

Reviving Cultural Heritage: A Novel Approach for Comprehensive Historical Document Restoration

Yuyi Zhang^{1,3} Peirong Zhang^{† 1} Zhenhua Yang^{† 1} Pengyu Yan^{1,3} Yongxin Shi¹
Pengwei Liu^{2,3} Fengjun Guo^{2,3} Lianwen Jin^{* 1,3,4}

¹South China University of Technology

²Intsig Information Co., Ltd.

³INTSIG-SCUT Joint Lab on Document Analysis and Recognition

⁴SCUT-Zhuhai Institute of Modern Industrial Innovation

yuyi.zhang11@foxmail.com eelwj@scut.edu.cn

Abstract

Historical documents represent an invaluable cultural heritage, yet have undergone significant degradation over time through tears, water erosion, and oxidation. Existing Historical Document Restoration (HDR) methods primarily focus on single modality or limited-size restoration, failing to meet practical needs. To fill this gap, we present a full-page HDR dataset (FPHDR) and a novel automated HDR solution (AutoHDR). Specifically, FPHDR comprises 1,633 real and 6,543 synthetic images with character-level and line-level locations, as well as character annotations in different damage grades. AutoHDR mimics historians' restoration workflows through a three-stage approach: OCR-assisted damage localization, vision-language context text prediction, and patch autoregressive appearance restoration. The modular architecture of AutoHDR enables seamless human-machine collaboration, allowing for flexible intervention and optimization at each restoration stage. Experiments demonstrate AutoHDR's remarkable performance in HDR. When processing severely damaged documents, our method improves OCR accuracy from 46.83% to 84.05%, with further enhancement to 94.25% through human-machine collaboration. We believe this work represents a significant advancement in automated historical document restoration and contributes substantially to cultural heritage preservation. The model and dataset are available at <https://github.com/SCUT-DLVCLab/AutoHDR>.

1 Introduction

Historical documents, encompassing books, rubbings, scrolls, and inscriptions, stand as a vital window into ancient civilizations and wisdom. Through the ages, they have sustained deterioration from various environmental factors, such as

[†]Equal contribution

^{*}Corresponding authors.

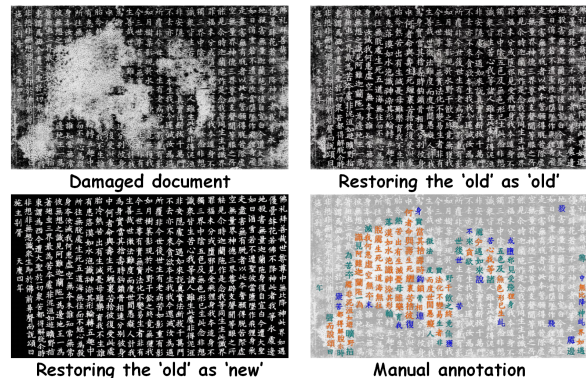


Figure 1: Restoration results of our AutoHDR. Orange, green, and blue indicate severe, medium, and light damage, respectively.

improper storage, transportation, and wartime upheavals, resulting in physical damage, water erosion, and oxidation. Therefore, restoring these ancient treasures is crucial to preserving their cultural and historical significance.

Yet, the task of Historical Document Restoration (HDR), remains complex and time-consuming. Traditional manual restoration involves three key stages: (1) identifying damaged regions through specialized knowledge and historical literature, (2) reconstructing damaged content based on literature references, and (3) applying delicate conservation techniques to restore the documents' original appearance. To alleviate the huge labor cost, various automated HDR techniques have been proposed. For example, Assael et al. (2022) employs a Transformer to predict damaged text, geographic origins, and dates. Yang et al. (2025) restores historical appearances through manually provided annotations. However, existing methods confront several critical limitations. (1) Most methods are confined to single-modal restoration (text- or image-only). (2) While some multimodal approaches are proposed, they are restricted to processing damage in very small regions, such as single image patches or a few characters. (3) The limited perceived region leads to two cascading problems: i) models fail to lever-

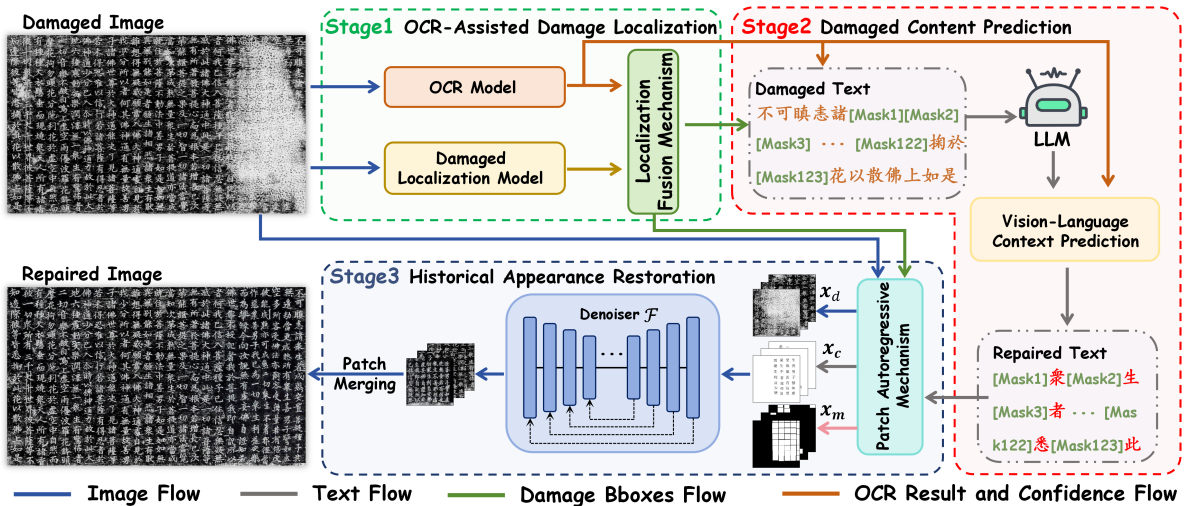


Figure 2: Overall workflow of the proposed AutoHDR. The framework contains three distinct yet interconnected stages: OCR-Assisted Damage Localization for character recognition and damage localization, Damaged Content Prediction for text restoration, and Historical Appearance Restoration for pixel-level reconstruction.

age broader contextual semantic information, and ii) restoration often fails when the damage extends beyond the model’s processing patch size. (4) Most importantly, current methods only address single stages of the restoration process, failing to provide a fully automated solution for all stages. Thus, manual intervention is still required in the restoration procedure, preventing the complete liberation of human experts from this demanding workload.

To address these challenges, we propose **AutoHDR**, a novel fully Automated solution for full-page HDR. As shown in Fig. 2, AutoHDR mirrors the workflow of historians, jointly restoring text and historical appearance in three stages: OCR-assisted damage localization, damaged content prediction, and historical appearance restoration. By combining OCR-assisted visual information with the language understanding capabilities of LLMs, AutoHDR achieves precise localization and restoration of damaged text. Then, adhering to the principle of "restoring the old as the old"¹ (Du, 1999; Wang, 2021), we design a patch autoregressive restoration approach to reconstruct the original document appearance at the page level, which conducts progressive restoration from simple to complex cases to ensure high fidelity. To our surprise, as depicted in Fig. 1, AutoHDR not only excels at "restoring the old as old", but also extends to "restoring the old as new"², providing higher flexibility to users. Note that the whole procedure is completely automatic, eliminating the need for

¹This means that the repaired text and background appear consistent with the condition of the ancient document.

²This means that the repaired text and background appear as if they are in a new and pristine condition.

human intervention. Furthermore, the modular architecture of AutoHDR enables seamless human-machine collaboration, allowing for flexible intervention and optimization at each restoration stage.

Subsequently, given the scarcity of HDR datasets and the limited focus on patch-level restoration of existing ones (Zhu et al., 2024; Yang et al., 2025), we introduce **FPHDR**, a pioneer dataset for Full-Page HDR. It includes 1,633 expertly annotated real samples and 6,543 high-quality synthetic samples, each providing character- and line-level location as well as character annotations in different damage grades, serving as a comprehensive benchmark for HDR model training and evaluation.

Extensive experiments are conducted to evaluate AutoHDR’s performance, which reveals its remarkable advantages over existing methods in both text restoration accuracy and historical appearance preservation. For severely damaged documents where OCR recognition accuracy starts at merely 46.83%, AutoHDR substantially improves the accuracy to 84.05%. Moreover, when combined with expert collaboration, the accuracy further rises to 94.25%. These compelling results not only validate AutoHDR’s effectiveness as a standalone system but also underscore its potential as a powerful assistive tool for historians in practical applications.

