BEYOND PIXELS: SVG TOKEN PREDICTION FOR HISTORICAL DOCUMENT RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Historical document restoration (HDR) aims to recover damaged content while preserving the original appearance of historical documents. Diffusion-based and MLLM-based methods have made promising progress in HDR. However, they still provide limited control over character geometry and unintentionally alter preserved character structures. Scalable Vector Graphics (SVG) provide a vector-based representation of character geometry through explicit paths and coordinates. To leverage this representation, we reformulate HDR as **SVG token prediction**, where surviving character structures are retained and only the missing parts are recovered in vector space. To support this formulation, we introduce VECODEX, the first large-scale vector-based HDR dataset, containing 35.7M aligned restoration instances across four realistic damage types. Its construction combines memory-guided multi-agent quality control with Boolean shape splitting for precise raster–vector alignment. We further propose SCRIPT, which jointly utilizes pixel-level visual evidence and vector-level character structure to predict missing SVG tokens while preserving observed paths. Experiments on VECODEX show that SCRIPT achieves state-of-the-art performance in both character restoration and final document quality, substantially improving Mask IoU and Mask B-F_1 by 23.74% and 38.36% over the strongest baseline, respectively. These results demonstrate the potential of explicit vector-space structure recovery as a new direction for HDR beyond direct pixel generation.

1 INTRODUCTION

Historical document restoration (HDR) aims to recover damaged content while preserving the original appearance of historical documents (Yang et al., 2025; Zhang et al., 2025). It plays an important role in preserving valuable linguistic, historical, and cultural records, while supporting applications such as document digitization and historical text analysis (Nikolaidou et al., 2022). Recent HDR methods mainly rely on diffusion-based generation to reconstruct damaged regions (Yang et al., 2023; 2025). However, such methods focus on pixel-level visual reconstruction without explicitly modeling character structure, which can lead to broken strokes or incomplete recovery, as shown in Fig. 1. More recently, MLLM-based restoration methods have attracted increasing attention due to their strong image understanding and generation capabilities, enabling effective recovery of damaged content (Zhang et al., 2025). However, their generative process may unintentionally alter preserved strokes, other characters, or surrounding backgrounds. Although these methods have achieved promising progress, their restoration results can still be unsatisfactory.

These limitations motivate us to explore a representation that provides more explicit control over character structure during restoration. Scalable Vector Graphics (SVG) provide a vector-based representation of visual content through explicit paths, drawing commands, and coordinates (Carlier et al., 2020; Rodriguez et al., 2025). This representation naturally motivates us to hypothesize that HDR can be reformulated as **SVG-based missing-structure prediction**. Following this hypothesis, our core idea is to preserve the surviving SVG paths and predict only the **missing SVG tokens**, as illustrated in Fig. 1(c). The predicted paths are then combined with the preserved ones to reconstruct the complete character.

Realizing this formulation requires addressing two challenges. **(1) Lack of high-quality restoration data.** Existing HDR datasets, such as TKH (Yang et al., 2018), MTH (Yang et al., 2018), and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

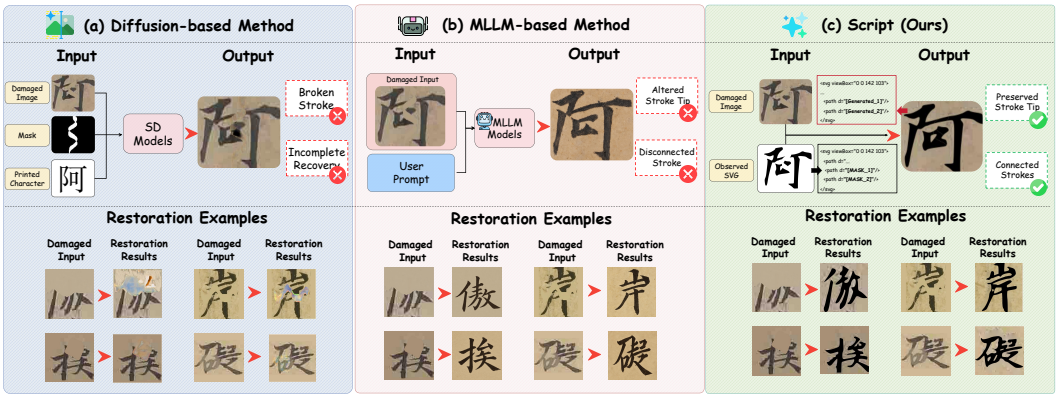


Figure 1: Comparison of SCRIPT with other restoration methods and additional examples.

