The associated pictures with the story help me understand and remember the storyline, enabling more efficient story retelling practices" (S7). Also, all learners in the evaluation study of the *RetAssist* prototype expressed their favor for the generative images. "I like to incorporate images to understand the story" (L8). Both English teachers believe that generative images are practical materials to promote story comprehension.

DP2: During each round of the repeated retelling stage, the system should offer the generative images to help users recall the storyline yet not prompt the next sentence when users get stuck. As suggested by the Cognitive Theory of Multimedia Learning [53], visual elements (*e.g.,* figures) associated with the story can facilitate recall of words and their contextualized usage [19, 49]. The ESL learners participating in both the formative study and the evaluation study indicate that the images can assist them recall the story and organize their retelling flow. "I can easily connect the pictures back to the story plot when retelling the story" (S2). "I connect the images provided with the story, which helps me reflect on the storyline in a short time" (L3). Previous learning support system *EnglishBot* [61] offers Chinese prompts to users in their conversations with a chatbot, and our participants in the formative study raise similar expectations that the intended system could provide in-situ prompts about the story when they get stuck in the repeated retelling stage. "Rather than re-reading the full story, I'd like to get hints from the

system about what's next when I get stuck in the retelling" (S3). Nevertheless, as indicated by eight participants in the evaluation study of *RetAssist* prototype, the proactive sentence prompts during the repeated retelling stage often interrupt their retelling process and may result in their dependence on the prompts to finish the retelling. Additionally, E1 and E2 agree that generative images can promote users' recall in the retelling, while sentence prompts are not necessary or even unhelpful.

DP3: To help users align the images and story content, the system should enable the users to select and enlarge an image while highlighting the related story sentence. As one of the twelve multimedia instructional principles [41], the spatial contiguity principle suggests that users could be more focused on the learning tasks when related text and image are visually close to each other. With the *RetAssist* prototype, five learners in the evaluation study also suggest that the images should align with the story content in a more clear way. "I have to consciously remind myself to combine the images to understand the text. Showing all the images simultaneously and fixedly makes it difficult to focus on text and images at the same time" (L1). Our English teachers help us identify the proper design to visually align the images and story content. "Interaction design for displaying images should strike a balance between individual images and the overall narrative. We could use an image slider that helps learners focus on one image at a time while having an overview of the image sequence" (E1). "Highlighting the corresponding story sentence when the users enlarge one of the images could be an intuitive way" (E2).

DP4: In the story retelling stage, the system should provide a speech transcription function to record the users' retelling content and help them keep track of their progress. As a similar feature with previous retelling-based English learning systems like *CoSpeak* [1] and *EnglishBot* [61], our participants in the formative study express their wish to check what they just spoke in the retelling exercise. "I want to see what I have said so far when retelling, which can help me organize what I will say next" (S2). In general, all ESL learners in the prototype evaluation and both teachers favor the component of speech transcription. "With speech transcription, I could pay attention to the pronunciations when speaking" (L14).

DP5: After each round of repeated retelling, to help learners review their performance, the system should offer feedback on the incorrect usage of target words, together with the story and generative images. Providing feedback on users' task performance is a common and effective feature in learning support tools like *ArgueTutor* [69] and *EnglishBot* [61]. Five participants in the formative study suggest that they want to get feedback on their performance in practice, *e.g.,* about the correctness of words' expressions. "It will be better if the system could indicate whether I was using the target word correctly in my retelling practice, which can help me make progress in the next retelling" (S1). More importantly, eight participants in the evaluation study account the feedback from *RetAssist* prototype for their perceived improvement in the learning outcome. "Unlike the baseline system, *RetAssist* tells me how well I did in the last exercise, which helps me recheck the target words' meanings and make progress in the next round of retelling" (L11). E1 and E2 concur on the role of assessing the correctness of semantic usage through similarity measures and agree that it enables learners to verify the accuracy of their semantic expressions. However, in the evaluation study, seven learners suggest that *RetAssist* would better present the feedback together with the story and images, so that they can better review their performance in the current round of repeated retelling before proceeding to the next round. "I hope to review the story and images again before starting the next round of retelling since it helps me fill in some of the details for the retelling" (L13). E1 and E2 also agree that the review of the story and images between two rounds of repeated retelling is helpful.

4 *RETASSIST* SYSTEM DESIGN AND IMPLEMENTATION

Based on the identified design principles and proposed story text-to-image generation workflow in the last section, we develop *RetAssist* to facilitate vocabulary learners in story retelling practices. We develop *RetAssist* as a web app that can be easily accessed by learners on their computers. As shown in Figure 6, in the structured procedure of a repeated retelling practice, *RetAssist* provides users with generative images aligned to the story sentences (DP3) to assist story comprehension (DP1) and repeated retelling (DP2), speech transcription during repeated retelling (DP4), and adaptive feedback after each round of retelling practice (DP5). We describe how vocabulary learners can use *RetAssist* in a story retelling practice as follows.

Story comprehension. At the beginning of a story retelling practice, users need to first acquaint themselves with the target words' meanings and read the story that contains the target words (Figure 7-A). At this stage, they can look up the bilingual definitions and pronunciation of each target word in the left part of the interface (A1). They can read the story with the targeted words marked in bold (A2). Meanwhile, users can click the "Play" button to listen to the audio of the story and click the "Translation" icon to check its Chinese meanings (A2). Furthermore, users can click each image preview to switch the enlarged image (A3). Such a "sliding" interaction design with the images could help users focus on processing one image at a time and could be engaging [66]. Users can also see a highlighted sentence in the story (A2) that corresponding to the enlarged image. Such a design follows the spatial contiguity principle, which states that users can learn better when related text and image are close to each other [41].

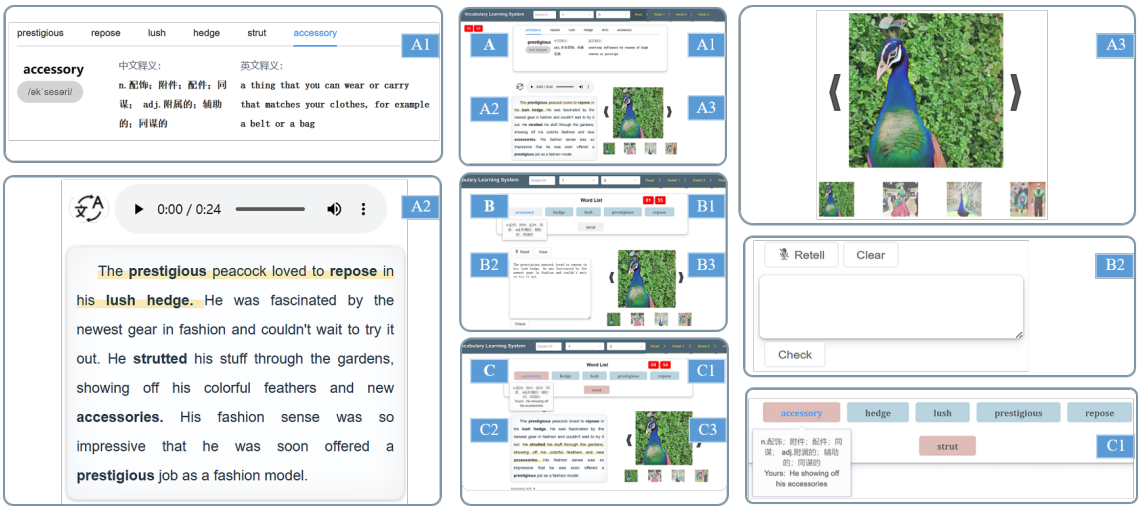

Fig. 7. User interface design of *RetAssist*. (A) In the story comprehension stage, users can 1) check the target words' meanings, 2) read the story, and 3) see the relevant images for each story sentence. (B) In the repeated retelling stage, users can retell the story with 1) the target words, 2) the retelling transcription, and 3) the generative images. (C) After each round of retelling, users can check feedback on their performance and review 1) the target words with incorrect marks, 2) the story, and 3) the generative images.

