Journal of KIIT. Vol. 24, No. 7, pp. 1-21, Jul. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.7.1

Explainable Multimodal Visual-Contextual Framework for Restoration of Severely Damaged Korean Historical Rubbings

Yejin Jhang*, Donghyun Hwang**, Junhyoung Choi***, Seongmin An****, and Gilsang Yoo*****

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2023-00246191). These authors contributed equally to this work.

Abstract

Historical rubbings are valuable for historical and philological research, but severe erosion, cracks, and ink smudging often reduce readability. Manual deciphering's high cost and limited expert throughput hinder large-scale digitalization. To address these limitations, this paper proposes a robust multimodal visual-contextual reasoning framework for damaged character restoration. The system classifies damaged characters into Totally Lost and Partially Lost categories and combines a Swin Transformer V2-based visual restoration module with a SikuRoBERTa-based contextual reasoning module for recovery according to the damage type. To improve trustworthiness, XAI-driven Attention Maps and confidence breakdown tables visualizes visual and contextual evidence for human expert verification. Experiments achieved Top-5 accuracies of 99.38% for visual restoration and 82.56% for contextual prediction.

요 약

탁본은 역사학 및 문헌학 연구에서 역사적 가치를 지니지만 침식, 분열 등 복합적 훼손으로 인해 원문 판독에 많은 어려움이 있다. 수동 판독은 막대한 비용과 전문가의 한정된 처리량으로 인해, 대규모 디지털화의 병목 현상으로 작용하고 있다. 이에 본 논문에서는 이러한 한계를 해결하기 위해 훼손 문자 복원을 위한 강건한 멀티모달 시각-문맥 추론 프레임워크를 제안한다. 제안한 시스템은 훼손 문자를 Totally Lost와 Partially Lost로 분류한 뒤, 훼손 유형에 따라 Swin Transformer V2 기반 시각 복원 모듈과 SikuRoBERTa 기반 문맥 추론 모듈을 결합하여 복원을 수행한다. 또한 신뢰성 향상을 위해 Attention Map과 기여도 표시 테이블을 비롯한 XAI를 적용해 시각, 문맥 근거를 시각화함으로써 전문가가 AI의 판단 근거를 검증할 수 있도록 하였다. 실험 결과, 제안한 방법은 시각 복원에서 Top-5 정확도 99.38%, 문맥 예측에서 82.56%를 달성하였다.

Keywords

historical rubbings restoration, visual-context fusion, swin transformer v2, sikuRoBERTa, explainable AI

* MS candidate, Dept. of Culture Technology, Graduate School of Culture Technology, KAIST
- ORCID: https://orcid.org/0009-0003-5827-4960
** BS degree, Dept. of Nursing, Seoul Women's College of Nursing
- ORCID: https://orcid.org/0009-0001-3585-7536
*** BS degree, Dept, of Industrial and Systems Engineering, Dongguk University
- ORCID: https://orcid.org/0009-0000-9109-0977
**** BS candidate, Dept. of IT Convergence Engineering, Jeonbuk National University
- ORCID: https://orcid.org/0009-0004-3862-1045
***** Professor, Dept. of Creative Informatics & Computing Institute, Korea University
- ORCID: https://orcid.org/0009-0002-1085-5355

· Received: Mar. 19, 2026, Revised: Apr. 14, 2026, Accepted: Apr. 17, 2026
· Corresponding Author: Gilsang Yoo
Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, S.Korea
Tel.: +82-2-3290-1674, Email: ksyoo@korea.ac.kr

Ⅰ. Introduction

Rubbings, which serve as reproductions of inscriptions from steles or monuments onto paper, possess irreplaceable value as primary sources in historical and philological research[1]. They play a pivotal role in verifying ancient calligraphic styles, linguistic evolution, and the societal contexts of their respective eras. Consequently, interpreting and systematically organizing these records holds significant academic value[2].

However, the majority of extant rubbings have suffered complex and severe degradations over centuries, including surface erosion due to weathering, cracks, and ink smudging that occurred during the rubbing process. In particular, background noise distributed throughout the rubbing blurs the boundaries between character and non-character regions, significantly increasing the difficulty of character recognition. These physical damages destroy the morphological information of characters, making decipherment extremely difficult.

To date, the decipherment process has relied entirely on manual methods dependent on the experience and intuition of experts. This reliance results in significant time and cost inefficiencies, serving as a major bottleneck in large-scale digital archiving projects[3]. Moreover, in severely damaged areas, subjective interpretations and inconsistencies arising from manual decipherment limit the objective reliability of the results.

Against this backdrop, research exploring digitization methods for rubbings has been actively pursued to provide high-quality digital replicas that researchers can utilize while maximally preserving the visual and physical characteristics embedded in the rubbings[4][5]. AI research for the preservation and restoration of ancient documents has largely developed around two main axes: the restoration of visual features and context-based information inference.

In terms of visual restoration, research utilizing Generative Adversarial Networks (GANs) is predominant. Souibgui et al.[6] improved denoising and binarization performance for degraded documents using Document Enhancement GAN (DE-GAN), while Su et al.[7] and Zheng et al.[8] proposed dual GAN and Example Attention GAN (EA-GAN) architectures, respectively, to intricately restore global and local features as well as structural details of damaged Chinese characters. Additionally, Shobha Rani et al.[9] improved the readability of physically worn documents using a semi-adaptive thresholding technique. However, while existing GAN-based models demonstrate excellent visual restoration capabilities, they have limitations in controlling morphological distortions or "hallucinations" in sections where character strokes are missing.

To overcome this, research on context-based restoration using language models has been conducted in parallel. Shen et al.[10] proposed a mechanism to predict missing parts using a Blank Language Model (BLM), and Fetaya et al.[11] utilized Recurrent Neural Networks (RNNs) to statistically reconstruct text from fragmented Babylonian clay tablets. Furthermore, Assael et al.[12] restored damaged characters in Greek inscriptions using deep neural networks. Kang et al.[13] systematized the contextual restoration of vast Korean classical literature, such as the Annals of the Joseon Dynasty (AJD), by proposing a Transformer-based multi-task learning approach. While these methods enable inference using the semantic structure of sentences, they carry the risk of leading to incorrect estimations when visual cues are excluded.

Recently, attempts have been made to integrate visual information and linguistic context to overcome the limitations of single models. Wang et al.[14] proposed the first inscription restoration model combining natural language processing and computer vision. However, this method is limited by being a semi-automatic approach requiring the intervention of

imaging experts, leaving the need for a fully automated and interpretable restoration framework.

Despite continued efforts to preserve and restore rubbings, existing studies have generally performed visual restoration and contextual inference in a fragmented manner, which reduces restoration reliability in environments with extreme damage. In summary, GAN-based approaches suffer from hallucination and morphological distortion; language model-based approaches are susceptible to incorrect estimation when visual cues are absent; and existing multimodal attempts either require manual intervention or lack interpretability mechanisms for expert validation. To bridge this gap between modalities, this study proposes an integrated restoration framework that detects damaged regions through ensemble recognition and organically fuses visual-contextual information.

The main contributions of this paper are summarized as follows:

- End-to-end automation: We propose the first recognition-guided multimodal framework that automates the entire rubbing restoration pipeline, from character detection and damage classification to visual restoration and contextual inference without the intervention of imaging experts.
- Multimodal fusion: We maximize restoration accuracy by combining Swin Transformer V2 and SikuRoBERTa through harmonic mean-based score fusion, enabling complementary use of stroke-level visual features and bidirectional linguistic context.
- Domain-specific performance: We demonstrate superior performance over existing single-modality baselines on a severely damaged Korean historical rubbing dataset, achieving Top-5 accuracies of 99.38% (visual) and 82.56% (contextual).
- Explainability and human oversight: We integrate Explainable AI (XAI) mechanisms by Attention Maps and confidence breakdown tables, designing a Human-in-the-loop interface that enables domain experts to verify and refine AI restoration decisions in real time.

Based on these contributions, the remainder of this paper is organized as follows. Section II reviews related work focusing on image and context-based restoration technologies, covering major models including Swin Transformer and SikuRoBERTa. Section III introduces the proposed system for recognizing and restoring damaged Ancient Korean characters. Section IV analyzes the experimental results, and Section V discusses the expected effects of this study and future research directions.

II. Related Work

2.1 Image recognition

Restoring rubbing images is a meticulous task that requires the simultaneous reconstruction and interpretation of local details of fine strokes and the global structure of character glyphs. Early research was dominated by Convolutional Neural Network (CNN)-based methodologies[15]. However, due to their fixed receptive fields, CNNs exhibited limitations in capturing global features. The Vision Transformer (ViT)[16], introduced to address this issue, excels at grasping global context but suffers from quadratic computational complexity with respect to input resolution. This inefficiency renders ViT impractical for rubbing restoration, where high-resolution processing is essential to preserve fine stroke details.