HDR28K (Yang et al., 2025), provide valuable restoration data but remain limited in scale, structural fidelity, or damage diversity. Constructing high-quality SVG-based HDR data is non-trivial, as character foregrounds must be accurately separated and vectorized from complex historical backgrounds, while pixel-level damage masks cannot be directly mapped to the corresponding missing SVG structures. **(2) Lack of methods for missing SVG prediction.** Existing HDR methods mainly focus on pixel reconstruction, while current SVG models are designed for complete vector generation (Carlier et al., 2020; Rodriguez et al., 2025). The key challenge is how to exploit two complementary but incomplete sources of evidence: damaged raster images may retain residual stroke cues, while observed SVG paths provide precise but partial character geometry. These heterogeneous cues must be jointly aligned to infer the missing paths without regenerating the preserved structures.

In this paper, we introduce VECODEX, the first large-scale vector-based dataset for historical document restoration, containing over 2.2M source character image–SVG pairs and 35.7M restoration instances across four simulated damage types that mimic realistic document degradation, including edge damage, holes, tears, and water stains. To ensure data quality, we develop a memory-guided multi-agent construction system that adaptively processes historical document scans according to sample quality. To overcome the mismatch between pixel-level damage masks and vector-level character structures, we introduce Boolean shape splitting, enabling each raster damage region to be precisely mapped to its corresponding observed and missing SVG geometry. These designs result in a large-scale, diverse, and structurally aligned benchmark that provides high-quality data for advancing historical document restoration under the new paradigm of SVG token prediction.

To address the second challenge, we propose **S**tructured **C**ross-modal **R**estoration via **I**mage–**P**ath **T**okens (SCRIPT), a model for historical document restoration that utilizes both pixel-level visual evidence and vector-level character structure. SCRIPT establishes their cross-modal correspondence so that visual cues from damaged images, such as residual ink under water stains, can guide the recovery of missing character structures, while the observed vector paths provide explicit geometric constraints. By integrating these complementary representations, SCRIPT recovers the missing structures without unnecessarily modifying the preserved character content.

Experiments on VECODEX show that SCRIPT achieves state-of-the-art performance across both character restoration and final document quality. For character restoration, SCRIPT improves IoU and B-F_1 by 12.45% and 13.10% over DiffHDR (Yang et al., 2025), respectively, while further improving Mask IoU and Mask B-F_1 by 23.74% and 38.36% over LaMa (Suvorov et al., 2022). For final document restoration, SCRIPT increases $PSNR_M$ from 15.99 to 17.16 dB compared with DiffHDR (Yang et al., 2025), and reduces LPIPS by 14.29% compared with the second-best LaMa (Suvorov et al., 2022). We further analyze the contribution of joint pixel–vector reasoning and observe that it is particularly beneficial when damaged strokes remain partially observable, such as under water stains. These results demonstrate that recovering missing character structures in SVG space provides a promising new paradigm for HDR.

Our contributions are fourfold:

- We reformulate HDR as prediction of missing SVG tokens, shifting restoration from direct pixel generation toward explicit recovery of missing character structures and enabling more controllable character-shape restoration.

- We introduce VECODEX, the first large-scale vector-based dataset for HDR, containing over 2.2M source samples and 35.7M restoration instances across four damage types.
- We propose SCRIPT, which jointly utilizes pixel-level visual evidence and vector-level structural information to recover damaged character structures.
- Extensive experiments on VECODEX show that SCRIPT achieves state-of-the-art performance in both character restoration and final document quality, improving Mask IoU and Mask B-F_1 by 23.74% and 38.36% over the strongest baselines, respectively.

2 RELATED WORK

Historical Document Restoration. Document image restoration is commonly formulated as raster-to-raster reconstruction from degraded observations (Souibgui & Kessentini, 2020; Yang et al., 2023). Existing methods explore adversarial learning for document clean-up (Souibgui & Kessentini, 2020), diffusion models for recovering textual details (Yang et al., 2023), and prompt- or instruction-based restoration for diverse document degradations (Zhang et al., 2024; Li et al., 2026). For historical documents, DiffHDR introduces semantic and spatial guidance into diffusion-based restoration (Yang et al., 2025), while AutoHDR combines OCR-assisted localization with vision–language prediction and autoregressive generation (Zhang et al., 2025). Data construction has also been explored through synthetic degradation pipelines such as DocCreator (Journet et al., 2017) and Augraphy (Groleau et al., 2023), together with historical document datasets including TKH and MTH (Yang et al., 2018), as well as HDR28K (Yang et al., 2025). Despite these advances, existing HDR methods remain largely centered on pixel-level reconstruction, where character structure is modeled only implicitly and fine-grained control over character geometry remains difficult. Meanwhile, the limited scale and quality of existing HDR data continue to restrict the development of more reliable restoration methods.

