Learning Where to Restore: Character-Safe Oracle-Free Diffusion Restoration of Japanese Classical Documents

Hanzhe Gao, Honghui Yuan, Junwen Chen, Peilin Xiong, and Keiji Yanai*

The University of Electro-Communications, Japan
{gao-h,yuan-h,chen-j,xiong-p}@mm.inf.uec.ac.jp
yanai@cs.uec.ac.jp

Abstract. Diffusion restoration of historical documents commonly assumes that damaged regions are known at test time. We remove this assumption by learning where editing is permitted. DamageMaskNet predicts a support with Dice 0.9678; a mask-robust diffusion restorer produces a candidate inside that support, while exact compositing copies all pixels outside that support unchanged. We evaluate 144 full-page images from 16 source pages using image, stroke, Unicode-recognition, and public-OCR probes. The final hard-support method reaches 27.89 dB in damaged regions and 0.9372 stroke-aware damaged-character accuracy, while editing less than 1% of oracle-intact pixels. Matched oracle-only continued training and full-frame robust-restorer controls separate the effects of localization, additional optimization, and editable area. The system is intended as a conservative reading aid, not as proof of historically correct reconstruction.

Keywords: Historical document restoration · diffusion models · damage localization · kuzushiji · character fidelity

1 Introduction

Japanese classical documents preserve literary, administrative, and historical records, yet scans often contain stains, mold-like spots, smudges, bleed-through, or missing ink. Restoration should reveal obscured writing while leaving every readable stroke untouched. The second requirement is unusually important for cursive kuzushiji: a small generated mark may change one Unicode identity into another even when full-page PSNR remains high.

Recent document restoration models use diffusion priors to recover text-like structure [8,15,16,19]. Conditional diffusion and masked sampling can constrain generation to a chosen region [7,10], but this shifts a crucial decision to the input mask. A synthesis mask is available when constructing a controlled benchmark, not when processing an archival scan. Treating the whole page as editable avoids

* Corresponding author.

that oracle input but exposes every intact stroke to unnecessary generation. The deployment problem is therefore not only how to restore a region, but first deciding which pixels may be changed.

We study restoration when this mask is unavailable. A compact U-Net predicts a probability p_i that pixel i is damaged. Its calibration-set-selected hard support gives an auditable answer to the deployment question "which pixels may change?" and forces an exact copy everywhere else. The restorer architecture and image objective are initially frozen, separating localization quality from restoration capacity. A subsequent mask-error curriculum adapts the restorer to dilation, erosion, shifts, missed components, false-positive blobs, false-negative holes, and learned probability maps. Continuous probability blending is retained as a controlled hard-versus-soft trade-off, not assumed to be superior.

Pixel similarity alone cannot establish whether the result remains readable. We therefore train a fixed character recognizer on existing CODH coordinates with strictly disjoint source pages. It provides a clean-image ceiling, macro Unicode accuracy, and intact- and stroke-damaged-character accuracy. We also measure intact-stroke deletion, insertion, and total change directly outside the oracle mask. Every method uses the same 144-image final test manifest and restoration seeds; the other 72 images are restricted to calibration. Confidence intervals resample source pages rather than correlated variants.

Here, "character-safe" means conservative editing designed and evaluated to minimize unintended changes to intact character strokes; it is not a formal guarantee of philological correctness. Our contributions are:

- an oracle-free conservative edit-support formulation that makes learned localization an explicit, auditable permission to generate;
- a mask-error curriculum that adapts an existing diffusion restorer to realistic localization errors without changing its architecture or loss;
- character- and stroke-aware auditing of recovery, deletion, insertion, total stroke change, and OCR consistency; and
- a source-page-disjoint calibration/test benchmark with matched oracle and full-frame controls.

Figure 1 summarizes the method as a single permission-to-edit pipeline and separates test-time inference from training-only mask sampling.

2 Related Work

2.1 Historical Document Restoration

Classical document enhancement includes binarization, deblurring, stain removal, inpainting, and recognition-aware processing. LaMa [11] offers a strong convolutional inpainting prior. DocDiff applies residual diffusion to several document enhancement tasks [16], while DiffHDR [15] and AutoHDR [19] target damaged historical pages using domain-specific supervision and modular localization or context components. DocWaveDiff combines wavelet U-Nets with a predict-and-refine diffusion design [8]. Japanese document research has mainly emphasized

