npj | heritage science

Article

https://doi.org/10.1038/s40494-025-01621-1

CATS: cultural-heritage classification using LLMs and distribute model

Check for updates

Hyerin Hwang[1], Chan-Woo Park[1,2], Hee-Kwon Kim[1,2] & Jae-Ho Lee[1]

In this paper, we introduce ***CATS (Cultural-heritage Advanced Translation Systems)***, a generative model applied to creating images for classification and exploring relationships among cultural heritage. We aimed to address the issue where large language models (LLMs) fail to generate appropriate sentences due to the limited training on classical Korean language, and the problem where text-to-image models trained on Korean language do not produce accurate sentences when using Korean words as they are. To solve this problem, a large language model was used to translate historical content containing classical Korean words into English sentences, which were then used as input for the text-to-image generation model. We found that the generation model using the translated English text produced more accurate and consistent images compared to the model using the original Korean text. Consequently, this approach offers highly convenient visual information for users and administrators at a low cost through the use of open-source models. Therefore, we propose the potential of a system that leverages generated images to facilitate the search and extraction of relevant information.

Recent rapid advancements in artificial intelligence and machine learning technologies have dramatically improved the performance of language and image generative models. Particularly, generative models have greatly expanded their utility by learning from vast amounts of digital data and transforming it into actionable forms. These models play a crucial role not only in classifying data but also in understanding data distributions beyond specific labels. In cases where information is extremely scarce making classification through text alone difficult, generative models can provide visual materials as outputs. Furthermore, in scenarios where textual information is absent entirely, automatic classification can still be achieved solely through images. We attempted to measure the correlation and similarity between cultural heritages through images using the latest technique, Stable Diffusion. However, we encountered a problem where Stable Diffusion struggles to recognize the classical language of Korean. Despite the Stable Diffusion being fine-tuned in Korean, it has difficulty understanding and processing classical Korean sentences properly. This issue complicates the analysis of the correlation between cultural heritages through generated images using Stable Diffusion. This is because the model needs to generate images similar to actual cultural heritages for it to learn effectively. The process described is illustrated in Fig. 1, which depicts the workflow from classical Korean text input to image generation. Interestingly, we found that when we translated classical Korean sentences into English using LLM and input them into the Stable Diffusion, the model generated images similar to the original. This suggests that the Stable Diffusion model has a higher recognition rate for English sentences. Therefore, it implies that even for classical Korean sentences, we can visualize the meaning of the original similarly through English translation.

The application of generative models can be extended to systems managing extensive digital data sets such as cultural heritage. Currently, efforts are underway to digitize accumulated cultural heritage, necessitating systematic classification and management systems due to the vast and varied information these heritages contain. Depending on the characteristics and significance of each heritage, management approaches vary, from basic information as shown in Fig. 2a to more detailed descriptions as seen in (b). Even for cultural heritages with different amounts of information, a system that applies various modern techniques is needed to automate the classification of cultural heritages.

Figure 3 is an archive-based visualization of associations. Currently, data is manually organized and organized without the help of AI which needs to be improved to organize a large amount of digital cultural heritage. This process is complex and time-consuming and has limitations in accuracy and efficiency. Therefore, it is essential to build an automated system using existing organized information for an accurate classification system.

As a result, cultural heritage data created by making the most of text information can be used in various ways. An example of 3D modeling is shown in Fig. 4, which converts visual data into 3D. Digitized cultural heritage data can be used to arrange exhibit spaces in museums analyze data

[1]Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea. [2]These authors contributed equally: Chan-Woo Park, Hee-Kwon Kim.
e-mail: jhlee3@etri.re.kr

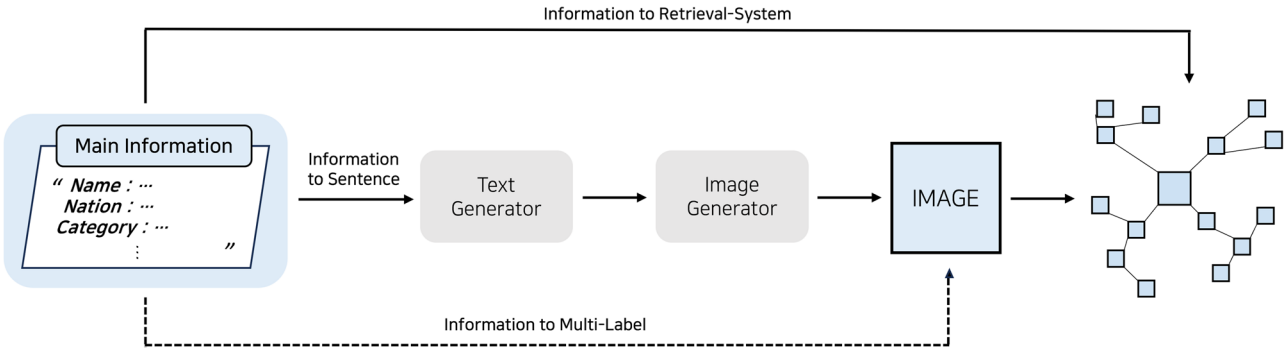


Fig. 1 | Workflow for Image Generation from Classical Korean Text.

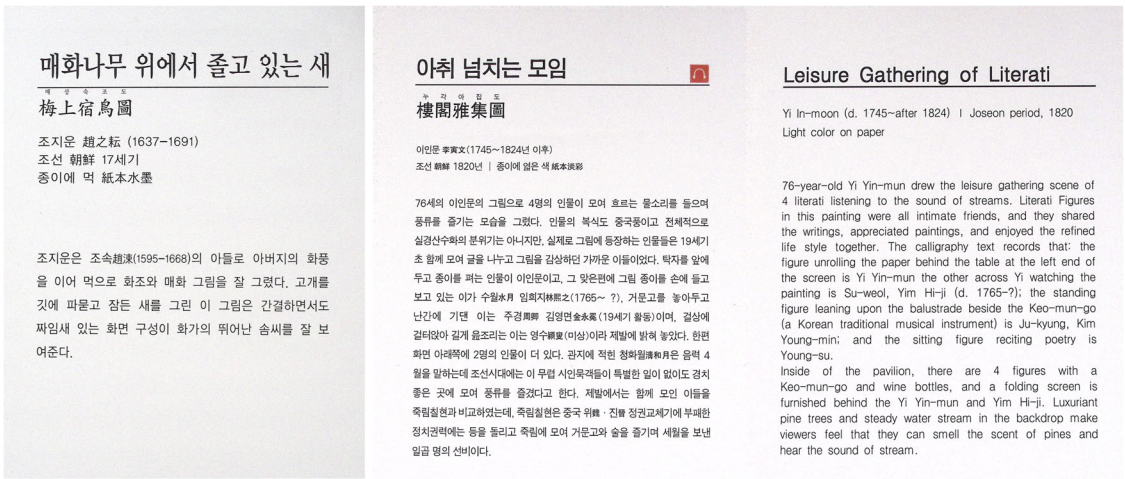


Fig. 2 | Information panel[56] comparison. a Basic information panel. A photo of an information panel displaying essential details, including the title *"Bird Resting on a Plum Tree"*, artist Jo Ji-un (1637–1691), period Joseon, 17th century, and medium Ink on paper. The panel also provides a brief description: *Jo Ji-un was the son of Jo Sok (1595–1668) and inherited his father's painting style, excelling in painting flowers, birds, and plum blossoms. This painting, depicting a bird dozing while perched on a branch, demonstrates a composition that, despite its simplicity, conveys a refined sense of balance and artistic sophistication, highlighting the painter's exceptional brush-work.* **b** Detailed information panel (Korean-English bilingual). A photo of an information panel containing a more detailed description of the cultural heritage artifact. The left side presents the original Korean text, while the right side provides its English translation.

using 3D materials, and create visual information for users. Therefore, this paper provides the following contributions:

1. We present a method for generating and using visuals using generative models in the digitization process of cultural heritage. We utilize Stable Diffusion, a generative model, to visually represent and reconstruct cultural heritage textual data. This generates augmented data that can be used to improve the performance of learning models for association analysis.