Vector Graphics and SVG Reconstruction. Learning-based vector graphics methods represent shapes through drawing commands and coordinates, enabling explicit modeling of geometric structure. Representative methods include DeepSVG for hierarchical vector generation and reconstruction (Carlier et al., 2020), Im2Vec for raster-to-vector learning through differentiable rasterization (Reddy et al., 2021), and StarVector for predicting SVG code from images or text (Rodriguez et al., 2025). Vector-font methods further combine raster appearance with vector outlines for glyph generation and reconstruction, including DeepVecFont (Wang & Lian, 2021), DeepVecFont-v2 (Wang et al., 2023), DualVector (Liu et al., 2023), and VecGlypher (Huang et al., 2026). SketchRevive additionally explores partial-shape recovery through pixel completion and vectorization (Zuo et al., 2026). Despite their ability to represent geometry explicitly, existing SVG methods mainly focus on general graphics or font generation and reconstruction. Their potential for HDR, where pixel-level visual evidence and vector-level character structure can be jointly utilized, remains largely unexplored. We therefore introduce SVG-based representation into HDR and formulate restoration as **prediction of missing SVG tokens**.

3 VECODEX: DATASET CONSTRUCTION

Overview. Fig 2 summarizes the overall construction process of our dataset VECODEX. Due to uneven scan conditions and complex historical backgrounds, we propose a memory-guided multi-agent construction pipeline for reliable character extraction and conversion. Furthermore, we propose Boolean shape splitting to precisely align damaged image regions with their corresponding character structures, producing consistent observed and target representations for restoration.

3.1 MULTI-AGENT CONSTRUCTION WITH MEMORY

Historical document scans often exhibit uneven quality: enhancement may recover faint strokes but can also amplify stains and background contamination. Fixed preprocessing strategies therefore struggle to adapt to such sample-specific variations. To address this issue, we propose a memory-guided multi-agent construction framework that combines complementary quality assessment with human-validated references to guide adaptive processing and produce quality-controlled source image–SVG pairs.

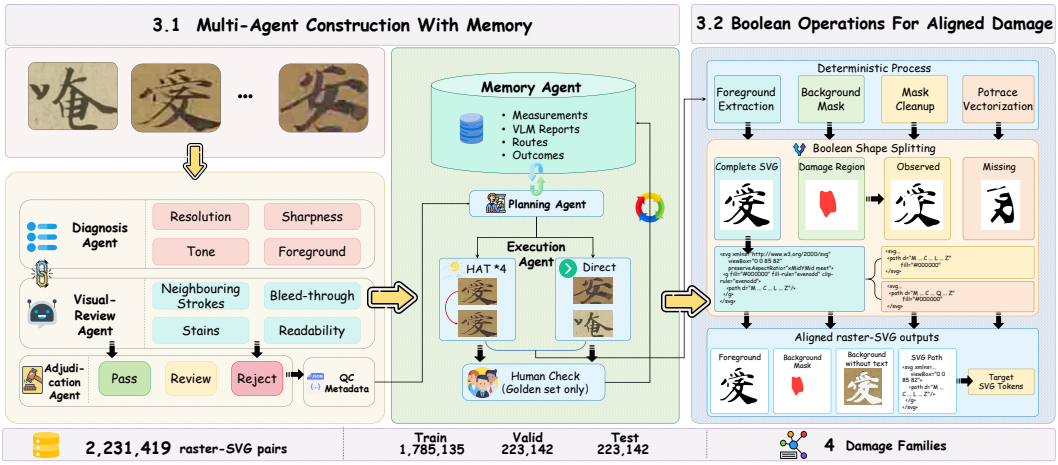


Figure 2: Overview of VECODEX construction.

For each candidate image I_i, the **diagnosis agent** measures resolution, sharpness, contrast, foreground separation, and fragmentation, while the **visual-review agent**, built on Qwen3.5-9B (Qwen Team, 2026), assesses visually complex degradations such as stains, bleed-through, neighbouring strokes, and character completeness. The **adjudication agent** then consolidates these complementary signals into a unified quality assessment, reducing the risk that numerical measurements alone mistake contamination for useful character detail.

To incorporate human quality preferences, we construct a human-annotated golden set of 14,000 images, including 11,200 memory references and an independent 2,800-image audit subset. Each reference records its quality measurements, visual assessment, processing route, and human-validated outcome. For a new image, the **memory agent** retrieves quality-similar references and provides them with the current assessment to the **planning agent**, which determines whether HAT $\times 4$ super-resolution (Chen et al., 2023) is required before vectorization, allowing decisions to draw on both the current sample and similar human-validated cases.

Finally, the **execution agent** applies the selected processing strategy, using HAT $\times 4$ super-resolution (Chen et al., 2023) when required and otherwise preserving the original resolution. We then extract and clean the character foreground and perform Potrace vectorization (Selinger, 2003) to obtain SVG V_i. The resulting quality-controlled image–SVG pairs provide reliable source data for the aligned damage construction described in the next stage.

3.2 BOOLEAN OPERATIONS FOR ALIGNED DAMAGE

Based on the quality-controlled image–SVG pairs obtained above, we construct aligned restoration instances by applying controlled damage to both raster and vector representations. However, a pixel-level damage mask cannot directly identify the missing SVG structure, since a single SVG path may span damaged and intact regions. To address this issue, we propose **Boolean shape splitting** to fill the gap between pixel-level damage masks and vector-level character structures.