To resolve this computational complexity, Liu et al.[17] proposed the window-based Swin Transformer V1. By computing attention only within local windows (W-MSA), Swin Transformer V1 combines the local feature extraction strengths of CNNs with the global information capability of Transformers, progressively expanding the receptive field. The difference in computational complexity between the two models is expressed in (1):

$$\begin{cases}\Omega(ViT) = 4hwC^2 + 2(hw)^{2C} \\ \Omega(SwinT) = 4hwC^2 + 2M^2hwC\end{cases} \quad (1)$$

where h and w denote the height and width of the image, C is the channel dimension, and M represents the window size. While ViT's complexity is determined by the $(hw)^2$ term, Swin Transformer V1 fixes M as a constant, resulting in a linear increase in complexity with respect to resolution hw. However, as the model depth increases, activation values may diverge, leading to training instability. Furthermore, performance degradation occurs when processing images with resolutions different from those used during training[18]. In rubbing data characterized by severe noise, such instability significantly hinders restoration performance. Swin Transformer V2 introduces several improvements to address these limitations. First, it employs a Post-Normalization technique to normalize the output of each layer, ensuring training stability in deep networks[18]. Second, it adopts Scaled Cosine Attention to replace the dot-product attention used in conventional Transformers, which is sensitive to pixel value magnitudes[19]. As shown in (2), this method utilizes cosine similarity to normalize attention values, enabling robust feature extraction even in rubbing images with low contrast between strokes and the background.

$$Attention(Q, K, V) = Softmax\left(\frac{\cos(Q, K)}{\tau} + B\right)V \quad (2)$$

In (2), Q, K, and V represent the Query, Key, and Value matrices, respectively, and τ is a learnable scalar parameter. Third, to minimize positional information distortion in rubbing images of varying resolutions, Log-spaced Continuous Position Bias (Log-CPB) is introduced. As shown in (3), Log-CPB transforms relative coordinates into a log space to flexibly adapt to changes in window size[19].

$$\begin{cases}\widehat{\Delta x} = sign(\Delta x) \cdot log(1 + |\Delta x|) \\ \widehat{\Delta y} = sign(\Delta y) \cdot log(1 + |\Delta y|)\end{cases} \quad (3)$$

Consequently, Swin Transformer V2 offers faster inference speeds compared to diffusion-based generative models and eliminates the risk of hallucination, which is critical in historical restoration. Simultaneously, it provides superior training stability and high-resolution processing capabilities compared to Swin Transformer V1, making it highly effective for the precise restoration of rubbing characters[14][20].

2.2 Context recognition

In natural language processing, simultaneously considering context from both directions is essential for resolving sentence ambiguity and understanding complex relationships between words. Bidirectional Encoder Representations from Transformers (BERT)[21], a representative model that established the foundation for such bidirectional context representation, is designed to grasp the deep semantic meaning of text based on the Transformer encoder architecture. However, the static masking approach adopted by the original BERT fixes masked tokens during the data generation phase. As training repeats, the model is exposed to identical patterns, which reduces efficiency and limits the ability to understand diverse contexts due to a lack of data variety.

To overcome these limitations and enhance model robustness, Robustly Optimized BERT Approach (RoBERTa)[22] was designed. RoBERTa optimizes BERT's hyperparameters and training strategies to learn more sophisticated bidirectional context representations. Specifically, it maximizes data efficiency by introducing dynamic masking, where masking positions are randomly changed each time training data is fed into the model. Furthermore, by removing the Next Sentence Prediction (NSP) task and significantly increasing the batch size, the model is

optimized to learn the internal semantic structure of sentences more deeply.

The core training mechanism of RoBERTa, Masked Language Modeling (MLM), operates by maximizing the probability of the actual token appearing for a specific masked token x_i in a sentence, by synthesizing the surrounding bidirectional context information $x_{\setminus M}$. This induces the model to understand advanced contextual dependencies by minimizing the loss function L_{MLM} defined in (4).

$$L_{MLM} = -\sum i \in M \log P\left(x_i | x_{\setminus M}; \theta\right) \quad (4)$$

RoBERTa's advanced training mechanism and powerful context representation capabilities provide an optimal foundation for reflecting the characteristics of specific languages or specialized domains. SikuRoBERTa[23] is a model proposed to implement performance optimized for the special domain of classical literature while inheriting the structural advantages of RoBERTa. Based on Chinese-RoBERTa-wwm-ext, this model effectively reflects the vast vocabulary system and grammatical characteristics of ancient Chinese through continuous pre-training using the Siku Quanshu corpus.

2.3 XAI for restoration

XAI[24]-[26] techniques can be broadly categorized into model-agnostic post-hoc explanation methods, which provide explanations independent of the model structure, and model-specific interpretation methods, which derive evidence from the internal architecture of the model. While earlier XAI research mainly focused on visualizing spatial saliency through techniques such as Grad-CAM and attention visualization, recent studies have expanded toward quantitatively presenting the reliability and uncertainty of model predictions.

This trend is particularly important in damaged data restoration. In restoration tasks, multiple plausible solutions may coexist depending on the degree of data loss; therefore, simple visual importance explanations may not provide sufficient grounds for selecting a specific result. Accordingly, an interpretable confidence-presentation method that enables quantitative comparison of the degree of support for each candidate is necessary. Such an explanatory approach allows users to intuitively examine the relative support and uncertainty of candidate outputs, thereby facilitating comparison among subtly different restoration candidates and providing a clearer basis for final decision-making.

III. Proposed Methods

The overall configuration of the proposed system for restoring and interpreting damaged characters in Korean historical rubbings is illustrated in Fig. 1. The proposed framework takes an original rubbing image as input and sequentially performs damaged character restoration and semantic interpretation. First, the input rubbing image undergoes an Image Pre-processing stage to improve character legibility and prepare character regions for subsequent analysis. Next, the pre-processed image is passed to a Damaged Character Classification module, which determines the damage status of each character at the character level. For character regions identified as damaged, Swin Transformer V2-based visual restoration and SikuRoBERTa-based contextual reasoning are performed in parallel, and each module produces restoration candidates with corresponding confidence scores. These outputs are then integrated through a probability fusion process to generate the final restored character. After all damaged characters are restored, the reconstructed text is finally translated into modern Korean using a generative model API.

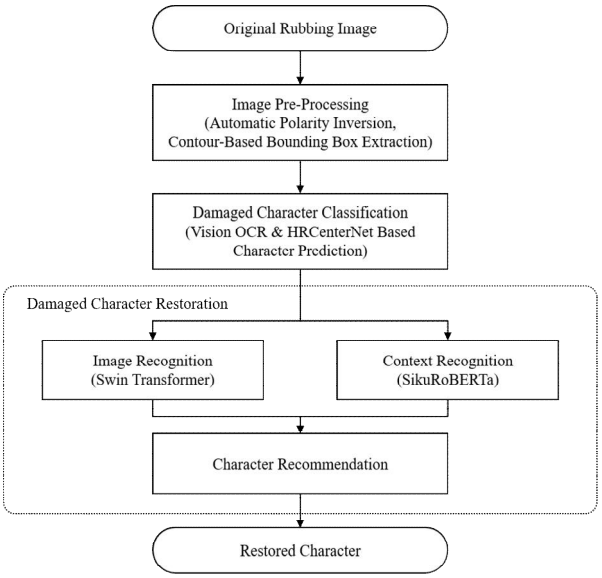

Fig. 1. Overall architecture of the proposed restoration and interpretation system

3.1 Image preprocessing

To preserve the unique characteristics of rubbing images and optimize the performance of restoration and recognition models, a four-stage preprocessing pipeline is employed, as shown in Fig. 2.

First, the input RGB rubbing image, shown in Fig. 2(a), is converted into a grayscale image to simplify the data dimensionality. Second, Otsu's binarization is applied to separate characters from the background. This effectively minimizes background noise and clearly delineates character regions, even within complex Ancient Korean character stroke structures. Third, automatic polarity adjustment is performed during the training of the character recognition engine and visual restoration model. This process ensures visual consistency across the dataset, as demonstrated in Fig. 2(b). Fourth, to remove non-character elements, a contour-based text bounding box extraction is performed to isolate the pure character region, as defined in (5).