Repeated retelling. After comprehending the target words and associated story, users can click "Retell" in the upper menu bar to proceed to the repeated retelling stage (Figure 7-B). They need to complete three rounds of retelling practices within decreasing time limits, *e.g.*, 120, 90, and 60 seconds based on our trials in the formative study and evaluation study of *RetAssist* prototype. The

design of decreasing time limits in the learning practices could help users develop language fluency [46]. Users can click "Retell" to start each round of retelling trials (B2). As they speak, *RetAssist* will transcribe their speech in real-time using Chrome's speech recognition API [59]. During each round of repeated retelling, users can access the pronunciation and definition of target words in the word list any time they want (B1). The background color of the word will turn "blue" when *RetAssist* detects that the user speaks it. Meanwhile, users can switch the image slider and click each generative image to enlarge it whenever they want (B3). Users can stop the current round of story retelling by clicking "Retelling" again. Then, they can edit the transcribed sentences to correct speech recognition errors in the text box if they want (B2).

Review after each round of repeated retelling. When users finish one round of repeated retelling, they can click the "Check" button (Figure 7-B2) to view *RetAssist*'s feedback on the their performance and review the story material with generative images (C). Users can check which target words have been correctly contextualized (marked in blue in C1) and which words are not used or incorrectly used in the repeated retelling (marked in red). They can click each red word to view its meanings and the associated sentence that the user spoke. Users can also read the story with the highlighted sentences that contain the target words they incorrectly use (C2). Meanwhile, they can check the associated generative images (C3). Users can click the "Retell' button in the upper bar to start the next round of repeated retelling.

We use semantic similarity to judge whether the user correctly uses the target words, inspired by the study of Cao *et al.* [8], which verifies the correctness of machine translation by checking semantic similarity between the original and the translated sentences. Specifically, we consider a target word is not correctly used if the spoken sentence that should contain this word is semantically different from the original story sentence that contains this word [62]. We calculate the semantic similarity (ranging from 0 to 1) between the expression of each target word in the story and that in the users' retelling. First, we identify the sentence containing each target word in the user's retelling and calculate their sentence embeddings by Sentence-BERT [58]. Then, we compute the cosine similarity (ranging from 0 to 1) between this identified sentence and the corresponding sentence from the original story. If the user mentions the target word in multiple sentences, the similarity is recorded as the maximum of similarity between the multiple sentences and the corresponding sentence from the original story. If the user does not mention the target word, the similarity is recorded as 0. To decide the thresholds of similarity scores that differentiate the correct and incorrect use of target words, three authors mark the correctness (*i.e.,* correct or incorrect) of the word usage in each sentence of the recorded retelled content of the participants in our formative study. After marking, we obtain two sets of similarities separately representing the correct use of word meanings and the incorrect use of word meanings by calculating the semantic similarities between story sentences and spoken sentences. Finally, the threshold is determined to be 0.7 based on the ROC curve for different similarity scores [18].

5 USER STUDY

To evaluate how the generative images in *RetAssist* impact users' vocabulary learning outcome and experience in the story retelling practices, we conduct a within-subjects (*RetAssist* vs. baseline) study with 24 ESL (English-as-the-Second-Language) university students in China. Our research questions are:

RQ1. How would *RetAssist*'s generative images affect users' learning outcomes regarding the retention and verbal expression of target words in their story retelling practices?

RQ2. How would *RetAssist*'s generative images affect users' a) learning experience and b) behaviors in their story retelling practices?

RQ3. How would users perceive the usefulness of *RetAssist*'s generative images in their story retelling practices?

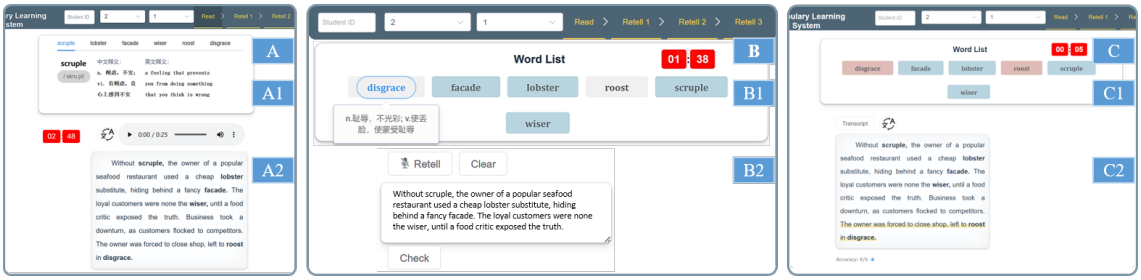

Fig. 8. User interface design of the baseline system. (A) In the story comprehension stage, users can 1) check the target words' meanings, and 2) read the story. (B) In the repeated retelling stage, users can retell the story with 1) the target words, and 2) the retelling transcription. (C) After each round of retelling, users can check feedback on their performance and review 1) the target words with incorrect marks, and 2) the story. The baseline system differs from *RetAssist* in that it does not provide generative images.

5.1 The Baseline System

The baseline system (Figure 8) supports the same user workflow (Figure 6) in story retelling practices as *RetAssist*. However, it does not offer generative images during both the story comprehension stage and the repeated retelling stage. The baseline system simulates the scenario in which the user is required to learn target words via story retelling practices without generative images. Specifically, the baseline system offers the word list (Figure 8-A1) and story (Figure 8-A2) in the story comprehension stage, and it provides decreasing time limits as well as word list (Figure 8-B1) and speech transcript (Figure 8-B2) in users' three rounds of retelling practices. Also, users can check adaptive feedback regarding the accuracy of the target word usage (Figure 8-C1) in such rounds and review the story text (Figure 8-C2). Such a baseline system satisfies all design principles without the involvement of generative images, specifically referring to DP4 and DP5. In summary, the only difference between *RetAssist* and the baseline system lies in the incorporation or exclusion of generative images, while all other functionalities are present in both conditions to meet users' demands.