For each source SVG V_i, we map its filled shape to a geometric region as follows:

$$G_i = \Gamma(V_i) \subseteq \Omega, \tag{1}$$

where Ω denotes the shared canvas and $\Gamma(\cdot)$ maps an SVG to its filled geometric region. We generate four damage types, including wormholes, edge loss, tears, and water stains, at ratios of 10%, 20%, 30%, and 40% of the character area, yielding 16 configurations per source. Here, $B_i^k \subseteq \Omega$ denotes the damage region of configuration $k \in \{1, \ldots, 16\}$.

We then partition the character geometry using Boolean difference and intersection, which can be formulated as follows:

$$G_{\text{obs},i}^k = G_i \setminus B_i^k, \qquad G_{\text{miss},i}^k = G_i \cap B_i^k. \tag{2}$$

The two regions form a complete partition of the original character geometry, satisfying:

$$G^k_{\mathrm{obs},i} \cup G^k_{\mathrm{miss},i} = G_i, \qquad G^k_{\mathrm{obs},i} \cap G^k_{\mathrm{miss},i} = \varnothing. \tag{3}$$

We encode them as the observed SVG $V^k_{\mathrm{obs},i}$ and missing SVG $V^k_{\mathrm{miss},i}$ in the same coordinate frame. This allows intact character structures to remain preserved even when the damage boundary cuts through an individual SVG path.

To establish pixel–vector correspondence, we apply the same damage region to the raster image. Specifically, the binary damage mask is obtained as:

$$M^k_i = \mathbf{1}\left[\mathcal{R}(B^k_i) > 0\right], \tag{4}$$

where $\mathcal{R}$ maps the geometric damage region to pixel coverage. Applying the corresponding degradation within M^k_i produces the damaged image $\widetilde{I}^k_i$. For wormholes, edge loss, and tears, the affected regions are replaced with an estimated paper colour. For water stains, we instead fade the affected strokes so that residual visual evidence remains available even though the corresponding structure is removed from the observed SVG.

Each restoration instance is therefore defined by the aligned inputs $(\widetilde{I}^k_i, M^k_i, V^k_{\mathrm{obs},i})$, with the missing SVG serialized as the target token sequence, which can be mathematically expressed as:

$$S^k_{\mathrm{miss},i} = \mathrm{Tok}_\Omega\left(V^k_{\mathrm{miss},i}\right), \tag{5}$$

where Tok_Ω converts SVG drawing commands and coordinates into a token sequence. The original image I_i and SVG V_i are retained as references for evaluating the restored document and recovered character structure.

We additionally construct background-restoration pairs by sampling damage regions outside the character area $T_i = \mathbf{1}[\mathcal{R}(G_i) > 0]$ and applying the same degradation process to obtain $I^{\mathrm{in}}_{\mathrm{bg},i}$, while retaining I_i as the restoration target. This provides supervision for recovering the surrounding paper without altering the original character structure.

Overall, Boolean shape splitting converts each quality-controlled source pair into 16 precisely aligned restoration instances, together with corresponding background-restoration pairs. By using the same damage geometry across raster and vector spaces, the resulting data establish consistent pixel–vector correspondence and provide aligned supervision for training and evaluating SCRIPT.

3.3 DATA STATISTICS

VECODEX contains 2,231,419 source samples covering 9,525 character types and 35,702,704 aligned restoration instances. We split the source samples into training, validation, and test sets with an 8:1:1 ratio. Table 1 summarizes the dataset statistics and comparison to existing datasets.

Each source image has four simulated damage types and four mask ratios (10%, 20%, 30%, and 40%), yielding 16 restoration variants. All variants from the same source remain in the same split, while background restoration uses an 80:20 train–validation split.

We evaluate the framework on 2,800 independent audit images, with the remaining 11,200 human-annotated images used as memory references. Among audit images assigned *pass*, 96.13% require no additional enhancement before vectorization, supporting the reliability of the quality-control process.

Table 1: Statistics and comparison of VECODEX.

VECODEX *statistics*				
Source samples	2,231,419			
Character types	9,525			
Restoration instances	35,702,704			
Source split	1,785,135 / 223,142 / 223,142			
Comparison with existing datasets				
Dataset	Characters	Damage	SVG	Align.
TKH (Yang et al., 2018)	323,491	–	×	×
MTH (Yang et al., 2018)	197,886	–	×	×
HDR28K (Yang et al., 2025)	–	3	×	×
VECODEX	2,231,419	4	✓	✓

4 SCRIPT: STRUCTURED CROSS-MODAL RESTORATION VIA IMAGE–PATH TOKENS

Overview. We propose SCRIPT, a historical document restoration framework that performs mask-aware joint reasoning over image and SVG path tokens. It jointly exploits residual visual evidence