$$BBox_{text} = BoundingRect(c_{i \in J}), \quad (5)$$

$$J = i | w_i h_i \geq A_{\min}$$

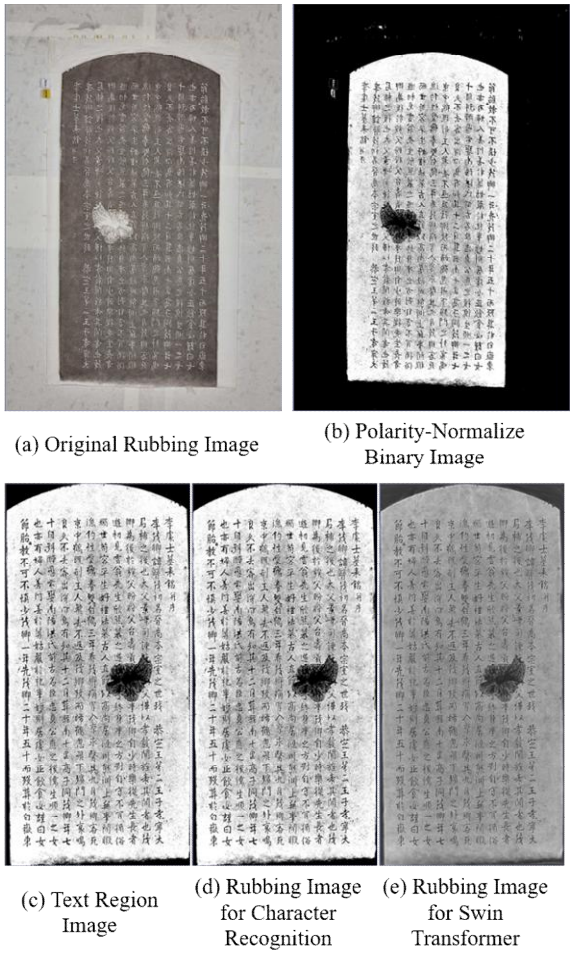

Fig. 2. Image preprocessing pipeline

In (5), C denotes the set of all contours extracted from the binarized image, where c_i represents the i-th individual contour. The area of each contour, calculated as the product of its width (w_i) and height (h_i), is compared with a minimum area threshold ($A_{\min}$) to eliminate non-character noise. J represents the set of indices for valid contours that satisfy this condition. The final text bounding box, $BBox_{text}$, is defined by calculating the minimum axis-aligned rectangle ($BoundingRect$) that encloses all contours belonging to the set J.

Subsequently, morphological operations are used to detect the rubbing area and select only the pure character region, as described in (6).

$$ROI_{char} = BBox_{text} \cap BBox_{rub} \quad (6)$$

In (6), ROI_{char} represents the final Region of Interest (ROI) used for character recognition. $BBox_{text}$ is the character region bounding box extracted via contour-based text detection, and $BBox_{rub}$ denotes the bounding box for the entire rubbing area detected through morphological operations. Defined by the intersection ($\cap$) of these two regions, ROI_{char} constitutes a pure character region with non-character backgrounds and unnecessary margins removed, as shown in Fig. 2(c). This refined output serves as the primary input for the character recognition model.

Through this four-stage preprocessing pipeline, input representations that allow for the stable learning of unique stroke thickness patterns are secured, establishing a foundation for the accurate classification and restoration of full-width Ancient Korean characters. Ultimately, the binarized and normalized results serve as inputs for Optical Character Recognition (OCR), with image Fig. 2(d) specifically optimized for character recognition. Meanwhile, images preserving the original visual texture, such as image Fig. 2(e), are utilized as inputs for the Swin Transformer V2-based visual restoration model.

3.2 Damaged character classification

To intuitively analyze character recognition results and damage conditions, this system utilizes a hybrid character recognition module combining Google Vision OCR and a Custom HRCenterNet. As shown in Fig. 3(a), the original rubbing image is used as the baseline for analysis. Furthermore, character regions are categorized into Fig. 3(b) 'Totally Lost,' Fig. 3(c) 'Partially Lost,' and Fig. 3(d) 'Undamaged' based on the severity of degradation.

For this purpose, the ink density is calculated for each detected character bounding box b, as defined in (7).

$$D(b) = N_{ink}(b)/N_{total}(b) \quad (7)$$

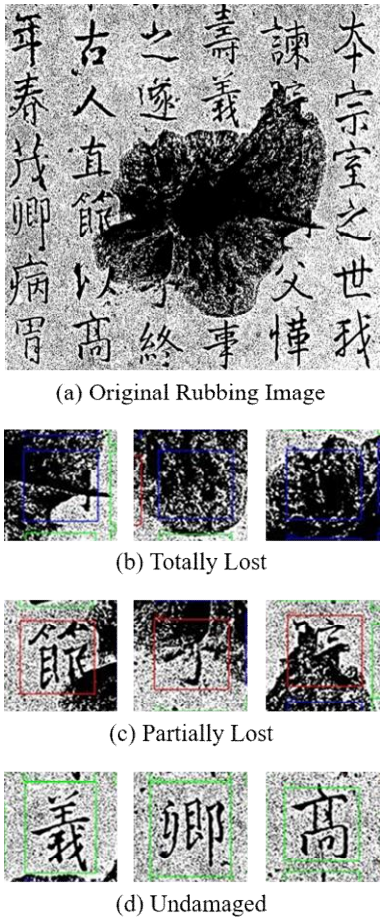

Fig. 3. Classification of character damage types: (a) Original rubbing image, (b) Totally lost, (c) Partially lost, and (d) Undamaged

Here, $N_{ink}(b)$ represents the number of foreground (black) pixels within the bounding box, and $N_{total}(b)$ denotes the total number of pixels in the bounding box. Based on the calculated ink density $D(b)$, the damage status of each character region is classified using an experimentally determined threshold of $\tau = 0.60$.

The classification criteria are as follows: First, Totally Lost characters are defined as regions where D\left(b\right)≥ τ. In these cases, the character outline has completely collapsed due to ink smudging or severe erosion, rendering the visual information unidentifiable. Second, Partially Lost characters are regions where $0.38 \leq D(b) < \tau$ but recognition confidence is low. These areas retain some strokes and partial morphological information, yet normal character recognition remains difficult. Third, the

remaining regions are classified as Undamaged, where the stroke structure is sufficiently preserved for the OCR engine to recognize the character normally.

By employing the proposed damage classification technique, the system preserves the structural position and damage status of unrecognized characters rather than simply discarding them. Specifically, Totally Lost regions, which lack significant visual information, are deemed suitable for context-based restoration and are routed exclusively to the SikuRoBERTa-based language model. Conversely, Partially Lost regions, which retain residual visual cues, are fed into both the Swin Transformer V2-based visual restoration model and the SikuRoBERTa-based language model to perform multimodal restoration, leveraging both morphological and contextual information.

3.3 Damaged character restoration

3.3.1 Image & context recognition

Based on the classified data, this study proposes a multimodal restoration architecture that fuses a Language Model and a Computer Vision (CV) model. To achieve precise restoration of damaged characters, the system integrally utilizes both the contextual information of the text and the structural features of the image. In this paper, 'Visual Restoration' does not imply the physical generation of new character forms; rather, it refers to the process of identifying the correct character by extracting structural visual cues from the damaged input. The detailed process is as follows.

The visual restoration model utilizes the Swin Transformer V2 as a backbone to restore damaged character regions. The entire inference process consists of three stages: input preprocessing, feature extraction, and probability calculation. First, the character patch extracted from the original image I and the bounding box b is resized to a fixed resolution 256 × 256 and normalized to generate the input tensor x_{in}, as shown in (8).

$$x_{in} = T_{norm}(T_{resize}(Crop(I,b))) \tag{8}$$

The generated tensor x_{in} is then input into the backbone function Φ_{Swin}, parameterized by θ_{v2} as described in Section II, and encoded into a high-dimensional feature map Z containing structural information of the character, as expressed in (9).

$$z = \Phi_{Swin}(xin;\theta_{v2}) \tag{9}$$

Finally, the feature map z passes through a classification head to be converted into a restoration probability distribution P_v is over all character classes, as shown in (10). Here, GAP denotes Global Average Pooling, which aggregates the spatial information of the feature map into a single feature vector by computing the average value of each channel. This vector is then projected into the class space via a linear layer with weights W_h and bias b_h. Based on this distribution, the system selects the top-k restoration candidates and forwards them to the ensemble module.

$$P_v = Softmax(W_h \cdot GAP(z) + b_h) \tag{10}$$

The subsequent language model component performs context analysis based on SikuRoBERTa. By analyzing the bidirectional context surrounding the position of missing characters, the model calculates the probability distribution of Ancient Korean characters that fit the historical and grammatical context, thereby contributing to the derivation of the final restoration result.

3.3.2 Character recommendation

To organically integrate the visual inference probability P_v analyzed by the CV model and the contextual suitability probability P_n derived by the

language model, this architecture adopts the harmonic mean-based F1-score[14] as the final restoration metric, as defined in (11).

$$Score_{final} = (2 \cdot P_n \cdot P_v)/(P_n + P_v) \quad (11)$$

Based on the calculated combined score ($Score_{final}$), the system integrally reflects the probabilistic evidence from both models to propose the optimal candidate character most likely to occupy the damaged region. Once the original text is finalized and restored through this process, it proceeds to the final translation stage to convey the textual meaning. The finally restored source text is translated into modern Korean using the Gemini-2.5-flash-lite model. To ensure professional quality and historical accuracy, an 'Expert in Translating Rubbings' persona was established for the model. The specific prompt configuration and constraints are detailed in Table 1. Furthermore, to ensure consistency in translation style and enhance the processing of archaic language, a 3-shot In-context Learning technique was applied, as shown in Table 2.