2. We propose a method for automatically classifying and managing large-scale cultural heritage data by applying Multi-label Classification to a cultural heritage management system. It combines class names and image features to distinguish various attributes of cultural heritage, reducing the cost and time of manual classification and improving management efficiency.

3. We use a large-scale language model to supplement the missing textual information of cultural heritage with a large-scale language model for cultural heritage that lacks textual information. This solves the problem of lack of textual data and adds diversity to visualization generation.

Related work

Tasks with Cultural Heritage

Existing cultural heritage data suffers from unstructured applications and insufficient information, which is a challenge for various institutions managing cultural heritage. Recent research utilizing cultural heritage data has been solving various problems based on deep learning algorithms. To address this[1], proposed a method to accelerate the process of recognizing historical architectural elements in detailed features and developing historical architectural information modeling (HBIM)[2] from data using semantic segmentation[3] of 3D point clouds[4]. DeFi[5] acquires 3D data through 3D sensors or multi-view reconstruction for the preservation of digital cultural heritage and proposes a method for learning hole boundary detection by generating synthetic datasets to form accurate 3D point clouds. ConvSRGAN[6] proposes a method for the restoration and enhancement of traditional Chinese paintings by applying Enhanced High-Frequency Retention Modules which utilize residual blocks and a high-frequency emphasis loss function to maintain high-frequency components and restore detailed features in the images. MROPM[7] uses photo-realistic style transfer (PST) to style an old photo using multiple references and then enhances the result to create a modern image. ArchGPT[8] proposes a method that focuses on customized tasks for the repair and preservation of traditional buildings by utilizing large-scale language models (LLMs). Specifically, it proposes to apply RAG (Retrieval Augmented Generation) technology to analyze the damage status of buildings through image recognition, information retrieval, and image rendering, and to derive appropriate restoration methods based on the results. While various studies have been conducted to apply these deep learning techniques to the cultural heritage field, no specific

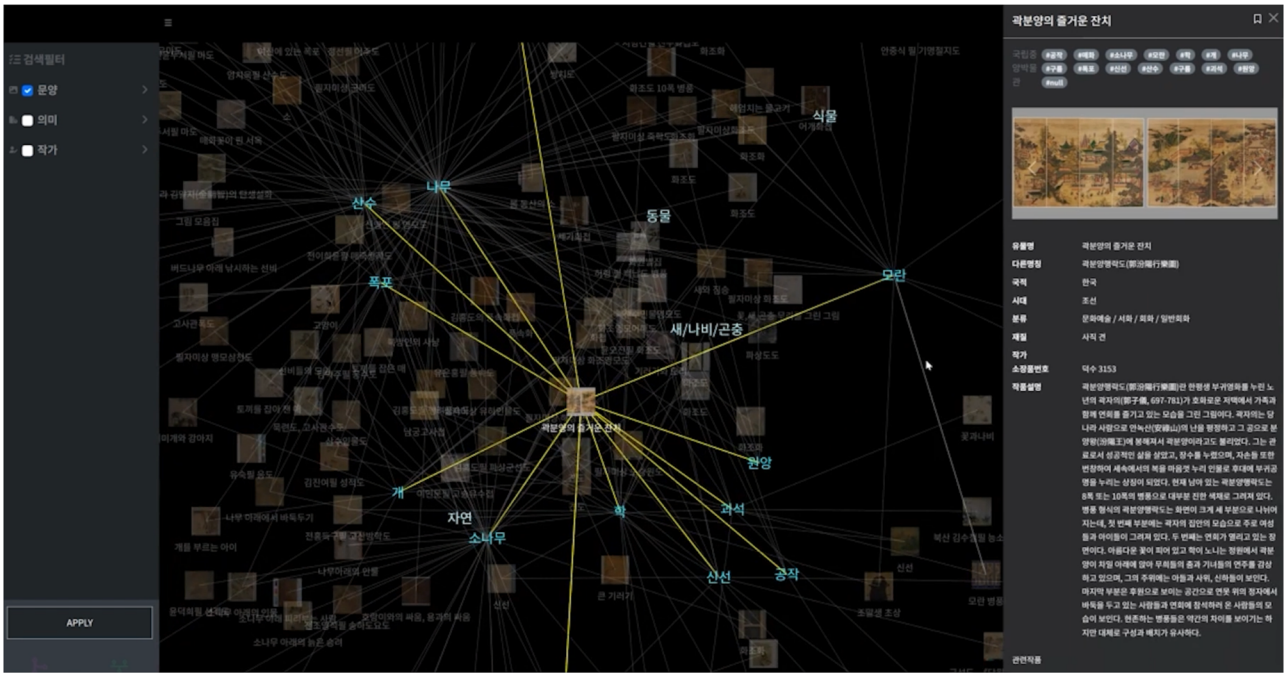
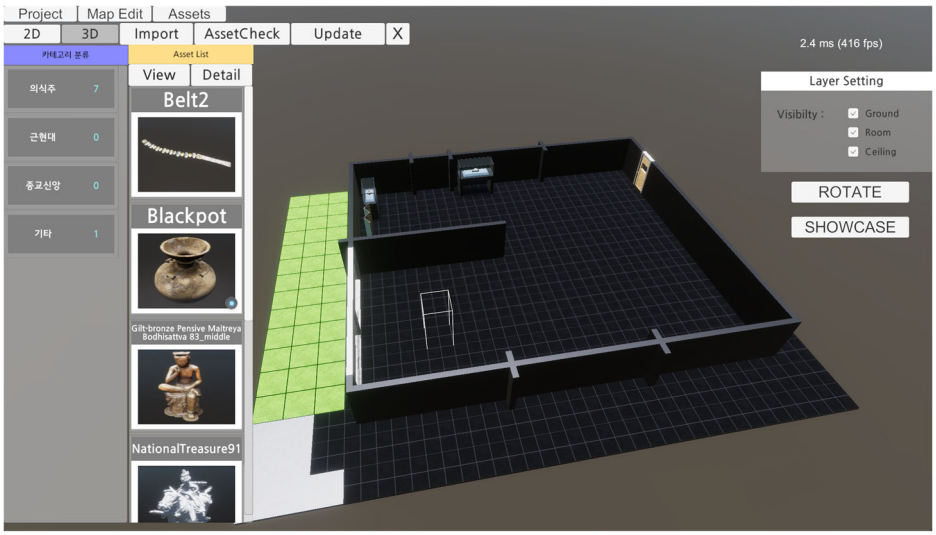


Fig. 3 | Archive-based Visualization.

Fig. 4 | Example of use when cultural heritage is processed into 3D images.



results have been reported using the text-to-image generative model Stable Diffusion[9]. Therefore, we demonstrate how Stable Diffusion can be used to process cultural heritage data and extract meaningful patterns and relationships from it.