We outline our main contributions as follows:

- We propose a novel fully Automated solution for HDR (AutoHDR), inspired by mirroring the workflow of expert historians.
- We introduce a pioneer Full-Page HDR dataset (FPHDR), which supports comprehensive HDR model training and evaluation.

- Extensive experiments demonstrate the superior performance of our method on both text and appearance restoration.
- The modular design enables flexible adjustments, allowing AutoHDR to collaborate effectively with historians.

2 Related Work

Historical document restoration primarily involves two modalities (Sommerschield et al., 2023), i.e., text and visual appearance.

Historical Text Restoration: Traditional historical text restoration relies heavily on expert labor, while recent advances in natural language processing (NLP) techniques offer promising solutions for this field. Pythia (Assael et al., 2019) pioneered Greek text restoration at both character and word levels, inspiring text restoration research across various languages (Fetaya et al., 2020; Bamman and Burns, 2020; Lazar et al., 2021; Papavassileiou et al., 2023). Notably, Ithaca (Assael et al., 2022) employs a transformer to jointly predict damaged texts, geographic origins, and dates, leveraging multi-task learning for enhanced performance.

Historical Appearance Restoration: Early methods in historical document appearance restoration depended on traditional image processing, such as Hedjam and Cheriet (2013) using ink’s properties under spectra to restore documents. Other methods focused on improving historical documents legibility (Raha and Chanda, 2019; Cao et al., 2022; Wadhvani et al., 2021). For instance, Cao et al. (2022) introduced adaptive binarization to isolate text from degraded backgrounds. Deep learning advances have enabled GAN-based (Huang et al., 2022; Shi et al., 2022) and Diffusion-based (Li et al., 2024) methods in historical appearance restoration, though mainly for single-character restoration. Given these limitations, Yang et al. (2025) developed DiffHDR, a patch-level restoration method that preserves the original style but needs manual guidance.

Joint Restoration: Recent research has transitioned to restoring historical texts and images jointly. Han et al. (2024) introduced a text-appearance restoration method via crowdsourcing, which requires labor input. Duan et al. (2024) introduced a model that jointly restores degraded texts and images by integrating contextual information but is limited to processing small-scale regions (a few characters). Zhu et al. (2024) proposed

a restoration framework that performs global appearance restoration followed by text correction through corpus retrieval, then conducts local refinement. However, this approach relies heavily on corpus coverage and is limited to patch-level binarized images.

3 FPHDR Dataset

Current open-source HDR datasets are severely scarce. While datasets like HDR28K (Yang et al., 2025) and CIRI (Zhu et al., 2024) exist, their restriction to patch-level images prevents the effective utilization of leverage full-page contextual information. To fill this gap, we introduce FPHDR, a page-level dataset with 1,633 labor-annotated samples for model evaluation and 6,543 synthetic samples for training.

3.1 Data Collection

The Fangshan Stone Sutras (FSS) is China’s largest surviving stone Buddhist canon³. However, extensive damage has hindered research on many sutras, making their restoration both an academic and social imperative. To address this, we invest substantial effort in collecting 1,633 typical damaged samples from the FSS and manually annotate both their damage locations and damage contents. However, these data cannot meet the training requirements of HDR models, since diffusion-based appearance restoration models demand pixel-level ground truths, which are impractical to generate manually. Therefore, we curate 6,543 well-preserved samples from the FSS, MTHv2 (Ma et al., 2020), and M⁵HisDoc (Shi et al., 2023) to synthesize pixel-level damaged-restored image pairs as training data.

The collected data exhibit the following characteristics: **(1) Semantic Integrity:** All samples maintain complete page-level context, preserving complete contextual semantic information. **(2) Degradation Diversity:** The data features a wide range of typical historical damages, such as surface erosion, radical loss, and character blur, presenting great challenges to HDR models. **(3) Dynasty Diversity:** The collected degradation samples span nearly a millennium from the Sui (581AD-618AD) to the Ming Dynasty (1368AD-1644AD), capturing both character evolution and degradation patterns across history. **(4) Source Diversity:** Various forms of historical documents are considered, in-

³wikipedia-Chinese Buddhist canon

cluding manuscripts, rubbings, and scrolls, representing diverse materials.

3.2 Manual Annotation for Damage

Due to long-term deterioration, historical documents have sustained varying degrees of damage, rendering their textual content partially or completely illegible. To ensure high-quality annotations of these damaged characters, we curate a professional annotation team consisting of ten experts with over five years of experience in HDR. Specifically, our annotation process consists of three main steps. **(1) Character Localization:** We annotate bounding boxes for all clearly visible characters and determine the positions of damaged characters based on the layout. **(2) Damage Assessment and Grading:** Given the inconsistency of character damage degrees, as shown in Fig. 3, we categorize the damage of characters into three levels:

- Severe damage: Characters exhibit complete loss of structural integrity, rendering them illegible even to expert examination.
- Medium damage: Characters show significant structural damage but remain identifiable through careful examination.
- Light damage: Characters maintain most structural features, enabling reliable identification despite visible damage.

(3) Content Annotation: We employ a differentiated approach for annotation depending on the condition of the characters. For light damage, direct visual annotation is performed. For medium damage, we attempt visual identification, and then verification using historical literature. For severe damage, annotation is conducted through the examination of multiple historical sources. Our annotation processes are based on authoritative historical literature, such as [CBeta](#), [National Library of China](#), and [Rushi Tripitaka Collection](#). Through this process, we construct a comprehensive dataset that includes character-level and line-level bounding box annotations, character content labels, and damage grades. Notably, to ensure dataset quality, every image in our dataset was independently annotated by at least two experts through a rigorous validation process. The entire manual workflow, including collection, annotation, and validation, requires approximately 2,400 person-hours.

3.3 Damaged-Restored Pairs Data Synthesis

As depicted in Sec. 3.1, we create synthetic training data by applying deterioration to well-preserved

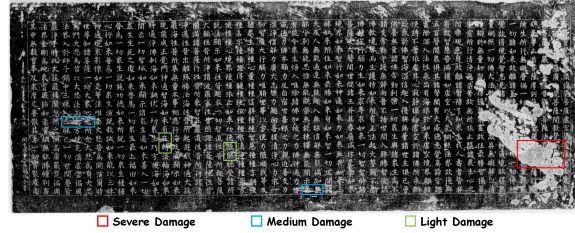


Figure 3: Illustration of damage grades in FPHDR. All damages are annotated at the character level, though only typical cases are highlighted here for clarity.

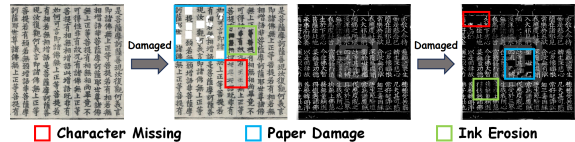


Figure 4: Illustration of different damage types in the FPHDR dataset. Please zoom in for a better view.

samples. Based on the approach of (Yang et al., 2025), we construct pixel-level damaged-restored pairs samples comprising three types of deterioration, as listed in Fig. 4: (1) Character Missing: Content removal is performed using LAMA (Suvorov et al., 2022) on randomly generated masks. (2) Paper Damage: Random areas in image patches are masked in black or white to simulate deterioration. (3) Ink Erosion: Water erosion and fading effects are simulated by applying genalog’s (Gupte et al., 2021) diverse degradation modes and kernels.

3.4 Dataset Analysis

As shown in Tab. 1, the statistical analysis of the data indicates that the average number of characters per sample is similar between the training and test sets. However, the training set includes a larger number of character categories, which helps the model learn more diverse character representations. In contrast, the average number of damaged characters per sample in the test set is higher, presenting a greater challenge to the restoration model’s robustness. For more details, please refer to Appendix A.

4 Methodology

4.1 Overall Framework

The proposed AutoHDR architecture is illustrated in Fig. 2, containing three distinct yet interconnected stages: OCR-Assisted Damage Localization (OADL) for character recognition and damage localization, Damaged Content Prediction (DCP) for text restoration, and Historical Appearance Restoration (HAR) for pixel-level reconstruction. The modular design enables independent training and

Subset	Images	Dam./im	Dam.	Char/im	Char classes
Training set	6,543	51.40	293,195	452.37	15,208
Test set	1,663	99.51	165,489	494.02	5,223

Table 1: Statistics of the FPHDR dataset. "Dam." denotes the number of damaged characters.

execution while maintaining seamless integration.