Table 1. Prompt strategy: instructional framework

Role: An expert translator proficient in the grammar and historical context of Classical Korean Rubbings.

Task: Analyze the provided **Hanja source text from a Korean rubbing** according to the following three steps

1. **[Korean Phonetic Reading]**: Convert every Hanja character into its Korean phonetic reading (Eumdok) without any spaces.
2. **[Proper Noun Extraction]**: Identify and extract **Korean personal names, official titles, geographic locations, and era name**s in the format: "Korean(Hanja)".
3. **[Final Translation]**: Perform a literal translation into an archaic Korean literary style (e.g., endings like ~하니라, ~하다), incorporating the extracted proper nouns.

Constraints:
- Do not attempt to guess or interpolate missing characters represented by the symbol "□".
You must preserve the exact number of "□" symbols and include them in every stage of the analysis and translation.

Table 2. In context learning: 3 shot example

Input:
惟昔始祖鄒牟王之創基也 出自北夫餘…
Output:
- **[Korean Phonetic Reading]**: 유석시조추모왕지창기야 출자북부여…
- **[Proper Noun Extraction]**: 시조(始祖), 추 모왕(鄒牟王), 북부여(北夫餘)
- **[Translation Result]**: 옛적 시조(始祖) 추모왕 (鄒牟王)이 나라를 세웠는데 북부여(北夫餘)에서 태어났으며…

The overall translation process begins with syntactic structure analysis to grasp the context of the source text, followed by key entity extraction to identify major proper nouns such as person names and location names. Subsequently, in the final translation stage, the system synthesizes the preceding information to generate a Korean translation grounded in the historical context. Notably, to preserve the damage state of the original text, strict constraints were imposed to maintain the exact count and positions of illegible symbols (e.g., □) without arbitrary estimation or deletion. This effectively suppresses hallucination phenomena caused by the model's arbitrary interpretation and secures the reliability of the translation.

IV. Experimental Result

4.1 Image recognition

4.1.1 Dataset and preprocessing

The dataset used in this study is based on the AI-HUB Old Book Hanja OCR dataset[27], a public

database released by the South Korean government. This dataset includes rubbing images along with bounding box coordinates and character labels for individual Hanja characters. A total of 11,416 rubbing images were utilized, from which approximately 2.02 million individual character instances were extracted. To ensure the reliability of the experiments, the dataset was partitioned into training, validation, and test sets at an 8:1:1 ratio. Furthermore, to guarantee reproducibility, a fixed seed value was used throughout all data splitting and training processes.

The rubbing data exhibits extreme variations in quality depending on the material of the stone monuments and the capturing environment. Additionally, damage to the character regions occurs frequently due to prolonged physical wear and environmental factors. Because it is difficult to obtain the original ground-truth images prior to the damage, synthetic training data was generated by applying advanced augmentation techniques to simulate realistic damage patterns. Specifically, six types of augmentations (three types of character damage and three types of background degradation) were integrated into the training data. However, to ensure that the reported performance metrics accurately reflect the model's ability to handle real-world degradations, all evaluations were performed solely on the original test set without any augmentations applied.

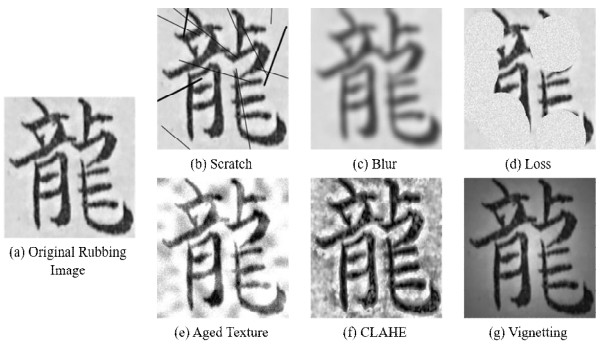

Fig. 4. Examples of the proposed degradation-based augmentation applied to stone-rubbing character images: (a) Original, (b) Scratch, (c) Blur, (d) Loss, (e) Aged texture, (f) CLAHE, and (g) Vignetting

Furthermore, to address these domain-specific challenges and ensure the consistency of the model inputs, ImageNet-based normalization and intensity polarity normalization were applied. These processes are presented in Equations (12) and (13).

$$I_{norm} = \frac{I - \mu_{ImageNet}}{\sigma_{ImageNet}} \tag{12}$$

$$I_{polar} = \begin{cases} 1 - I_{norm}, \text{if } mean(I_{norm}) > 0.5 \\ I_{norm}, otherwise \end{cases} \tag{13}$$

Here, $\mu_{ImageNet}$ and $\sigma_{ImageNet}$ denote the mean and standard deviation of the ImageNet dataset, used to standardize the data distribution. In the training phase, to go beyond preventing simple overfitting and to precisely simulate the complex degradation patterns found in actual rubbings, we applied Advanced Augmentation techniques that integrate physical damage and environmental distortion to the original images shown in Fig. 4(a)[28].

First, the types of Character Damage are as follows:

1. Scratch: As shown in Fig. 4(b), complete black (pixel value 0) straight lines with random angles are overlaid on the original image. This simulates a "Deep Cut" where stroke information is completely lost, rather than a transparent overlay, thereby enhancing the model's ability to restore disconnected strokes[29]. This is expressed in (14), where $\odot$ denotes the element-wise product, 1 is a matrix of ones, and M_{cut} is a binary mask representing the trajectory of the cut stroke.

$$I_{scratch} = I \odot (1 - M_{cut}) \tag{14}$$

2. Blur: As shown in Fig. 4(c), Gaussian Blur is applied to simulate the phenomenon where stroke boundaries become indistinct due to prolonged weathering[29]. This is defined in (15).

$$I_{blur}(x, y) = (I * G_{\sigma})(x, y) \tag{15}$$

3. Loss: As shown in Fig. 4(d), this simulates paper tearing or omission. The missing areas are filled not with a simple solid color but with High-frequency Noise to blend naturally with the surrounding texture[29].

$$I_{loss} = I \odot (1 - M_{miss}) + N_{hf} \odot M_{miss} \quad (16)$$

In (16), M_{miss} is a binary mask representing the irregular missing region, and N_{hf} denotes the high-frequency texture noise term used to fill the void.

Next, the types of Background Degradation are as follows: 4. Ancient Document Texture: As shown in Fig. 4(e), an aged texture synthesized with Gaussian noise and irregular spots is applied to reproduce the rough surface texture characteristic of traditional Korean paper (Hanji)[29].

$$I_{texture} = (1 - \lambda) \cdot I + \lambda \cdot (T_{ancient} + \eta) \quad (17)$$

In (17), $T_{ancient}$ represents the ancient document pattern, \eta is noise, and \lambda is weighting coefficients controlling the synthesis intensity.

5. Contrast Enhancement: As shown in Fig. 4(f), local contrast is forcibly maximized using Contrast Limited Adaptive Histogram Equalization (CLAHE) to simulate adverse conditions where background noise becomes more prominent than the strokes[30].

$$I_{\text{CLAHE}} = J_{\text{HE}}(I, limit, grid) \quad (18)$$

In (18), J_{HE} refers to the contrast-limited adaptive histogram equalization function, with $limit$ and $grid$ representing the contrast limit threshold and grid size, respectively.

6. Vignetting: As shown in Fig. 4(g), this applies optical distortion where the image periphery darkens due to uneven illumination[31].

$$I_{vignette}(x,y) = I(x,y)(1-\alpha \cdot \sqrt{(x-c_x)^2 + (y-c_y)^2}) \quad (19)$$

In (19), (c_x, c_y) are the center coordinates of the image, and \alpha is the vignetting intensity coefficient determining the degree of light fall-off at the edges. This integrated augmentation strategy ensures the model robustly extracts unique structural features of characters even when stroke information is physically severed or interfered with by background texture and lighting.

4.1.2 Hyperparameter tuning

The optimization settings adopted to ensure stable convergence and reproducibility of the fine-tuned Swin Transformer in a resource-constrained environment are detailed in Table 3.

Table 3. Hyperparameter settings for swin transformer fine-tuning

Category	Hyperparameter	Value
Input	Resolution	256 x 256
Train	Steps	19,720
	Effective batch	384
Optimizer	Type	AdamW
	Weight decay	0.05
Learning rate	Backbone LR	1e-5
	Head LR	3e-4
Schedule	Warm-up	5 step, start factor 0.1
	Cosine	T0=15, Tmult=2
Stability	Backbone freeze	first 3 epochs
	AMP	On
	Gradient clipping	1.0
Regularization	Drop-path	0.2

Additionally, to mitigate the weakening of training signals due to the long-tail class distribution, Weighted Cross-Entropy Loss was applied. Class weights were based on the inverse of the sample frequency for each class, with an exponential smoothing coefficient α=0.5 applied to prevent excessive weight bias. For stable training, weights were normalized by the mean of all class weights. The loss function and weight definition are given in (20).

$$L_{WCE} = -\frac{1}{N}\sum_{n=1}^{N} w_{y_n} \log(p_{y_n}^{(n)}), \quad (20)$$

$$w_c = \frac{(n_c)^{-\alpha}}{\frac{1}{c}\sum_{j=1}^{C}(n_j)^{-\alpha}}, \alpha = 0.5$$

4.1.3 Recognition results

The Macro-F1 performance of the Swin Transformer V2 model across training steps is presented in Fig. 5 and Table 4. The validation Macro-F1 score converges rapidly in the early stages and peaks at 19,720 training steps.