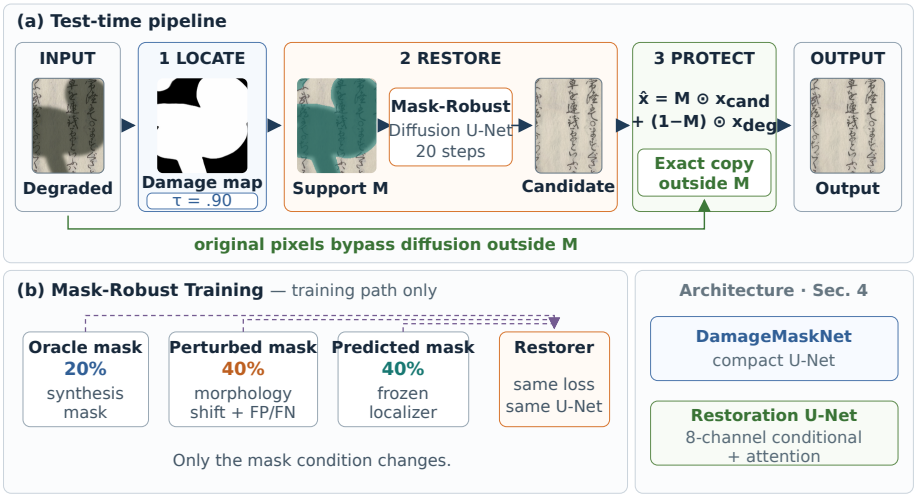


Fig. 1. LOCATE → RESTORE → PROTECT. (a) At test time, DamageMaskNet thresholds a probability map into hard edit support M. The diffusion model produces a restoration candidate under that support, and exact compositing copies degraded-input pixels everywhere outside M. (b) During fixed 50-epoch adaptation only (dashed paths), oracle, perturbed-oracle, and predicted masks are sampled with probabilities 20/40/40%; only the mask condition changes. Architecture details are given in Sec. 4.

kuzushiji recognition and character generation [3,4,5,17]. KuroNet jointly predicts character locations and identities from a complete page [4], while the recent NDL Koten OCR-Lite combines layout detection, line recognition, and reading-order estimation in a CPU-capable public system [1]. We focus on conservative full-page restoration and explicitly audit character identity.

Our narrower claim treats localization as an explicit permission mechanism for generation. DiffHDR [15] accepts user-specified damaged content and location, whereas AutoHDR [19] learns OCR/DINO-based localization and autoregressively restores masked patches. Neither system enforces an exact copy outside the predicted support or trains the restorer against a curriculum of mask errors. Our distinction is therefore the coupling of oracle-free localization, exact hard-support compositing, mask-error robustness, and collateral-edit auditing, rather than automatic localization alone.

2.2 Conditional Diffusion and Masked Sampling

Palette formulates image-to-image tasks as conditional diffusion [10]. RePaint imposes known pixels during reverse diffusion without retraining [7], and DPM-Solver++ enables short sampling trajectories [6]. Binary inpainting masks encode a hard decision, although damage boundaries and detector errors are uncertain. Our soft-support ablation generalizes the known/generated switch to a continuous probability while preserving the original model interface.

2.3 Diffusion-Based Anomaly Localization

Reconstruction differences have been used to localize anomalies with diffusion models [14,18]. MDPS averages difference maps from multiple posterior samples [13]. These studies motivate our frozen edit-consensus baseline: repeated restorer edits can signal suspect pixels. We do not claim that multi-sample difference maps are new. Instead, edit consensus is a training-free comparator and a possible auxiliary cue for document damage, whereas our main contribution is learned localization coupled to preservation-aware sampling and character evaluation.

3 Controlled Benchmark and Evaluation

3.1 Data and Split

We use one 238-page CODH digitized work, `brsk00000` [9]. Source pages are split before synthesis into 214 training and 24 evaluation-pool pages. Three independent variants are generated at each of the weak, medium, and strong *evaluation difficulty levels*, yielding 1,926 training and 216 evaluation triplets. Before final experiments, a salted SHA-256 ranking of source-page identifiers assigns eight pages (72 triplets) to calibration and 16 pages (144 triplets) to the final test set. All variants and severities from one source page remain together. Each triplet contains a clean page x_0, degraded page x_d, and binary synthesis mask m with $m = 1$ for damaged pixels. The mask is training supervision and an evaluation reference, not an input to the proposed test-time method. These levels are not a pure degradation-strength ablation: strong pages also retain twice the maximum spatial dimension used by weak and medium pages. We therefore avoid attributing every level difference to corruption severity alone.