5.2 Participants

We recruit 24 undergraduate students (P1-24, 15 females, 9 males, mean age: 20 (SD = 1.67)) from a university in mainland China via a post in the social media.They major in various domains such as Computer Science, Historiography, Philosophy, Physics, Finance, Literature, and International Relations. Twenty-three participants have passed the national English exam CET-4 in China, with an average score of 575 (SD = 48.04) [4]. Seventeen participants additionally have passed a higher-level national exam CET-6 in China (Mean score: 523 (SD = 48.47)). None of our participants have taken the IELTS exam. However, they exhibit a strong interest in learning their unknown IELTS vocabulary via the story telling practices (M = 5.75, SD = 1.05; 1 - not interested at all, 7 - very interested).

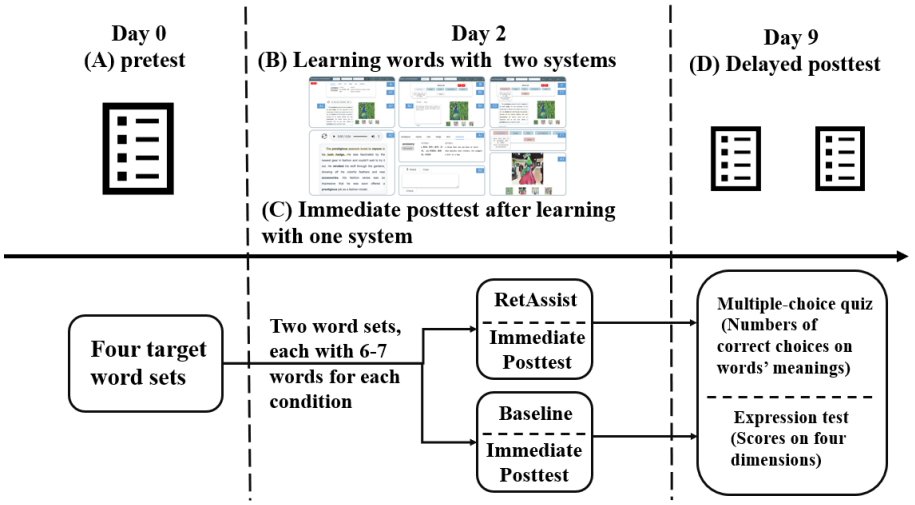

Fig. 9. Procedure of the within-subjects (*RetAssist* vs. the baseline system) user study. In each task, participants learn two sets of target words with either the *RetAssist* or the baseline system.

5.3 Procedure and Tasks

Figure 9 shows the procedure and task of our user study conducted remotely. Following [3, 55], on Day 0, participants fill in a consent form and a background survey and take a vocabulary pretest. We randomly select 4 stories from the 20 prepared stories mentioned in Section 3.2.2 as the learning materials for all participants, each containing six or seven target words. In total, the pretest consists of the 26 target words that participants will learn in our learning sessions. For each target word in the pretest, participants are required to select one option from five choices, including one that gives the correct meanings of the word in Chinese, three distractors, and an "I don't know" option. According to the results of the pretest, the average number of correctly chosen options among 24 participants is 8.62. In other words, on average, participants do not know the meanings of 17.38 words prior to the learning sessions. We inform them not to learn the words that appear in the pretest before the learning sessions.

On Day 2, participants first watch our pre-recorded video that describes the learning task and introduces the interfaces of *RetAssist* and baseline systems with blind names. They then use their laptops to log in to our systems. Each participant has two learning sessions. In each session, participants have two story retelling practices with either *RetAssist* or baseline system to learn two target word sets that they encountered in the pretest. Based on the pilot study with two participants, we allocate 30 minutes for each learning session. After each learning session, participants rate their engagement, enjoyment, task workload, and perceptions of the system in a questionnaire. In the questionnaire, we also ask them to write down responses to some short questions so as to make sense of the ratings. Additionally, they need to conduct an immediate posttest that examines their learning outcome on remembering the target words' meanings and being able to verbally use them to retell a story. Upon completion of two learning sessions, participants fill in a questionnaire that asks them to write down their preferences on the interfaces, comments on the generative images, and suggestions for improving *RetAssist*. We counterbalance the order of the used systems and word sets using Latin Square, *i.e.,* six participants experience "set 1 and 2 with *RetAssist* –> set 3

[4]710 is the full mark of both CET-4 and CET-6, and 425 is the minimum score to pass the exams.

and 4 with Baseline", six "set 1 and 2 with Baseline –> set 3 and 4 with *RetAssist*", six "set 3 and 4 with *RetAssist* –> set 1 and 2 with Baseline", and the rest six "set 3 and 4 with Baseline –> set 1 and 2 with *RetAssist*".

On Day 9, they conduct a delayed posttest that has the same format as the immediate posttest to examine their retention and verbal expression of target words learned on Day 2. The procedure on Day 2 and Day 9 is video- and audio-recorded for further data analyses. Overall, each participant spends approximately one hour and a half in our study and receives 80 RMB as compensation.

5.4 Measurements

5.4.1 RQ1. Learning outcomes. We measure participants' vocabulary learning outcomes through performance on an immediate posttest right after each learning session and a delayed posttest one week later. Specifically, both posttests include a multiple-choice quiz and an expressive test. The multiple-choice quiz is the same as the pretest that requires users to select one of five options that is the correct Chinese meaning of the target word. We calculate the number of correct answers to the multiple-choice questions to capture the learning outcome on the meanings of target words.

In the expression test, participants need to verbally retell each story in their learning sessions based on the story synopsis in Chinese and the target word set. As suggested by our two English teachers in the design process, we choose to present the synopsis instead of presenting nothing or providing the full Chinese translation of the original story to balance the difficulty of the expression test. We adapt the marking scheme of the IELTS speaking test [12] but have a focus on the verbal expressions of target words. With confirmation from our two English teachers in the design process, for each expression test of two stories within a learning session, we capture:

- **Number of target words used** (range 0 - 13 [5]).
- **Number of target words pronounced correctly**, *i.e.,* the number of target words correctly pronounced.
- **Number of target words used correctly**, *i.e.,* the number of target words that have been used semantically correctly.
- **Fluency**, which is determined by the expression of individual clauses and the lag between sentences, ranging from 0 to 9 on a scale referenced to the IELTS marking scheme.

Three authors of our research team first independently score six randomly selected audio samples, each consisting of two retelling stories in a learning session. They then meet and discuss together with one of our two English teachers (male, age: 29) to refine their rating scheme. For example, to focus on the usage and expression of target words, the rating scheme excludes factors like the participants' volume of voice, intonation, or accent. The three authors then apply the rating scheme to all 192 (24×2 systems $\times$ 2 stories per system $\times$ 2 posttests) audio samples in a shuffled order. For each dimension of the measured performance on the verbal expressions of target words, we average the three authors' scores (ICC = 0.939) as the final score in each retelling story. For each of the first three dimensions, we add the scores of two stories within one learning session as user performance in verbally expressing target words learned in that session, while for the last dimension of fluency, we average the scores of the two stories as the final score.