Large Language Models

LLM models play an important role in processing and generating textual data. First, the field of natural language processing can be categorized into three main types. They are encoder type, decoder type, and encoder-decoder type based on the transformer structure. The first model of the encoder type Bidirectional Encoder Representations from Transformers (BERT)[10] was proposed to predict masked tokens in text and to determine the probability of one text passage following another. It includes methods for masked language modeling (MLM)[11] and next sentence prediction (NSP). The innovation of BERT has influenced the development of many other models one of which is the Robustly optimized BERT approach RoBERTa[12]. RoBERTa builds on BERT emphasizing that modifying the pre-training method by training for longer in larger batches with more training data leads to better performance. It significantly improved performance over the original BERT model by not including the NSP task. DistillBERT[13] proposed a faster and less memory-intensive method by applying knowledge distillation to improve the drawback that BERT is difficult to deploy in environments that require low latency due to the size of the model. ALBERT[14] made three major modifications to the encoder structure to make it more efficient. They reduced the embedding dimensionality by separating the token embedding dimension from the hidden dimension, reduced the actual number of parameters by having all layers share the same parameters, and changed the NSP goal to sentence order prediction. ELECTRA[15] was proposed to address the constraint that the standard MLM pre-training necklace only updates the masked token representation at each step, while the other input tokens are not updated. It uses two models, one to predict the masked tokens and the other to predict through a discriminator, to improve

training efficiency. DeBERTa[16] proposed that by separating token content and relative position, the self-attention layer better models the dependence of pairs of adjacent tokens. By adding absolute position embeddings just before the softmax layer of the token decoding head, it became the first model to outperform humans on the SuperGLUE[17] benchmark. The next type of decoder is represented by GPT models. Decoder models excel at predicting the next word in a sentence and can therefore be used for most text-generation tasks. GPT was trained to predict the next word based on the previous word by combining a transformer decoder architecture with transfer learning. Its successor, GPT-2[18], is a model built by extending the training set of the original model, and is excellent at creating coherent long sequences of text. CTRL[19] is an improvement on GPT-2 in that it offers little control over the style of the sequences it generates. By adding a control token at the beginning of the sequence, we can control the style of the generated sentence, allowing us to generate different sentences. GPT-3[20], which grew out of the success of GPT-2, analyzed the behavior of different language models to discover rules for computation, dataset size, model size, etc. This model not only produces highly realistic text passages, but also performs well in supervised learning. The last type of transformer architecture, the encoder-decoder type, is represented by T5[21]. T5 unifies all the work of NLU[22] and NLG[23] in a text-to-text conversion. It uses the original architecture of the transformer and has many versions with different parameters. BART combines pre-trained models from BERT and GPT in an encoder-decoder architecture. The input sequence is transformed and passed through an encoder, and the decoder reconstructs the original text.

LLaMA[24], a recent high-performing model, follows the same technique used in GPT-3 by normalizing the inputs of each Transformer sublayer to increase the stability of the training. Performance is optimized by removing absolute position embeddings and instead applying Rotary Positional Embeddings (RoPE)[25]. Alpaca[26] is a finetuned version of LLaMA that can generate large instruction datasets by applying self-Instruct[27], a method that helps LLMs improve their ability to follow human instructions. We show that the combination of a pre-trained LLM can create a small instruction-following model that can match the performance of LLMs at a low cost. In this paper, we chose the model of GPT 3.5, which is both search-based and conversational, to create appropriate prompts for cultural heritage metadata. We found that it is easy to refine the desired prompts and provides a highly accurate English description of the original prompts.

Diffusion Models

Generative models are techniques that sample a dataset into a trained model to create new realistic images that do not exist in the original dataset. Among the methods of generative models, VAE[28], a structure of autoencoders[29] that utilizes latent variables, has emerged. It was proposed as a mapping to a multivariate normal distribution around a point in latent space. The later GAN[30] uses a constructor that serves the same purpose as the decoder in a VAE, converting vectors in the latent space into an image. The discriminator predicts whether the image is real or fake and produces an output through a convolutional layer. The idea is that the constructor converts random noise into samples that appear to be sampled from the original dataset, and the discriminator predicts whether they are from the original dataset or a forgery of the constructor. Rather than training directly on full-resolution images, ProGAN[31] trains the generator and discriminator on low-resolution images of 4x4 pixels and then incrementally adds more to the training process to increase the resolution. StyleGAN[32], which is based on ProGAN, injects style vectors into the neural network at various points to improve the difficulty of distinguishing latent space vectors corresponding to high-level attributes. StyleGAN2[33] used a method of removing the AdaIN[34] layer of the constructor and replacing it with a weighted modulation and demodulation step to improve the quality of the generated output. The paper showed how to eliminate artifact problems while maintaining control over the style of the image. SAGAN[35] applied the self-attention mechanism of transformers to solve the problem that convolutional feature maps only process local information. BigGAN[36] proposed a method that uses the same distribution of latent vectors for training but uses a truncated normal distribution for sampling to increase the reliability of the generated samples. VQ-GAN[37] creates a codebook, a list of trained vectors, and uses the codebook vectors associated with the corresponding indexes. ViT VQ-GAN[38], an extension of VQ-GAN, proposed a neural network structure applied to image data by replacing the convolutional encoder and decoder with transformers. The image is divided into a series of patches, which are tokenized and used as input, and the resulting embedding is quantized according to a learned codebook. The decoder-transformer is then processed with an integer code and the image is formed from the sequence of patches.

Generative models evolved with the introduction of the diffusion model, DDPM[39]. DDPM is a method for training a deep-learning model to remove noise from an image in a continuous fashion. In a forward diffusion process, the image is progressively corrupted to make it indistinguishable from standard Gaussian noise. The backward diffusion process removes the noise from the random noise to produce the output image. DDIM[40], a refinement of DDPM, is an image generation method that allows for much faster sampling than traditional diffusion-based generation models. It is a variant of the traditional diffusion-based generation model that uses implicit sampling to accelerate the image generation process. This model has the advantage of generating high-quality images quickly, while maintaining the performance of the original diffusion model. Text-to-Image generative model is a method of generating images from a given text prompt. A representative of the text-to-image generative model, LDM[9], is a proposed method that wraps a diffusion model with an autoencoder so that the diffusion process operates on a latent space representation of the image rather than the image itself. The autoencoder encodes the image details into the latent space and decodes the latent space back into the high-resolution image, which is not good for speed and performance. LDM significantly improves the speed and performance of the training process because it only works on the conceptual latent space.

DALL-E[41] proposes to model the text-to-image generation task autoregressively by utilizing transformers from a single data source. DALL-E2[42] does not train a text encoder, but uses a pre-trained CLIP[43] as an encoder and generates the final image by embedding the image output by text prompts. Flamingo[44] is proposed to use a visual encoder that encodes visual input features into a small number of visual tokens using VLM[45] techniques and a persistent resampler to pass the visual information to the transformer for use. ControlNet[46] proposes a way to control the output image using a canny edge map on the input image. It is configured to fine-tune Stable Diffusion[9] using a small number of images. Stable Diffusion is a latent diffusion model technique that operates in the latent space of the autoencoder rather than the image itself. Rather than directly predicting pixel values, the diffusion model predicts the compressed latent embedding from the autoencoder, which has the advantage of reconstructing the image based on a more abstract representation of the data rather than a pixel-by-pixel prediction. SDXL[2] uses a U-Net structure[47] that is three times larger than the original Stable Diffusion, significantly enhancing high-resolution image generation performance. It effectively handles various resolutions and aspect ratios by incorporating image size and crop conditions into model training. Additionally, it refines generated images to greatly improve their detail and resolution. Openjourney[48] was developed with the background of enabling complete control over the generated content. It is an open-source model adjusted from Stable Diffusion version 1.5. Imagen[49] shows that using T5, where the encoder is pre-trained on text only, has a greater impact on overall performance than scaling the decoder of the diffusion model. They also show that the generated image is output in super-resolution using a super-resolution upsampler model. Prompt-to-Prompt[50] proposes a method for semantically editing specific regions of an image in the text-to-image generative model by simply manipulating text prompts. By controlling which pixels are attended to which tokens in the prompt text during the diffusion step, it is possible to inject a cross-attention map during the diffusion process to edit the image. BK-SDM[51] A model that removes residual blocks and attention blocks from Stable Diffusion and reduces the model size through knowledge distillation. It can be used at low cost by