Synthetic degradation. Each variant sequentially alpha-composites ink (I), water-ring (W), mold-cluster (M), and smudge (S) layers. Centers follow a four-fold-downsampled text-density map blurred at radius 50, with .03 probability floor. For weak/medium/strong, uniform opacity ranges are I .28–.45/.40–.65/.78–.95, W .18–.32/.30–.50/.55–.80, M .18–.36/.30–.55/.60–.85, and S .14–.26/.20–.40/.45–.70; counts are I 2–3/2–4/3–6, W 1/1–2/2–3, M 1–2/1–3/2–4 clusters, and S 1/1–2/2–4. Short-side size fractions are I .12–.28/.15–.40/.20–.50, W .12–.28/.12–.30/.15–.35, and S .18–.35/.20–.45/.25–.55; M spot radii are 3–7/4–9/5–13 pixels. Independent RGB channels darken monotonically. Weak-to-strong half-open boxes are $(10{:}45, 8{:}35, 4{:}25) \rightarrow (0{:}20, 0{:}12, 0{:}8)$ for I and $(110{:}165, 70{:}115, 35{:}75) \rightarrow (70{:}130, 35{:}80, 15{:}50)$ for W. They are $(40{:}90, 25{:}60, 15{:}40) \rightarrow (15{:}55, 8{:}35, 3{:}20)$ for M and $(80{:}140, 60{:}120, 40{:}90) \rightarrow (40{:}100, 30{:}80, 20{:}55)$ for S; medium uses the component-wise intermediate intervals in the configuration. Accepted damaged-area ratios are .06–.10/.16–.25/.35–.50; after 12 failed draws the closest is retained. The mask is the union of layer alpha $> .01$. Stored maximum dimensions are 1024/1024/2048 with aspect ratio preserved. The generation command and seed log were not retained, so we do not claim bit-exact regeneration.

3.2 Image and Localization Metrics

We report global PSNR, SSIM, and LPIPS, plus $\text{PSNR}_{\text{mask}}$ and $\text{PSNR}_{\text{known}}$. Localization metrics are Dice/F1, IoU, precision, recall, false-positive rate on intact pixels, and boundary F1 with a two-pixel tolerance.

To measure collateral editing, we apply the same page-level Otsu threshold to clean and restored images outside m. Deletion is the fraction of clean foreground changed to background; insertion is inserted foreground divided by clean background; foreground symmetric difference and stroke F1 summarize both directions. We additionally report the fraction of oracle-intact pixels whose RGB value is changed by the restoration operation relative to its degraded input. This operation-based statistic is distinct from disagreement with the clean target, which can include unresolved corruption. All restoration methods share seed $2027 + j$ for evaluation-pool index j, retained after the calibration/test split. Ninety-five-percent confidence intervals for image metrics use 2,000 paired bootstrap draws over the 16 final-test source pages, retaining all levels and variants for each selected page.

3.3 Character-Fidelity Protocol

The CODH coordinates contain 75,462 character crops from 2,171 Unicode labels. The restoration evaluation-pool pages are held out as the recognition test set. The other pages are split into 193 training and 21 development pages with no page overlap. We fix a minimum of 20 training crops per class before training, retaining 288 classes and 56,258/5,762/5,885 train/development/pool crops. This covers 89.9% of annotations on held-out pages. Final paper metrics use the 12 annotated source pages in the 16-page test split: 3,573 coordinate boxes, repeated over nine level/variant conditions (32,157 instances), with 254 observed classes. A grayscale ResNet-18 is chosen by development macro accuracy and then frozen. Its clean final-test ceiling is 0.9726 micro and 0.9677 macro accuracy. A character is considered stroke-damaged when at least 5% of the clean Otsu foreground inside its box overlaps m. We report macro Unicode accuracy, intact-character accuracy, and stroke-damaged-character recovery. Character confidence intervals instead bootstrap the 12 annotated final-test source pages; no level/variant crop is treated as an independent page. Labels below the 20-crop training cutoff are excluded from recognizer metrics, but remain in the full Unicode transcription used by the full-page OCR probe.

As a second, independently trained semantic probe, we freeze NDL Koten OCR-Lite [1] and report full-page character error rate (CER). OCR runs on the complete image; scoring retains lines centered inside the hull of CODH-annotated characters so unannotated headers do not count as insertions. The reference follows the coordinate file's reading order and is normalized with Unicode NFC plus whitespace removal only. We treat this as an auxiliary paired metric: overlap between the external OCR training resources and related public historical collections cannot be ruled out, so absolute CER is not evidence of generalization by itself.

3.4 Paired Decisions and Claim Rules

For a comparison $A \to B$, we first compute the metric difference for each matched sample and then resample source pages, so repeated degradation variants never act as independent pages. We retain soft rather than hard predicted masks only if the soft-minus-hard point estimates increase macro Unicode accuracy and decrease intact-stroke deletion, while the lower 95% bound of $\Delta\text{PSNR}_{\text{mask}}$ stays above -0.5 dB. A directional interval that crosses zero is described as inconclusive rather than as an improvement. FEC $K = 8$ is called better than $K = 1$ only if the paired Dice interval excludes zero. These rules were fixed before the fixed-checkpoint test reruns and FEC calibration were completed.