5.4.2 RQ2. Learning process. In each learning session with either *RetAssist* or baseline system, we measure participants' engagement and enjoyment in the learning process using items adapted from [69, 72]: "I was absorbed in using this interface to learn vocabulary" and "It is enjoyable to learn vocabulary with this interface". Besides, we measure the perceived task workload of learning

[5]In each learning session, participants learn vocabulary based on two stories. One contains 6 target words, and the other contains 7 target words. The maximum score for one learning session is therefore 6 + 7 = 13.

sessions using items adapted from NASA Task Load Index [25] (*e.g.,* "I require much mental and perceptual activity such as thinking and remembering in the process of the story retelling practice"). In addition to the questionnaire data, we also measure how learners perform in each of the three rounds of repeated retelling in each practice. For each round of repeated retelling, we measure: 1) spent time, *i.e.,* the time period between clicking the "Check" and the "Retell" button in this round of repeated retelling; 2) performance in practice, *i.e.,* how well users can retell the story content, reflected on the semantic similarity between learners' retold content and original story (ranging from 0 to 1, detailed in **Review after each round of repeated retelling** in Section 4). For each learning session with two story retelling practices, we average the spent time in two practices as the mean time spent in one round of repeated retelling in that session. Similarly, we average the semantic similarity scores of two practices to reveal user performance in one round of repeated retelling in each learning session.

5.4.3 RQ3. Perceptions towards RetAssist. For each system interface, we adapt the technology acceptance model [67, 70] to measure the perceived *usefulness* (four items, e.g., "I find the vocabulary learning support system useful in my vocabulary learning process by story retelling"; Cronbach's α = 0.921); *easiness to use* (four items, e.g., "My interaction with the vocabulary learning support system is clear and understandable"; α = 0.830); and *intention to use* (two items, e.g., "I intend to be a heavy user of the vocabulary learning support system when I want to learn vocabulary"; α = 0.901). We average the ratings of multiple questions as the final score for each aspect. All statements in the questionnaires are rated on a standard 7-point Likert Scale, with 1 - strongly disagree and 7 - strongly agree.

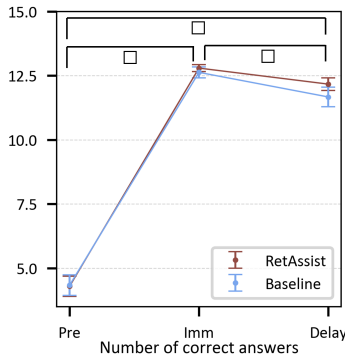

Fig. 10. RQ1 results regarding the number of correct choices on target words' meanings. □ : $p < .05$ for time factor (pretest vs. immediate posttest vs. delayed posttest) using repeated measures ANOVA with Bonferroni post-hoc test.

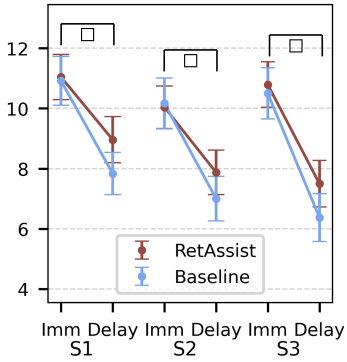

Fig. 11. RQ1 results regarding the number of target words used in expression (S1), the number of target words pronounced correctly in expression (S2), and the number of target words used correctly in expression (S3). □ : $p < .05$ for time factor (immediate posttest vs. delayed posttest) using repeated measures ANOVA.

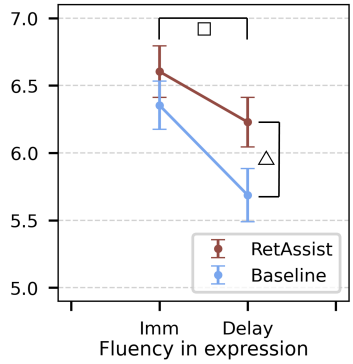

Fig. 12. RQ1 results regarding the fluency in expression. □ : $p < .05$ for time factor (immediate posttest vs. delayed posttest), △ : $p < .05$ for system factor (*RetAssist* vs. Baseline) using repeated measures ANOVA.

6 ANALYSES AND RESULTS

For the rated items, we first conduct a set of mixed ANOVA tests to check whether the order of system usage or the learned word sets associated with the systems affected our results (order and

word sets as between-subjects, systems as within-subjects). The results indicate that neither the main effects of the order and word sets nor their interaction with the systems are significant. For the measurements for RQ1, we perform two-way (time and system) repeated measures ANOVA to account for the dependencies in time. As for the measurements for RQ2 and RQ3, we perform Shapiro-Wilk normality tests before running all the paired samples t-tests. If the hypothesis that the data satisfies a normal distribution is rejected, we use paired samples Wilcoxon signed rank tests instead. As a result, for the spent time and performance in each round of repeated retelling, we perform paired-sample t-tests to compare the *RetAssist* and the baseline system. For the rest measures, we perform paired samples Wilcoxon signed rank tests. Additionally, two authors conduct open coding on participants' comments and suggestions on both vocabulary learning systems. They have multiple rounds of discussions and finally reach an agreement on the codes, which are incorporated into the following result presentation.

6.1 Learning Outcomes (RQ1)

6.1.1 Multiple-choice Quiz. As shown in Figure 10, participants demonstrate comparable performance with *RetAssist* ($M = 12.792, SD = 0.644$) and baseline system ($M = 12.625, SD = 1.033$) regarding the number of correct answers to the multiple-choice questions in the immediate posttest. In the delayed posttest, participants have better performance on average with *RetAssist* ($M = 12.167, SD = 1.179$) than baseline system ($M = 11.667, SD = 1.863$) regarding the number of correct answers. The results of repeated measures ANOVA indicate that neither the system factor (*RetAssist* and Baseline) nor its interaction with the time factor (pretest vs. immediate posttest vs. delayed posttest) significantly affects participants' performance in the multiple-choice quiz ($p > 0.05$). However, the time factor has significant effects on participants' performance in the multiple-choice quiz ($p < 0.001, F = 596.792, \eta^2 = 0.912$), and the results of the Bonferroni post-hoc test ensure the significant difference among the three quizzes (pretest vs. immediate posttest: $p < 0.001$; pretest vs. delayed posttest: $p < 0.001$; immediate posttest vs. delayed posttest: $p < 0.05$).

6.1.2 Expression Test. As shown in Figure 11, participants generally perform well in using the target words, pronouncing them correctly, and using them correctly in the immediate expression posttest after the learning session with either *RetAssist* or baseline system; $M > 10$ and $p > 0.05$ in all the three dimensions. This finding suggests that the story retelling practice, either with or without the involvement of generative images, is an effective approach to learning the verbal expression of target words in the short term. In the delayed posttest after one week of the learning sessions, the user performance with both systems naturally decreases compared to that in the immediate posttest. We find that participants are able to use more target words learned with *RetAssist* ($M = 8.96, SD = 3.77$ and use them correctly ($M = 7.5, SD = 3.77$) compared to the baseline system (use target words: $M = 7.83, SD = 3.45$, use them correctly: $M = 6.375, SD = 3.89$) in average. The average number of correctly pronounced target words is also higher in the learning session with *RetAssist* ($M = 7.875, SD = 3.61$) than that in the session with baseline system ($M = 6.375, SD = 3.89$). With the repeated measures ANOVA, we find that neither the system factor (*RetAssist* and Baseline) nor its interaction with the time factor (immediate posttest vs. delayed posttest) significantly affects the number of target words used (S1), the number of target words pronounced correctly (s2), and the number of target words used correctly (S3) in expression ($p > 0.05$). Among the expression measurements of S1 - S3, the time factor has significant effects (S1: $p < 0.001, F = 49.127, \eta^2 = 0.699$; S2: $p < 0.001, F = 43.381, \eta^2 = 0.712$; S3: $p < 0.001, F = 76.741, \eta^2 = 0.827$).