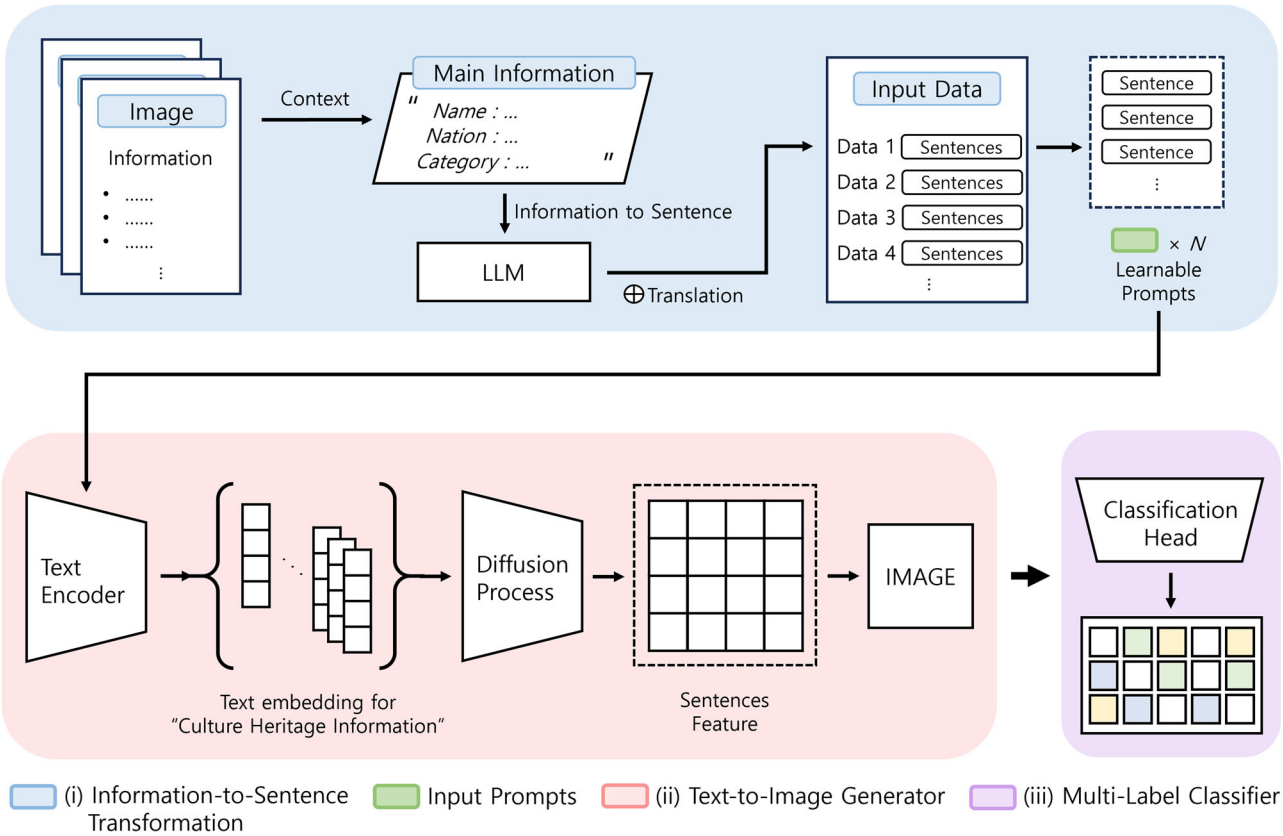


Fig. 5 | Main Process: (i) Translation of complex heritage texts from ancient languages into simplified English, (ii) Use of these English descriptions as inputs for a text-to-image generative model, (iii) Application of generated images in a multi-label image classification system to analyze cultural heritage associations.

compressing the existing stable diffusion and simplifying the model structure.

While these models have the ability to effectively process and generate text and image data, they present a number of challenges, especially when dealing with non-English languages such as Korean. We found that models for generating images from Korean text perform poorly compared to English models. This is because even models trained on Korean have less training on historical content, and the text-to-image conversion is less complete when using Korean words verbatim. Korean also has a different grammatical structure and vocabulary than English, which requires additional training for models to handle it effectively. One of the main problems with processing Korean is the mix of classical words and foreign words (e.g., Chinese characters, Japanese, and English). It is very difficult to accurately understand and process such text, which can lead to poor model performance. To improve this, in this study, we tried an approach to convert Korean text to English. This approach helps the model perform better while preserving the meaning of the text. The main goal of this paper is to compare and analyze various LLMs and image generation models, and to present the problems they encounter in processing Korean and classical words and their solutions. By doing so, we hope to provide users with better performance and convenience, and explore the possibility of implementing them at a lower cost. In particular, we propose that the English conversion approach can effectively solve the problems encountered in Korean processing.

Methodology

In the case of cultural heritage with little textual information, it is difficult to classify using text alone, so it is possible to utilize a generative model represented visually. As such, the management system of cultural heritage data has been classified and managed mainly by text, but this study proposes a new method of generating images and classifying them through them.

Figure 5 shows the overall process of this study. This process enables managers and users of difficult cultural heritage information to visually understand and utilize cultural heritage information through images. In addition, even if the amount of information is small, the generated images can be used to find new classification associations.

Problem Setting

Prompt Refining. Each cultural heritage item commonly possesses information such as name, material, and historical context, and this information was used to construct prompts. In this study, the results of using Korean and English were compared and analyzed; the English descriptions generated images with more consistency and higher utility as visual materials. Specifically, using Korean directly produced very similar results even between unrelated meanings. As the historical materials included foreign languages besides Korean, better results could not be obtained. Therefore, an attempt was made to unify these into English prompts. This process can be seen in (i) of Fig. 5. Traditional Korean languages posed challenges for the latest LLMs, which could not understand them. Therefore, the GPT 3.5 is used, as it can easily translate word corpus into English descriptions.

Text-to-Image Generator

In this study, we analyzed various text-to-image generative models including Openjourney, DALL-E2, DALL-E3, SD 1.5, SD 2.1, and SDXL 1.0. Initially, images were generated using models fine-tuned in Korean to directly use Korean text. However, we observed that these models generated images that were not variations of similar patterns but entirely different content for the same input. This issue includes the fact that models trained in Korean lack sufficient learning of historical content to generate appropriate sentences. The Korean fine-tuned Stable Diffusion model demonstrated fast processing speed and cost efficiency but tended to generate unrelated images when expressing classical words, making it unsuitable for multi-label classification tasks. Therefore, we translated cultural heritage information containing classical words into English and compared text-to-image

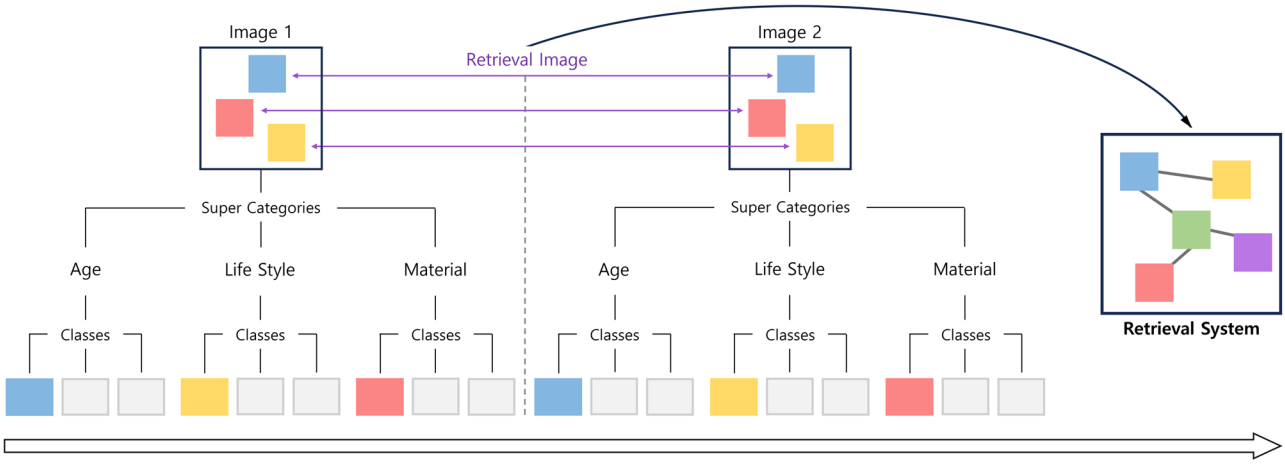


Fig. 6 | Overview of the Image Search System using Multi-Label Classification: This figure illustrates how the transformer-based multi-label classification model integrates within the image retrieval system, demonstrating the workflow from image categorization to retrieval based on 'super categories.'

generative models based on this input. Among them, DALL-E 3 and SDXL 1.0 particularly excelled in generating images suitable for English descriptions. However, DALL-E 3, despite its excellent performance, faces difficulties in mass image generation due to its closed-source nature. In contrast, SDXL 1.0, being open-source, successfully generated visually distinct and excellent feature images.