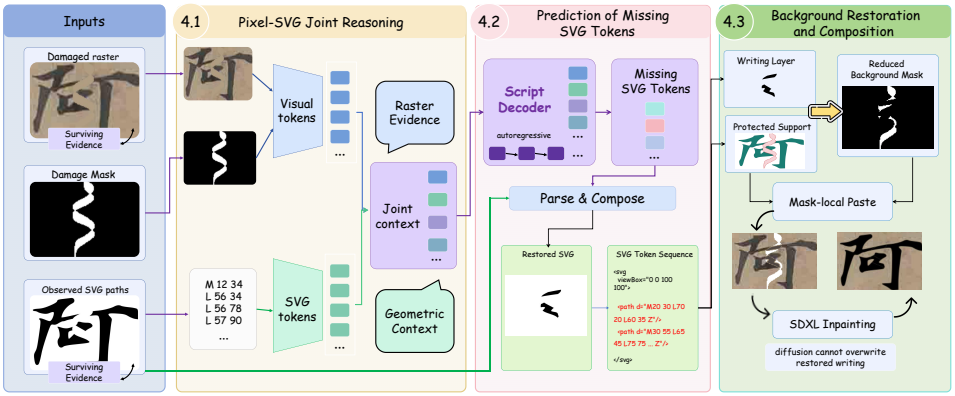


Figure 3: Overview of SCRIPT. The damaged raster, damage mask, and observed SVG paths form the context for missing SVG token prediction. The decoded paths are combined with the observed paths. The rendered characters are protected while SDXL repairs the remaining background region.

and observed SVG geometry to predict missing SVG tokens while preserving existing paths. The recovered SVG is then combined with background restoration to produce the final document image, as illustrated in Fig. 3.

Following Sec. 3.2, the inputs are a damaged image $\widetilde{I} \in \mathbb{R}^{H\times W\times 3}$, a binary mask $M \in \{0,1\}^{H\times W}$, and observed SVG paths V_{obs}. Images, masks, and SVG coordinates share the canvas Ω, with $M = 1$ indicating regions to restore. We omit the source and damage-configuration indices below. The outputs are a restored SVG $\widehat{V}$ and final image $\widehat{I}$.

4.1 PIXEL–SVG JOINT REASONING

We exploit two complementary sources of information for character recovery. The observed SVG provides explicit geometry of the surviving strokes, while the damaged raster image may retain visual evidence unavailable in the partial vector structure, such as residual ink under water stains. We therefore propose mask-aware pixel–SVG joint reasoning to integrate these cues, with the damage mask explicitly indicating where restoration is required.

Specifically, we encode the damaged image and mask as $Z_I = E_{\text{vis}}(\widetilde{I})$ and $Z_M = E_{\text{vis}}(M)$, where E_{vis} projects both inputs into the decoder embedding space and the binary mask is represented at the same resolution as the raster image. We serialize the observed paths as $S_{\text{obs}} = \text{Tok}_\Omega(V_{\text{obs}})$ and obtain their token embeddings $Z_V = E_{\text{tok}}(S_{\text{obs}})$. We then concatenate the three representations along the sequence dimension:

$$C = [Z_I; Z_M; Z_V]\,. \tag{6}$$

The resulting context jointly captures residual visual evidence, damage location, and surviving character geometry for subsequent missing SVG token prediction. A fixed task instruction and chat delimiters are included in every input but omitted from C for clarity, and no transcription, character label, or sample-specific semantic description is provided.

4.2 PREDICTION OF MISSING SVG TOKENS

Following the Boolean splitting in Eq. 5, we preserve the observed SVG paths and predict only the missing SVG tokens rather than regenerating the complete character. This keeps restoration focused on the missing structure while retaining the existing geometry. We define the target sequence as:

$$S_{\text{miss}} = \text{Tok}_\Omega(V_{\text{miss}}) = (s_1, \ldots, s_L), \tag{7}$$

where V_{miss} contains the missing SVG paths and each $s_\ell \in \mathcal{U}$ represents a drawing command, coordinate value, or separator. Given the joint context C, we autoregressively model the target sequence as:

$$p_\theta(S_{\text{miss}} \mid C) = \prod_{\ell=1}^{L} p_\theta(s_\ell \mid s_{<\ell}, C), \tag{8}$$

where $s_{<\ell}$ denotes the preceding target tokens. During training, we provide the ground-truth prefix and optimize only the target SVG tokens:

$$\mathcal{L}_{\text{svg}} = -\frac{1}{L}\sum_{\ell=1}^{L} \log p_\theta(s_\ell \mid s_{<\ell}, C). \tag{9}$$

We instantiate the SVG predictor with Qwen3.5-9B (Qwen Team, 2026) and adapt it with LoRA (Hu et al., 2022) while freezing the base model.

