EpiAgent: An Agent-Centric System for Ancient Inscription Restoration

Supplementary Material

A. Implementation Details

A.1. Evaluation Protocol

Our evaluation framework comprehensively assesses inscription restoration quality across two dimensions:

Visual Fidelity. We quantify low-level fidelity using established full-reference metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [18], and Learned Perceptual Image Patch Similarity (LPIPS) [28]. To address the absence of ground-truth references in real-world test sets (R-I and R-II), we further incorporate four No-Reference (NR) metrics, i.e., CLIP-IQA [16], MUSIQ [9], MANIQA [22], and NIMA [14]. These metrics utilize pre-trained assessment models to evaluate perceptual quality, ensuring the restored images maintain stylistic consistency and aesthetic fidelity.

Textual Authenticity. This dimension prioritizes the structural integrity of restored characters and the downstream textual semantic accuracy. In particular, we employ a pre-trained YOLOv5 model [2] for character detection, which yields an F1 score exceeding 0.98 at an Intersection-over-Union (IoU) threshold of 0.5. End-to-end spotting performance is measured via Normalized Edit Distance (NED), obtained by applying a ResNet-50 recognizer [6] to the detected character regions. Furthermore, to decouple recognition performance from detection variance, we evaluate standalone recognition accuracy (Top-1, Top-5, and Macro Accuracy) by applying the recognizer to ground-truth cropped character regions on the restored images. Notably, Macro Accuracy is computed on a per-category basis, providing a stricter evaluation of performance across the long-tailed distribution inherent in ancient Chinese inscription datasets.

It is noteworthy that the Test Set S in CIRI dataset [30] provides both intact clean images and character-level labels (text and locations), enabling full evaluation of visual presentation and textual integrity. In contrast, the real test sets R-I and R-II lack ground-truth images and character-level location annotations, so their evaluation is restricted to no-reference IQA metrics and end-to-end 1-NED.

A.2. Training Details of EpiAgent

Although our fine-grained restoration strategy is inherently agnostic to input image resolution, we conduct our implementation on the Chinese Inscription Rubbing Image (CIRI) dataset [30] due to copyright restrictions.

We follow the standard data partitioning protocols defined in CIRI dataset for both training and evaluation. The training procedure is divided into two distinct phases:

1. **Tool Pre-training:** First, the three specialized generative tools (Background Denoising, Stroke Completion, and Font Imitation) are individually trained on the CIRI training split until convergence.
2. **Agent Optimization:** Subsequently, we freeze the weights of these specialized tools. The high-level EpiAgent framework then undergoes an iterative optimization phase on the same training set. This phase focuses on accumulating planning experience and refining decision-making logic through expert feedback loops.

Following this process, comprehensive evaluations are conducted on the synthetic Test set (S) and real-world benchmarks (R-I and R-II). To ensure a fair comparison, all baseline methods are trained and evaluated using this identical data configuration.

Table 1. Ablation studies on the accuracy of inscription character recognition for representative multimodal large language models using the end-to-end 1-NED metric. The best and the second-best results are **highlighted** and underlined.

MLLM	Test S	Test R-I	Test R-II
PaddleOCR-VL-0.9B [1]	0.3985	0.3191	0.2605
Qwen3-Thinking-VL-Plus [21]	<u>0.5727</u>	<u>0.5235</u>	<u>0.4147</u>
Seed-1.5-Thinking-Vision-Pro [4]	**0.6481**	**0.5509**	**0.4336**

A.3. Multimodal Perception Tools

In the **Observe** stage, we deploy both general and specialized modules to analyze and disentangle the coupled visual-textual cues within inscriptions. Concretely, this stage integrates a Multimodal Large Language Model (MLLM) for holistic multimodal perception, a Degradation Assessment Model (DAM) to quantify degradation severity, a Corrective Language Model (CLM) for predictive script correction, and Retrieval-Augmented Generation (RAG) for querying external Chinese corpora.

We instantiate the **MLLM** using Seed-1.5-Thinking-Vision-Pro [4], prompting it to extract both textual content and spatial bounding box layouts. Tab 1 presents a comparative analysis of mainstream MLLMs on Chinese inscription Optical Character Recognition (OCR). As evidenced by the results, Seed-1.5-Thinking-Vision-Pro demonstrates superior robustness, consistently outperforming baseline models across all three benchmarks.

DAM. We implement the Degradation Assessment Model using a U-Net architecture [11]. The model is trained on the synthetic set of the CIRI dataset, which provides ground-truth degradation masks, to perform pixel-level segmentation of degraded regions. Subsequently, we assign discrete

severity levels based on the ratio of degraded pixels within each detected character bounding box. We set thresholds at [0.02, 0.20, 0.50, 1.0] to categorize the degradation levels as "none", "slight", "middle", and "severe", respectively.

CLM. For the Corrective Language Model, we adopt Qwen-2-7B [15] as the foundation model, utilizing a Parameter-Efficient Fine-Tuning (PEFT) strategy via LoRA [7] to adapt the model for understanding and restoring damaged historical text. To facilitate this, we curate a large-scale dataset of paired damaged-restored historical texts from public classical literature repositories, including Daizhige[1] and CBeta[2]. These corpora span diverse genres, including literature, poetry, inscriptions, and scriptures. We simulate real-world degradation by applying character-level deletions, additions, and replacements to the original text. During training, damaged characters are represented by consecutive mask tokens, guiding the model to reconstruct the missing semantics.