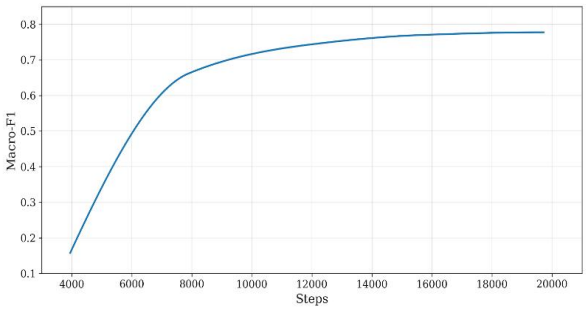

Fig. 5. Validation Macro-F1 score curve during training steps.

The experimental results show a rapid performance improvement from 0.1576 at the initial 3,944 steps to 0.6619 at 7,888 steps. Subsequently, the performance steadily increased, recording 0.7421 at 11,832 steps and 0.7705 at 15,776 steps, finally achieving a peak Macro-F1 of 0.7777 at 19,720 steps. This indicates that the model entered a stable convergence phase around the 0.77 level after approximately 15,776 steps. Furthermore, the final recognition performance for damaged character patches recorded a Top-1 Accuracy of 96.63% and a Top-5 Accuracy of 99.38%. Considering the difficulty of rubbing data mixed with complex physical damage and environmental degradation, the Top-5 accuracy of 99.38% demonstrates that the proposed model generates candidate groups containing the ground truth character with high reliability and stability.

Table 4. Validation Macro-F1 score by training steps

Steps	Macro-F1 score
3,944	0.1576
7,888	0.6619
11,832	0.7421
15,776	0.7705
19,720	0.7777

4.2 Context recognition

4.2.1 Dataset and preprocessing

This study utilized a total of 6,751 rubbing decipherment records (approximately 3.84 million characters) collected from four major institutions: the National Institute of Korean History, Korea Heritage Service, Kyujanggak Institute for Korean Studies, and National Research Institute of Cultural Heritage. To ensure data consistency, preprocessing steps such as normalizing unknown character markers (e.g., □), removing unnecessary annotation symbols, and excluding non-Hanja data (Hangul, Hiragana) were performed.

The average length of the collected rubbing texts is 693 characters, which significantly exceeds the maximum input length of 512 tokens for BERT-based models. To prevent information loss in long sequences and preserve the full context, the text was reconstructed using a SikuRoBERTa-based punctuation restoration model[32], which segments the unpunctuated Hanja text into logical units. Given a document D composed of a character sequence $D = \{c_0, c_1, \ldots, c_{L-1}\}$, we applied a Sliding Window technique with window size $W = 400$, overlap $O = 100$, and stride $S = 300$. The i-th window segment $input_i$ is defined as in (21).

$$Input_i = \{c_k | i \cdot S \le k < \min(i \cdot S + W, L)\} \quad (21)$$

The model M predicts the punctuation tag sequence

for each window c_k. Since a character $Input_j$ within the overlap region is predicted by multiple windows, we collect the set of predictions P_k for that position as shown in (22). The final punctuation y_k is defined in (23) by applying majority voting to the collected set over the label set L, which consists of none, comma(,), period(。), question mark(?).

$$P_k = \{M(Input_j)_k | window\ j\ covers\ c_k\} \quad (22)$$

$$y_k = \underset{l \in L}{\arg max} |\{p \in P_k | p = l\}| \quad (23)$$

By segmenting sentences based on restored periods and question marks, the average sentence length was adjusted to approximately 25 characters, resulting in a dataset of 228,630 sentences. This distribution not only satisfies the BERT input constraints but also forms an optimal structure for efficient context learning.

4.2.2 Comparative evaluation of context models

To select the optimal base model capable of effectively reflecting the characteristics of the rubbing domain, we conducted comparative experiments on three models pre-trained on different corpora: SikuRoBERTa[23], HUE[33], and SillokBERT[34]. Notably, SillokBERT was trained on the AJD, and HUE utilized both the AJD and The Diaries of the Royal Secretariats (DRS). Both models share the commonality of being built on Korean Hanja corpora from the Joseon Dynasty. First, analyzing the number of Unknown Tokens to compare the vocabulary coverage of each model's tokenizer revealed that SikuRoBERTa recorded the lowest count of 1,903 tokens, indicating its superior suitability for processing rubbing characters(Table 5).

The results of evaluating MLM performance in a Zero-shot environment prior to fine-tuning are shown in Table 6. While HUE and SillokBERT showed high performance around 90% on the AJD dataset, their performance dropped sharply to the 40% range on rubbing data, revealing vulnerability to domain shift. In contrast, SikuRoBERTa recorded the highest Top-5 accuracy on rubbing texts among the comparison models.

Table 5. Number of unknown tokens by model

Model	Number of unknown tokens
SikuRoBERTa	1,903
HUE	2,927
SillokBERT	6,055

Table 6. Top-5 accuracy comparison across models (zero-shot)

Dataset	SikuRo BERTa	HUE	SillokBERT
AJD	40.8%	90.3%	87.6%
Rubbing transcription	55.6%	48.2%	47.4%

To optimize the performance of SikuRoBERTa, we conducted stepwise comparative experiments setting Input Unit and Data Duplication as key variables. Input units were categorized into Document level (using the entire text) and Sentence level (based on punctuation). Detailed results are presented in Table 7. Document-level training (Exp 1, 2) showed minimal performance improvement due to truncation caused by the input limit (512 tokens). Conversely, when segmented into sentences with an optimized input length of 128 tokens and allowing data duplication (Exp 3), performance significantly improved to 76.05%. Furthermore, applying the sliding window technique for context preservation (Exp 4) achieved the optimal restoration performance with a Top-5 Accuracy of 82.56%. This improvement is attributed to the sliding window technique effectively mitigating the 'Boundary Effect'—where contextual continuity is severed during simple sentence segmentation—through window overlapping. Specifically, this approach ensures that tokens located at the segment boundaries are

processed within sufficient bidirectional context. Furthermore, the majority voting mechanism applied to the redundantly predicted regions induces an ensemble effect that cancels out local prediction errors from individual windows, thereby significantly enhancing the reliability of the final restoration. These results demonstrate that the combination of length optimization and context reinforcement is critical for enhancing model performance.

Table 7. Comparative analysis of model performance based on preprocessing strategies

Exp	Input Unit	Data Duplication	Max Length	Top-5 Acc(%)
1	Document	Excluded	512	70.44
2	Document	Included	512	71.03
3	Sentence	Included	128	76.05
4	Windowed Sentence	Included	128	82.56

4.2.3 Hyperparameter tuning

To enable the model to sufficiently grasp the context surrounding the target sentence (S_{tgt}), the input sequence I was constructed by concatenating ($\oplus$) the previous context (C_{prev}) and the next context (C_{next}). The final input vector configuration including special tokens ([CLS], [SEP]) is defined in (24).

$$I = [CLS] \oplus C_{prev} \oplus S_{tgt} \oplus C_{\neq xt} \oplus [SEP] \quad (24)$$

Here, the total sequence length $|I|$ is limited to a maximum of 128 tokens ($|I| \leq 128$). The dataset consists of 228,630 sentences, randomly split into Training, Validation, and Test sets in an 8:1:1 ratio to ensure experimental reliability. The number of sentences in each dataset is shown in Table 8.

To process unique characters specific to the rubbing dataset, new tokens were added to the existing vocabulary, expanding the tokenizer to 31,885 tokens, and the model's embedding layer was reconstructed. The specific hyperparameter settings used for training are detailed in Table 9.

Table 8. Strategies statistics of the rubbing dataset split

Dataset	Number of sentence
Training set (80%)	182,904
Validation set (10%)	22,863
Test set (10%)	22,863
Total set (100%)	228,630

Table 9. Hyperparameters for SikuRoBERTa training

Hyperparameter	Value
Learning rate	2e-5
Weight decay	0.01
Scheduler type	Linear warmup
Warmup ratio	0.06
Batch size	32
Max epochs	20
Early stopping	3 Epochs (Patience)

4.2.4 Context recognition results

The changes in training and validation loss throughout the learning process are illustrated in Fig. 6. While the initial loss exceeded 6.0, it exhibited a sharp decline, recording 2.3441 (train) and 2.3102 (validation) by the end of the first epoch. The losses continued to converge consistently, finally reaching 1.6065 and 1.5973 at step 114,320. As training progressed, the rate of loss reduction plateaued, and the gap between the two metrics remained narrow, confirming that the model was stably optimized without overfitting. The context inference performance of the final model is presented in Table 10. The experimental results show that the model achieved a Top-1 Accuracy of 68.22%, representing single prediction accuracy, and a Top-5 Accuracy of 82.56%, indicating the probability that the correct answer is included within the top 5 predicted candidates.