4.2 OCR-Assisted Damage Localization

The OCR-assisted damage localization stage is primarily responsible for recognizing legible characters and detecting the locations of damaged characters. To achieve this, we develop a character-level OCR model using data from various Chinese historical datasets, including MTHv2 (Ma et al., 2020), M⁵HisDoc (Shi et al., 2023), AHCDB (Xu et al., 2019), and HisDoc1B (Shi et al., 2025). This model demonstrates excellent performance on the test set, achieving a character localization F1 score of 98.35% and a character recognition accuracy of 96.93% under an Intersection over Union (IoU) threshold of 0.7. We then develop a model to localize severely damaged characters based on DINO (Zhang et al., 2023). After training the two models, we implement a localization fusion mechanism to merge the localization boxes from both models. Specifically, characters with an OCR confidence score below 0.1, indicating ambiguity, are designated as damaged, and their corresponding localization boxes B_o are extracted. In parallel, we extract the localization boxes B_s from the damage localization model. Then, we calculate the IoU between all B_o and B_s . If $b_o \in B_o$ has an IoU greater than 0.5 with any $b_s \in B_s$, b_o is removed. Conversely, if b_o does not overlap with any B_s (IoU lower than 0.5), b_o is retained:

$$B = B_s \cup \{b_o \in B_o \mid \max_{b_s \in B_s} \text{IoU}(b_o, b_s) \leq 0.5\}, \quad (1)$$

We evaluate damage localization extensively in Sec. 5.2, demonstrating it achieves human-comparable performance. By arranging the character and damage bounding boxes in the natural reading order, we generate a sequence that specifies the positions of damaged characters, serving as input for the subsequent content prediction module.

4.3 Damaged Content Prediction

Typically, historians first identify legible content from visual perception before restoring the incomplete or missing portions. Inspired by this paradigm, we combine both OCR’s visual recognition and LLM’s linguistic expertise to predict

damaged content in the Damaged Content Prediction (DCP) stage.

We first adopt Qwen2 (Yang et al., 2024), an advanced LLM, as our backbone model, specializing it in historical text prediction ability with a two-stage fine-tuning strategy. In the first stage, inspired by Cao et al. (2024), we conduct incremental pre-training using data from Daizhige (Garychowmu, 2019) and HisDoc1B (which encompass historical documents, poetry, art, Buddhist text, etc.) to enhance the model’s comprehension of classical Chinese. In the second stage, we fine-tune the model on pairwise damaged-restored historical texts from CBeta (an authoritative Buddhist text repository) to enhance its content prediction ability. We employ sequential mask tokens to represent damaged characters, directing the model to predict the corresponding contents. The forms of the LLM’s input and output are illustrated in Fig. 2 (Stage 2). A persisting issue is the inclusion of variant characters in classical Chinese texts, i.e., characters sharing identical meanings but differ in written form. Their rare occurrence challenges the model to recognize their equivalence to standard characters, hindering overall understanding. To tackle this, we augment the data using character variants (detailed in Appendix B.3). After training, the model acquires the capability to restore damaged content effectively.

While the trained LLM shows impressive content restoration performance, we discover that predicting the damaged content remains challenging due to the inherent complexity of classical Chinese, where multiple reasonable results could fit naturally in the same position. Therefore, relying solely on this LLM cannot guarantee the accuracy of text restoration. From the perspective of visual perception, we observe that OCR methods can recognize lightly damaged characters. This could serve as valuable auxiliary information to reduce the volume of damaged content requiring prediction and alleviate LLM’s prediction burden. Motivated by this insight, we propose Vision-Language Context Prediction (VLCP), which leverages OCR for lightly damaged content recognition while allowing the LLM to focus on severely damaged content.

The procedure of VLCP is detailed in Algorithm 1. For each character, we first recognize its content through OCR. When OCR confidence exceeds a pre-defined threshold, we adopt its prediction directly. Otherwise, we score Top-k predictions from both OCR and LLM through the follow-

ing strategies. For each candidate character (from the union of OCR and LLM prediction results), we compute a composite score incorporating: **(1) Base Score:** A weighted sum of OCR and LLM probability scores. OCR achieves high confidence for lightly damaged characters but low confidence for severe damage, while LLMs excel in the latter case. Such complementarity allows our system to adaptively select predictions based on damage level. **(2) Ranking Score:** A score is derived from characters’ ranking positions in both models’ predictions. Specifically, we rank the probabilities output by the LLM and OCR model separately, with each model generating its own ranking score based on the order of predictions according to their probabilities. This ranking criterion helps distinguish similar characters when their probability scores are close. **(3) Matching Bonus:** Characters appearing in both models’ predictions receive a bonus score, indicating visual and semantic plausibility. Finally, we sum the above scores to obtain the composite score. The candidate character with the highest composite score is selected as the final prediction. At this point, the damaged content has been restored.

Discussion. DCP stands as a crucial step to enable the full automation of the proposed AutoHDR. Since existing methods either necessitate manually inputting damaged content (Yang et al., 2025) or retrieving text from a limited database (Zhu et al., 2024), the DCP firstly transcends these limitations by automatic prediction, achieving high restoration performance without human efforts. So far, we have obtained the coordinates of the damaged positions and their corresponding content, which will be used for the next stage.

4.4 Historical Appearance Restoration

Adhering to the "restoring the old as old" principle, we develop a diffusion model to restore the damaged historical appearance at the pixel level, built based on DiffHDR (Yang et al., 2025), as depicted in Fig. 2 (Stage3). The model takes a damaged image x_d as input and generates a restored image x_r under the guidance of a mask image x_m (indicating damaged regions) and a content image x_c (specifying damaged content). Specifically, we corrupt the x_r by adding Gaussian noise to obtain the noised image x_n . Then, the model input consists of four concatenated components: $x_n, x_d \in \mathbb{R}^{3 \times H \times W}$, and $x_c, x_m \in \mathbb{R}^{1 \times H \times W}$. These form an 8-channel tensor that is processed by a denoiser \mathcal{F} to generate the x_r . The training objective of the

Algorithm 1 Vision-Language Context Prediction

Require: Input text \mathcal{T} ; OCR model \mathcal{O} , Language model \mathcal{L} ; OCR threshold τ ; OCR, LM weights w_o, w_l ; Ranking score weight α ; Matching bonus β ; TopK k
** s_{ocr} : OCR score; s_c : final candidate score*

- 1: **for** $p \in \mathcal{T}_{damaged}$ **do**
- 2: $s_{ocr} \leftarrow \mathcal{O}(p)$
- 3: **if** $s_{ocr}.conf > \tau$ **then**
- 4: $pred_p \leftarrow s_{ocr}.pred$
- 5: **else**
- 6: $P_o \leftarrow \mathcal{O}(p).topk, P_l \leftarrow \mathcal{L}(p).topk$
- 7: **for** $c \in P_o \cup P_l$ **do**
- 8: $r_o \leftarrow \text{rank of } c \text{ in } P_o, \text{ else } k$
- 9: $r_l \leftarrow \text{rank of } c \text{ in } P_l, \text{ else } k$
- 10: $s_c \leftarrow w_o p_o + w_l p_l + \alpha(2k - r_o - r_l)$
- 11: $s_c \leftarrow s_c \cdot (\beta \text{ if } c \in P_o \cap P_l \text{ else } 1)$
- 12: **end for**
- 13: $pred_p \leftarrow \arg \max_c (s_c)$
- 14: **end if**
- 15: **end for**
- 16: **return** $pred$

model is as follows:

$$\mathcal{L} = \|\mathbf{x}_g - \mathcal{F}(\mathbf{x}_n; \mathbf{x}_d, \mathbf{x}_c, \mathbf{x}_m)\|^2, \quad (2)$$

where x_g denotes the ground truth image. After training, the model performs pixel-level restoration that maintains character style consistency and background feature similarity by leveraging the intact regions in damaged image x_d .

While this model performs well, it is limited to patch-level restoration. To extend it to page-level, we introduce a Patch-AutoRegressive (PAR) mechanism during inference. PAR begins by dynamically selecting the starting patch from the four corners of the damaged image, choosing the one with the least number of damaged characters to ensure the model has the most intact characters for reference. The selected patch is restored and placed back in its original location. Then, an overlap sliding window operation extracts the next patch, leveraging previously restored regions as references for further restoration. To avoid split characters caused by the sliding window, we apply a mask to these regions during processing, ensuring all restored characters are complete. The process iterates until full-page restoration is complete. By leveraging references from previously restored patches, the PAR ensures visual consistency across the full page.