4 Method

4.1 Frozen Palette Restorer

The existing Palette-style U-Net predicts the clean image x_0. Its eight input channels concatenate noisy target RGB x_t, degraded RGB x_d, a constant content channel, and one mask channel. The content channel is an all-one map (value 1 at every pixel): it occupies the model's optional rendered content slot while indicating that no OCR/content hint is supplied. We write $R(x_d, M)$ for the restoration candidate produced by restorer R from degraded document image x_d under editable-support mask M. Training uses $\mathcal{L}_{\text{MSE}} + 0.05\mathcal{L}_{\text{Sobel}}$. Unless stated otherwise, we freeze this checkpoint and use 20 DPM-Solver++ steps on 512×512 patches with stride 256.

4.2 DamageMaskNet

DamageMaskNet is a four-level U-Net with 32 base channels and 7.24M parameters. It maps normalized degraded RGB to logits z and probabilities $p = \sigma(z)$. Synthetic masks provide direct supervision by combining binary cross-entropy (BCE) with Dice loss:

$$\mathcal{L}_{\text{mask}} = \text{BCEWithLogits}(z, m) + \left(1 - \frac{2\sum_i p_i m_i + \epsilon}{\sum_i p_i + \sum_i m_i + \epsilon}\right). \tag{1}$$

Here m is the binary synthesis mask, i indexes its pixels, $m_i \in \{0, 1\}$ is the target label, p_i is the predicted damage probability, and ϵ is a small numerical stabilizer. BCE supplies pixel-wise supervision, whereas the Dice term directly promotes damaged-region overlap under foreground–background imbalance. We use the fixed epoch-40 checkpoint. The decision threshold is selected once on the eight-page calibration manifest for Dice and then fixed for every hard-mask experiment. Full-resolution inference is fully convolutional.

4.3 Calibration-Set-Selected Hard Edit Support

The primary interface thresholds p into $M_i = \mathbf{1}[p_i \geq \tau]$ and passes M to the frozen or mask-robust restorer R. The delivered image is

$$\hat{x} = (1 - M) \odot x_d + M \odot R(x_d, M). \tag{2}$$

Thus the localizer controls both generation and the maximum editable support; pixels outside M are copied byte-for-byte. This hard support is M0 with the frozen restorer and M2-H (Ours) with the adapted restorer. We refer to the latter simply as Ours hereafter.

4.4 Soft-Support Ablation

The retained method uses the calibration-set-selected binary support in Eq. (2). To test whether uncertain boundaries benefit from continuous rather than binary support, an ablation replaces the usual masked sampling switch with the learned probability at every reverse step:

$$x_{t-1} = (1 - p) \odot x_{t-1}^{\text{known}} + p \odot x_{t-1}^{\text{pred}}. \tag{3}$$

The same p is supplied as the U-Net mask-condition channel. Eq. (3) recovers exact preservation at $p = 0$ and ordinary generation at $p = 1$. A symmetric Hann overlap window combines sliding patches; pixels with exactly zero probability are copied byte-for-byte from x_d. Both terms carry index $t - 1$ because the scheduler first performs the reverse update from t to $t - 1$; compositing then combines the predicted previous state x_{t-1}^{pred} with x_d re-noised to the same timestep, x_{t-1}^{known}.

4.5 Mask-Robust Fine-Tuning

The second stage initializes the same restorer checkpoint and leaves the architecture and image loss unchanged. During training, its mask channel is sampled from oracle masks (20%), curriculum-perturbed oracle masks (40%), or frozen DamageMaskNet probabilities (40%). Perturbation strength increases linearly with $p = (e + 1)/50$: the morphology radius is at most round$(12p)$ pixels and translation at most round$(10p)$ pixels per axis. Dilation and erosion are each selected with probability $.4p$, translation with $.6p$, and component deletion with $.5p$ (per-component probability $\min(.5, .25p)$). False-positive blobs and false-negative holes are each applied with probability $.7p$, with up to round$(4p)$ sampled ellipses. This isolates robustness to realistic mask errors from architectural changes. We train for a fixed 50 epochs using the fixed epoch-40 DamageMaskNet checkpoint; the training schedule is unchanged across the matched fine-tuning controls.

4.6 Frozen Edit-Consensus Baseline

For a degraded image, the frozen restorer produces K candidates. Pixel i is marked as edited in sample k when a calibration-selected RGB distance exceeds

τ_r; $q_i = K^{-1}\sum_k e_i^{(k)}$ and $M_i = \mathbf{1}[q_i \geq \tau_q]$. We compare $K \in \{1, 4, 8\}$ with thresholds selected only on the calibration split. Spatial permutation preserves the histogram of q, and a same-area random mask preserves mask size, testing whether spatial localization rather than editable area explains restoration quality. Here $k \in \{1, \ldots, K\}$ indexes a candidate sample, and $e_i^{(k)} \in \{0, 1\}$ indicates whether candidate k changes pixel i by more than τ_r.