As for the fluency of participants' spoken English in the immediate posttest (Figure 12), participants can tell the story significantly more fluently after the learning session with *RetAssist*

($M = 6.604, SD = 0.935$) that that with the baseline system ($M = 6.354, SD = 0.872$). Similarly, in the delayed posttest, participants' spoken English in telling the story with target words is significantly more fluent after learning with *RetAssist* ($M = 6.229, SD = 0.901$) compared to the baseline system ($M = 5.688, SD = 0.966$). The results of repeated measures ANOVA indicate that both the system factor (*RetAssist* and Baseline) and the time factor (immediate posttest vs. delayed posttest) significantly affect the fluency of participants' spoken English (system factor: $p < 0.05, F = 4.01, \eta^2 = 0.041$; time factor: $p < 0.001, F = 21.552, \eta^2 = 0.238$). These results suggest that ***RetAssist*'s generative images can significantly improve the learners' fluency in using target words to tell a story after the story retelling practices compared to the baseline system**.

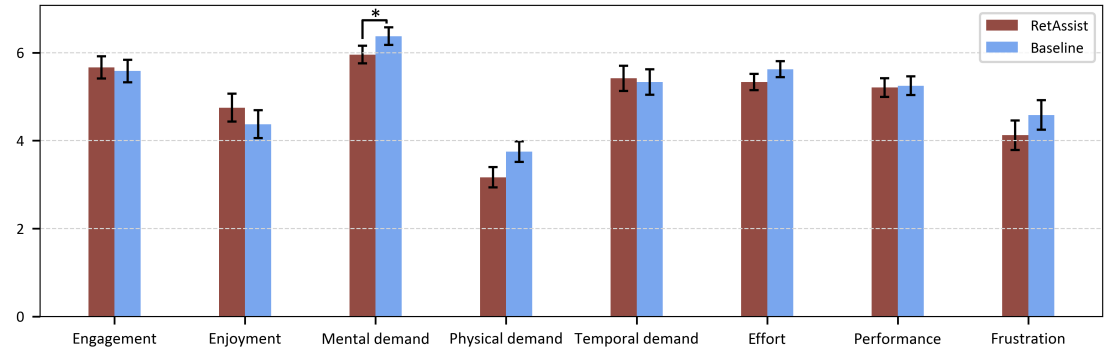

Fig. 13. RQ2 results regarding engagement, enjoyment, and workload in vocabulary learning sessions. $* : p < 0.05$ using paired samples Wilcoxon signed rank tests.

6.2 Learning Process (RQ2)

6.2.1 Engagement, enjoyment and workload. As shown in Figure 13, participants report a slight increase in engagement and enjoyment during the vocabulary learning process with *RetAssist* compared to the baseline system, though the difference was not statistically significant. However, 22 (out of 24) participants commend the quality of the images and feel that the images are closely aligned with the text. "The pictures are appealing, and I can interact with them by switching the picture and checking its related sentence" (P12). Furthermore, participants report a lower level of mental demand ($p < 0.05, z = 2.066$, Cohen's d = 0.455) during the vocabulary learning process with *RetAssist* than that with the baseline system. 21 participants perceive that practicing story retelling with the baseline system is notably more challenging, as they need to mentally visualize and construct the scene of the story. "Recalling the story's details and scenarios (with the baseline system) takes up a lot of my mental effort. In contrast, *RetAssist* helps me to recall the story in a visual way" (P6). Five participants further report that the baseline system is monotonous compared to *RetAssist*. "I do not like the (baseline) interface as it is monotonous and inflexible" (P3).

6.2.2 Performance in each round of repeated retelling. Figure 14 shows the spent time in each round of repeated retelling with *RetAssist* and the baseline system. Participants spend less time with *RetAssist* in the second ($M(SD) : 113(45.07)$ vs. $137(61.02); p < 0.05, t = -2.227$, Cohen's d = 0.321) and third ($110(47.23)$ vs. $132(55.28); p < 0.05, t = -2.18$, Cohen's d = 0.315) rounds of repeated retelling compared to the cases with the baseline system. "While using the baseline system, I frequently run out of the limited time before finishing retelling the story; however,

when using *RetAssist*, I am more comfortable in the repeated retelling stage and can complete the retelling on time" (P20). Meanwhile, as shown in Figure 15, the semantic similarity between users' retelling content and original story significantly increases during the second ($p < 0.05, t = 2.397$, Cohen's d = 0.346) and the third rounds ($p < 0.001, t = 3.793$, Cohen's d = 0.547) of repeated retelling. Nineteen participants attribute their improvement during the practice to the provided images in *RetAssist*. "Images provided in *RetAssist* help me better connect my native language expression and English expression of the target words, which helps me reflect on the details and storyline in a short time" (P4). **These results indicate that compared to the baseline system, *RetAssist* can reduce learners' workload and improve their efficiency and performance of each round of repeated retelling during the story retelling practices.**

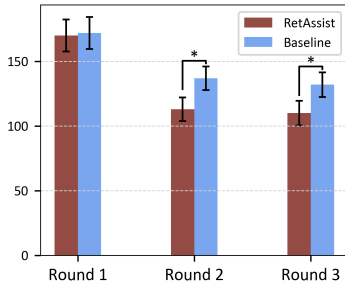

Fig. 14. RQ2 results regarding time spent by users in three rounds of retelling. $* : p < 0.05$ using paired-sample t-tests.

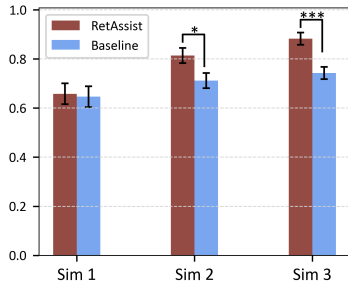

Fig. 15. RQ2 results regarding the semantic similarity between users' retelling content and story. $* : p < 0.05, * * * : p < 0.001$ using paired-sample t-tests.

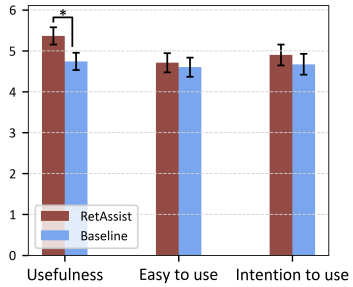

Fig. 16. RQ3 results regarding user perceptions of each interface. $* : p < 0.05$ using paired samples Wilcoxon signed rank tests.