PAR exhibits significant practical value in engineering applications by addressing common challenges in HDR, such as the limitation to patch-level restoration and the difficulty in maintaining consistency across the entire page. We demonstrate its

Method	Venue	Top1 w/o VLCP	Top1 w/ VLCP	Top5 w/ VLCP
SikuBERT (Wang et al., 2022a)	HuggingFace'22	40.49%	83.57% (+43.08%)	87.28%
Ithaca (Assael et al., 2022)	Nature'22	39.78%	86.73% (+46.95%)	91.15%
GujiBERT (Wang et al., 2023b)	arXiv'23	45.57%	83.58% (+38.01%)	87.23%
AutoHDR-MegatronBERT-1.3B	This work	46.21%	83.42% (+37.21%)	86.59%
AutoHDR-Qwen2-1.5B	This work	50.49%	92.55% (+42.06%)	96.83%
AutoHDR-Qwen2-7B	This work	64.80%	95.15% (+30.35%)	97.75%
OCR-Only	This work	-	82.13%	-

Table 2: Comparison of damaged content prediction results with existing methods. Our model variants are built upon Erlangshen-MegatronBERT(Wang et al., 2022b) and Qwen2(Yang et al., 2024).

Method	Precision	Recall	F1 score
YOLOv7 (Wang et al., 2023a)	87.1	86.4	86.5
Co-DETR (Zong et al., 2023)	80.8	87.4	83.7
DINO (Zhang et al., 2023)	97.0	91.4	94.1
Historian*	98.9	95.6	97.2

Table 3: Comparison of damage localization results across different methods. * indicates that only a subset of the data is evaluated.

Method	Accuracy
Historian	44.08%
AutoHDR-Qwen2-7B	76.38%
Historian + AutoHDR	85.05%

Table 4: Evaluating AutoHDR’s collaboration.

effectiveness in Sec. 5.3, with detailed pseudocode provided in Appendix (Algorithm 2).

5 Experiments

5.1 Evaluation Metrics

For damage localization, performance is evaluated using the F1 score, precision, and recall at an IoU threshold of 0.5. For damaged content prediction, Top-1 and Top-5 accuracy metrics are adopted. For appearance restoration, since obtaining pixel-level ground truth from real data is extremely difficult, we evaluate the restoration quality through character recognition accuracy. Specifically, we train a text-line OCR using AHCDDB, MTHv2, and M⁵HisDoc to recognize the restored data, and adopt the commonly used Accurate Rate (AR) (Zhang et al., 2025) as our evaluation metric. The formula for AR is as follows:

$$AR = (N_t - D_e - S_e - I_e)/N_t, \quad (3)$$

where N_t is the total number of characters in annotations, while D_e , S_e , and I_e denote deletion, substitution, and insertion errors, respectively. For pixel-level evaluation on synthetic data, we use LPIPS (Zhang et al., 2018) as the evaluation metric, since Yang et al. (2025) demonstrated that PSNR

and SSIM are unsuitable for historical document restoration tasks.

5.2 Comparison with Existing Method

Damage Localization: We train DINO (Zhang et al., 2023), Co-DETR (Zong et al., 2023), and YOLOv7 (Wang et al., 2023a) on the FPHDR dataset. As shown in Tab. 3, DINO achieves the best performance with an F1 score of 94.1%. Additionally, we invite historians to evaluate a randomly selected set of 30 samples, achieving an F1 score of 97.2%. These historians (external to our annotation team) were not familiar with our strict annotation process, and for some characters with light damage that were still recognizable, they deemed restoration unnecessary, leading to a less than 100% F1 score in human evaluation. Overall, using DINO as the localization model is already comparable to human performance. In addition, when working collaboratively with historians, it can provide high-quality initial localization, allowing historians to make minor adjustments to achieve better detection results, thereby significantly reducing manual workload.

Damaged Content Prediction: We compare our method with SikuBERT (Wang et al., 2022a), GujiBERT (Wang et al., 2023b), and Ithaca (Assael et al., 2022). For a fair comparison, we retrain them following the approach in Sec.4.3. As shown in Tab. 2, AutoHDR models outperform other methods in both Top-1 and Top-5 accuracy, with larger models achieving better performance. Notably, the VLCP significantly improves Top-1 accuracy across all methods by an average of 39.61%. Moreover, as shown in the fourth column, all models outperform the OCR-Only baseline after incorporating VLCP. These results show that VLCP enables the model to classify damage grades automatically, using OCR for recognizable characters and LLM for unrecognizable ones, highlighting the effectiveness of the proposed VLCP.

Furthermore, our method achieves a maximum

Method	AR (%)			User Study ↑		LPIPS ↓
	Light	Medium	Severe	Style Consistency	Overall Quality (%)	
Damaged Documents	77.42	68.98	46.83	-	0.00	-
NAFNet (Chen et al., 2022)	78.56	73.56	61.07	2.721	2.57	<u>0.0585</u>
Uformer (Wang et al., 2022c)	78.03	72.72	62.94	<u>3.153</u>	<u>8.75</u>	0.0633
Restormer (Zamir et al., 2022)	87.27	84.40	75.98	2.877	3.90	0.0691
AutoHDR (Ours)	<u>91.80</u>	<u>90.01</u>	<u>84.05</u>	3.934	84.78	0.0541
Historian + AutoHDR (Ours)	93.63	93.81	94.25	-	-	-

Table 5: Comparison of historical appearance restoration results with existing methods.

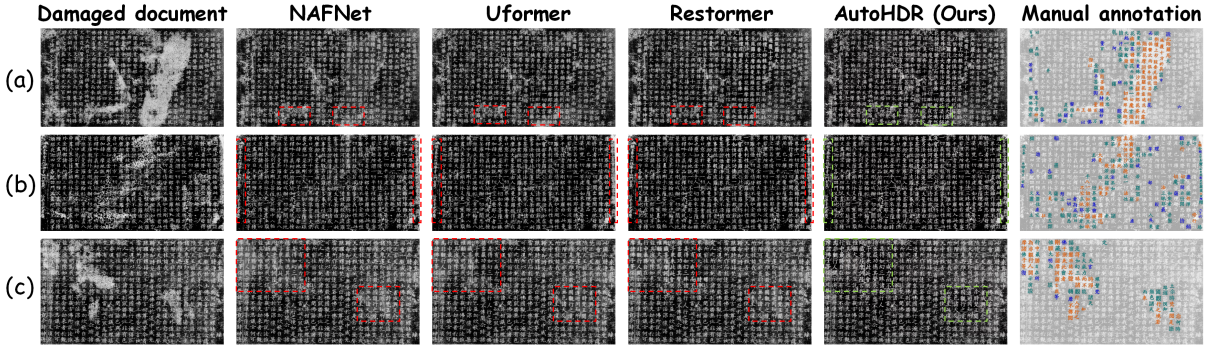


Figure 5: Qualitative comparison. We visualize the results of some evaluated methods. Red highlights regions with varying degrees of restoration inaccuracies, while green denotes areas with satisfactory restoration quality.



Figure 6: Restoring different types of documents.

Top-5 accuracy of 97.75%, demonstrating that AutoHDR can provide valuable suggestions for historians. To validate its collaborative potential, we test 23 severely damaged documents in three scenarios (Tab. 4): historian-only (44.08%), AutoHDR-only (76.38%), and collaborative predictions where historians select from AutoHDR’s Top-5 suggestions (85.05%). These results highlight AutoHDR’s collaborative capability, offering critical support for restoring and studying historical texts.