Considering that semantic visualization results are more important than visual quality in search systems, we employed the open-source and high-performance SDXL 1.0 for image generation. The process for generating images from textual descriptions is formulated as follows:

$$I = f(D;\theta) \tag{1}$$

where I denotes the generated image, D represents the input sentences, and θ are the parameters of the model. We use SDXL 1.0, which is open source and has good performance and processing speed, and find that it produces visually distinguishable and good features. Since the performance of the search system is important in this study, the semantic visualization results are more important than the visual quality of the images.

Multi-Label Classifier

Multi-label classification was used to enable efficient classification and search in the image search system. The dataset used is structured in the MSCOCO[52] format, with each image being assigned multiple labels. These labels are categorized according to super categories. As shown in Fig. 6, a model for multi-label classification can be applied to a retrieval system by simultaneously predicting classes belonging to each supercategory. Specifically, when imaging only the textual information of a cultural property, each image is multi-label classified to reveal the subclasses predicted by the model in each category. From these results, the model learns which features are similar between cultural heritage sites whose textual information has been imaged, and this allows it to find and connect similar cultural heritage sites. As a result, the system is able to analyze the associations between images and provide recommended images, which can be used to visually understand the relevance of cultural heritage information.

Figure 7 shows the structure of multi-label classification with a transformer-based architecture[53] for each super-category. This structure consists of two main streams: Spatial Stream and Semantic Stream. The Spatial Stream uses a Vision Transformer (ViT) to process the cultural heritage image data generated in the previous step. It divides the image into patches, extracts the visual features of each patch, converts them into vectors, and encodes them into a learnable form. The visual information in the image is processed through a transformer model, which is important for the classification process. The Spatial Stream uses a Vision Transformer (ViT) to process the cultural heritage image data generated in the previous step. It divides the image into patches, extracts the visual features of each patch, converts them into vectors, and encodes them into a learnable form. The visual information in the image is processed through a transformer model, which is important for the classification process. Semantic Stream uses BERT, a large-scale language model, to process textual information. The textual data as input contains the actual multi-label classes of the cultural heritage to be matched with the generated images. The stream learns by matching images generated from cultural heritage textual information with the actual categories of the cultural heritage. Specifically, it is trained by matching the generated images with the original textual information (category information: material, age, lifestyle). It analyzes the information between text and images, and understands the context and semantic relationships between labels. The two streams function independently, collecting visual and textual features respectively, which are then merged through a convolutional layer.

During this process, visual and text features are combined to prepare the data for final classification. The transformer encoder-decoder architecture then processes this combined data to produce the final classification result. This architecture is used to predict labels by considering the visual features of the image patches together with the semantic information in the text data. The process of connection and classification is formulated as follows:

$$P = Softmax(\mathrm{W_h} * (Concat(\mathrm{f_s}, \mathrm{f_t})) + b) \tag{2}$$

$\mathrm{f_s}$ and $\mathrm{f_t}$ and represent the feature vectors extracted from the spatial and semantic streams, respectively. $\mathrm{W_h}$ and b denote the trainable weights and biases. The *Concat* function concatenates the feature vectors from the two streams. The *Softmax* function converts each element of the final output vector into probabilities, presenting the label with the highest probability as the final prediction result.

Dataset

The dataset we will introduce is **MUCH (Multi-purpose Universal Cultural Heritage)**, which is augmented with 9600 images of Korean cultural heritage, totaling 96,000 images. Of these, 86,400 images were used for training and 9600 images were used for validation. The dataset was created by processing data provided by the National Museum of Korea. The dataset is categorized into three super-categories: age, life, and material, with a total of 32 classes. We used name, era, and material information to conduct experiments with minimal information from the metadata. The difficulty of categorizing non-distinct objects comes from supercategories such as age,

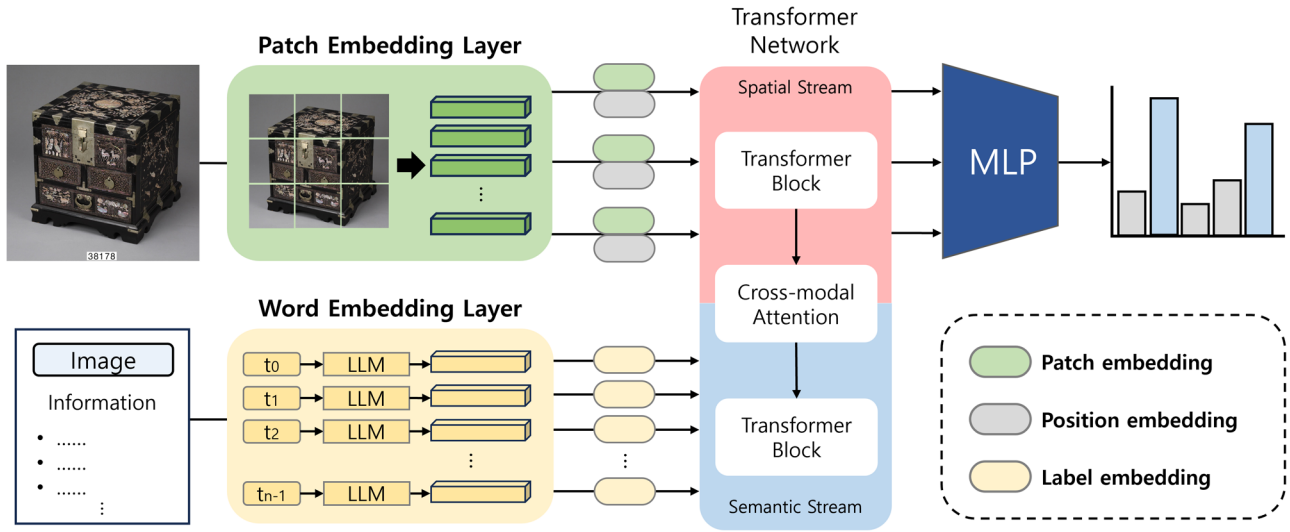


Fig. 7 | Transformer-based Multi-label Classification Architecture.

Table 1 | Classification Based on Korean Age, Materials, and Lifestyles Using the MUCH Dataset

Super Categories	Class Name	Number
Age	Bronze Age, Early Iron Age, Proto-Three Kingdoms, Baek-je, Silla, Three Kingdoms, Unified Silla, Go-ryeo, Late Joseon, Joseon, Japanese colonial period	11
Material	Wood, stone, soil, paper, mineral, fossil, seed, chilgi, leaf, leather, bone, fiber, ceramic, rubber	14
Life style	Transportation/communication, culture/art, social life, industry/livelihoods, dietary life, clothing life, daily life	7

lifestyle, etc. Age has 11 classes, Lifestyle has 7 classes, and Material has 14 classes.