5 Experiments

5.1 Compared Methods and Implementation

We group controls by the question they answer. *Oracle references* (B0-F/B0-R) reveal the frozen/robust restorers' capacity when support is known. *Unrestricted restoration* uses an all-one mask with either the frozen (Full-frame Frozen) or the same robust checkpoint as Ours (Full-frame Robust). *Ours and ablations* comprise predicted hard support with the frozen restorer (M0), soft support (M1/M2-S), oracle-only continued fine-tuning (M2-O), and **M2-H (Ours)**, the only final method. *Alternative localization* uses frozen edit consensus (FEC $K = 1/4/8$). *External references* are Telea [12], Navier–Stokes [2], pretrained LaMa [11], and the official cross-domain DiffHDR checkpoint [15] under the same predicted support. LaMa and DiffHDR are used without fine-tuning on our training set, so they provide external context rather than matched training comparisons; the controlled claims instead use B0-R, M2-O, and Full-frame Robust. Finally, C0 permutes the predicted probability map, whereas C1 uses a same-area random mask; both serve as *spatial controls*. Blank copying is a metric diagnostic, not a restoration method.

DamageMaskNet uses 512^2 inputs, 40 epochs, batch 8, and AdamW (2×10^{-4} learning rate, 10^{-4} weight decay) with cosine scheduling, mixed precision, and seed 0. The initial restorer uses 500 epochs, batch 4/GPU with four-step accumulation, AdamW at 10^{-4}, 10-epoch warm-up plus cosine decay, unit clipping, EMA .9999, mixed precision, and seed 42. Robust and oracle-only controls start from that checkpoint and share 50 epochs, batch 2/GPU with four-step accumulation, AdamW at 2×10^{-5}, two-epoch warm-up, loss, and seed policy; only the 20/40/40 curriculum versus 100% oracle masks differs. The recognizer uses 15 epochs, batch 256, AdamW at 10^{-3} with 10^{-4} weight decay, .05 label smoothing, cosine scheduling, mixed precision, seed 2027, and inverse-square-root class weighting. Its checkpoint is independent of restoration outputs. Numeric macros are generated from frozen per-page/per-character JSON files and source-page bootstraps.

5.2 Damage Localization

RQ1. Can damage be localized without an oracle mask? Table 1 summarizes localization performance.

Table 1. DamageMaskNet localization on 144 final-test images. Confidence intervals are source-page bootstrap intervals.

Method	Dice	IoU	Precision	Recall	FPR	Boundary F1
DamageMaskNet	0.9678	0.9387	0.9757	0.9608	0.0071	0.7491

Region overlap is high: Dice has a 95% source-page interval [0.9602, 0.9735], while intact-pixel FPR is 0.0071 [0.0051, 0.0093]. Boundary F1 is lower than region Dice, especially at the strong difficulty level, which motivates evaluating a soft-support ablation rather than assuming every predicted boundary is certain.

5.3 Oracle vs. Predicted Support

RQ2. What is lost without oracle support? The matched robust pair changes only the support. Replacing B0-R's oracle support with predicted hard support changes $\text{PSNR}_{\text{mask}}$ by -0.15 dB (95% CI [-0.25, -0.05]). The paired change in stroke-aware damaged-character accuracy is 0.0002 (CI [-0.0014, 0.0019]), and OCR CER changes by 0.0021 (CI [-0.0021, 0.0057]); both semantic differences are inconclusive. Thus the regional gap is measurable but small, while neither semantic metric shows a statistically supported loss.

Table 2 places this matched comparison beside the two other controls central to the paper. Figure 2 then compares the same methods under a method-independent page and crop rule.

5.4 Effect of Mask-Robust Fine-Tuning

RQ3. Does robust training help with localization errors? M0→M2-O isolates 50 additional epochs under oracle masks; M2-O→Ours then isolates the mask-error curriculum with all other training choices matched. Additional oracle-only training changes $\text{PSNR}_{\text{mask}}$ by 0.13 dB (95% CI [-0.05, 0.30]), stroke-damaged accuracy by 0.0076 (CI [0.0022, 0.0129]), and OCR CER by 0.0015 (CI [-0.0034, 0.0077]); only the semantic gain is supported. The curriculum further changes these metrics by 0.32 dB (CI [0.08, 0.56]), 0.0172 (CI [0.0094, 0.0260]), and -0.0054 (CI [-0.0106, -0.0009]), supporting benefit beyond extra epochs. We therefore attribute M0→Ours to both continued training and the curriculum. Figure 3 contrasts false-positive and false-negative support with a case benefiting from continued training with the curriculum. These diagnostic, non-representative cases do not estimate failure or benefit frequencies.