Historical Appearance Restoration: We compare our method with three state-of-the-art methods: NAFNet (Chen et al., 2022), Uformer (Wang et al., 2022c), and Restormer (Zamir et al., 2022). All methods are trained using the same procedure as DiffHDR (Yang et al., 2025). They receive identical input from the first two stages of AutoHDR and are required to output the restored images. As shown in Tab. 5, AutoHDR achieves SOTA performance. Compared to the original damaged images, it improves the recognition accuracy by 14.38%,

21.03%, and 37.22% on light, medium, and severe damage grades, respectively, demonstrating the strong restoration capability of our solution. Due to the difficulty of obtaining pixel-level annotations for real images, we conduct two user studies (Style Consistency and Overall Quality) with 20 participants to evaluate restoration quality. For style consistency, participants are asked to score the font style similarity between restored and original regions on a 1-5 scale (5 = completely consistent, 1 = completely inconsistent), focusing solely on style while ignoring other factors like image clarity. For overall quality, participants are asked to consider font style similarity, background integration, and character accuracy, then select the best result from the above four models. As presented in Tab. 5, our method achieves the highest score in the user study, indicating its superior capability in faithfully restoring the original appearance of historical documents. Additionally, we select 100 intact images of the Fangshan Stone Sutra and degrade them according to the method described in Sec. 3.3 to evaluate pixel-level restoration performance. As shown in the last column of Tab. 5, our model achieves the best performance with the lowest LPIPS. Furthermore, we invite historians to collaborate with AutoHDR by reviewing and modifying the intermediate results at each stage of the process. As shown in the last row of Tab. 5, this

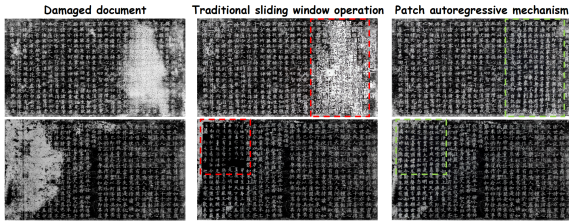


Figure 7: Effectiveness of patch autoregressive.

collaboration significantly improves performance, particularly on severe damage grade, achieving a 10.20% improvement over AutoHDR only. This further underscores AutoHDR’s strong collaborative capability.

The qualitative results are visualized in Fig. 5. NAFNet often shows character distortion and stroke loss, while Uformer and Restormer produce blurry regions in restored areas (see Fig. 5(a)(c)). They also struggle with small or complex characters (see Fig. 5(b)). In contrast, AutoHDR achieves superior performance. Additionally, Fig. 6 illustrates AutoHDR’s generalization by restoring various historical documents.

The outstanding performance of AutoHDR, coupled with its fully automated capabilities, underscores its practicality and potential for widespread application. Collaborative efforts with historians further reinforce the effectiveness and utility of AutoHDR, making it a valuable tool for the restoration and study of historical documents.

5.3 Ablation Study

To validate the effectiveness of our proposed patch-autoregressive mechanism, we compared it with traditional sliding window operations. As shown in Fig. 7, the traditional sliding window operation may lead to restoration failure or incomplete character restoration. Conversely, our patch-autoregressive mechanism achieves high-quality page-level restoration. Additionally, we conduct ablation studies on the LLM’s input/output formats, the data augmentation method, and the VLCP algorithm in Appendix C. These studies demonstrate that our current format design and proposed methods are effective.

6 Conclusion

In this paper, we propose AutoHDR, a novel solution for HDR that mimics historians’ restoration practices through a three-stage approach: OCR-assisted damage localization, vision-language context text prediction, and patch autoregressive ap-

pearance restoration. AutoHDR’s modular architecture enables seamless collaboration between AI and historians, allowing for flexible intervention and enhancement at various stages of the restoration process. Furthermore, we present FPHDR, a pioneering full-page HDR dataset containing 6,543 synthetic samples for training and 1,633 annotated real samples for evaluation. Extensive experimental results demonstrate AutoHDR’s outstanding performance in HDR tasks and its effectiveness in supporting historians’ work. We anticipate that this research will significantly advance AI-assisted HDR and make a substantial contribution to cultural heritage preservation.

Acknowledgements

This research is supported in part by the National Natural Science Foundation of China (Grant No.: 62476093, 62441604).

Limitations

AutoHDR utilizes a three-stage process for the restoration of historical documents, which inherently introduces certain limitations in processing speed. In our experiments, inference on a single NVIDIA A10 GPU requires an average of approximately five minutes per image. Furthermore, as indicated in Tab. 5, although our method has achieved promising performance, the restoration results may still exhibit inaccuracies, particularly in scenarios involving severe document degradation. Therefore, collaboration with historians emerges as a more robust and reliable strategy for document restoration and research. In future work, we will explore the feasibility of utilizing large vision-language models (such as Qwen2.5-VL (Team, 2025) and InternVL 2.5 (Chen et al., 2025)) to perform end-to-end restoration of historical documents.

Ethical Statements

This research is dedicated to advancing historical document restoration, ensuring positive contributions to cultural preservation. While AutoHDR offers significant advantages in restoring damaged documents, we remain mindful of the potential risks of misuse, such as the generation of falsified historical records. To address these risks, we apply strict licensing agreements that limit the dataset and code to academic research and non-commercial use, ensuring the technology is applied ethically and responsibly.

References

- Y Assael, T Sommerschild, and J Prag. 2019. Restoring ancient text using deep learning: a case study on greek epigraphy. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.
- Yannis Assael, Thea Sommerschild, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.
- David Bamman and Patrick J Burns. 2020. Latin bert: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*.
- Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, Yang Liu, Kai Ding, and Lianwen Jin. 2024. TongGu: Mastering classical Chinese understanding with knowledge-grounded large language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 4196–4210.
- Songxiao Cao, Zichao Shu, Zhipeng Xu, Dailiang Xie, and Ya Xu. 2022. Character segmentation and restoration of qin-han bamboo slips using local autofocus thresholding method. *Multimedia Tools and Applications*, 81(6):8199–8213.
- CBeta. <https://www.cbeta.org/>.
- Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. 2019. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.
- Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. 2022. Simple baselines for image restoration. In *European conference on computer vision (ECCV)*, pages 17–33.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhui Wang. 2025. [Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling](#). *Preprint*, arXiv:2412.05271.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). *Preprint*, arXiv:2010.11929.
- Wei-Sheng Du. 1999. “restoration as old” and “restoration as new” in the repair of ancient books. *Journal of Beijing Library*, (04):99–102.
- Siyu Duan, Jun Wang, and Qi Su. 2024. Restoring ancient ideograph: A multimodal multitask neural network approach. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 14005–14015.
- Ethan Fetaya, Yonatan Lifshitz, Elad Aaron, and Shai Gordin. 2020. Restoration of fragmentary babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences*, 117(37):22743–22751.
- Garychowcmu. 2019. Daizhigev20. <https://github.com/garychowcmu/daizhigev20>.
- Amit Gupte, Alexey Romanov, Sahitya Mantravadi, Dalitso Banda, Jianjie Liu, Raza Khan, Lakshmanan Ramu Meenal, Benjamin Han, and Soundar Srinivasan. 2021. Lights, camera, action! a framework to improve nlp accuracy over ocr documents. *Document Intelligence Workshop at KDD 2021*.
- Kaixin Han, Weitao You, Huanghuang Deng, Lingyun Sun, Jinyu Song, Zijin Hu, and Heyang Yi. 2024. Lant: finding experts for digital calligraphy character restoration. *Multimedia Tools and Applications*, pages 1–24.
- Rachid Hedjam and Mohamed Cheriet. 2013. Historical document image restoration using multispectral imaging system. *Pattern Recognition*, 46(8):2297–2312.
- Hongxiang Huang, Daihui Yang, Gang Dai, Zhen Han, Yuyi Wang, Kin-Man Lam, Fan Yang, Shuangping Huang, Yongge Liu, and Mengchao He. 2022. Aggan: Unpaired image translation for photographic ancient character generation. In *Proceedings of the ACM international conference on multimedia (MM)*, pages 5456–5467.
- Koren Lazar, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. Filling the gaps in ancient akkadian texts: a masked language modelling approach. *arXiv preprint arXiv:2109.04513*.
- Haolong Li, Chenghao Du, Ziheng Jiang, Yifan Zhang, Jiawei Ma, and Chen Ye. 2024. Towards Automated Chinese Ancient Character Restoration: A Diffusion-Based Method with a New Dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 3073–3081.

- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.
- Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, and Yongpan Wang. 2020. Joint layout analysis, character detection and recognition for historical document digitization. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 31–36. IEEE.
- National Library of China. <https://www.nlc.cn>.
- Katerina Papavassileiou, Dimitrios I Kosmopoulos, and Gareth Owens. 2023. A generative model for the mycenaean linear b script and its application in in-filling text from ancient tablets. *ACM Journal on Computing and Cultural Heritage*, 16(3):1–25.
- Poulami Raha and Bhabatosh Chanda. 2019. Restoration of historical document images using convolutional neural networks. In *IEEE region 10 symposium (TENSYMP)*, pages 56–61.
- Rushi Tripitaka Collection. *Jin Shan Zang*. <https://www-test.tripitakas.net/js>.
- Daqian Shi, Xiaolei Diao, Hao Tang, Xiaomin Li, Hao Xing, and Hao Xu. 2022. Rcrn: Real-world character image restoration network via skeleton extraction. In *Proceedings of the ACM international conference on multimedia (MM)*, pages 1177–1185.
- Yongxin Shi, Chongyu Liu, Dezhi Peng, Cheng Jian, Jiarong Huang, and Lianwen Jin. 2023. M5HisDoc: A Large-scale Multi-style Chinese Historical Document Analysis Benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 78483–78495.
- Yongxin Shi, Dezhi Peng, Yuyi Zhang, Jiahuan Cao, and Lianwen Jin. 2025. A large-scale dataset for chinese historical document recognition and analysis. *Scientific Data*, 12(1):169.
- Thea Sommerschildt, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747.
- Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159.
- Qwen Team. 2025. [Qwen2.5-vl](#).
- Mayank Wadhvani, Debapriya Kundu, Deepayan Chakraborty, and Bhabatosh Chanda. 2021. Text extraction and restoration of old handwritten documents. *Digital Techniques for Heritage Presentation and Preservation*, pages 109–132.
- Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2023a. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 7464–7475.
- Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, and Xiyu Wang. 2023b. [GujiBERT and GujiGPT: Construction of Intelligent Information Processing Foundation Language Models for Ancient Texts](#). *Preprint*, arXiv:2307.05354.
- Dongbo Wang, Chang Liu, Zihe Zhu, and et al. 2022a. Sikubert and sikuroberta: Construction and application of pre-training models for digital humanities based on siku quanshu. *Library Forum*, 42(6):31–43.
- Guo-Qiang Wang. 2021. Minimal intervention principle for ancient book conservation in china :techniques and application strategy. *Library Forum*, 41(07):141–148.
- Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaying Zhang. 2022b. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.
- Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. 2022c. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 17683–17693.
- Yue Xu, Fei Yin, Da-Han Wang, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. 2019. Casia-ahcdb: A large-scale chinese ancient handwritten characters database. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 793–798. IEEE.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge,

- Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. *Qwen2 Technical Report*. *Preprint*, arXiv:2407.10671.
- Zhenhua Yang, Dezhi Peng, Yongxin Shi, Yuyi Zhang, Chongyu Liu, and Lianwen Jin. 2025. Predicting the Original Appearance of Damaged Historical Documents. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 5728–5739.
- Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. 2023. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.
- Yuyi Zhang, Yuanzhi Zhu, Dezhi Peng, Peirong Zhang, Zhenhua Yang, Zhibo Yang, Cong Yao, and Lianwen Jin. 2025. Hiercode: A lightweight hierarchical codebook for zero-shot Chinese text recognition. *Pattern Recognition*, 158:110963.
- Shipeng Zhu, Hui Xue, Na Nie, Chenjie Zhu, Haiyue Liu, and Pengfei Fang. 2024. Reproducing the past: A dataset for benchmarking inscription restoration. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 7714–7723.
- Zhuofan Zong, Guanglu Song, and Yu Liu. 2023. Detsr with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 6748–6758.

A Dataset Details

In this section, we present the details of our dataset. Tab. 6 shows the distribution of degradation grades (light, medium, and severe) in the FPHDR test set, which contains 1,663 images with a total of 165,489 degraded characters. Furthermore, we provide additional visualizations from our FPHDR dataset, including real samples in Fig. 9 and synthetic samples in Fig. 10.

B Implementation Details

B.1 Character-Level OCR Model

The OCR model used in the OCR-Assisted Damage Localization stage first utilizes YOLOv7 (Wang et al., 2023a) for character detection, followed by ViT-Base (Dosovitskiy et al., 2021) for character recognition. This model is trained on the MTHv2 (Ma et al., 2020), M⁵HisDoc (Shi et al., 2023), AHCDB (Xu et al., 2019), and HisDoc1B (Shi et al., 2025) datasets, with the division of training and testing data strictly following the official splits of these datasets.

B.2 Damage Localization

We implement and train two localization models (DINO⁴ and Co-DETR⁵) based on the framework of MMDetection (Chen et al., 2019). For both models, we employ SwinTransformer-Large (Liu et al., 2021) as the backbone, and all other configurations follow the default settings of MMDetection. For YOLOv7 (Wang et al., 2023a), we use the official source code⁶ for implementation. All models are trained using the pre-trained weights provided by the official sources. The training is conducted on 6 NVIDIA A800 GPUs. The image size is 1333×1333 . During training, we first pre-train the model using synthetic data, then randomly select 1,163 real images for fine-tuning, and evaluate the model on the remaining 500 real images.

B.3 Damaged Content Prediction

The hyper-parameter settings of incremental pre-training and content prediction fine-tuning are shown in Tab. 7. All experiments are completed on 8 NVIDIA A800 GPUs.

⁴https://github.com/open-mmlab/mmdetection/blob/main/configs/dino/dino-5scale_swin-l_8xb2-12e_coco.py

⁵https://github.com/Sense-X/Co-DETR/blob/main/projects/configs/co_dino/co_dino_5scale_swin_large_1x_coco.py

⁶<https://github.com/WongKinYiu/yolov7>

Subset	Images	Damage			Total
		Light	Medium	Severe	
Test set	1,663	27,231	96,114	42,144	165,489

Table 6: The distribution of different damaged grades in the FPHDR test set.

During content prediction fine-tuning, we simulate damaged texts using sequential mask tokens ([mask1], [mask2]...) to randomly replace characters, with masking ratios varying from 5% to 90%. To address the challenge caused by variant characters in classical Chinese texts, we propose a Variant-based Data Augmentation method (VDA). Specifically, we compile a reference table of 32,260 variant characters and randomly replace standard characters with their variants during data construction to improve the model’s comprehension of variant characters. Additionally, to enhance the model’s robustness, we randomly remove characters with a 3% probability during training.

In the Vision-Language Context Prediction (VLCP) algorithm, we set the OCR threshold τ to 0.9, OCR and LM weights (w_o , w_l) to 0.6 and 0.4 respectively, Ranking score weight α to 0.05, Matching bonus β to 1.5, and TopK k to 5. These values have proven robust across a wide range of documents and damage conditions in our experiments. However, users can adjust these parameters based on the degree of document damage. For instance, for severely damaged documents, higher weights should be assigned to the language model to leverage contextual information, while for less damaged documents, higher weights can be assigned to the OCR model to prioritize visual recognition accuracy.

B.4 Historical Appearance Restoration

We train the appearance restoration model with a batch size of 16 and a total epoch of 195 and adopt an AdamW optimizer with $\beta_1 = 0.95$ and $\beta_2 = 0.999$. The learning rate is set as 1×10^{-4} with the linear schedule. The image size is 512×512 . The training is conducted on 4 NVIDIA A6000 GPUs. Additionally, we adopt the DPM-Solver++ as our sampler with the inference step of 20.

The detailed procedure of the Patch Autoregressive mechanism (PAR) is presented in Algorithm. 2. For our PAR implementation, we configure the patch size P to 448 and the stride S to 224.

C Ablation Study

The ablation study is designed to investigate how different input-output formats affect the perfor-

Hyperparameter	Incremental Pretraining	Content prediction fine-tuning
Precision	bf16	bf16
Epoch	1	5
Batch size	288	240
Learning rate	1e-5	6e-6
Weight decay	0	0
Warmup ratio	0.03	0.03
LR scheduler type	cosine	cosine
Optimizer	AdamW	AdamW
β_1	0.9	0.9
β_2	0.999	0.999
Max length	3072	3072

Table 7: Hyper-parameter settings in incremental pre-training and content prediction fine-tuning.

	Input	Output
(1)	不可读志谱[Mask][Mask][Mask] ... [Mask]韵於[Mask] 花以数德上知是	蒙生身...志此
(2)	不可读志谱[Mask1][Mask2][Mask3] ... [Mask122]韵於 [Mask123]花以数德上知是	不可读志谱[Mask1]蒙[Mask2]生[Mask3]身 ... [Mask122] 志韵於[Mask123]此花以数德上知是
(3)	不可读志谱[Mask1][Mask2][Mask3] ... [Mask122]韵於 [Mask123]花以数德上知是	[Mask1]蒙[Mask2]生[Mask3]身 ... [Mask122]身[Mask123]此

Figure 8: Examples of input and output formats.

mance of text restoration and validate the effectiveness of our proposed Variant-based Data Augmentation (VDA) method. As shown in Fig. 8, we design three types of input-output formats for damaged content prediction. Format (1) uses a single mask token to represent damaged characters, outputting the prediction sequentially. Format (2) uses sequential mask tokens to represent damaged characters and outputs the restored text with mask tokens indicating damaged positions. Format (3) is the same as introduced in Sec. 4.3. Then, we conduct experiments using AutoHDR-Qwen2-1.5B. The experimental results in Tab. 8 indicate that formats (2) and (3) achieve comparable and better performance. However, format (3) provides a shorter output sequence, thus leading to faster inference speed, making it the preferred choice. Furthermore, as shown in the last two columns of Tab. 8, the proposed variant-based data augmentation method and VLCP demonstrate significant effectiveness.

