Draft, Verify, Restore: Self-Refining Historical Inscription Restoration with a Unified MLLM

Yuyi Zhang [†1] **Junle Liu** [†1] **Peirong Zhang** [†1] **Jianliang Liu**[1]
Zhenhua Yang[1] **Lianwen Jin** [*1]

[1]South China University of Technology
yuyi.zhang11@foxmail.com junle_liu@foxmail.com eelwjin@scut.edu.cn

Abstract

Inscriptions are invaluable cultural heritage, yet centuries of degradation (e.g., fractures, erosion, oxidation) have rendered many partially illegible. Existing Historical Inscription Restoration (HIR) methods rely on task-separated pipelines with irreversible error accumulation and patch-based generation that sacrifices page-level consistency. Therefore, we present **UniHIR**, the first unified MLLM for end-to-end historical inscription restoration. It integrates two novel designs, **Draft-Guided Localization** and **Hierarchical Self-Refinement**, to enable accurate damage localization and illegible-content prediction via iterative reasoning and self-correction. This unified approach enables true page-level restoration with consistent typography and style. To support training under high-resolution inputs and long sequences, we design **UHIRFactory** and construct **HIR-Bench**, enabling step-wise, memory-efficient instruction tuning with step-aware annotations for intermediate drafts and refinements. Experiments demonstrate that UniHIR achieves superior performance in both text restoration accuracy and appearance restoration quality, validating that HIR can be effectively tackled by a standalone model in a unified manner. The model and code are available at https://github.com/ZZXF11/UniHIR.

1 Introduction

Standing as an enduring and irreplaceable lithic archive, ancient steles anchor the continuity of human culture and history. Unfortunately, countless steles have suffered degradation from manual activities or natural forces, such as war, water erosion, and oxidation, rendering their inscriptions partially or entirely illegible. Thus, Historical Inscription Restoration (HIR) holds immense value for preserving these cultural heritages. Yet, accumulated deterioration makes restoration increasingly complex and renders traditional manual approaches inefficient (taking days or months for one stele) and difficult to sustain at scale. These constraints underscore the pressing need for automated HIR methods. Recently, rapid advances in artificial intelligence (AI) have spurred growing interest in AI-assisted restoration of historical documents (Yang et al., 2025; Zhu et al., 2026; Zhang et al., 2026). For example, Assael et al. (2025) combines a Transformer with a vision backbone to predict damaged text, geographic origin, and date. Zhang et al. (2025d) proposes a three-stage pipeline with decoupled large language and diffusion-based models to recover the historical appearance.

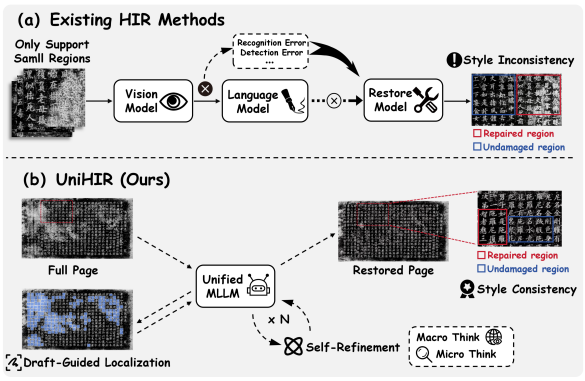

Figure 1: Comparison between existing historical inscription restoration (HIR) methods with the proposed UniHIR.