6.3 Perceptions towards the Systems (RQ3)

As shown in Figure 16, participants feel that our *RetAssist* ($M = 5.365, SD = 1.031$) is significantly more useful than the baseline system ($M = 4.74, SD = 0.996$; $p < 0.05, z = 2.29$, Cohen's d = 0.604). Nineteen participants implied that the images in *RetAssist* are the reason for rating it more useful. "Without the images, I find it difficult to go through the repeated retelling stage. The images are especially useful when I am stuck" (P13). There is no significant difference between *RetAssist* ($M = 4.708, SD = 1.156$) and the baseline system ($M = 4.604, SD = 1.141$) regarding easiness of use. We have comments from twenty-one participants that praise the interaction design of *RetAssist*. "The interface of *RetAssist* is intuitive, and the interaction flow is clear. I can listen to the story while easily reading the story with aligned images" (P11). Lastly, participants generally have a higher intention to use *RetAssist* ($M = 4.9, SD = 1.249$) for vocabulary learning in the future compared to baseline system ($M = 4.6, SD = 1.337$). Twenty participants comment that they prefer the *RetAssist* for future vocabulary learning. "With *RetAssist*, I can express the learned words more correctly with less pressure. I want to have it as my weekly used vocabulary learning system" (P16). However, four participants prefer the baseline system, because they feel it is time-consuming to view the images and mentally connect them with the story.

7 DISCUSSION

In this work, we develop the *RetAssist* system that aims to facilitate vocabulary learners in their story retelling practices. Its core features are the generative images relevant to the story in the

story comprehension and repeated retelling stages. Our study shows that participants using either *RetAssist* or the baseline system can master the meanings and expressions of the target words right after a story retelling practice, supporting that story retelling is an effective approach to vocabulary learning [16, 23, 42, 44]. However, one week after the practices, participants better recall and verbally express the target words learned with *RetAssist* than those with the baseline system. This proves the value of our generative images for supporting vocabulary learning and provides empirical evidence for the benefits of visual aids for language learning stated in the Cognitive Theory of Multimedia Learning [53].

7.1 Design Considerations

Based on our findings, we provide three design considerations for story-based vocabulary learning tools.

Provide more types of visual aids. Participants generally favor *RetAssist*'s images for helping them comprehend and recall stories. However, three participants comment that they still have difficulty in recalling the expression of target words with the generative images and suggest that it would be better to visualize stories through mind maps or flow charts [45, 56]. Moreover, participants expect *RetAssist* to incorporate short videos [4] or motion graphics [36] into vocabulary learning. "Understanding the images themselves is an additional burden for me. I would prefer a more explainable form of visual aids to help me understand some abstract storylines in the story comprehension stage" (P18). We, therefore, suggest that the generative technique could offer other forms of visual aids such as an extracted mind map and a relevant video clip, and allow users to customize them based on their interests.

Offer prompts that are adaptive to users' performance. *RetAssist* currently provides fixed image prompts and word prompts. However, two participants suggest that they need more personalized and interactive prompts. For example, the system can "recognize my stuckness and give me corresponding hints based on the progress of my current retelling." To provide timely and personalized support during each round of repeated retelling, it would require future researchers to label a set of story retelling audio clips for training a model to predict users' difficult timing based on their tone, speed, and pauses in the current practice.

Provide suggestions to improve. The feedback offered by *RetAssist* includes the correctness of target words' semantic usage as well as highlighting the incorrectly used target words and the corresponding sentences. Four participants expect that it can also explicitly tell them how to improve in the next round of practice. For instance, the system can "correct mispronunciations of words, list the target word's grammatical usage and provide additional example sentences" to enhance the comprehension of the target vocabulary. We suggest that future vocabulary learning tools should offer not only feedback on what and why a target word is misused but also suggestions on how to deepen the understanding of this word, *e.g.,* with more example sentences.

7.2 Broader Impact to Generative AIs for Education

Our design and development of *RetAssist* offers a feasible example of leveraging generative AIs to support learning tasks. First, generative models can offer meaningful and flexible learning materials, *e.g.,* ChatGPT [5] that generates a story given any target words in our case. It is also promising to apply these models to prepare listening materials [59] and provide contextually personalized learning materials [15] for language learners. Second, generative models can support additional modalities of learning activities used in traditional instruction on a large scale. In addition to serving as visual aids as in our case, text-to-image AI can be integrated into 3D Design Workflow to produce reference images, prevent design fixation, and inspire design considerations [39]. Also,

they can empower a conversational agent, which acts like a lecturer, to socially converse with the learners to practice their spoken language [61] on any topic.

However, utilizing generative content as learning materials may have the potential to hinder learning gains in certain scenarios. One concern is the risk of generative AIs in terms of accuracy and reliability. Learners need to take precautions against generating errors or false information when adopting generated content as learning material, and the generative content may be one-sided and outdated because of the limitations of the training data for generative AIs [40]. Another concern is that the assistance of generative AIs may discourage users from putting in enough effort in the learning process. For instance, Peng *et al.* 's study suggests that learners could experience reduced gains in vocabulary acquisition when engaged in writing exercises with generative AIs compared to those without AI assistance [55]. This could be attributed to the fact that participants invested less time in writing and wrote significantly fewer words in the story, as it indicates a preference for dependence on the generative model for assistance [55]. To mitigate these potential negative impacts, we suggest that the developers of learning support systems should examine the quality of generated content beforehand and work with targeted learners and educators to identify proper design principles (Figure 1).

Although *RetAssist* is initially designed for independent learning outside of the classroom, it can be useful in diverse educational settings beyond individual study. For example, teachers in traditional classrooms can use *RetAssist* to enrich vocabulary instruction around story reading or story retelling. In addition to assisting ESL learners in vocabulary acquisition using story retelling, the system's story text-to-image generation workflow is expected to be useful in general education scenarios that combine images with stories. For example, our workflow can generate sentence-level illustrations for children's storybooks to help them better understand the meaning of textual descriptions. In addition, for cultivating children's expressive language skills, our story text-to-image generation workflow can be used as an interactive and creative way for teachers or parents to practice expressive language in children's education. By retelling stories with their illustrations, children can develop the ability to clearly organize their verbal expressions, make associations between visual materials and textual materials, and creatively conceptualize the plot with the illustration details. Despite the enormous potential of our proposed system and workflow in education, we must approach potential risks cautiously to ensure that they bring positive and sustainable impacts. We must ensure that the generative images and stories conform to widely accepted educational standards and ethical norms to avoid conveying incorrect information or inappropriate content.