D More Visualization Results

As shown in Fig. 11, we provide more visualization of restoration results from Restormer (Zamir et al., 2022), NAFNet (Chen et al., 2022), Uformer (Wang et al., 2022c), and AutoHDR. The visual comparison demonstrates that AutoHDR achieves superior performance.

Furthermore, we present additional restoration results of AutoHDR in Fig. 12, which demonstrate

Method	Input/Output formats			+VDA	+VLCP
	(1)	(2)	(3)		
Top1 Acc	35.72%	40.43%	40.32%	50.49%	92.55%

Table 8: Ablation study on input/output formats and Variant-based Data Augmentation (VDA) method (with VDA and VLCP based on format 3).

its dual restoration capabilities. On the one hand, AutoHDR can effectively adhere to the principle of "restoring the old as old", maintaining font style consistency and background feature similarity. On the other hand, it can extend to "restoring the old as new", thus accommodating diverse user requirements for both heritage preservation and modern restoration.

E The Impact of Patch Size in Patch Autoregressive Mechanism

Based on our observations, the optimal patch size should be determined relative to the character size in the document being restored. Specifically, the length and width of the patch should accommodate at least three characters (to ensure there are enough intact characters for reference). When this requirement is met, the patch size has minimal impact on page-level restoration. Typically, our default setting (patch size = 448) is sufficient to meet the needs of most practical applications.

F Potential Risks and Human-AI Collaboration Solutions

Although the LLM is fine-tuned with historical corpora, and our model significantly improves the accuracy of damaged content predictions by combining OCR visual information with the semantic understanding of the LLM, there remain certain special cases where the model may still generate plausible but incorrect results.

Therefore, we recommend that the optimal use case for our model is in collaboration with historians. Through extensive experimental validation (Tab. 3, Tab. 4, and Tab. 5), we found that allowing historians to review and modify the intermediate results of our model significantly enhances the accuracy and reliability of historical document restoration. This collaborative approach not only addresses the model's potential errors in ambiguous cases but also leverages domain expertise to ensure historical accuracy.

Algorithm 2 Patch-Autoregressive Mechanism

Require: Damaged image X_d with damage detection boxes $B = \{b_1, b_2, \dots\}$, Patch size P , Stride S

Ensure: Restored image X_r

```
1:  $X_r \leftarrow \text{Copy}(X_d)$ 
2: Initialize each  $b_i \in B$  with  $\text{restored\_flag}(b_i) \leftarrow \text{False}$ 
3: while there exists an unrestored box  $b_i \in B$  with  $\text{restored\_flag}(b_i) = \text{False}$  do
4:    $\mathcal{U} \leftarrow \{b_i \mid \text{restored\_flag}(b_i) = \text{False}\}$  ▷ Collect unrestored boxes
5:    $(x_{\min}, y_{\min}, x_{\max}, y_{\max}) \leftarrow \text{ComputeExtent}(\mathcal{U})$ 
6:    $\mathcal{C} \leftarrow \text{DefineCorners}(x_{\min}, y_{\min}, x_{\max}, y_{\max})$  ▷ Four corners for patch placement
7:   for corner  $c \in \mathcal{C}$  do
8:      $\text{cnt}(c) \leftarrow \text{CountUnrestoredInPatch}(c, \mathcal{U}, P)$  ▷ Compute the number of unrestored boxes in the patch at corner  $c$ 
9:   end for
10:   $c^* \leftarrow \arg \min_{c \in \mathcal{C}} (\text{cnt}(c))$  ▷ Select corner with minimal damage
11:  for (startX, startY) in  $\text{SlidingWindow}(c^*, S, P)$  do
12:     $(x_s, y_s, x_e, y_e) \leftarrow \text{ClipToBounds}(startX, startY, X_r, P)$ 
13:     $B_{\text{inside}} \leftarrow \text{FindFullyContainedBoxes}(x_s, y_s, x_e, y_e, \mathcal{U})$ 
14:     $x_c, x_m \leftarrow \text{RenderContentMask}(X_r, B_{\text{inside}})$ 
15:     $x_r \leftarrow \text{InpaintPatch}(X_r, x_c, x_m)$  ▷ Restore this patch
16:    Paste( $x_r$  into  $X_r$  at  $(x_s, y_s, x_e, y_e)$ )
17:    Mark each box in  $B_{\text{inside}}$  as restored ( $\text{restored\_flag}(b) \leftarrow \text{True}$ )
18:  end for
19: end while
20: return  $X_r$ 
```

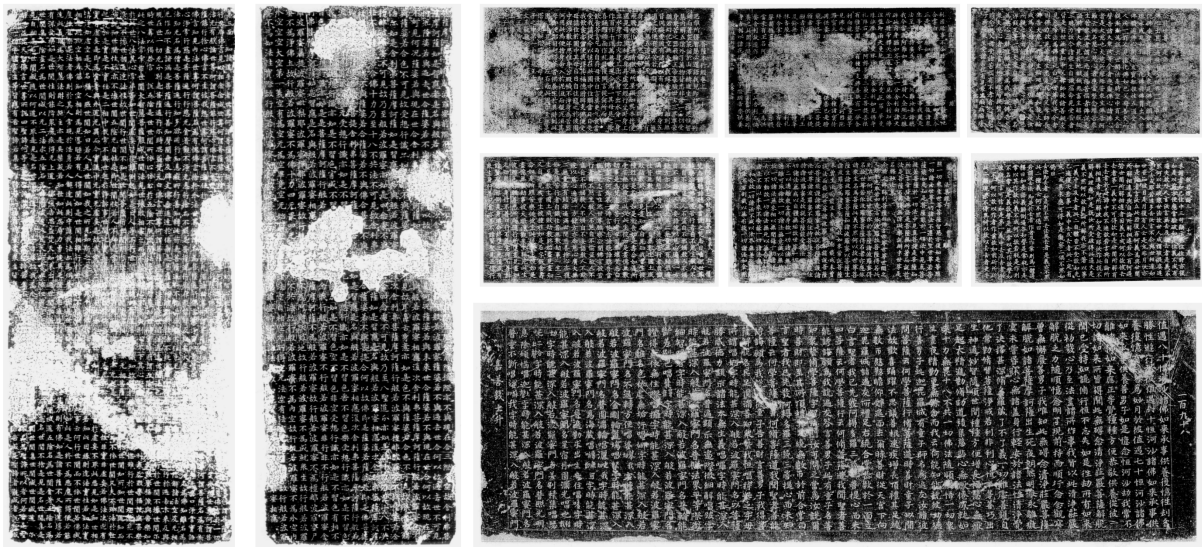


Figure 9: Examples of real samples in the FPHDR dataset.