RAG. In the Retrieval-Augmented Generation module, we leverage BGE-Large-zh [20] as the embedding model to encode classical Chinese texts into high-dimensional semantic vectors. We construct a comprehensive multi-source corpus by integrating the above corpora (Daizhige, CBeta) and compilations of historical documents and stele inscriptions. To manage the resulting database, FAISS [8] is deployed as the retrieval engine, enabling efficient and precise similarity search across the large-scale vector database.

A.4. Details of User Study

To ensure a comprehensive and unbiased subjective evaluation, we recruited a diverse cohort of participants stratified into three distinct groups: **domain experts** (epigraphers and historians), **humanities students** (with background knowledge in ancient texts), and **general volunteers** (laypersons). The study comprised 100 individual evaluation cases. In each case, participants were presented with a tuple consisting of the degraded input image, the ground-truth text, and the restoration outputs from 9 comparative methods. Participants were tasked with ranking these 9 results based on their visual perceptual quality and restoration fidelity.

A.5. Prompt for MLLM and Central Planner

Tab. 2 shows the designed prompt for the MLLM in the **Observe** stage and the central planner π.

A.6. Hierarchical Restoration Tools

In the **Execution** stage, we deploy specialized restoration tools tailored for complex degradation patterns. We detail their specific implementations below:

Background Denoising: This tool f_{den} specializes in eliminating background clutter and noise while rigorously preserving stroke topology. Unlike existing diffusion-based models [23] that operate within a continuous feature space, our tool adopts a discrete diffusion paradigm [27, 30]. This is implemented via an attention-based U-Net [11] equipped with three pairs of symmetric residual blocks, effectively preventing information loss during feature transitions. During **training**, we adhere to the discrete diffusion formulation by successively adding noise sampled from a Bernoulli distribution ϵ to the ground-truth image I_{gt} based on time step t. In the denoising phase, we leverage the degradation segmentation mask $\mathcal{S}_d$ provided by the DAM. The original degraded inscription I is concatenated with $\mathcal{S}_d$ to serve as the conditional input to estimate the intact image. The process is supervised by the KL divergence loss between the predicted and ground-truth distributions. During **inference**, the pre-trained model iteratively generates the restored result over T steps, conditioned on the degraded input and its mask.

Stroke Completion: This tool f_{inp} targets the inpainting of missing or severely degraded regions (as indicated by $\mathcal{S}_d$) to avoid deforming intact strokes. We implement this using a vanilla U-Net [11] with five pairs of symmetric residual blocks. During **training**, we bridge the modality gap by rendering the corrected text sequence $\hat{\mathcal{H}}$ into a canonical glyph image I_s according to the predicted layout $\hat{O}$. This explicit structural guidance ensures the model aligns with the correct textual content. Subsequently, we combine the degraded character regions, degradation masks, and the canonical glyph images as conditioning inputs. The model generates restored images through an iterative denoising process, supervised by an MSE loss against the ground-truth intact images. During **inference**, we employ a non-Markovian sampling process (DDIM) [13] to accelerate generation and improve sampling efficiency.

Font Imitation: This tool f_{imi} focuses on synthesizing stylistically consistent glyphs by learning style priors from high-quality exemplars within the same stele, thereby ensuring calligraphic harmony. We follow the architecture from [24], incorporating two decoupled encoders: a style encoder E_s and a content encoder E_c. The style encoder E_s extracts calligraphic features from well-preserved or restored exemplars in the same sequence, while the content encoder E_c captures semantic structural features from the canonical glyph image I_s. Ultimately, the conditional diffusion process generates font-consistent imitations by fusing these extracted style and content representations.

Character Retrieval: Designed to compensate for potential imperfections in the preceding generative restoration phases, this tool (f_{ret}) retrieves high-fidelity exemplars to substitute for degraded regions. It searches for well-preserved or successfully restored image patches that share identical semantic content and consistent calligraphic style, thereby ensuring the final output maintains visual coherence and historical authenticity.

[1] https://github.com/garychowcmu/daizhigev20

[2] https://www.cbeta.org

Table 2. Prompts for the MLLM and the central planner of EpiAgent. $\{\cdot\}$ denotes a placeholder to be filled based on the context.

Prompt for the MLLM to perceive visual and textual cues within inscription images
`Here's an ancient Chinese text image. The characters in the image are corrupted with varying degradations, such as scatter noise, grid etching, irregular absence, and large spalling, making the text hard to recognize. Please recognize the textual content and locate the bounding box for each character within the inscription image. For each character that is recognizable, provide the recognized text and coordinates. For each character that is completely unrecognizable, please predict the coordinates of the degraded character according to the coordinates of recognizable characters. Generally, each bounding box is aligned with the bounding boxes of the surrounding characters. The sizes of the bounding boxes should be consistent. Note that do not attempt to recognize areas without text. Please output strictly in JSON format like [{"text": "", "coordinate": [x1,y1,x2,y2]}].`
Prompt for the central planner to schedule restoration plans based on multimodal cues and distilled experience
`Here's an inscription image with text content, coordinates, and degradations {Cues}. We will invoke specialized restoration tools to eliminate degradations as well as restoring textual content and visual appearance. There's the distilled historical restoration experience: {Experience}. Please refer to these experiences and schedule a correct invocation order of these tools for each degraded character: background denoising, stroke completion, font imitation, and character retrieval.`
Prompt for the central planner to distill referenceable experience from historical restoration statistics
`We are studying inscription image restoration with varying degradations. The degradation degree is categorized into four levels: L0, L1, L2, and L3. Characters at level L0 are intact and noise-free. Characters with L1 degradations are slightly degraded with scatter noise. Characters with L2 degradations are moderately degraded with local character occlusion. Characters with L3 degradations are severely degraded with extensive character occlusion, making the text hard to recognize. We have specialized tools to address these degradations: background denoising, stroke completion, font imitation, and character retrieval. The problem is, given a degraded character, we need to determine the invocation order of the tools according to the degradation status. Note that it is complex and complicated, as different tasks can deliver different restoration effects, and unsuitable tools may cause side effects. Therefore, the correct tool combination and invocation order based on the degradation levels are significantly important. We have conducted many trials and collected the following statistics about the success rates of different tool invocation orders: {Statistics}. Please distill useful restoration experience from these statistics to help us select the correct tools and determine the tool invocation order.`

Table 3. Restoration performance of SeedEdit-3.0, LucidFlux, and EpiAgent on Test Set S.