At inference, we decode the missing SVG tokens and parse them into the corresponding vector paths as follows:

$$\widehat{S}_{\text{miss}} = \text{Decode}_\theta(C), \qquad \widehat{V}_{\text{miss}} = \text{Parse}(\widehat{S}_{\text{miss}}). \tag{10}$$

We then combine the predicted missing paths with the preserved observed paths:

$$\widehat{V} = \text{Compose}(V_{\text{obs}}, \widehat{V}_{\text{miss}}). \tag{11}$$

The resulting $\widehat{V}$ recovers the complete character geometry without regenerating the observed structure and is subsequently used for background restoration.

4.3 Background Restoration and Composition

We decouple character recovery from background synthesis so that the recovered character structure is not regenerated during pixel-level inpainting. We first render the restored SVG $\widehat{V}$ onto the damaged image, producing an RGB layer F and opacity map A. The recovered character is inserted only within the original damage region:

$$I_{\text{paste}} = (1 - M \odot A) \odot \widetilde{I} + (M \odot A) \odot F, \tag{12}$$

where $\odot$ denotes element-wise multiplication.

We then protect the recovered character region by defining $\widehat{T} = \text{Dilate}(\mathbf{1}[A > 0])$ and removing it from the original damage mask as follows:

$$M_{\text{bg}} = M \odot (1 - \widehat{T}). \tag{13}$$

This restricts subsequent inpainting to the remaining damaged background. We fine-tune an SDXL 1.0 inpainting model (Podell et al., 2024) with LoRA on the background-restoration pairs constructed in Sec. 3.2, using I_{paste} and M_{bg} to generate the missing background content.

Finally, we compose the generated background with I_{paste} only inside M_{bg} to obtain the restored document image $\widehat{I}$. In this way, SCRIPT completes the surrounding paper appearance while preserving the recovered character geometry, yielding the final restoration output of the overall framework.

5 Experiments

We organize the experiments around four questions: **Q1: Restoration formulation.** How does SCRIPT's missing SVG token prediction compare with raster-based restoration methods in recovering character shapes? **Q2: Joint reasoning.** Does residual visual evidence improve missing SVG token prediction beyond the observed SVG context? **Q3: Background restoration.** Does placing restored characters before background inpainting improve final document quality? **Q4: Robustness.** How does SCRIPT perform across different damage ratios compared with existing methods?

5.1 Experimental Setup

Data. We evaluate all methods on the VECODEX test set, containing 223,142 source samples and 3,570,272 restoration instances across four damage types and four ratios of 10%, 20%, 30%, and 40%. For character restoration, we use rasterized source SVGs as references to evaluate damaged character recovery without background texture. For document restoration, we compare the final restored images with the original source images.



Figure 4: Qualitative comparison of document restoration and character restoration. Each row shows a damaged input and its mask.

Table 2: Restoration results on VECODEX.

	Character restoration		Masked-region restoration		Final document image	
Method	IoU↑	B-F_1↑	Mask IoU↑	Mask B-F_1↑	$PSNR_M$ (dB)↑	LPIPS↓
LaMa (Suvorov et al., 2022)	69.96%	64.10%	56.39%	41.51%	13.83	0.21
ZITS (Dong et al., 2022)	69.38%	63.80%	52.21%	33.59%	14.84	0.21
DiffHDR (Yang et al., 2025)	75.75%	69.90%	27.78%	26.28%	15.99	0.25
SDXL (Podell et al., 2024)	65.58%	60.20%	39.27%	19.54%	10.71	0.34
SCRIPT	**88.20%** (+12.45%)	**83.00%** (+13.10%)	**80.13%** (+23.74%)	**79.87%** (+38.36%)	**17.16** (+1.17)	**0.18** (-0.03)

Baselines. We compare SCRIPT with four representative restoration methods: LaMa (Suvorov et al., 2022), ZITS (Dong et al., 2022), DiffHDR (Yang et al., 2025), and SDXL inpainting (Podell et al., 2024). LaMa, ZITS, and SDXL use the damaged image and mask, while DiffHDR additionally requires a printed-character prior. SCRIPT uses the damaged image, mask, and observed SVG paths.

Metrics. We evaluate character restoration using Character IoU and Character B-F_1 (2 px) over the complete character, and Mask IoU and Mask B-F_1 (2 px) within the damaged region. For final document restoration, we report $PSNR_M$ and LPIPS (Zhang et al., 2018), together with qualitative comparisons of stroke continuity and restoration quality.

5.2 COMPARISON WITH EXISTING METHODS

SCRIPT achieves state-of-the-art performance across character restoration, masked-region restoration, and final document quality, outperforming all evaluated baselines on every reported metric, as shown in Table 2. On complete characters, SCRIPT achieves 88.20% IoU and 83.00% B-F_1, exceeding DiffHDR by 12.45% and 13.10%, respectively. The gains are even larger within damaged regions, where SCRIPT reaches 80.13% Mask IoU and 79.87% Mask B-F_1, improving over LaMa by 23.74% and 38.36%. For final document restoration, SCRIPT further improves $PSNR_M$ from 15.99 to 17.16 dB and reduces LPIPS from 0.21 to 0.18. Qualitative results in Fig. 4 show a consistent pattern: several baselines produce broken or incomplete strokes and may alter nearby preserved structures, whereas SCRIPT recovers more complete and continuous character shapes while better preserving the original geometry.