5.5 Collateral Editing

RQ4. Does local support reduce collateral editing? Full-frame Robust and Ours share checkpoint, sampler, steps, patching, stride, and seed; only support changes. Ours-minus-Full-frame edit rate is -0.9921, deletion is -0.12772 (CI

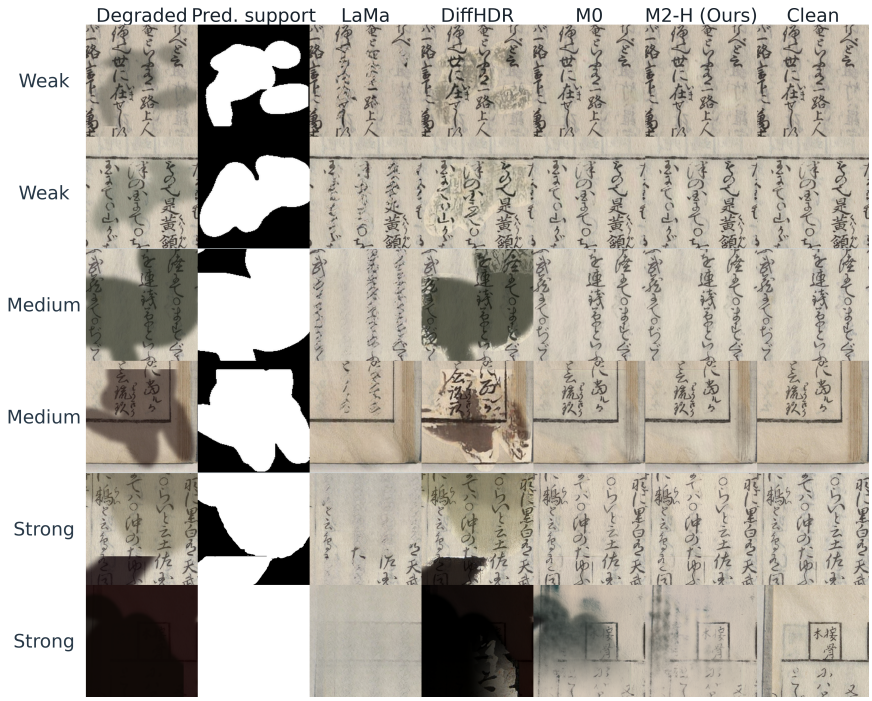


Fig. 2. Method-independent representative comparison using two examples per difficulty. Selection uses neither model outputs, PSNR/OCR, nor visual inspection. Columns show the degraded input, predicted support, LaMa, cross-domain DiffHDR, M0, Ours, and the clean reference.

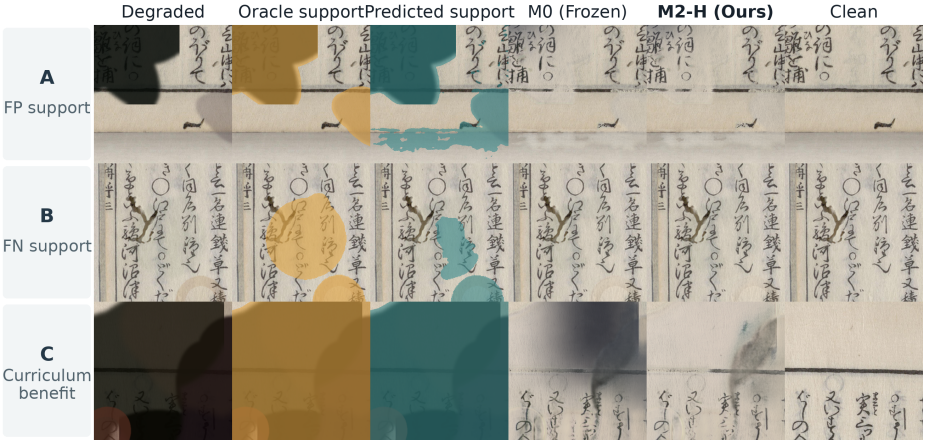


Fig. 3. Diagnostic, non-representative cases selected by fixed quantitative rules: (A) the highest false-positive support rate, (B) the highest false-negative support rate, and (C) the largest Ours-minus-M0 $\text{PSNR}_{\text{mask}}$ gain among eligible annotated final-test pages. These cases expose localization failure modes rather than estimate their frequency.