Strategy / Metric	Quality			End-to-End
	PSNR ↑	SSIM ↑	LPIPS ↓	1-NED ↑
SeedEdit-3.0 [17]	15.18	0.8414	0.2908	0.3870
LucidFlux [3]	17.73	0.8665	0.2890	0.3828
EpiAgent (Ours)	**22.14**	**0.9684**	**0.0254**	**0.9069**

Table 4. Restoration performance comparison between SeedEdit-3.0, LucidFlux, and EpiAgent on Test Set R-I.

Strategy / Metric	Quality				End-to-End
	CLIP-IQA ↑	MUSIQ ↑	MANIQA ↑	NIMA ↑	1-NED ↑
SeedEdit-3.0 [17]	0.9111	47.81	0.3817	0.4921	0.2758
LucidFlux [3]	0.9167	48.25	0.3857	0.5022	0.3462
EpiAgent (Ours)	**0.9393**	**50.29**	**0.4179**	**0.5414**	**0.5766**

Table 5. Restoration performance comparison between SeedEdit-3.0, LucidFlux, and EpiAgent on Test Set R-II.

Strategy / Metric	Quality				End-to-End
	CLIP-IQA ↑	MUSIQ ↑	MANIQA ↓	NIMA ↑	1-NED ↑
SeedEdit-3.0 [17]	0.9136	47.25	0.3803	0.4725	0.2280
LucidFlux [3]	0.9021	46.36	0.3815	0.4955	0.2984
EpiAgent (Ours)	**0.9388**	**49.94**	**0.4157**	**0.5381**	**0.5546**

B. More Results

B.1. Extended Comparative Analysis

Comparison with Foundation Models. Acknowledging the rapid advancements in the field of Large Multimodal Models (LMMs), we extend our evaluation to include recent representative foundation models designed for image editing and restoration tasks. Specifically, we benchmark against SeedEdit-3.0 [17] and LucidFlux [3], *both of which explicitly claim capabilities for text image editing and restoration*. The quantitative results are detailed in Tab. 3, Tab. 4, and Tab. 5. This disparity highlights a critical insight: general-purpose foundation models, despite their generative power, struggle with the complex, coupled degradations of ancient inscriptions. They lack the *comprehensive strategic planning and domain-specific expertise* required to disentangle noise from semantics. This validates the necessity of our agent-centric approach, which can systematically orchestrate multi-stage specialized tools.

Evaluation of Qwen-Image-Edit. We further explored the feasibility of replacing our restoration toolkit with the pow-

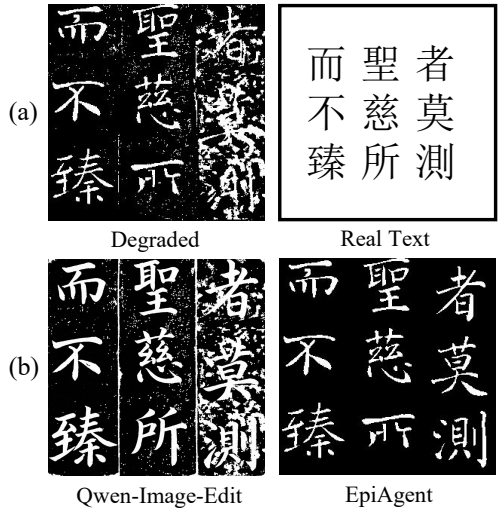

Figure 1. Restoration results of Qwen-Image-Edit-2509 for inscription images using ground-truth character bounding boxes and text prompts.

erful Qwen-Image-Edit-2509 [19], a recent image editing model that explicitly claims proficiency in modifying ancient Chinese text. However, as illustrated in Fig. 1, a critical limitation emerges: while the model can successfully modify target regions based on textual prompts, it fails to effectively eliminate the underlying degradation. Furthermore, the generated text exhibits **significant glyph distortion**, frequently deviating from correct calligraphic structures. This indicates that while the model functions as a capable semantic editor, it lacks the fine-grained structural fidelity and denoising capabilities required for authentic restoration.

Qualitative Superiority. Fig. 2 presents a broader qualitative comparison against state-of-the-art methods introduced in the main paper. The visual results highlight the strong stability and generalization capabilities of EpiAgent, particularly when handling the complex, non-uniform degradations typical of real-world inscription rubbings.

Metric Insensitivity Analysis. To address the potential limitations of standard image quality metrics in this specific domain, Fig. 3 displays restored samples alongside their corresponding PSNR values. It can be observed that while PSNR may reflect minimal numerical variation between methods, the perceptual differences are substantial, particularly in terms of structural integrity and stylistic consistency. This discrepancy underscores the necessity for specialized evaluation protocols beyond generic pixel-level metrics.