7.3 Limitations and Future Work

Our study has several limitations that urge future work. First, as our primary focus is on vocabulary learning support, we did not examine *RetAssist*'s impact on learners' story retelling skills. Learners may have overreliance on generative images in the story retelling practices, while in the English exams that test story retelling performance, they would not have such assistance. Future work can extend *RetAssist* for training story retelling skills. Second, we evaluate *RetAssist* with twenty-four English-as-second-language undergraduates learning IELTS words, who could not represent all target user groups. We would like to extend the study to include learners of different age groups or proficiency levels in our future work, and we also encourage future researchers to customize our system and evaluate it to support vocabulary learners of different ages, expertise, and cultures (e.g., middle or high-school students and English students learning Chinese). Third, we conducted a short-term user study that can reveal *RetAssist*'s user experience and effectiveness in our proposed learning tasks. To examine its usage in the wild, we need a long-term field study in which users can specify any target words and take story retelling practices at any time they want. Fourth,

we design our computational workflow of generating multiple image prompts relevant to each story. In our formative study, we indicate that the stories are short ones with approximately 60 words. However, this study design may not apply to all user groups. As the story gets longer, our computational workflow will generate more sentence-level images that may decrease the coherence among images and increase users' cognitive workload to process them in the story retelling practices. To alleviate this cognitive load, future work could consider ways to generate images based on the semantic segments of the story (*i.e.,* one or multiple sentences that describe one image). Fifth, future design iterations of *RetAssist* could incorporate more advanced AI features like adaptive learning algorithms that tailor image selection or presentation based on individual learner performance. Sixth, our study exclusively utilized generative images as visual aids, yet alternative media formats might yield different outcomes. In order to understand the affordances of static or dynamic images for learning, we will consider comparing the efficacy of generative images with other media types (*e.g.,* videos or interactive graphics) in our future work.

8 CONCLUSION

In this paper, based on educational literature and working with teachers as well as ESL learners, we iteratively design and develop an interactive system, *RetAssist*, to facilitate vocabulary learners in story retelling practices. *RetAssist* equips our proposed computational workflow that generates images relevant to the story to foster users' understanding and recall of the story that contains a set of target words. We conduct a within-subjects study with 24 participants in comparison to the baseline system without generative images. Our results show that learning with *RetAssist* leads to significantly better learning outcomes on mastering meanings and expressions of target words than learning with the baseline system. Our work demonstrates the feasibility and effectiveness of generative models to support language learning tasks and offers implications for future learning support tools.

ACKNOWLEDGMENTS

This work is supported by the Young Scientists Fund of the National Natural Science Foundation of China (NSFC) with Grant No.: 62202509 and NSFC Grant No.: U22B2060. Also, this work is supported in part by HKUST 30 for 30 with Grant No.: 3030_003. We are grateful to the anonymous reviewers for their insightful suggestions.

REFERENCES

[1] Richa Agrawal and Ravi Poovaiah. 2021. CoSpeak: Peer Feedback on Voice Stories to Inform Learning Spoken English. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '21)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3462204.3481750

[2] Munassir Alhamami. 2016. Vocabulary learning through audios, images, and videos: Linking technologies with memory. *Call-Ej* 17, 2 (2016), 87–112.

[3] Riku Arakawa, Hiromu Yakura, and Sosuke Kobayashi. 2022. VocabEncounter: NMT-powered Vocabulary Learning by Presenting Computer-Generated Usages of Foreign Words into Users' Daily Lives. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21.

[4] Betül Bal-Gezegin. 2014. An investigation of using video vs. audio for teaching vocabulary. *Procedia-Social and Behavioral Sciences* 143 (2014), 450–457.

[5] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[6] Frank Boers. 2014. A reappraisal of the 4/3/2 activity. *RELC Journal* 45, 3 (2014), 221–235.

[7] Arthur Caetano, Alyssa Lawson, Yimeng Liu, and Misha Sra. 2023. ARLang: An outdoor augmented reality application for portuguese vocabulary learning. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1224–1235.

[8] Jialun Cao, Meiziniu Li, Yeting Li, Ming Wen, Shing-Chi Cheung, and Haiming Chen. 2022. SemMT: a semantic-based testing approach for machine translation systems. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31, 2 (2022), 1–36.

[9] Chih-Ming Chen and Ching-Ju Chung. 2008. Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle. *Computers & Education* 51, 2 (2008), 624–645.

[10] Shizhe Chen, Bei Liu, Jianlong Fu, Ruihua Song, Qin Jin, Pingping Lin, Xiaoyu Qi, Chunting Wang, and Jin Zhou. 2019. Neural storyboard artist: Visualizing stories with coherent image sequences. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2236–2244.

[11] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. 2018. Cartoongan: Generative adversarial networks for photo cartoonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9465–9474.

[12] Pauline Cullen. 2012. *Cambridge Vocabulary for IELTS Advanced Band 6.5+ with Answers and Audio CD*. Vol. 6. Cambridge University Press.

[13] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 8780–8794.

[14] H Douglas and LEE BROWN. 2001. *Teaching by principles: An interactive approach to language pedagogy*. P ED AUSTRALIA.

[15] Fiona Draxler, Albrecht Schmidt, and Lewis L Chuang. 2023. Relevance, Effort, and Perceived Quality: Language Learners' Experiences with AI-Generated Contextually Personalized Learning Material. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2249–2262.

[16] Carl J Dunst, Andrew Simkus, and Deborah W Hamby. 2012. Children's story retelling as a literacy and language enhancement strategy. *Center for Early Literacy Learning* 5, 2 (2012), 1–14.

[17] Ferda Gül Aydın Emekligil and İlkay Öksüz. 2022. Game character generation with generative adversarial networks. In *2022 30th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 1–4.

[18] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[19] Diamanto Filippatou and Peter D Pumfrey. 1996. Pictures, titles, reading accuracy and reading comprehension: a research review (1973-95). *Educational Research* 38, 3 (1996), 259–291.

[20] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 289–293.

[21] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

[22] Mohammad Reza Ghorbani. 2014. Story Retelling and the EFL Vocabulary Learning Process. *The Iranian EFL Journal* 11 (2014), 398.

[23] Akimi Gibson, Judith Gold, and Charissa Sgouros. 2003. The power of story retelling. *The tutor* (2003), 1–11.

[24] Yongqi Gu and Robert Keith Johnson. 1996. Vocabulary learning strategies and language learning outcomes. *Language learning* 46, 4 (1996), 643–679.

[25] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[26] Mohammad Nehal Hasnine, Masatoshi Ishikawa, Yuki Hirai, Haruko Miyakoda, and Keiichi Kaneko. 2017. An algorithm to evaluate appropriateness of still images for learning concrete nouns of a new foreign language. *IEICE TRANSACTIONS on Information and Systems* 100, 9 (2017), 2156–2164.

[27] Mohammad Nehal Hasnine, Kousuke Mouri, Brendan Flanagan, Gokhan Akcapinar, Noriko Uosaki, and Hiroaki Ogata. 2018. Image recommendation for informal vocabulary learning in a context-aware learning environment. In *Proceedings of the 26th International Conference on Computer in Education*. Asia-Pacific Society for Computers in Education Philippines, Asia, 669–674.

[28] Mohammad Nehal Hasnine and Junji Wu. 2021. Wordhyve: A context-aware language learning app for vocabulary enhancement through images and learning contexts. *Procedia Computer Science* 192 (2021), 3432–3439.