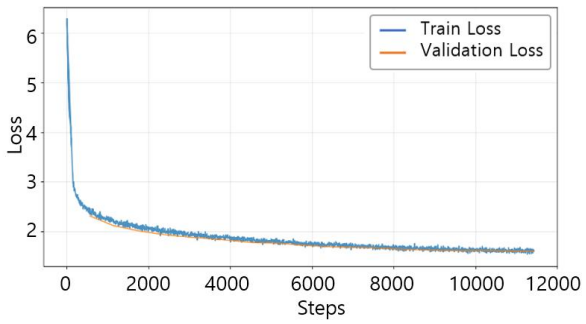

Fig. 6. Context recognition results

Table 10. Final performance results of the SikuRoBERTa model

Metric	Value
Test loss	1.5714
Test perplexity	4.8163
Top-1 accuracy	68.22%
Top-5 accuracy	82.56%

4.3 Damaged character restoration performance

4.3.1 Damaged character prediction

To quantitatively verify the performance of the multimodal restoration framework, a multimodal simulation was conducted in a controlled environment where the ground truth is clearly defined. Specifically, a target character was selected from a pristine rubbing image, and an actual damaged region extracted from another rubbing was overlaid on that position to construct an artificially damaged environment. At this juncture, to minimize the discrepancy between the artificially induced damage and real-world conditions, we applied the degradation-based augmentation techniques detailed in Section IV-A. By rigorously simulating not only stroke disconnections but also the texture distortions and noise patterns characteristic of actual rubbings, we effectively compensated for the limitations of artificial manipulation and ensured the realism of the simulation data.

This approach allowed for the objective evaluation of the model's restoration performance while reflecting the actual stroke loss and texture distortion of rubbings, utilizing known ground truth characters. The overall workflow of the simulation is illustrated in Fig. 7.

The simulated damaged rubbing image was segmented into characters, and the same damaged position was fed into both the context-based and visual-based models. Each model output the top-5 restoration candidates along with their probability values for the damaged character. The comparison of the predictions confirmed the existence of common candidate characters. Subsequently, utilizing the F1-score-based probability fusion method adapted from previous studies like Wang et al. [14], the final confidence score integrating both contextual and stroke information was calculated.

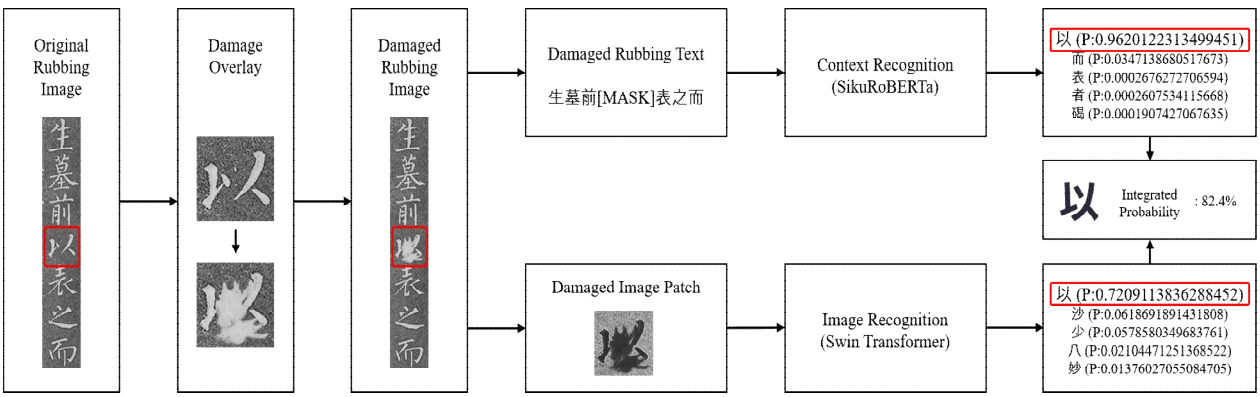

Fig. 7. Workflow of the integrated restoration simulation

As a result, both context-based and visual-based recognition predicted the correct character as the top candidate. The combined final probability was calculated at 82.4%, matching the actual ground truth character. This demonstrates that, unlike single-modality approaches, the multimodal approach—which combines complementary information based on damage types—operates effectively even in complex damage environments. Furthermore, these simulation results suggest that the proposed framework can provide stable character restoration performance in actual cultural heritage restoration scenarios.

4.3.2 Translation performance evaluation

To verify the performance of the translation prompts from multiple perspectives, quantitative evaluations were performed using Noun-focused BLEU to measure morphological consistency and BERTScore to evaluate contextual similarity. First, to measure the transmission of key information such as person names, place names, and official titles, nouns were extracted using the Kiwi morphological analyzer, and the Corpus BLEU was calculated. The result showed a high accuracy of 44.42. This indicates that the "explicit entity extraction" step within the prompt strategy effectively prevented the distortion or omission of key information during the translation process. Next, to verify the preservation of contextual meaning, BERTScore was calculated using mDeBERTa-v3-base and KLUE-BERT as backbones. BERTScore is calculated based on the cosine similarity between token embeddings of the source text and the generated text, as represented by the recall metric R_{BERT} in (25).

$$R_{BERT} = \frac{1}{m}\sum_{j=1}^{m} \max_{1 \le i \le n} \cos(x_i, y_j) \tag{25}$$

The final performance evaluation utilizes F_{BERT}, which is the harmonic mean of precision P_{BERT} and recall R_{BERT}, as defined in (26). This metric provides a balanced assessment of the generated translation by simultaneously accounting for both accuracy (precision) and completeness (recall) regarding the source text.

$$F_{BERT} = 2 \cdot (P_{BERT} \cdot R_{BERT}) / (P_{BERT} + R_{BERT}) \tag{26}$$

The evaluation resulted in high scores of 89.04 with mDeBERTa and 81.35 with KLUE-BERT. These figures indicate that the generated translations maintain a high level of structural and contextual meaning relative to the original text.

4.4 Implementation of explainable restoration system

The implementation results of the proposed system for restoring and interpreting damaged characters in rubbings are shown in Fig. 8. The system is designed considering practical utility in actual rubbing research environments. By visually presenting the damage ratio, distribution of restoration targets, and overall character status for the input rubbing image, researchers can intuitively grasp the document's degradation state. The interface showing the distribution of characters requiring restoration is presented in Fig. 8. Restoration target characters are highlighted along with their positional information in the text. Users can efficiently inspect damaged regions via the main text view or the damaged character list button. For each damaged character, context-based and stroke-based candidate characters are presented in order of priority according to the damage type, and the restoration result is finalized based on the user's selection. Simultaneously,

inspection progress and confidence statistics are updated in real-time, providing immediate feedback on the workflow.