B.2. Time Consumption Analysis

Fig. 5 illustrates the average time required for EpiAgent to restore degraded inscriptions and the time proportion of each stage throughout the restoration process. It reveals that iterative tool invocation and execution account for 48.8% of the time. Additionally, perception, planning, and experience distillation contribute 12.5%, 8.3%, and 10.7%, respectively. As observed, the **Execute** stage is the most computationally intensive, accounting for 48.8% of the total duration due to the iterative inference of specialized generative tools (e.g., imitation and completion). In contrast, the agent's cognitive processes remain relatively efficient: the General Perception module (within the Observe stage) utilizes 12.5%, while the planning logic in the **Conceive** stage occupies only 8.3%. Furthermore, the distill process, which handles experience accumulation, contributes 10.7%. Meanwhile, the average runtime increases with degradation severity, from 76.33 s/image for slight cases to 238.53 s/image for severe cases. On average, EpiAgent requires 1.54 iterations per image and 2.72 LLM calls per image. Such results demonstrate that EpiAgent effectively balances the heavy computational load of low-level visual restoration with low-latency high-level reasoning. *Moving forward, we aim to further optimize the temporal efficiency of EpiAgent, specifically by accelerating the generative inference processes, to facilitate broader-scale applications.*

B.3. Failed Cases

As shown in Fig. 6, the inscription restoration task confronts a critical bottleneck when handling extreme degradation scenarios. These cases are characterized by massive cross-character structural deficits and **atypical degradation with dense diffuse corruption** that is inextricably coupled with character strokes. Such severe degradations make it difficult for existing restoration methods to *distinguish genuine character strokes from degradation artifacts*, thereby rendering them ineffective. It is worth noting that, while achieving a flawless restoration remains challenging under these adversarial conditions, our method stands out as the *only method in our comparison that preserves textual authenticity* at the visual level. This contrast highlights the inherent difficulty of restoring ancient inscriptions and underscores the necessity for continued research in this domain.

B.4. Robustness Analysis for Degradation Severities

To further evaluate the robustness of EpiAgent against varying degrees of degradation, we stratified Test Set S into three distinct subsets: "slight", "middle", and "severe". The quantitative comparisons are detailed in Tab. 6. As evidenced by the results, our approach establishes a consistent performance lead over competing methods across all metrics and degradation levels. This comprehensive superiority validates the exceptional generalization capability and adaptability of our agent-centric restoration paradigm.

Qualitative results corresponding to these three severity levels are visualized in Fig. 7, Fig. 8, and Fig. 9. These examples visually confirm the advantages of EpiAgent in maintaining both calligraphic style fidelity and textual se-

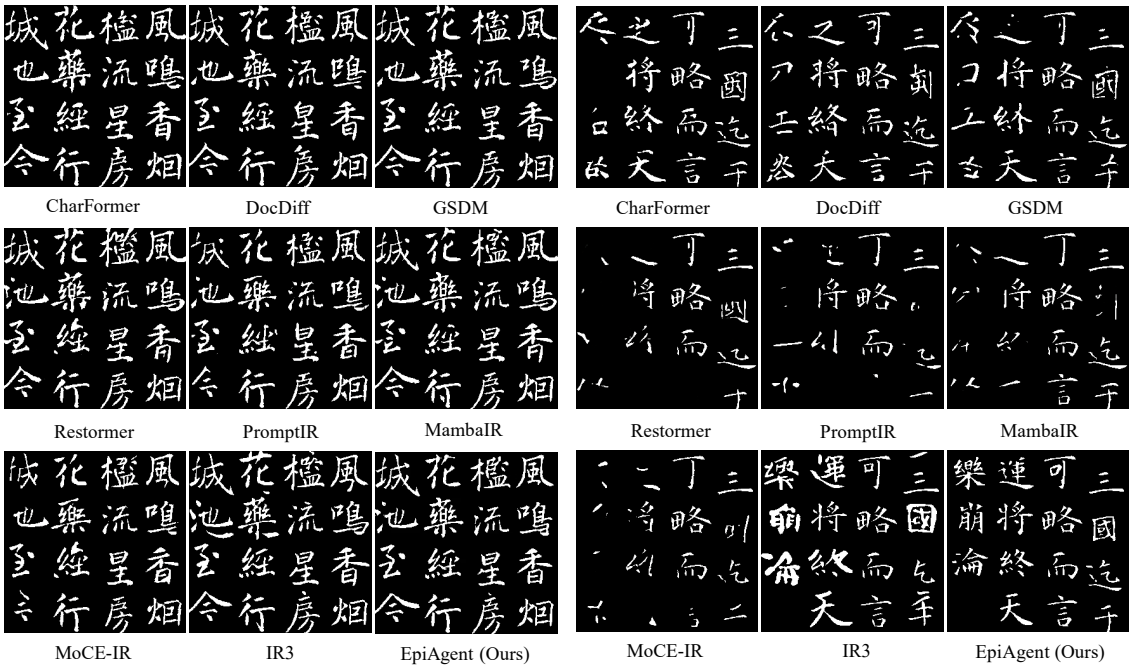

Figure 2. Restoration result comparison between EpiAgent and current state-of-the-art methods. (a), (b) are respectively from Test Set R-I and R-II.

mantic authenticity. Collectively, these results not only validate the technical efficacy of EpiAgent but also highlight its profound potential in advancing the digital preservation of cultural heritage.