While it is viable to directly transfer existing historical document restoration methods to HIR, they suffer from many critical limitations (Fig. 1). (1) Current approaches often understand and generate separately in a pipeline manner (e.g., understanding text with a language model, then generating restored images using diffusion models). This prevents joint optimization of both capabilities and limits their synergistic potential. (2) Most models generate predictions in a single pass without explicit verification, making errors introduced early in the generation process irreversible. (3) Current

[†]Equal contribution
[*]Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27216–27231

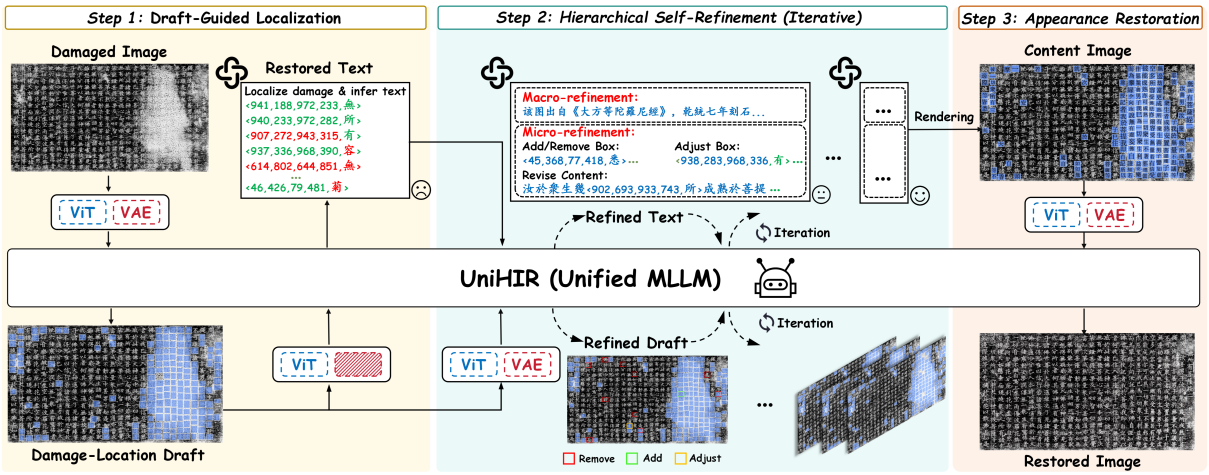

Figure 2: Framework of UniHIR. It contains three steps: Draft-Guided Localization, Hierarchical Self-Refinement, and Appearance Restoration.

methods predominantly perform patch-based pixel restoration, either restoring only a few characters or dividing a full page into patches for processing. They thus fail to leverage the global layout and long-range character styles cues, leading to jarring or inconsistent restored text.

Concurrently, unified multimodal large language models (MLLMs) have fueled increasing attention in the community, which seamlessly integrate visual understanding and generation with a single backbone (Zhang et al., 2025b). HIR demands both decoding missing or illegible characters and generating pixel-level restorations that preserve original glyphs and style, which are capabilities that naturally align with unified models' complementary strengths. Moreover, these models can restore entire pages with global content and style coherence, a capability difficult to achieve within current HIR solutions. Such compelling congruence raises a natural question: **Can the HIR task be truly tackled by unified MLLMs in an end-to-end manner, or are task-separate pipelines still necessary?**

In response to this question, we propose **UniHIR**, a pioneering Unified MLLM for end-to-end HIR. To address the challenge of HIR and the limitations of current approaches, UniHIR introduces two novel designs, **Draft-Guided Localization (DGL)** and **Hierarchical Self-Refinement (HSR)**, aiming at accurate damage localization and illegible-content prediction. Since steles often contain numerous damaged regions, MLLMs typically tend to hallucinate in such dense detection, producing shifted and overlapping boxes. Therefore, UniHIR first generates a damage-location draft—a visual "scratchpad" that marks damaged regions and guides the inference of illegible content. It then performs hierarchical self-refinement at two levels: macro-level refinement predicts global attributes (e.g., source text, date, and engraver), and micro-level refinement adjusts the damage regions (Bounding box) and verifies the inferred content using contextual cues. Finally, UniHIR restores the image from the finalized damage locations and predictions.

To make UniHIR trainable under high-resolution inscription images and long restoration sequences, we propose UHIRFactory, a unified training framework featuring step-isolated optimization and memory-adaptive data sampling, and construct HIRBench to support step-aware training.