For Q1: These results show that SCRIPT recovers character structures more accurately than the evaluated raster-based methods. In particular, the larger gains within damaged regions indicate that the improvement mainly comes from more accurate reconstruction of missing strokes rather than from preserved character regions. Together with the qualitative evidence, this supports our hypothesis that missing SVG token prediction provides a more structurally stable formulation for character restoration.

5.3 ABLATION STUDIES

We examine two design choices: combining raster and SVG evidence, and using recovered characters to guide background restoration. Table 3 summarizes the results.

For joint reasoning, we compare the full model with an SVG-only variant that removes raster evidence from Eq. 6 while retaining the observed SVG and damage mask. Both variants predict the same missing SVG tokens under identical training settings. Table 3(a) reports the gain from adding raster evidence, averaged across the four damage ratios for each damage type.

For Q2: Raster evidence improves both character metrics across all damage types, with the largest gains under water stains: +6.67% in IoU and +6.94% in B-F_1. Since water stains retain faint stroke evidence in the raster while the corresponding structure is absent from the observed SVG, these results show that residual visual cues help recover missing SVG tokens when strokes remain partially observable.

For background restoration, we compare placing the recovered SVG before and after background inpainting. Both variants use the same predicted SVG, SDXL checkpoint, and sampling settings.

Table 3: Ablation studies.

(a) Raster evidence

Damage	IoU Δ	B-F_1 Δ
Edge loss	+2.14	+2.33
Wormholes	+2.31	+1.98
Tears	+2.41	+2.28
Water stains	+6.67	+6.94

(b) Character before inpainting

Setting	$PSNR_M$	LPIPS
Direct	10.56	0.24
Vector-first	17.16	0.18

For Q3: placing the recovered character first improves $PSNR_M$ from 10.56 to 17.16 dB and reduces LPIPS from 0.24 to 0.18, as shown in Table 3(b). This indicates that preserving the recovered character before inpainting provides a more reliable context for completing the surrounding background.

5.4 Comparison across Damage Ratios

We compare SCRIPT with LaMa, ZITS, DiffHDR, and SDXL across damage ratios of 10% to 40%.

As shown in Fig. 5, SCRIPT achieves the best IoU and B-F_1 across all damage types and ratios. Averaged across damage types, its IoU decreases from 94.0% at 10% damage to 82.0% at 40%, while B-F_1 decreases from 88.5% to 76.5%. Compared with DiffHDR, the strongest baseline on average, the corresponding margins narrow from 17.0 and 17.2 points to 6.8 and 7.2 points as damage increases.

DiffHDR varies less across ratios, partly reflecting its additional printed-character prior, whereas SCRIPT relies on the observed SVG. Nevertheless, SCRIPT consistently maintains higher restoration accuracy across all settings.



Figure 5: Character IoU and B-F_1 by damage type and masked-area ratio.

For Q4: SCRIPT consistently outperforms existing methods across all tested damage ratios, demonstrating robust character restoration under increasing damage severity.

6 Conclusion

In this work, we reformulate historical document restoration as **prediction of missing SVG tokens**, enabling explicit recovery of missing character structures beyond direct pixel generation. We introduce VECODEX, the first large-scale vector-based HDR dataset with over 2.2M source samples and 35.7M aligned restoration instances. Beyond the dataset contribution, we further propose SCRIPT, which jointly exploits pixel-level visual evidence and vector-level character structure for restoration. Experiments show that SCRIPT achieves state-of-the-art performance in both character restoration and final document quality, with particularly strong gains in damaged regions, while further analyses verify the benefit of joint pixel–SVG reasoning. These results demonstrate the potential of explicit vector-space recovery for HDR, while future work may address extremely severe damage with limited surviving evidence.

AI USE STATEMENT

In this work, we used generative AI tools only to refine the wording of selected sentences. We did not use generative AI to generate experimental results, perform data analysis, or formulate scientific claims, and no other disclosure categories apply. The majority of the manuscript was written by the authors. All AI-assisted text was carefully reviewed and revised for accuracy, clarity, originality, and appropriateness. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we will provide an anonymized supplementary artifact at https://anonymous.4open.science/r/SCRIPT-5147/. The artifact includes the implementation of SCRIPT, dataset construction and evaluation code, training and inference configurations, and instructions for reproducing the reported experiments. We instantiate the SVG predictor with Qwen3.5-9B (Qwen Team, 2026) and apply LoRA (Hu et al., 2022) to its linear layers while freezing the base weights. The LoRA rank and scale are set to 16 and 32, respectively, with a dropout of 0.05. We train the model for three epochs on eight NVIDIA A800 GPUs with 80 GB memory using a learning rate of 2×10^{-4}, bfloat16, gradient checkpointing, and eight gradient-accumulation steps. During decoding, we set the maximum sequence length to 16,384 tokens and disable thinking mode.