References

[1] Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. Paddleocr-vl: Boosting multilingual document parsing via a 0.9 b ultra-compact nision-language model. *arXiv preprint arXiv:2510.14528*, 2025. 1

[2] Tausif Diwan, G Anirudh, and Jitendra V Tembhurne. Object detection using yolo: challenges, architectural successors, datasets and applications. *Multimedia Tools and Applications*, 82(6):9243–9275, 2023. 1

[3] Song Fei, Tian Ye, Lujia Wang, and Lei Zhu. Lucidflux: Caption-free universal image restoration via a large-scale diffusion transformer. In *Proceedings of the International Conference on Learning Representations*, 2026. 3

[4] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 1

[5] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *Proceedings of the European Conference on Computer Vision*, pages 222–241, 2024. 7

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Proceedings of the International Conference of Learning Representation*, 2022. 2

[8] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 2

[9] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 1

[10] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36:71275–71293, 2023. 7

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. 1, 2

[12] Daqian Shi, Xiaolei Diao, Lida Shi, Hao Tang, Yang Chi, Chuntao Li, and Hao Xu. Charformer: A glyph fusion based attentive framework for high-precision character image de-

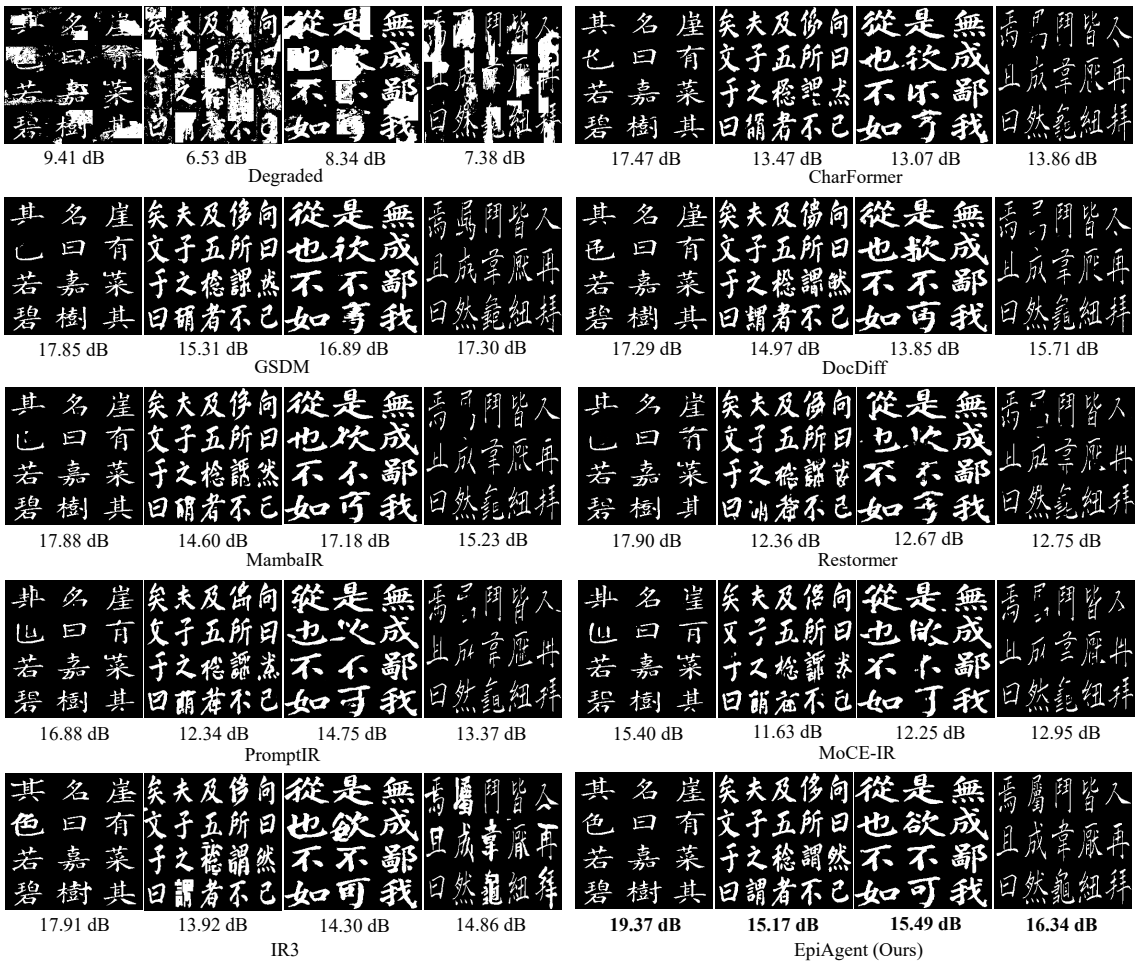

Figure 3. Case study on restored images and PSNR (dB) of different methods. The best results are **highlighted**.

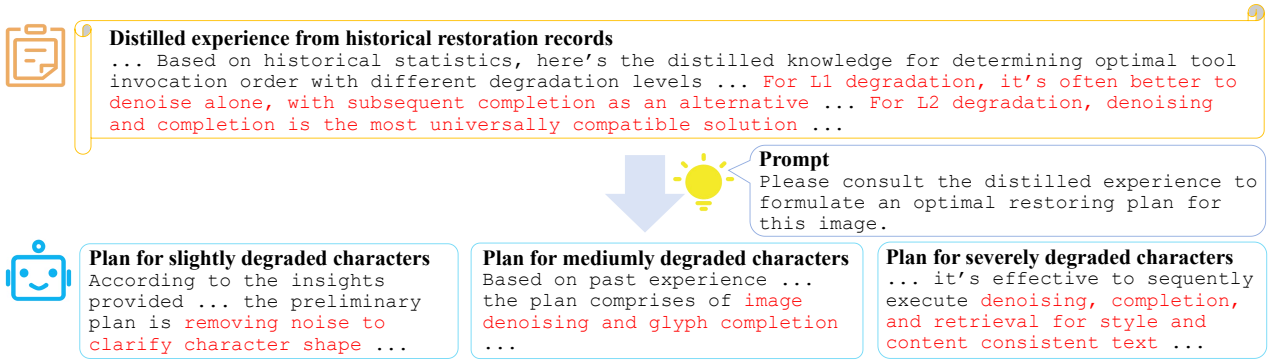

Figure 4. Examples of EpiAgent's planning response to the prompt that incorporates distilled experience.