In Table 1, the classes within the subcategories of super categories are shown. Age includes classes categorizing age in Korea: Bronze Age, Early Iron Age, Proto-Three Kingdoms, Baek-je, Silla, Three Kingdoms, Unified Silla, Go-ryeo, Late Joseon, Joseon, Japanese colonial period. Material refers to the surface or texture of substances: wood, stone, soil, paper, mineral, fossil, seed, lacquer, leaf, leather, bone, fiber, ceramic, rubber. Last, Life-style represents the background of industrial and other age: transportation/communication, culture and art, social life, industry/livelihoods, dietary life, clothing life, daily life.

Results

Implementation Details

Table 2 shows the key parameters used in the multi-label classification model. To evaluate the proposed approach, we cond ucted experiments using MUCH. We used the average precision (AP) for each category and the mean average precision (mAP) as evaluation metrics. We also measured the F1-measure (OF1), which indicates the performance of the model. We utilize a transformer-based multi-label classification model. To maintain a balance between detail capture and computational efficiency, all images are normalized to a resolution of 448×448. The AdamW optimizer[54] is employed with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-5}. For the loss function, binary cross entropy(BCE)[55] is adopted as it is well-suited for multi-label classification.

Table 2 | The summarize the key parameters used in Multi-Label Classification Model

Parameters	Value
Learning rate	1e−5
Batch size	128
Epoch	100
Resolution	448×448
(Train, Valid, Test)	(8: 1: 1)
Optimizer	AdamW
Weight Decay	1e−5
Loss Function	BCE

Table 3 | Experimental results on datasets showing a performance (mAP in %)

Dataset	Super Category	Class	mAP	OF1
MUCH(Ours)	All	32	56.3	50.1
MUCH(Ours)	Age	11	45.7	41.1
MUCH(Ours)	Material	14	78.8	69.5
MUCH(Ours)	Life style	7	44.4	39.7

Experiment Results

Quantitative results. Table 3 shows the performance results. The MUCH we used consisting of 9600 images before augmentation, yielded high-performance results. Analyzing the performance of each super category reveals the following characteristics.

Firstly, the mAP value in the 'Age' category is relatively low at 45.7%. This is likely due to the difficulty in classifying age in Korea primarily based on visual information alone. Clear visual clues for distinguishing between different eras may not be readily apparent, leading to decreased classification accuracy. The mAP value for the 'Life style' category is also low at 44.4%. This is because the concept of lifestyle itself is challenging to define with distinct visual objects. Moreover, with only 7 classes, fewer than other categories, and a limited dataset, there may not have been sufficient training. Conversely, the 'Material' category achieved the highest performance with an mAP value of 78.8%. This is attributed to the ease of visually distinguishing between different types of materials. Characteristics such as color

and texture provide clear visual clues, enabling high accuracy in image-based classification. Thus, the differences in mAP values across each super category highlight how visual clarity of the classification targets, dataset size, and number of classes can significantly influence performance. If criteria are established for abstract terms such as Age, Material, and Life style, and distinguishable object classes are added, MUCH can expect good performance. Therefore, the performance of mAP 56.3%, OF1 50.1% is not a low score by any means. If we increase the amount of datasets, it will outperform the comparison dataset.

To evaluate the effectiveness and user experience of the proposed system, we collected feedback from users. Users gave their opinions on the system, both positive and negative, which helped us to understand the system's strengths and what needs to be improved. Table 4 shows the feedback from cultural heritage professionals. Feedback has been anonymized by user number. For each item, usability evaluates the intuitiveness of the user interface and ease of use, while speed evaluates the response time or data processing speed of the system. Accuracy evaluates the accuracy or reliability of the results provided by the AI system, and scalability evaluates whether the system has the potential to evolve in the future and is sustainable.

Table 4 | User Feedback from Cultural Heritage Professionals on System

	Usability	Speed	Accuracy	Scalability
User 1	✓	✓	✗	✓
User 2	✓	✓	✓	✓
User 3	✓	✓	✓	✓

As shown in Table 4, the majority of users gave the system positive ratings for most of these evaluation criteria, but also identified the need for a high level of accuracy in order for the system to become popular. This feedback will be used as an important reference to set the direction for future system improvements.

Qualitative results. Figure 8 compares the output results from different LLMs. Alpaca, for instance, produces results that diverge significantly from the original meaning of the input prompt. While Llama demonstrates good performance, it suffers from slower speeds and occasional inconsistencies in output. On the other hand, GPT-3.5 consistently provides accurate results with rich English descriptions based on search queries. Given the importance of high accuracy for generating correct images in this study, GPT-3.5 was chosen.

Figure 9 illustrates the translation of Korean sentences containing classical words into English using GPT-3.5. The highlighted portions indicate where Korean has been translated into English. It demonstrates high performance in seamlessly translating Korean into English, even when classical words are included. The model adeptly grasps the context of sentences and generates appropriate English expressions. Thus, GPT-3.5 was refined using prompts to achieve these results.

Figure 10a represents a coin, and the result from the SDXL 1.0 model appears most similar to the actual image, closely matching its texture representation as well. While other models also generated similar images, SDXL 1.0 stands out for its fidelity in texture representation akin to the real image. Figure 10b depicts celadon, showing a creation closely resembling the

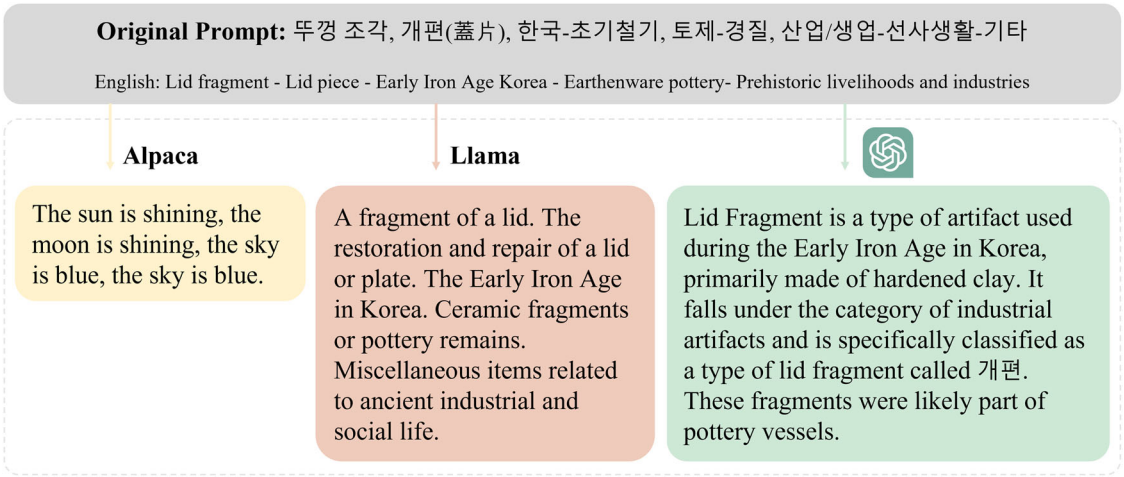


Fig. 8 | Comparison results of different LLMs for translating Korean into English sentences.

Fig. 9 | The result of translating Korean into English sentences using GPT-3.5.

Fig. 10 | Compare the visualization results of text-to-image generative models when using translated English sentences as input; (**a**) coins, (**b**) celadon bottles, (**c**) white porcelain lanterns.

Fig. 11 | Compare images generated using SDXL 1.0 (top) with actual cultural heritage images (bottom). a a duck sculpture, **b** fragments of a jar, **c** bead bracelets.

original overall. Figure 10c features a porcelain lamp, accurately capturing the characteristics of a lamp or porcelain. Overall, considering the generated image results alongside time and computational costs, SDXL 1.0 proves suitable for data generation and augmentation tasks. This model not only contributes to high performance but also offers practical convenience for users and administrators focusing on real-world applications.