REFERENCES

Alexandre Carlier, Martin Danelljan, Alexandre Alahi, and Radu Timofte. Deepsvg: A hierarchical generative network for vector graphics animation. *NeurIPS*, 33:16351–16361, 2020.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, pp. 22367–22377, 2023.

Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *CVPR*, pp. 11358–11368, 2022.

Alexander Groleau, Kok Wei Chee, Stefan Larson, Samay Maini, and Jonathan Boarman. Augraphy: A data augmentation library for document images. In *ICDAR*, pp. 384–401. Springer, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xiaoke Huang, Bhavul Gauri, Kam Woh Ng, Tony Ng, Mengmeng Xu, Zhiheng Liu, Weiming Ren, Zhaochong An, Zijian Zhou, Haonan Qiu, et al. Vecglypher: Unified vector glyph generation with language models. *arXiv preprint arXiv:2602.21461*, 2026.

Nicholas Journet, Muriel Visani, Boris Mansencal, Kieu Van-Cuong, and Antoine Billy. Doccreator: A new software for creating synthetic ground-truthed document images. *Journal of imaging*, 3 (4):62, 2017.

Heng Li, Xingyuan Wang, Yang Fan, Yunan Zhang, Xiangping Wu, and Qingcai Chen. Mmdir: Multimodal instruction-driven framework for mixed-degradation document image restoration. In *CVPR*, pp. 8387–8396, 2026.

Ying-Tian Liu, Zhifei Zhang, Yuan-Chen Guo, Matthew Fisher, Zhaowen Wang, and Song-Hai Zhang. Dualvector: Unsupervised vector font synthesis with dual-part representation. In *CVPR*, pp. 14193–14202. IEEE, 2023.

Konstantina Nikolaidou, Mathias Seuret, Hamam Mokayed, and Marcus Liwicki. A survey of historical document image datasets. *IJDAR*, 25(4):305–338, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Pradyumna Reddy, Michael Gharbi, Michal Lukac, and Niloy J Mitra. Im2vec: Synthesizing vector graphics without vector supervision. In *CVPR*, pp. 7342–7351, 2021.

Juan A Rodriguez, Abhay Puri, Shubham Agarwal, Issam H Laradji, Pau Rodriguez, Sai Rajeswar, David Vazquez, Christopher Pal, and Marco Pedersoli. Starvector: Generating scalable vector graphics code from images and text. In *CVPR*, pp. 16175–16186. IEEE, 2025.

Peter Selinger. Potrace: A polygon-based tracing algorithm. Technical report, September 2003. URL https://potrace.sourceforge.net/potrace.pdf.

Mohamed Ali Souibgui and Yousri Kessentini. De-gan: A conditional generative adversarial network for document enhancement. *TPAMI*, 44(3):1180–1191, 2020.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, pp. 2149–2159, 2022.

Yizhi Wang and Zhouhui Lian. Deepvecfont: synthesizing high-quality vector fonts via dual-modality learning. *TOG*, 40(6):1–15, 2021.

Yuqing Wang, Yizhi Wang, Longhui Yu, Yuesheng Zhu, and Zhouhui Lian. Deepvecfont-v2: Exploiting transformers to synthesize vector fonts with higher quality. In *CVPR*, pp. 18320–18328. IEEE, 2023.

Hailin Yang, Lianwen Jin, Weiguo Huang, Zhaoyang Yang, Songxuan Lai, and Jifeng Sun. Dense and tight detection of chinese characters in historical documents: Datasets and a recognition guided detector. *IEEE Access*, 6:30174–30183, 2018.

Zhenhua Yang, Dezhi Peng, Yongxin Shi, Yuyi Zhang, Chongyu Liu, and Lianwen Jin. Predicting the original appearance of damaged historical documents. In *AAAI*, volume 39, pp. 9382–9390, 2025.

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 *ACM MM*, pp. 2795–2806, 2023.

Jiaxin Zhang, Dezhi Peng, Chongyu Liu, Peirong Zhang, and Lianwen Jin. DocRes: A generalist model toward unifying document image restoration tasks. In *CVPR*, pp. 15654–15664, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595. IEEE, 2018.

Yuyi Zhang, Peirong Zhang, Zhenhua Yang, Pengyu Yan, Yongxin Shi, Pengwei Liu, Fengjun Guo, and Lianwen Jin. Reviving cultural heritage: A novel approach for comprehensive historical document restoration. In *ACL*, pp. 28876–28892, 2025.

Ran Zuo, Haoxiang Hu, Chenxi Pei, Yanxuan Liu, Wenwen Qiang, Fang Liu, Xiaoming Deng, Cuixia Ma, and Yong-Jin Liu. Sketchrevive: Fine-grained pixel-to-vector sketch completion with diffusion-prior-guided multimodal llms. In *CVPR*, pp. 43165–43174, 2026.