noising. In *Proceedings of the ACM International Conference on Multimedia*, pages 1147–1155, 2022. 7

[13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations*, 2021. 2

[14] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8): 3998–4011, 2018. 1

[15] Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 2

[16] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In

Table 6. Quantitative comparison with the state-of-the-art methods on Test Set S. Inscription images are categorized into "slight", "middle", and "severe" according to the degradation severity. The best and the second best results are **highlighted** and underlined.

Degradations	Method / Metric	Quality							Recognition			End-to-End
		PSNR ↑	SSIM ↑	LPIPS ↓	CLIP-IQA ↑	MUSIQ ↑	MANIQA ↑	NIMA ↑	Top-1 Acc. ↑	Top-5 Acc. ↑	Macro Acc. ↑	1-NED ↑
Slight	CharFormer [12]	21.56	0.9690	0.0316	0.8796	52.95	0.4358	0.5551	0.9544	0.9819	0.5820	0.8726
	DocDiff [23]	22.62	0.9761	0.0220	0.8951	53.37	0.4452	0.5560	0.9656	0.9863	0.6055	0.8814
	GSDM [29]	23.00	0.9779	0.0223	0.9039	53.30	0.4434	0.5556	0.9617	0.9831	0.5617	0.8761
	Restormer [26]	20.43	0.9629	0.0415	0.8988	53.11	0.4411	0.5533	0.8993	0.9479	0.4925	0.8088
	MambaIR [5]	23.13	0.9782	0.0219	0.8940	53.35	0.4458	0.5561	0.9585	0.9818	0.5900	0.8734
	PromptIR [10]	21.11	0.9672	0.0347	0.8904	53.15	0.4405	0.5551	0.9287	0.9681	0.5221	0.8394
	MoCE-IR [25]	20.03	0.9583	0.0450	0.8890	53.00	0.4421	0.5556	0.8943	0.9516	0.4971	0.8103
	IR3 [30]	22.82	0.9771	0.0213	0.9019	53.34	0.4443	0.5556	0.9815	0.9954	0.6769	0.9077
	EpiAgent (Ours)	**24.06**	**0.9817**	**0.0163**	**0.9064**	**53.38**	**0.4467**	**0.5598**	**0.9945**	**0.9992**	**0.7118**	**0.9094**
Middle	CharFormer [12]	19.45	0.9501	0.0483	0.8758	52.75	0.4342	0.5547	0.9150	0.9584	0.5402	0.8302
	DocDiff [23]	20.31	0.9587	0.0361	0.8925	53.28	0.4442	0.5559	0.9337	0.9683	0.5420	0.8500
	GSDM [29]	20.66	0.9611	0.0378	0.8872	53.14	0.4428	0.5542	0.9216	0.9595	0.4887	0.8368
	Restormer [26]	18.23	0.9370	0.0685	0.8910	52.83	0.4393	0.5510	0.8053	0.8839	0.4472	0.7197
	MambaIR [5]	20.75	0.9618	0.0375	0.8930	53.23	0.4443	0.5558	0.9161	0.9566	0.5476	0.8308
	PromptIR [10]	18.97	0.9449	0.0563	0.8896	52.94	0.4394	0.5543	0.8606	0.9263	0.4704	0.7749
	MoCE-IR [25]	18.08	0.9341	0.0691	0.8868	52.73	0.4403	0.5543	0.8093	0.8938	0.4413	0.7268
	IR3 [30]	20.43	0.9599	0.0355	0.8932	53.25	0.4440	0.5549	0.9690	0.9893	0.6087	0.8816
	EpiAgent (Ours)	**21.92**	**0.9696**	**0.0252**	**0.8971**	**53.30**	**0.4449**	**0.5581**	**0.9851**	**0.9951**	**0.6588**	**0.9080**
Severe	CharFormer [12]	15.43	0.8957	0.0942	0.8665	52.27	0.4269	0.5533	0.7469	0.8238	0.2769	0.6800
	DocDiff [23]	15.75	0.9047	0.0784	0.8844	52.22	0.4392	0.5535	0.7886	0.8657	0.3280	0.7073
	GSDM [29]	16.16	0.9100	0.0873	0.8815	52.63	0.4361	0.5523	0.8901	0.9414	0.5368	0.6583
	Restormer [26]	14.56	0.8749	0.1353	0.8759	51.92	0.4322	0.5431	0.5597	0.6806	0.2155	0.4731
	MambaIR [5]	16.39	0.9138	0.0855	0.8863	52.79	0.4411	0.5533	0.7346	0.8217	0.3130	0.6574
	PromptIR [10]	15.19	0.8898	0.1112	0.8760	52.41	0.4327	0.5514	0.6532	0.7623	0.2442	0.5750
	MoCE-IR [25]	14.70	0.8778	0.1271	0.8700	52.03	0.4318	0.5502	0.5987	0.7142	0.2303	0.5224
	IR3 [30]	15.80	0.9048	0.0804	0.8905	52.84	0.4346	0.5510	0.8948	0.9330	0.4227	0.7804
	EpiAgent (Ours)	**17.22**	**0.9238**	**0.0539**	**0.8932**	**53.17**	**0.4414**	**0.5563**	**0.9685**	**0.9776**	**0.4878**	**0.9007**