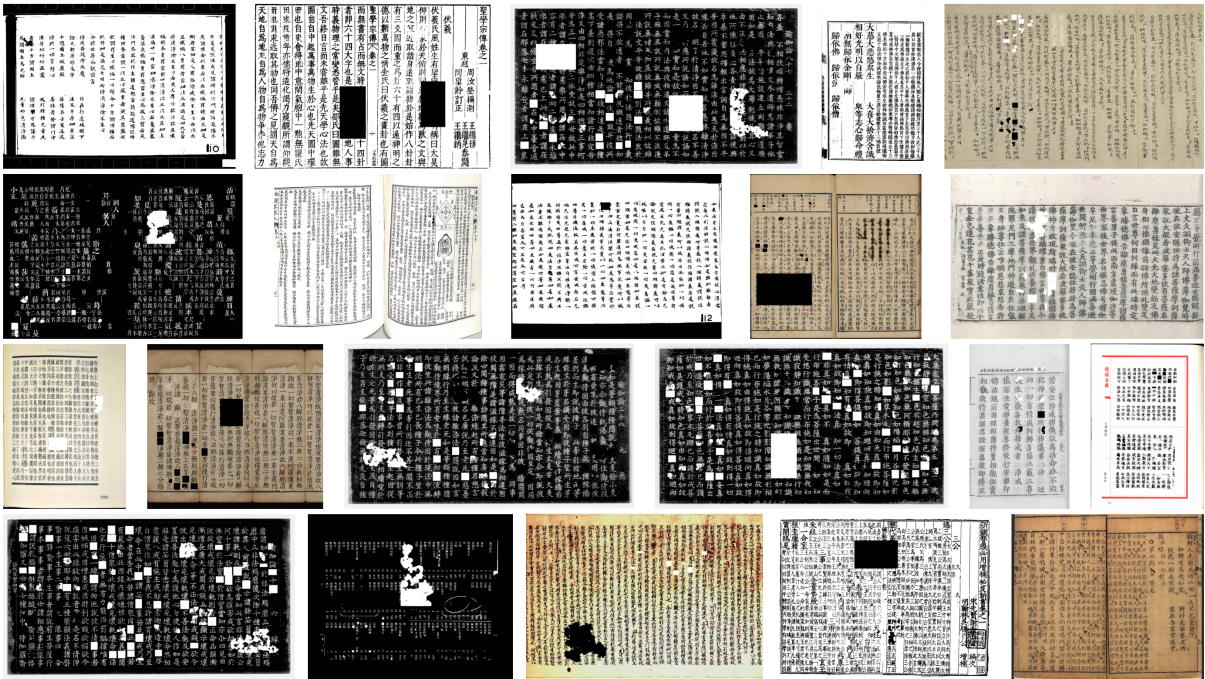


Figure 10: Examples of synthetic samples in the FPHDR dataset.

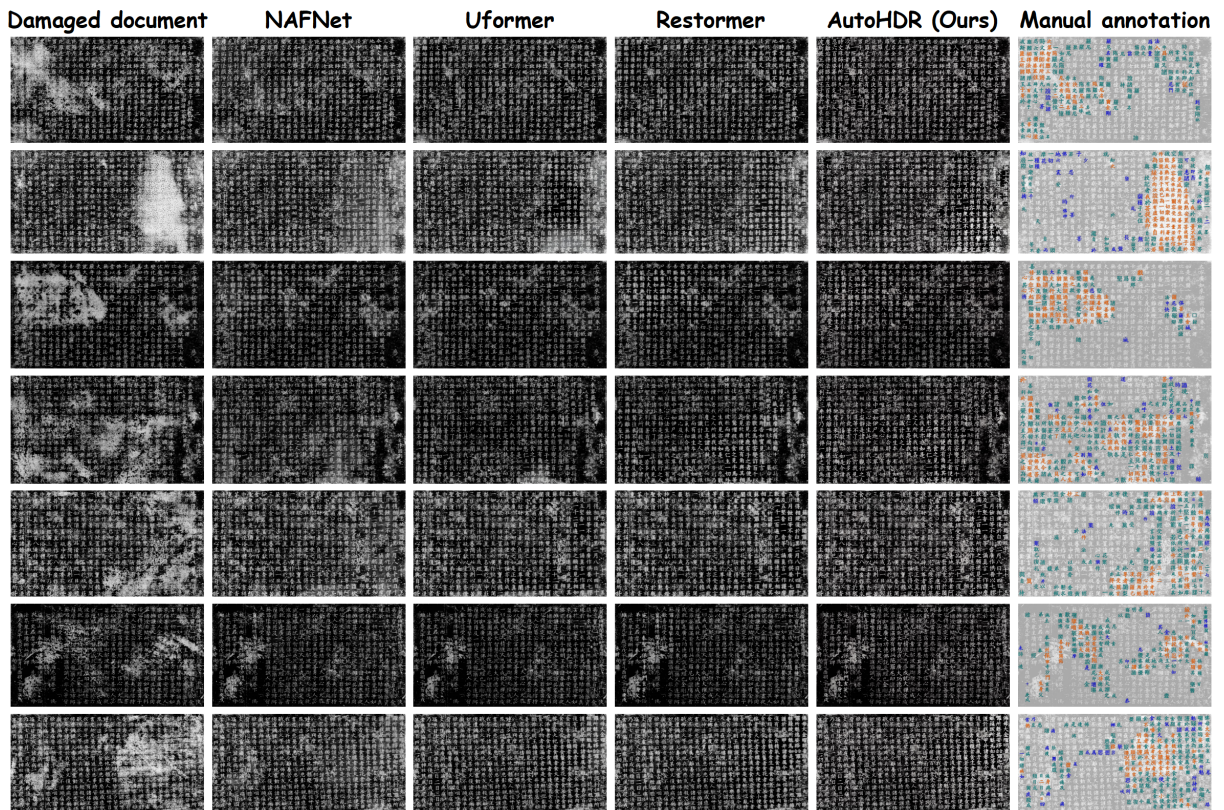


Figure 11: Additional qualitative comparison.

Damaged document

Restoring the 'old' as 'old'

Restoring the 'old' as 'new'

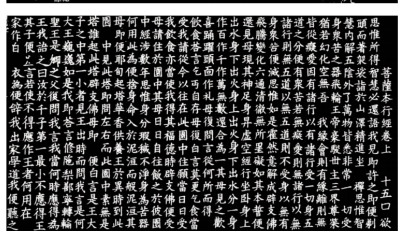
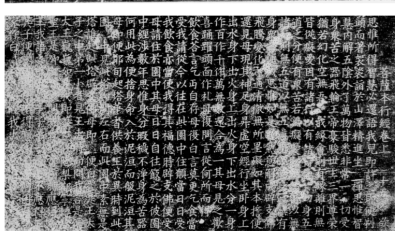
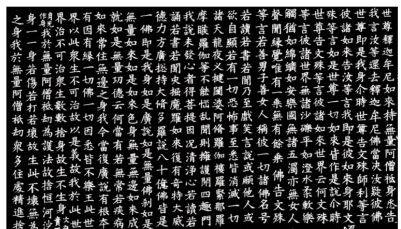
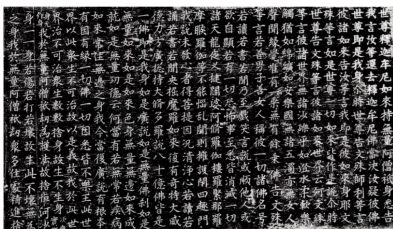
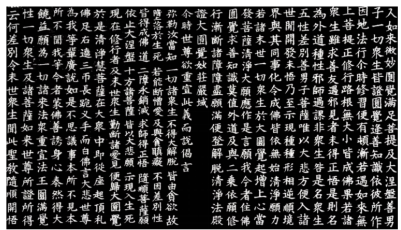
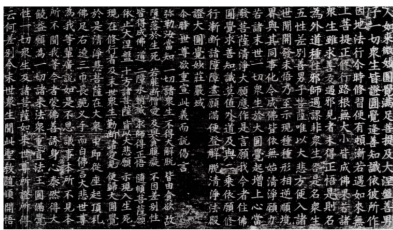
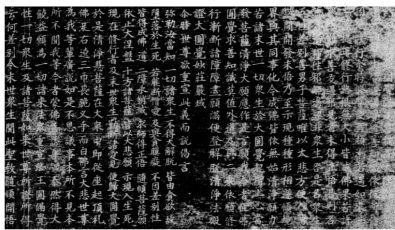
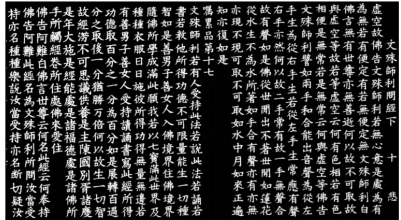
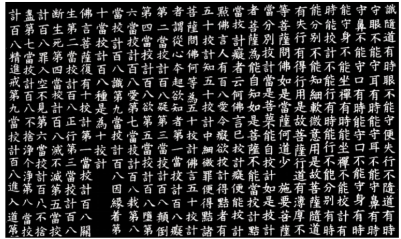
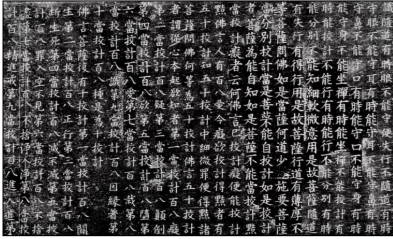
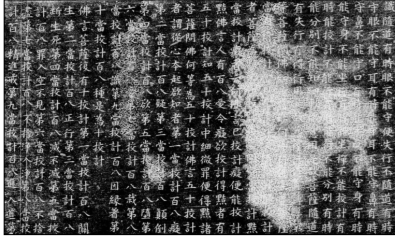


Figure 12: Additional restoration results of AutoHDR.