Figure 11 is a comparison between the image generated through the proposed method and the actual image. The amount of metadata information for each cultural heritage is very different. However, it can be seen that the image generated after switching to the English description is generated very similar to the actual image. It is generated so similarly that it is difficult to distinguish the generated images of (a), (b), and (c) of Fig. 11 from the original image. Using these images for the model to learn will also help automatically classify newly added cultural heritage.

Figure 12 shows a comparison of automated and manual processing of a cultural heritage search and recommendation system. Figure 12a shows a

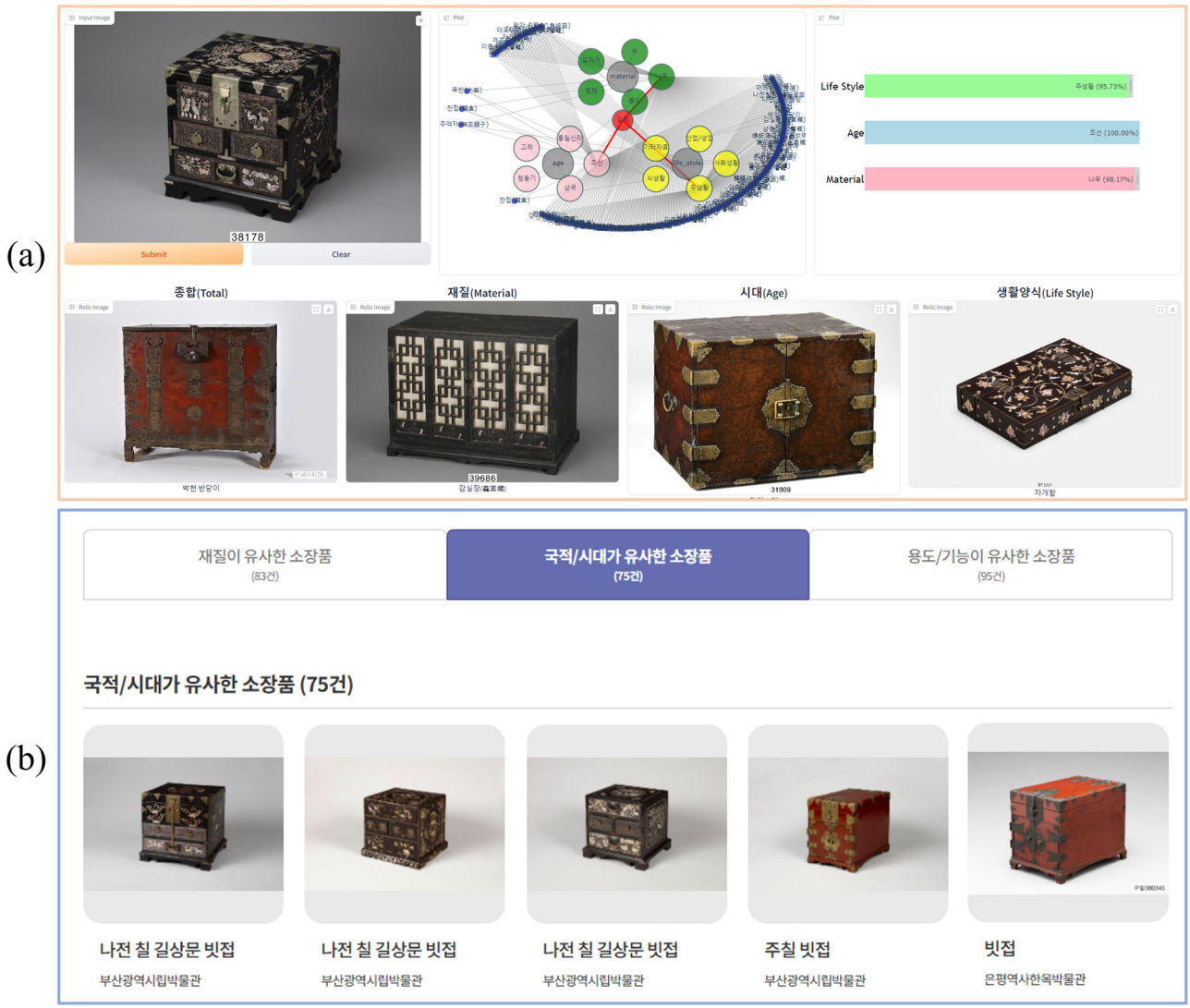


Fig. 12 | Examples of a search and recommendation system for cultural heritage. **a** Graphical user interface for automated classification and recommendation. Uploaded images of cultural heritage data are analyzed and classified based on three main criteria: material, age, and lifestyle. The classification results are visualized using a network graph and bar charts, illustrating the hierarchical structure of subcategories within each criterion. Both visualization methods present the predicted classification probabilities, accurately identifying the given cultural heritage data as *Life Style: Daily Life* (주생활), *Age: Joseon* (조선), *and Material: Wood* (나무). The "Total" score represents an aggregated confidence measure across all three classification categories. **b** Expert-curated recommendations on the National Museum of Korea's eMuseum platform. The same cultural heritage data used as input in (**a**) is shown on the recommendation image page within the age category of the National Museum of Korea's eMuseum site. This page is curated by experts who manually select similar cultural heritage data. As shown in (**a**), the AI-predicted recommendations closely align with the manually curated results, demonstrating the system's effectiveness in identifying relevant cultural heritage data.

GUI environment where a model trained on user input data analyzes images and automatically classifies and recommends appropriate categories based on the evaluation results. Figure 12b shows the cultural heritage information page of an e-museum site operated by the National Museum of Korea, showing the same items as in Fig. 12a. This page requires all images to be uploaded and organized manually. If an automated approach like the one in Fig. 12a were to be adopted, the time and effort spent on manual work could be significantly reduced, and costs could also be reduced. The system can provide results at a practical level because it automatically analyzes the various textual information of each cultural property and classifies them into appropriate categories.

Limitation

Figure 13 shows the results of inputting Korean directly into models trained on Korean. Korean language errors include these models' inability to generate appropriate sentences due to their limited historical context training. In text-to-image generative models, attempts to use clusters of Korean words directly resulted in low completion quality for Korean. Even when using Korean, the problem of randomly generating diverse features was discovered rather than consistently creating similar features for the same word. In Fig. 13a lacks images that closely resemble the original. The cultural heritage name of this image implies a tower made of stone, but except for DALL-E 3, images with different characteristics from the original image are generated. For Fig. 13b, the original image is a nameplate, but using Korean explanations results in completely unrelated images. Overall, DALL-E 3 generates images most similar to the original, but as it is not open-source, it cannot generate large quantities of images. Therefore, attempting to translate Korean into English sentences resulted in generating better image results.

One of the main limitations of the system is the categorization of abstract classes. This problem arises when defining classes for unclear concepts, such as dividing age or industries for association classification. Especially in the case of age classification, where the historical

Fig. 13 | Comparison of visualization results from text-to-image generative models using Korean text input. a Generated image of a stone tower. The input text used for generation: 옥개석, 한국 - 고려, 종교신앙 - 불교 - 예배 - 탑, 돌 - 화강암 (*translated*: *Roof stone, Korea - Goryeo, Religious belief - Buddhism - Worship - Pagoda, Stone - Granite*). **b** Generated image of a nameplate. The input text used for generation: 김한동 호패, 한국 - 조선, 사회생활 - 사회제도 - 신표 - 호패, 골각패갑 - 수각, 길이 *9.6 cm* (*translated: Kim Han-dong Nameplate, Korea - Joseon, Social life - Social system - Identification tag - Nameplate, Bone shell plate - Engraved, Length 9.6 cm*).

context of a country is intertwined with many other historical contexts, it is difficult to categorize them with clear criteria. Since the generated image does not contain a wide range of features, it is difficult to distinguish between them, so it is expected that increasing the dataset will significantly improve the performance of learning various features from the feature map. To improve performance, we plan to generate images with more complex features and apply methods such as negative prompts to improve performance.