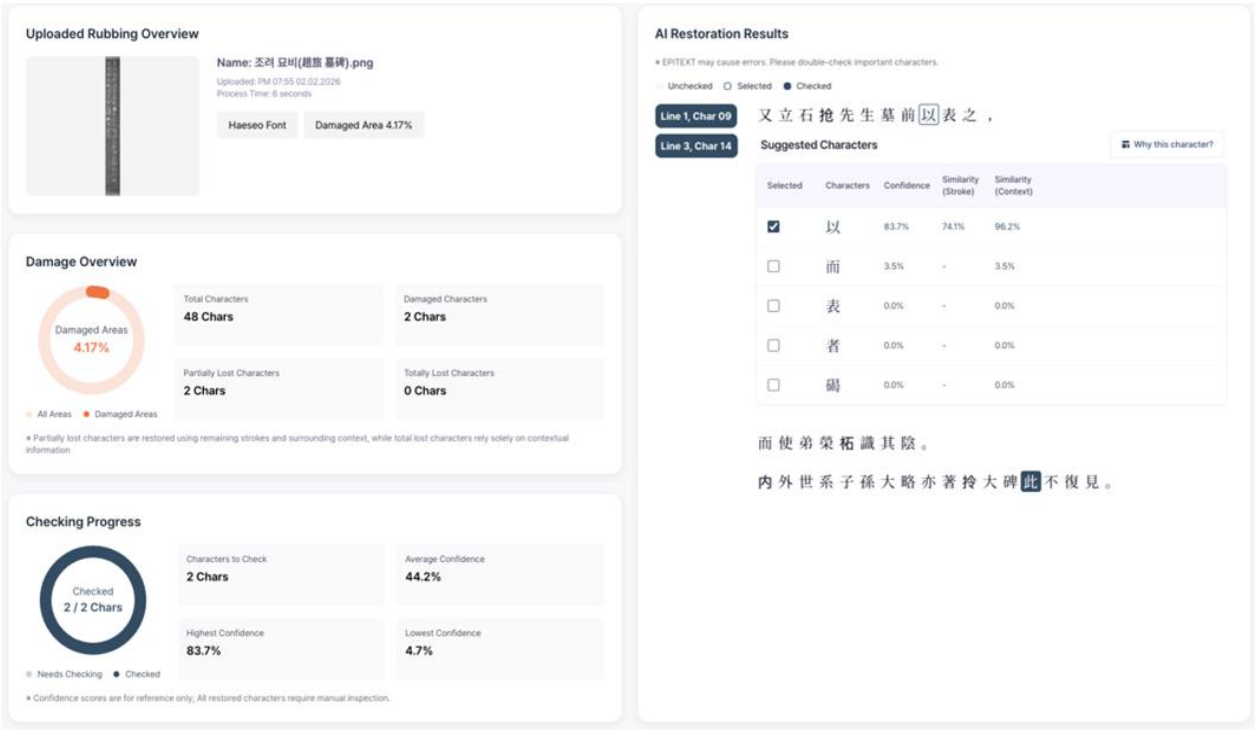

Fig. 8. System interface showing restoration distribution and status

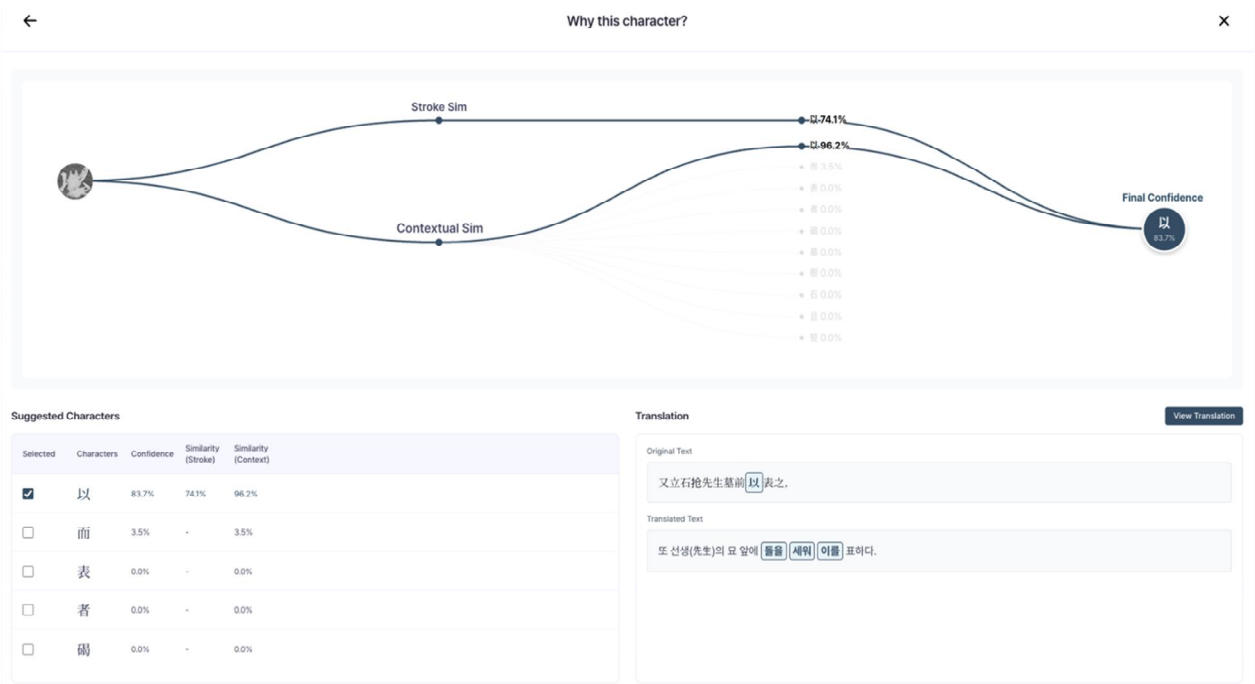

Fig. 9. System interface providing translation and restoration rationale

To clarify the rationale behind restoration decisions and ensure system transparency, this study integrates XAI features into the user interface[24]-[26]. Specifically, the probabilistic contributions of context consistency and stroke consistency are visualized using mechanisms such as Attention Maps and interpretable confidence breakdown tables, as shown in Fig. 9. This enables researchers to intuitively comprehend the AI's

decision-making process. Based on these interpretability cues, users can verify in real-time how the semantic interpretation changes depending on the selected candidate character. This establishes a 'Human-in-the-loop' architecture, allowing domain experts to identify and correct potential model errors in the final stage.

Ⅴ. Conclusion

In this paper, we proposed an integrated framework for restoring damaged rubbings based on image and context information. Unlike traditional manual restoration methods, the proposed framework utilizes OCR technology to automatically detect character regions and assess damage status. By organically combining visual restoration and contextual reasoning according to the damage type, the system provides restoration results and translation information applicable to the entire decipherment and interpretation process in an automated manner.

The main contributions of this study are as follows. First, we established an end-to-end pipeline that automatically recognizes and deciphers character regions from input rubbing images without manual preprocessing. Second, we ensured system transparency by integrating XAI features. For damaged regions detected during the decipherment process, the system presents suitable Hanja candidates along with visual rationales (e.g., Attention Maps) in addition to confidence scores. This implementation establishes a 'Human-in-the-loop' environment where researchers can verify and refine the AI's suggestions, thereby significantly enhancing the objectivity and academic utility of the restoration results.

In conclusion, the automated rubbing decipherment and restoration framework proposed in this paper is expected to have significant academic, economic, and social impacts. Academically, the restored high-resolution rubbing data can serve as reliable primary sources for analyzing historical backgrounds in history, tracing the evolution of calligraphic styles in calligraphy, and empirically studying vocabulary and orthography in Korean linguistics. Economically and socially, replacing repetitive manual tasks requiring specialized labor with automated technology can drastically reduce work time and labor costs. This contributes to lowering the costs of establishing large-scale rubbing archives and managing cultural heritage in the long term. Furthermore, as a core technology scalable to various cultural property restoration fields, it is expected to enhance the sustainability of cultural heritage preservation.

However, the proposed model may exhibit reduced restoration accuracy when encountering highly unique calligraphic styles or rare Hanja characters that are not sufficiently represented in the training dataset. In future work, to address these limitations, we plan to go beyond the current character-level (code-level) restoration and conduct research on utilizing generative models to sophisticatedly generate and restore the visual shape (glyph-level) of missing Ancient Korean characters, aiming to achieve the complete restoration of original rubbings

References

[1] Z. Meng, Z. Zhang, Z. Man, H. Tomiyama, and L. Meng, "Rubbing character recognition with machine learning", Proc. Int. Conf. Adv. Mechatron. Syst. (ICAMechS), Hanoi, Vietnam, pp. 290-295, Dec. 2020. https://doi.org/10.1109/ICAMechS49982.2020.9310165.

[2] B. Lyu, A. Tanaka, and L. Meng, "Computer-assisted Ancient Documents Re-organization", Procedia Comput. Sci., Vol. 202, pp. 295-300, Mar. 2022. https://doi.org/10.1016/j.procs.2022.04.039.

[3] N. Wang, W. Wang, B. Li, H. Zhang, Q. Jiao, and C. Liu, "Multi-modal ancient scripts

recognition via deep learning with data homogenization and augmentation", Heritage Sci., Vol. 13, No. 1, Art. no. 522, Oct. 2025. https://doi.org/10.1038/s40494-025-02095-x.

[4] I. N. Gotsikas, Y. Z. Tzifopoulos, and P. A. Mitkas, "Digital Ektypon: Using an RTI dome for digitizing squeezes and epigraphic research", IEEE Access, Vol. 13, pp. 25154-25162, Jan. 2025. https://doi.org/10.1109/ACCESS.2025.3531991.

[5] H. U. Heo, "Methodological research on digital rubbing of artifacts", Komunhwa, No. 102, pp. 131-153, Dec. 2023. https://doi.org/10.61130/kmh.2023.102.131.

[6] M. A. Souibgui and Y. Kessentini, "DE-GAN: A conditional generative adversarial network for document enhancement", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 44, No. 3, pp. 1180-1191, Mar. 2022. https://doi.org/10.1109/TPAMI.2020.3022406.

[7] B. Su, X. Liu, W. Gao, Y. Yang, and S. Chen, "A restoration method using dual generate adversarial networks for Chinese ancient characters", Visual Informatics, Vol. 6, No. 1, pp. 26-34, Mar. 2022. https://doi.org/10.1016/j.visinf.2022.02.001.

[8] W. Zheng, B. Su, R. Feng, X. Peng, and S. Chen, "EA-GAN: Ancient books text restoration model based on example attention", Research Square, pp. 1-14, Oct. 2022. https://doi.org/10.21203/rs.3.rs-2131172/v1.