[29] Ari Hautasaari, Takeo Hamada, Kuntaro Ishiyama, and Shogo Fukushima. 2020. VocaBura: A Method for Supporting Second Language Vocabulary Learning While Walking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 135 (sep 2020), 23 pages. https://doi.org/10.1145/3369824

[30] Shinichi Izumi. 2002. Output, input enhancement, and the noticing hypothesis: An experimental study on ESL relativization. *Studies in second language acquisition* 24, 4 (2002), 541–577.

[31] Slava Kalyuga, Paul Chandler, and John Sweller. 1999. Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 13, 4 (1999), 351–371.

[32] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. 2021. Variational diffusion models. *Advances in neural information processing systems* 34 (2021), 21696–21707.

[33] Walter Kintsch and Eileen Kintsch. 2005. Comprehension. In *Children's reading comprehension and assessment*. Routledge, 89–110.

[34] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Text-to-image generation grounded by fine-grained user attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 237–246.

[35] Raziye Kütük. 2007. The Effect of Mnemonic Vocabulary Learning Strategy and Story Telling on Young Learners' Vocabulary Learning and Retention. *Unpublished MA Thesis)(1-101)* (2007).

[36] Vahid Norouzi Larsari and Radka Wildová. 2020. The psychological effect of motion info graphics on reading ability of primary school students. (2020).

[37] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6329–6338.

[38] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.

[39] Vivian Liu, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2023. 3DALL-E: Integrating text-to-image AI in 3D design workflows. In *Proceedings of the 2023 ACM designing interactive systems conference*. 1955–1977.

[40] Chung Kwan Lo. 2023. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences* 13, 4 (2023). https://doi.org/10.3390/educsci13040410

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

[42] Donna DiSegna Merritt and Betty Z Liles. 1989. Narrative analysis: Clinical applications of story generation and story retelling. *Journal of Speech and Hearing Disorders* 54, 3 (1989), 438–447.

[43] Sara Miller and Lisa Pennycuff. 2008. The power of story: Using storytelling to improve literacy learning. *Journal of Cross-Disciplinary Perspectives in Education* 1, 1 (2008), 36–43.

[44] Lesley Mandel Morrow. 1985. Retelling stories: A strategy for improving young children's comprehension, concept of story structure, and oral language complexity. *The Elementary School Journal* 85, 5 (1985), 647–661.

[45] Shobana Musti, Jesslyn M Smith, and John C Begeny. 2022. A Virtual Tutoring Program to Increase Students' Text Reading Fluency. *Intervention in School and Clinic* (2022), 10534512221140474.

[46] Paul Nation. 2007. The four strands. *International Journal of Innovation in Language Learning and Teaching* 1, 1 (2007), 2–13.

[47] Chi-Duc Nguyen and Frank Boers. 2019. The effect of content retelling on vocabulary uptake from a TED talk. *Tesol Quarterly* 53, 1 (2019), 5–29.

[48] Aurélien Nioche, Pierre-Alexandre Murena, Carlos de la Torre-Ortiz, and Antti Oulasvirta. 2021. Improving artificial teachers by considering how people learn and forget. In *26th International Conference on Intelligent User Interfaces*. 445–453.

[49] Putu Santi Oktarina, Ni Putu Lila Sri Hari, and Ni Made Winda Ambarwati. 2020. The effectiveness of using picture book to motivate students especially young learners in reading. *Yavana Bhasha: Journal of English Language Education* 1, 1 (2020), 72–79.

[50] Rebecca L Oxford and Robin C Scarcella. 1994. Second language vocabulary learning among adults: State of the art in vocabulary instruction. *System* 22, 2 (1994), 231–243.

[51] Allan Paivio. 1990. *Mental representations: A dual coding approach*. Oxford university press.

[52] Allan Paivio. 2014. Bilingual dual coding theory and memory. *Foundations of bilingual memory* (2014), 41–62.

[53] Allan Paivio and Alain Desrochers. 1980. A dual-coding approach to bilingual memory. *Canadian Journal of Psychology/Revue canadienne de psychologie* 34, 4 (1980), 388.

[54] Hilal Peker, Michele Regalla, and Thomas Dwight Cox. 2018. Teaching and learning vocabulary in context: Examining engagement in three prekindergarten French classrooms. *Foreign Language Annals* 51 (2018), 472–483. https://api.semanticscholar.org/CorpusID:149920259

[55] Zhenhui Peng, Xingbo Wang, Qiushi Han, Junkai Zhu, Xiaojuan Ma, and Huamin Qu. 2023. Storyfier: Exploring Vocabulary Learning Support with Text Generation Models. arXiv:2308.03864 [cs.HC]

[56] Sasitorn Praneetponkrang and Malinee Phaiboonnugulkij. 2014. The use of retelling stories technique in developing English speaking ability of grade 9 students. *Advances in Language and Literary Studies* 5, 5 (2014), 141–154.

[57] 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.

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

[59] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. FastSpeech: Fast, Robust and Controllable Text to Speech. arXiv:1905.09263 [cs.CL]

[60] 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 (CVPR)*. 10684–10695.

[61] Sherry Ruan, Liwei Jiang, Qianyao Xu, Zhiyuan Liu, Glenn M Davis, Emma Brunskill, and James A Landay. 2021. Englishbot: An ai-powered conversational system for second language learning. In *26th international conference on intelligent user interfaces*. 434–444.

[62] Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM* 8, 10 (1965), 627–633.

[63] RunwayML. 2021. Stable-diffusion-v1-5. https://huggingface.co/runwayml/stable-diffusion-v1-5. Accessed on March 25, 2023.

[64] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

[65] Stephen D Sorden. 2012. The cognitive theory of multimedia learning. *Handbook of educational theories* 1, 2012 (2012), 1–22.

[66] S Shyam Sundar, Saraswathi Bellur, Jeeyun Oh, Qian Xu, and Haiyan Jia. 2014. User experience of on-screen interaction techniques: An experimental investigation of clicking, sliding, zooming, hovering, dragging, and flipping. *Human–Computer Interaction* 29, 2 (2014), 109–152.

[67] Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences* 39, 2 (2008), 273–315.

[68] Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

[69] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[70] Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. AL: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[71] Thomas Wolf, James Ravenscroft, Julien Chaumond, and Maxwell Rebo. 2018. Neuralcoref: Coreference resolution in spacy with neural networks.

[72] Ziming Wu, Yulun Jiang, Yiding Liu, and Xiaojuan Ma. 2020. Predicting and diagnosing user engagement with mobile ui animation via a data-driven approach. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[73] Liren Zeng and Ling Lin. 2011. An interactive vocabulary learning system based on word frequency lists and Ebbinghaus' curve of forgetting. In *2011 Workshop on Digital Media and Digital Content Management*. IEEE, 313–317.

[74] Yeshuang Zhu, Yuntao Wang, Chun Yu, Shaoyun Shi, Yankai Zhang, Shuang He, Peijun Zhao, Xiaojuan Ma, and Yuanchun Shi. 2017. ViVo: Video-Augmented Dictionary for Vocabulary Learning. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5568–5579. https://doi.org/10.1145/3025453.3025779

Received February 2024; revised April 2024; accepted May 2024