Conclusions

The method proposed in this study is to automatically manage cultural heritage materials with varying amounts of information by analyzing their associations with images generated using textual information. This study suggests that it can make a significant contribution to the classification of cultural heritage, not by competing for high performance but by its potential impact. We have adopted a method that moves beyond traditional approaches of classifying and managing cultural heritage solely based on textual information by integrating various algorithms. This approach will also be useful for quickly identifying the relevance of new information and automatically categorizing it as it is added. This is the first step towards reducing the cost and time of large-scale retrieval systems, and will lead to more sophisticated and efficient management.

In future work, we will include classes not covered in our experiments and improve the classification performance of abstract classes. Metadata not used in the research will also be further refined and continually updated to MUCH. We will apply these improvements to an automated system to validate their effectiveness. The expanded system will contribute to improving the digital archives of cultural heritage and serve as a new starting point for various interdisciplinary research efforts.

Data availability

Data is available from the corresponding author upon reasonable request.

Received: 23 July 2024; Accepted: 1 December 2024;
Published online: 13 March 2025

References

1. Pierdicca, R. et al. Point cloud semantic segmentation using a deep learning framework for cultural heritage. *Remote Sens.* **12.6**, 1005 (2020).
2. Podell, D. et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv Prepr. arXiv* **2307**, 01952 (2023).
3. Sun, W. & Wang, R. Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM. *IEEE Geosci. Remote Sens. Lett.* **15**, 474–478 (2018).
4. Guo, Y. et al. Deep learning for 3d point clouds: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 4338–4364 (2020).
5. Tabib, R. A. et al. DeFi: detection and filling of holes in point clouds towards restoration of digitized cultural heritage models. Proceedings of the IEEE/CVF International Conference on Computer Vision. (2023).
6. Hu, Q. et al. ConvSRGAN: super-resolution inpainting of traditional Chinese paintings Heritage. *Science* **12**, 176 (2024).
7. Gunawan, A. et al. Modernizing old photos using multiple references via photorealistic style transfer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2023).
8. Zhang, J. et al. "ArchGPT: harnessing large language models for supporting renovation and conservation of traditional architectural heritage.". *Herit. Sci.* **12.1**, 220 (2024).
9. Rombach, R. et al. High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2022).
10. Devlin, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1 (long and short papers) (2019).
11. Sinha, K. et al. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv Prepr. arXiv* **2104**, 06644 (2021).
12. Liu, Y. et al. Roberta: A robustly optimized bert pretraining approach. *arXiv Prepr. arXiv* **1907**, 11692 (2019).
13. Sanh, V. et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv Prepr. arXiv* **1910**, 01108 (2019).
14. Lan, Z. et al. Albert: A lite bert for self-supervised learning of language representations. *arXiv Prepr. arXiv* **1909**, 11942 (2019).
15. Clark, K. et al. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv Prepr. arXiv* **2003**, 10555 (2020).
16. He, P. et al. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv Prepr. arXiv* **2006**, 03654 (2020).

17. Sarlin, P.-E. et al. Superglue: Learning feature matching with graph neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020).
18. Alec Radford, J. W. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
19. Keskar, N. S. et al. Ctrl: A conditional transformer language model for controllable generation. *arXiv Prepr. arXiv* **1909**, 05858 (2019).
20. Ben Mann, N. et al. Language models are few-shot learners. *arXiv Prepr. arXiv* **2005**, 14165 (2020).
21. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2020).
22. Qiu, X. et al. Pre-trained models for natural language processing: A survey. *Sci. China Technol. Sci.* **63**, 1872–1897 (2020).
23. Dong, C. et al. A survey of natural language generation. *ACM Comput. Surv.* **55**, 1–38 (2022).
24. Touvron, H. et al. Llama: Open and efficient foundation language models. *arXiv Prepr. arXiv* **2302**, 13971 (2023).
25. Su, J. et al. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).
26. Taori, R. et al. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm.stanford.edu/2023/03/13/alpaca.html 3.6: 7 (2023)
27. Wang, Y. et al. Self-instruct: Aligning language models with self-generated instructions. *arXiv Prepr. arXiv* **2212**, 10560 (2022).
28. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv Prepr. arXiv* **1312**, 6114 (2013).
29. Baldi, P. Autoencoders, unsupervised learning, and deep architectures. Proceedings of ICML workshop on unsupervised and transfer learning. JMLR Workshop and Conference Proceedings, (2012).
30. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
31. Gao, H., Jian, P. & Huang, H. Progan: Network embedding via proximity generative adversarial network. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. (2019).
32. Karras, T., Laine, S. & T. A. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019).
33. Karras, T. et al. Analyzing and improving the image quality of stylegan. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020).
34. Huang, X. & Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. Proceedings of the IEEE international conference on computer vision. (2017).
35. Zhang, H. et al. *Self-attention generative adversarial networks. International conference on machine learning*. (PMLR, 2019).
36. Brock, A., Donahue, J. & Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *arXiv Prepr. arXiv* **1809**, 11096 (2018).
37. Esser, P., Rombach, R. & Ommer, B. Taming transformers for high-resolution image synthesis. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2021).
38. Yu, J. et al. Vector-quantized image modeling with improved vqgan. *arXiv Prepr. arXiv* **2110**, 04627 (2021).
39. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
40. Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020).
41. Ramesh, A. et al. *Zero-shot text-to-image generation. International conference on machine learning.* (Pmlr, 2021).
42. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with CLIP latents. Preprint at http://arxiv.org/abs/2204.06125 (2022).
43. Radford, A. et al. *Learning transferable visual models from natural language supervision. International conference on machine learning*. (PMLR, 2021).
44. Alayrac, J.-B. et al. Flamingo: a visual language model for few-shot learning*Adv. neural Inf. Process. Syst.* **35**, 23716–23736 (2022).
45. Xu, U. et al. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv Prepr. arXiv* **2105**, 09996 (2021).
46. Zhang, L., Rao, A. & Agrawala, M. Adding conditional control to text-to-image diffusion models. Proceedings of the IEEE/CVF International Conference on Computer Vision. (2023).
47. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. (Springer International Publishing, 2015).
48. Prompthero. Openjourney. Hugging Face, n.d. Web. 12 July 2024. https://huggingface.co/prompthero/openjourney.
49. Saharia, C. et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. neural Inf. Process. Syst.* **35**, 36479–36494 (2022).
50. Hertz, A. et al. Prompt-to-prompt image editing with cross attention control. *arXiv Prepr. arXiv* **2208**, 01626 (2022).
51. Kim, B.-K. et al. Bk-sdm: Architecturally compressed stable diffusion for efficient text-to-image generation. Workshop on Efficient Systems for Foundation Models@ ICML2023. (2023).
52. Lin, T.-Y. et al. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzer- land, September 6-12, 2014, Proceedings, Part V 13, pages 740–755 (Springer, 2014).
53. Xuelin Z., Jiawei G., Jiawei, C., Weijia, L. & Liu, B. Two-stream transformer for multi-label image classification. In Proceedings of the 30th ACM International Conference on Multimedia. 3598–3607 (2022).
54. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv Prepr. arXiv* **1711**, 05101 (2017).
55. Zhang, Z. & Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems, **31**, (2018).
56. National museum of korea e-museum. https://www.emuseum.go.kr/main.

Acknowledgements

This research is supported by the Ministry of Culture, Sports and Tourism and Korea Creative Content Agency (Project Number: RS-2023-00219579).

Author contributions

H.H. was responsible for writing the main manuscript, designing the experiments, creating the figures, and conducting the main model experiments. C-W.P. assisted with experimental design and performed data pre-processing. H-K.K. assisted with experimental design and setting up the experiments. J-H.L. reviewed the manuscript and contributed to the experimental design. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Jae-Ho Lee.