Verification and Evaluation of Eye Tracking and Memory in Generated Images with Emphasized Words for Vocabulary Learning

2410030 Ochiai Takuto

Recent advances in generative artificial intelligence, especially deep learning–based text-to-image generation, have made it possible to create highly realistic images that do not necessarily exist in the real world. Such generated images have attracted attention not only in entertainment and advertising but also in educational contexts, where they may serve as visual materials that support comprehension and memory. In vocabulary learning, the combination of verbal information and visual information has long been regarded as effective, because images can complement linguistic input and help learners build stronger associations between words and meanings. However, most conventional vocabulary learning support systems employ images that correspond to correct answers and present typical semantic representations of target words. By contrast, relatively little research has examined how images generated from learners' incorrect answers influence attention, comprehension, and memory.

This study focuses on error-emphasized generated images used for vocabulary learning. These images are generated from English sentences that include an incorrect word selected by the learner, and they are intended not to present the correct answer directly but to visualize the learner's error in a way that promotes reflection and re-interpretation. The images used in this study were generated through the L-VEIGe (Learning Vocabulary Error Image Generation) framework proposed in prior work. L-VEIGe is designed to support vocabulary learning by presenting images that reflect learners' incorrect answers and by controlling the degree to which the error is represented visually. Although previous studies have suggested that such images may help reduce repeated errors, the cognitive processes underlying this effect have not been sufficiently clarified. In particular, it remains unclear how learners visually explore these images and how such gaze behavior relates to the later recall and retention of the incorrect word and its contextual information.

The purpose of this study is to investigate how error-emphasized generated images affect learners' eye movements and memory retention in vocabulary learning. More specifically, the study examines whether differences in the saliency of generated images influence learners' visual exploration behavior, immediate recall, delayed recall, and subjective impressions. Unlike approaches that attempt to measure whether learners looked directly at a predefined error region, this study emphasizes how learners explore the overall visual context that contains the error. This perspective is based on the assumption that understanding an error may require not only looking at a specific object but also integrating surrounding contextual information represented in the generated image.

To analyze this issue, an eye-tracking experiment was conducted with university students who were native speakers of Japanese and learners of English as a second

language. Sixteen participants took part in the experiment. The stimulus set consisted of ten English fill-in-the-blank sentences, each associated with an incorrect word that had been assumed or selected in advance. For each sentence, sixteen error-emphasized generated images were prepared by manipulating four saliency-related parameters at high and low levels: error concept consistency, contextual consistency, surface ratio of the emphasized concept, and color difference of the emphasized concept. These four parameters correspond to two higher-order dimensions. The first is recallability, which concerns how accurately the image represents the incorrect concept and its context. The second is impressiveness, which concerns how strongly the emphasized concept stands out visually. By combining the high and low values of these four parameters, sixteen image conditions were created for each sentence, resulting in a total of 160 generated images.

In the experiment, each participant viewed ten images, one for each sentence, so that across sixteen participants all sixteen conditions were covered for each sentence. The order of image presentation was randomized across participants in order to reduce sequence effects. Before each image was displayed, the participant read the corresponding sentence containing the incorrect word and advanced to the image presentation by pressing a key. Each image was then displayed for ten seconds against a black background, with the square stimulus centered on the screen. Eye movements were recorded during image presentation using a Tobii Pro Fusion 250 eye tracker. After each trial, participants answered a short impression questionnaire concerning image readability and the ease with which they could notice where the emphasized incorrect word was reflected in the image. After the viewing phase, participants completed an immediate free-recall test in which they were asked to recall the emphasized word and contextual information from the image. A delayed recall test using the same procedure was conducted one week later to evaluate memory retention.

The analysis focused on gaze exploration over the entire image rather than direct fixation on a predefined semantic target. To achieve this, the displayed image area was divided into equal-sized Areas of Interest (AOIs). The main analysis used a 3×3 AOI grid, while a supplementary analysis used a 4×4 AOI grid in order to examine robustness and reduce the effect of central bias. Gaze samples that fell outside the image area, such as those on the black background, were excluded. Based on these AOIs, several indices were calculated, including AOI distribution entropy and average pupil diameter. AOI distribution entropy was used as a measure of how broadly or narrowly gaze was distributed across the image. A larger entropy value indicates a more exploratory viewing pattern in which attention is spread over multiple regions, whereas a smaller value indicates more concentrated viewing. Average pupil diameter was treated as a supplementary measure related to cognitive effort or processing load.

The results were organized from two perspectives: recallability and impressiveness. From the recallability perspective, images with high error concept consistency and high contextual consistency tended to yield more distributed gaze behavior, as reflected in higher AOI entropy. In particular, the condition in which both parameters were high showed the largest entropy values among the recallability conditions, suggesting that learners explored the whole image more actively when the emphasized concept and its

surrounding context were represented coherently. These conditions also tended to produce higher immediate recall, delayed recall, and retention rates. In addition, average pupil diameter was larger in the condition where both recallability-related parameters were high, and significant differences were observed between this condition and several other recallability conditions. This result suggests that learners may have engaged in deeper cognitive processing when the image provided both a clear representation of the incorrect concept and a meaningful contextual structure.

From the impressiveness perspective, images with a higher surface ratio of the emphasized concept and greater color difference generally led to stronger subjective impressions and better noticeability of the emphasized element. AOI entropy in these conditions also tended to be higher, indicating that visual emphasis did not merely attract gaze to one local point but could encourage broader exploration of the image. Moreover, immediate and delayed recall tended to be relatively higher for highly impressive conditions, suggesting that visually salient emphasis may function as a memory cue. At the same time, average pupil diameter also tended to be larger in conditions with stronger impressiveness, implying that visual emphasis may increase cognitive effort as learners attempt to interpret the emphasized content.

Taken together, the results suggest that error-emphasized generated images can support vocabulary learning not simply by drawing attention to a wrong element, but by encouraging learners to explore the broader visual context in which the error is embedded. Recallability appears to contribute mainly to semantic understanding and memory retention by making the incorrect concept and its context interpretable, whereas impressiveness appears to contribute mainly to noticeability and subjective salience by making the emphasized element easier to detect. The findings therefore indicate that effective image design for vocabulary learning should not rely solely on visual emphasis; rather, it should balance semantic interpretability and perceptual salience.

This study has several limitations. First, the number of participants was limited, and some results were interpreted as tendencies rather than conclusive effects. Second, AOI-based analysis provides a spatial approximation of gaze behavior but does not necessarily correspond to semantic units in the image. Third, pupil diameter is sensitive to multiple factors, including lighting, blinking, and missing data, and should therefore be interpreted cautiously. Future work should increase the sample size, conduct more comprehensive statistical testing, refine semantic region analysis, and examine interaction effects between recallability and impressiveness in greater detail.

Despite these limitations, this study contributes foundational knowledge about how learners visually and cognitively respond to error-emphasized generated images in vocabulary learning. By integrating generated images, eye-tracking analysis, impression evaluation, and memory tests, the study provides empirical evidence that the design of error-emphasized visual materials can influence not only what learners notice but also how they explore, interpret, and remember incorrect vocabulary. These findings offer a useful basis for the future design of adaptive vocabulary learning systems that employ generative images as reflective and cognitively effective educational materials..