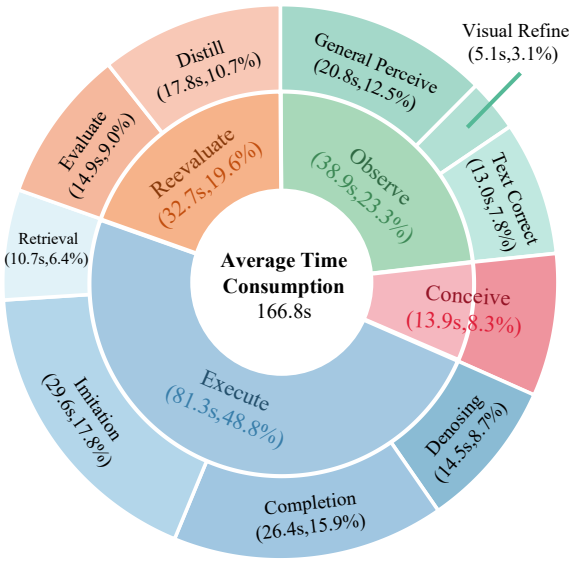

Figure 5. Average time consumption percentage for EpiAgent at different stages.

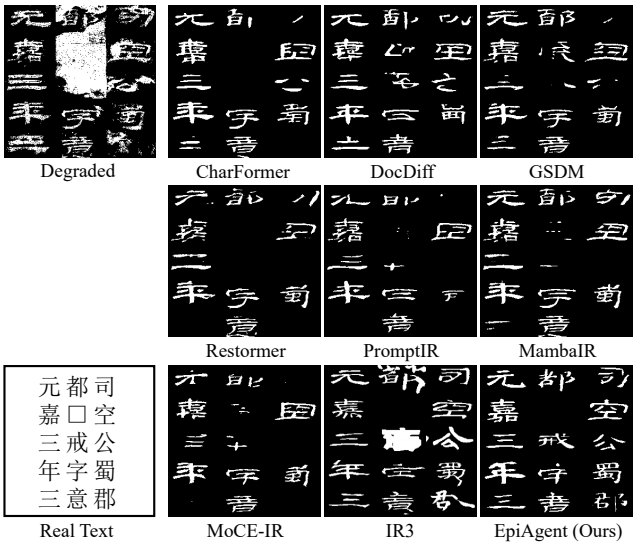

Figure 6. Failed restoration by different methods on inscription images with extreme degradations.

Proceedings of the Annual AAAI Conference on Artificial Intelligence, pages 2555–2563, 2023. 1

[17] Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. Seedeit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*, 2025. 3

[18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1

[19] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 4

[20] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muen-

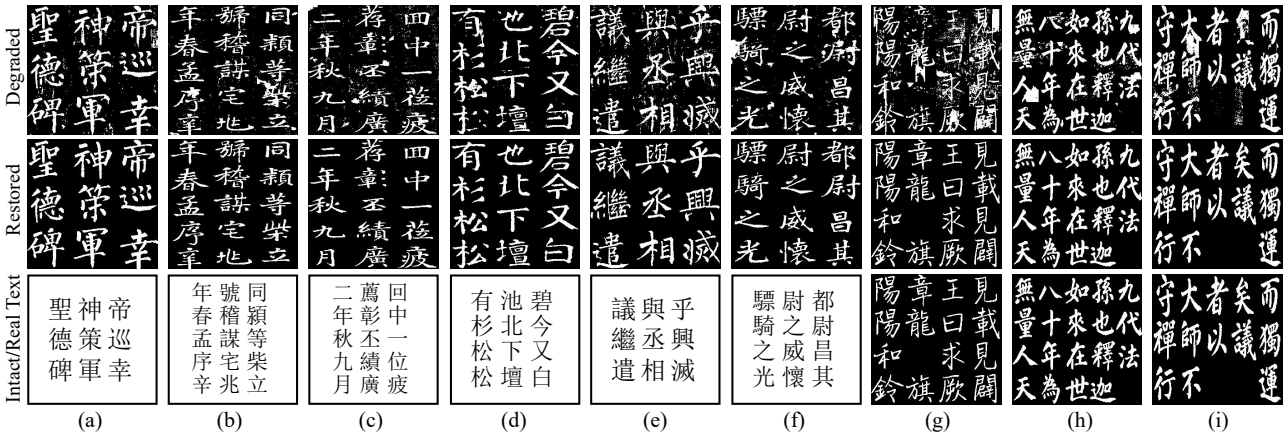

Figure 7. Restoration results of EpiAgent on slightly degraded inscription images. (a)-(c) are from Test Set R-I, (d)-(f) belong to Test Set R-II, and (g)-(i) belong to Test Set S.