Table 2. Selected 144-image final-test results from frozen evaluator JSON files. B0-R and FF-R (Full-frame Robust) use the robust restorer with oracle and full support. M0, M2-O, and Ours use predicted hard support with frozen, oracle-only fine-tuned, and robust fine-tuned restorers, respectively. LaMa and DiffHDR also use predicted hard support but are external pretrained references, not matched training comparisons. Del./Ins. denote intact-stroke deletion/insertion; paired uncertainty is reported in the text.

Method	PSNR (dB)			SSIM	LPIPS↓	Intact strokes		
	Global	Mask	Known			Del.↓	Ins.↓	F1↑
B0-R	36.00	28.05	65.84	0.9808	0.0369	0.00000	2.1×10^{-5}	0.9999
FF-R	24.20	21.22	25.60	0.9488	0.0943	0.13576	3.9×10^{-5}	0.9242
M0	35.17	27.45	52.06	0.9768	0.0429	0.00760	2.1×10^{-5}	0.9960
M2-O	35.28	27.58	52.03	0.9789	0.0412	0.00757	2.0×10^{-5}	0.9961
M2-H (Ours)	**35.56**	**27.89**	52.23	**0.9796**	**0.0395**	0.00804	2.1×10^{-5}	0.9958
LaMa [11]	29.21	21.36	51.30	0.9037	0.0830	0.00212	0.00038	0.9982
DiffHDR [15]	21.95	14.01	51.89	0.8896	0.1328	0.00166	6.8×10^{-5}	0.9989

[-0.17016, -0.09381]), and $\text{PSNR}_{\text{mask}}$ is +6.67 dB (CI [6.01, 7.45]). Macro, intact, and stroke-damaged accuracy change by -0.0035 (CI [-0.0062, -0.0012]), -0.0045 (CI [-0.0067, -0.0026]), and -0.0078 (CI [-0.0114, -0.0049]). Thus hard support exchanges a small recognition decrease for much lower edit exposure; OCR CER remains inconclusive. Otsu deletion is binarized pixel change, not semantic loss: Full-frame Robust has both greater deletion and higher recognition. Boundary undercoverage is only an unverified possible explanation. An audit confirms exact outside-support copying; Blank's residual insertion is renderer feathering beyond the stored mask.

5.6 Character Fidelity

Table 3 reports the matched character and OCR probes. Ours obtains 0.9372 recovery on 9,195 primary-threshold cases, below the fixed clean ceiling.

Table 3. Unicode accuracies and OCR CER for matched controls. Stroke-damaged characters have at least 5% of clean foreground affected.

Method	Macro	Stroke-dmg.	Intact	OCR CER↓
Clean reference	0.9677	–	–	0.1073
Degraded	0.8274	0.4886	0.9683	0.1925
B0-R	0.9525	0.9370	0.9683	0.1295
Full-frame Robust	0.9558	0.9451	0.9725	0.1359
M0	0.9445	0.9125	0.9683	0.1354
M2-O	0.9456	0.9201	0.9682	0.1369
M2-H (Ours)	0.9522	0.9372	0.9679	0.1315

5.7 Ablations and Controls

Hard versus soft support. With the frozen restorer, M0→M1 changes macro accuracy from 0.9445 to 0.9447; with the robust restorer, M2-H→M2-S changes it from 0.9522 to 0.9522. Soft-minus-hard differences computed before rounding are +0.00024 and −0.00006, with 95% paired source-page intervals [-0.00073, 0.00082] and [-0.00096, 0.00054]; neither supports a macro-accuracy gain. Deletion increases from 0.00760 to 0.00970 and from 0.00804 to 0.00910, respectively. The $\mathrm{PSNR}_{\mathrm{mask}}$ changes are +0.053 and +0.032 dB, with lower 95% bounds 0.001 and -0.002 dB, both above −0.5 dB. Neither soft variant meets all criteria in Sec. 3.4: both fail the deletion criterion, and M2-S also fails the macro point-estimate criterion. We therefore retain hard support.

FEC $K = 1 \rightarrow 8$ changes Dice 0.6632→0.6668 ($\Delta = 0.0037$, CI [-0.0001, 0.0076]), hence is inconclusive. Destroying location while preserving the probability histogram (C0) or area (C1) reduces $\mathrm{PSNR}_{\mathrm{mask}}$ to 12.69/12.71 dB versus M0's 27.45 dB, confirming that placement matters. Calibration selects $\tau = .90$ by Dice (.9669). On a fixed three-page NVIDIA A40 subset, localization takes 0.07 s/page, M2-H takes 41.98 s/page, and serial FEC $K = 8$ is estimated at 376.11 s/page (nine diffusion samples); diffusion calls dominate.