[9] N. S. Rani, B. J. B. Nair, M. Chandrajith, G. H. Kumar, and J. Fortuny, "Restoration of deteriorated text sections in ancient document images using a tri-level semi-adaptive thresholding technique", Automatika, Vol. 63, No. 2, pp. 378-398, Feb. 2022. https://doi.org/10.1080/00051144.2022.2042462.

[10] T. Shen, V. Quach, R. Barzilay, and T. Jaakkola, "Blank Language Models", arXiv preprint, arXiv:2002.03079, pp. 1-13, Nov. 2020. https://doi.org/10.48550/arXiv.2002.03079.

[11] E. Fetaya, Y. Lifshitz, E. Aaron, and S. Gordin, "Restoration of fragmentary Babylonian texts using recurrent neural networks", Proc. Natl. Acad. Sci. U.S.A., Vol. 117, No. 37, pp. 22743-22751, Sep. 2020. https://doi.org/10.1073/pnas.2003794117.

[12] Y. Assael, T. Sommerschield, and J. Prag, "Restoring ancient text using deep learning: A case study on Greek epigraphy", arXiv preprint, arXiv:1910.06262, pp. 1-9, Oct. 2019. https://doi.org/10.48550/arXiv.1910.06262.

[13] K. Kang, K. Jin, S. Yang, S. Jang, J. Choo, and Y. Kim, "Restoring and mining the records of the Joseon Dynasty via neural language modeling and machine translation", Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. (NAACL-HLT), Online, pp. 4031-4042, Jun. 2021. https://doi.org/10.18653/v1/2021.naacl-main.317.

[14] Z. Wang, Y. Li, and H. Li, "Chinese inscription restoration based on artificial intelligent models", npj Heritage Science, Vol. 13, Art. no. 326, Jul. 2025. https://doi.org/10.1038/s40494-025-01900-x.

[15] D. Cheng, X. Li, W.-H. Li, C. Lu, F. Li, H. Zhao, and W.-S. Zheng, "Large-scale visible watermark detection and removal with deep convolutional networks", Proc. 1st Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV), Guangzhou, China, pp. 27-40, Nov. 2018. https://doi.org/10.1007/978-3-030-03338-5_3.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", Proc. Int. Conf. Learn. Represent. (ICLR), Online, pp. 1-10, Jan. 2021. https://doi.org/10.48550/arXiv.2010.11929.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer:

Hierarchical Vision Transformer using Shifted Windows", arXiv preprint, arXiv:2103.14030, pp. 1-14, Aug. 2021. https://doi.org/10.48550/arXiv.2103.14030.

[18] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin Transformer V2: Scaling Up Capacity and Resolution", arXiv preprint, arXiv:2111.09883, pp. 1-15, Apr. 2022. https://doi.org/10.48550/arXiv.2111.09883.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need", arXiv preprint, arXiv:1706.03762, pp. 1-15, Jun. 2017. https://doi.org/10.48550/arXiv.1706.03762

[20] G. Sun, Z. Zheng, and M. Zhang, "End-to-End Rubbing Restoration Using Generative Adversarial Networks", arXiv preprint, arXiv:2205.03743, pp. 1-8, Oct. 2022. https://doi.org/10.48550/arXiv.2205.03743.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv preprint, arXiv:1810.04805, pp. 1-16, May 2019. https://doi.org/10.48550/arXiv.1810.04805.

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach", arXiv preprint, arXiv:1907.11692, pp. 1-13, Jul. 2019. https://doi.org/10.48550/arXiv.1907.11692.

[23] D. Wang, C. Liu, Z. Zhu, J. Liu, H. Hu, S. Shen, and B. Li, "Construction and application of pre-trained models of Siku Quanshu in orientation to digital humanities", Lib. Trib., Vol. 42, No. 6, pp. 31-43, Jun. 2022.

[24] T. Martins, A. M. de Almeida, E. Cardoso, and L. Nunes, "Explainable artificial intelligence (XAI): A systematic literature review on taxonomies and applications in finance", IEEE Access, Vol. 12, pp. 618-629, Dec. 2023. https://doi.org/10.1109/ACCESS.2023.3347028.

[25] K. Kalasampath, K. N. Spoorthi, S. Sajeev, S. S. Kuppa, K. Ajay, and A. Maruthamuthu, "A literature review on applications of explainable artificial intelligence (XAI)", IEEE Access, Vol. 13, pp. 41111-41140, Feb. 2025. https://doi.org/10.1109/ACCESS.2025.3546681.

[26] R.-K. Sheu, M. S. Pardeshi, K.-C. Pai, L.-C. Chen, C.-L. Wu, and W.-C. Chen, "Interpretable classification of pneumonia infection using eXplainable AI (XAI-ICP)", IEEE Access, Vol. 11, pp. 28896-28919, Mar. 2023. https://doi.org/10.1109/ACCESS.2023.3255403.

[27] AI-Hub, "Ancient Chinese Character Recognition Data", https://www.aihub.or.kr. [accessed: Jan. 30, 2026].

[28] W. Zeng, "Image data augmentation techniques based on deep learning: A survey", Math. Biosci. Eng., Vol. 21, No. 6, pp. 6190-6224, Jun. 2024. https://doi.org/10.3934/mbe.2024272.

[29] U. Saha, S. Saha, S. A. Fattah, and M. Saquib, "Npix2Cpix: A GAN-based image-to-image translation network with retrieval-classification integration for watermark retrieval from historical document images", IEEE Access, Vol. 12, pp. 95857-95870, Jul. 2024. https://doi.org/10.1109/ACCESS.2024.3424662.

[30] A. Khozaimi, I. Darti, S. Anam, and W. M. Kusumawinahyu, "Optimized Pap Smear Image Enhancement: Hybrid PMD Filter-CLAHE Using Spider Monkey Optimization", arXiv preprint, pp. 1-11, Feb. 2025. https://doi.org/10.48550/arXiv.2502.15156.

[31] B. Tian, F. Juefei-Xu, Q. Guo, X. Xie, X. Li, and Y. Liu, "AVA: Adversarial Vignetting Attack against Visual Recognition", arXiv preprint, arXiv:2105.05558, pp. 1-8, May 2021. https://doi.org/10.48550/arXiv.2105.05558.

[32] S. Song, H. Yoo, J. Jin, K. Cho, and A. Oh, "HERITAGE: An End-to-End Web Platform for Processing Korean Historical Documents in Hanja", arXiv preprint, arXiv:2501.11951, pp. 1-8, Jan. 2025. https://doi.org/10.48550/arXiv.2501.11951.

[33] H. Yoo, J. Jin, J. Son, J. Y. Bak, K. Cho, and A. Oh, "HUE: Pretrained Model and Dataset for Understanding Hanja Documents of Ancient Korea", Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Findings (NAACL), Seattle, WA, USA, pp. 1832-1844, Jul. 2022. https://doi.org/10.18653/v1/2022.findings-naacl.140.

[34] ddokbaro, "SillokBert: A Language Model for Veritable Records of the Joseon Dynasty", https://huggingface.co/ddokbaro/SillokBert. [accessed: Jan. 30, 2026].

Authors

Yejin Jhang

2025. 12 : Completed Korea University Intelligent Information SW Academy, 7th Cohort (640H)
2026. 2 : BS degree, Dept. of Korean History, Korea University
2026. 3 ~ Present: M.S. Candidate, Dept. of Culture Technology, Graduate School of Culture Technology, KAIST
Research Interests: Natural Language Processing, Machine Learning, Deep Learning

Donghyun Hwang

2018. 2 : BS degree, Dept. of Nursing, Seoul Women's College of Nursing
2025. 12 : Completed Korea University Intelligent Information SW Academy, 7th Cohort (640H)
Research Interests: Natural Language Processing, Medical Data Analysis, Machine Learning, Deep Learning

Junhyoung Choi

2025. 2 : BS degree, Dept. of Industrial Systems Engineering, Dongguk University
2025. 12 : Completed Korea University Intelligent Information SW Academy, 7th Cohort (640H)
Research Interests: AI-UX, Data-Driven UX, User Experience Design

Seongmin An

2025. 12 : Completed Korea University Intelligent Information SW Academy, 7th Cohort (640H)
2019. 3 ~ Present: BS Candidate, Dept. of IT Mechatronics Engineering, Jeonbuk National University
Research Interests: Data Analysis, Deep Learning, Natural Language Processing, Computer Vision

Gilsang Yoo

2010. 2 : Phd degree, Department of Imaging Engineering, Chungang University
2010. 3 ~ Present : Director, Korea Computer Game Society
2011. 3 ~ Present : Professor, Dept. of Creative Informatics & Computing Institute, / Intelligent Information Software Academy, Korea University
2023. 3 ~ Present : Senior Vice President, Korea Media Art Industry Association
Research interests : Data Science, 3D Content, Machine Learning, Deep Learning, Computer Education