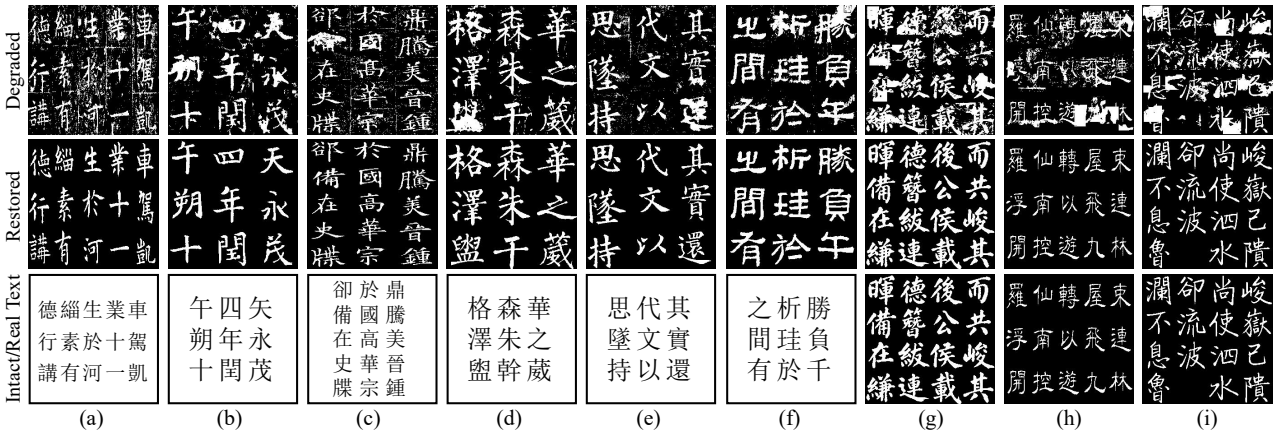

Figure 8. Restoration results of EpiAgent with mediumly degraded inscription images. (a)-(c) are from Test Set R-I, (d)-(f) belong to Test Set R-II, and (g)-(i) belong to Test Set S.

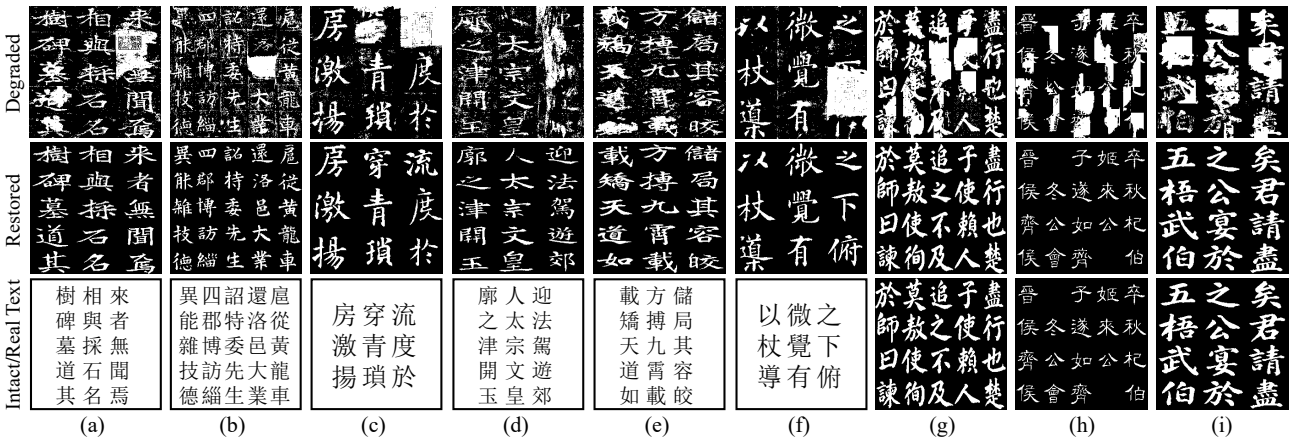

Figure 9. Restoration results of EpiAgent with severely degraded inscription images. (a)-(c) are from Test Set R-I, (d)-(f) belong to Test Set R-II, and (g)-(i) belong to Test Set S.

nighoff. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023. 2

[21] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1

[22] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 1

[23] Zongyuan Yang, Baolin Liu, Yongping Xxiong, Lan Yi, Guibin Wu, Xiaojun Tang, Ziqi Liu, Junjie Zhou, and Xing Zhang. Docdiff: Document enhancement via residual diffusion models. In *Proceedings of the ACM International Conference on Multimedia*, pages 2795–2806, 2023. 2, 7

[24] Zhenhua Yang, Dezhi Peng, Yuxin Kong, Yuyi Zhang, Cong Yao, and Lianwen Jin. Fontdiffuser: One-shot font generation via denoising diffusion with multi-scale content aggregation and style contrastive learning. In *Proceedings of the Annual AAAI Conference on Artificial Intelligence*, pages 6603–6611, 2024. 2

[25] Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12763, 2025. 7

[26] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 7

[27] Lukas Zbinden, Lars Doorenbos, Theodoros Pissas, Adrian Thomas Huber, Raphael Sznitman, and Pablo Márquez-Neila. Stochastic segmentation with conditional categorical diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1119–1129, 2023. 2

[28] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 1

[29] Shipeng Zhu, Pengfei Fang, Chenjie Zhu, Zuoyan Zhao, Qiang Xu, and Hui Xue. Text image inpainting via global structure-guided diffusion models. In *Proceedings of the Annual AAAI Conference on Artificial Intelligence*, pages 7775–7783, 2024. 7

[30] Shipeng Zhu, Hui Xue, Na Nie, Chenjie Zhu, Haiyue Liu, and Pengfei Fang. Reproducing the past: A dataset for benchmarking inscription restoration. In *Proceedings of the ACM International Conference on Multimedia*, pages 7714—–7723, 2024. 1, 2, 7