6 Limitations and Responsible Use

Results cover one CODH work under our synthetic generator and do not establish real-damage generalization. Hard-support false negatives leave damage, whereas false positives expose intact pixels; our design favors abstention. Character/OCR consistency is not philological correctness, so outputs should remain reversible aids beside source scans. A legacy localization run passively logged the 24-page pool before its 8/16 split; its fixed final epoch was not selected from those logs, and final thresholds/comparisons use the disjoint split. Diffusion's necessity against a matched non-diffusion model remains untested. LaMa and DiffHDR are external pretrained references; AutoHDR lacks a compatible quantitative run. The 41.98 s/page A40 measurement on three fixed pages is a deployment limitation, not a corpus-wide runtime estimate.

7 Conclusion

We learn hard edit support without a test-time oracle mask, train against mask errors, and copy outside-support pixels exactly. Predicted support approaches oracle support; M2-O isolates continued-training gains, while Full-frame Robust shows that localization reduces collateral editing. Outputs remain reading aids, not historically guaranteed reconstructions.

Acknowledgements. This work was supported by JST CRONOS JPMJCS24K4 and JSPS KAKENHI 26K23875.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aoike, T.: Development of NDLkotenOCR-Lite, a lightweight OCR that runs at high speed in a CPU environment. In: Proceedings of the Symposium on Humanities and Computers (Jinmonkon 2024). vol. 2024, pp. 181–186. Information Processing Society of Japan (2024), https://ipsj.ixsq.nii.ac.jp/records/241527, no. 1
2. Bertalmio, M., Bertozzi, A.L., Sapiro, G.: Navier–stokes, fluid dynamics, and image and video inpainting. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 1, pp. I–355–I–362. IEEE Computer Society (2001). https://doi.org/10.1109/CVPR.2001.990497
3. Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., Ha, D.: Deep learning for classical Japanese literature. arXiv:1812.01718 (2018)
4. Clanuwat, T., Lamb, A., Kitamoto, A.: KuroNet: Pre-modern Japanese Kuzushiji character recognition with deep learning. arXiv:1910.09433; presented at ICDAR 2019 (2019). https://doi.org/10.48550/arXiv.1910.09433
5. Kaneko, H., Ishibashi, R., Meng, L.: Deteriorated characters restoration for early Japanese books using enhanced CycleGAN. Heritage **6**(5), 4345–4361 (2023). https://doi.org/10.3390/heritage6050230
6. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv:2211.01095 (2022)
7. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: RePaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
8. Marulli, M., Bertini, M.: DocWaveDiff: A predict-and-refine approach for document image enhancement with wavelet U-nets and diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8511–8520 (2026)
9. ROIS-DS Center for Open Data in the Humanities: CODH Japanese Dataset. https://codh.rois.ac.jp/ (2026), National Institute of Informatics, Research Organization of Information and Systems
10. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-Image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022). https://doi.org/10.1145/3528233.3530757
11. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with Fourier convolutions. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2149–2159 (2022)
12. Telea, A.: An image inpainting technique based on the fast marching method. Journal of Graphics Tools **9**(1), 23–34 (2004). https://doi.org/10.1080/10867651.2004.10487596
13. Wu, D., Fan, S., Zhou, X., Yu, L., Deng, Y., Zou, J., Lin, B.: Unsupervised anomaly detection via masked diffusion posterior sampling. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. pp. 2442–2450 (2024). https://doi.org/10.24963/ijcai.2024/270
14. Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 650–656 (2022)

15. Yang, Z., Peng, D., Shi, Y., Zhang, Y., Liu, C., Jin, L.: Predicting the original appearance of damaged historical documents. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9382–9390 (2025). https://doi.org/10.1609/aaai.v39i9.33016
16. Yang, Z., Liu, B., Xiong, Y., Yi, L., Wu, G., Tang, X., Liu, Z., Zhou, J., Zhang, X.: DocDiff: Document enhancement via residual diffusion models. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2795–2806 (2023). https://doi.org/10.1145/3581783.3611730
17. Yuan, H., Yanai, K.: KuzushijiFontDiff: Diffusion model for Japanese Kuzushiji font generation. In: MultiMedia Modeling. Lecture Notes in Computer Science, vol. 15524, pp. 142–149. Springer Nature Singapore (2025). https://doi.org/10.1007/978-981-96-2074-6_15
18. Zhang, X., Li, N., Li, J., Dai, T., Jiang, Y., Xia, S.: Unsupervised surface anomaly detection with diffusion probabilistic model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6782–6791 (2023)
19. Zhang, Y., Zhang, P., Yang, Z., Yan, P., Shi, Y., Liu, P., Guo, F., Jin, L.: Reviving cultural heritage: A novel approach for comprehensive historical document restoration. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 28876–28892. Association for Computational Linguistics, Vienna, Austria (2025). https://doi.org/10.18653/v1/2025.acl-long.1402