Extensive experiments demonstrate that UniHIR consistently outperforms existing methods in both text restoration accuracy and appearance preservation. On severely damaged steles where OCR accuracy is only 55.68%, UniHIR boosts accuracy to 86.65%. These results validate UniHIR's effectiveness as a standalone model and highlight its potential as a practical assistive tool for historians.

We outline our main contributions as follows:

- We propose **UniHIR**, a pioneering Unified MLLM for end-to-end HIR, offering a new perspective on HIR.
- UniHIR incorporates two novel designs, **Draft-Guided Localization** and **Hierarchical Self-Refinement**, to support iterative reasoning and self-correction for HIR.
- We propose **UHIRFactory** and construct **HIRBench** to enable step-wise, memory-

efficient training of UniHIR.

- Extensive experiments demonstrate that our method achieves superior restored-text accuracy and generation quality.

2 Related Work

Historical Inscription Restoration. Traditional HIR largely relies on expert manual restoration. With recent advances in generative AI, progress has been made: Fetaya et al. (2020); Bamman and Burns (2020); Lazar et al. (2021); Papavassileiou et al. (2023) perform text restoration for various languages using a language model. Assael et al. (2022) then used a Transformer with multi-task learning to jointly predict missing text, geographic origin, and date; and they further improved performance by introducing a vision network in Aeneas (Assael et al., 2025). However, these methods remain mainly text-centric and cannot repair the visual appearance of damaged steles. To address this, recent work has begun to explore appearance restoration. Han et al. (2024) proposed a crowdsourcing-based text–appearance restoration framework, still demanding human labor. Zhu et al. (2024) and Zhang et al. (2025d) proposed three-stage pipelines that restore text via retrieval or LLM prediction and refine appearance with diffusion models. However, they mainly target local regions (e.g., a few characters or a small patch) and, when applied to an entire image, restore it patch by patch rather than holistically—motivating end-to-end full-page restoration.

Unified MLLMs. Multimodal large language models (MLLMs) that unifies visual understanding and generation, i.e., unified MLLMs, have emerged as a potent force in the community, enabling mutual enhancement between these two realms for comprehensive visual intelligence. Early attempts (Sun et al., 2024a; Ge et al., 2024; Tong et al., 2025; Pan et al., 2025) combine external diffusion models with MLLMs but prevent joint optimization of shared parameters. Researchers later unify both capabilities within single transformer models for better synergy. These approaches fall into three categories: (1) pure autoregressive models (Team, 2024; Wu et al., 2025b, 2024; Wang et al., 2024), (2) diffusion/flow matching-embedded models (Xie et al., 2025; Ma et al., 2025; Lin et al., 2025; Deng et al., 2025; He et al., 2025), and (3) models with lightweight diffusion heads (Fan et al., 2025; Sun et al., 2024b; Wu et al., 2025c). Currently, most works demonstrate that understanding facilitates generation (Deng et al., 2025; Tong et al., 2025; Pan et al., 2025), while recent studies reveal that generation can also enhance understanding (Chen et al., 2025a; Yan et al., 2025). However, they still struggle with difficult scenarios (Yan et al., 2025; Shi et al., 2025a). Historical Inscription Restoration is also a demanding task that requires precise understanding of degraded textual content and generating outputs with stylistic consistency. Thus, we propose UniHIR, seeking to address this task from a unified perspective.

3 Preliminaries

3.1 Task Definition

Historical Inscription Restoration aims to faithfully recover the original appearance of damaged stone-inscription rubbing images under the principle of "restoring the old as the old" (Du, 1999; Wang, 2021). This entails repairing only damaged regions while preserving the original typographic style of characters and maintaining background texture consistency with intact areas. Formally, the restoration objective is defined as follows:

$$\big(\boldsymbol{I}_r,\ \boldsymbol{C}\big) = \mathcal{F}(\boldsymbol{I}_d)\,, \tag{1}$$

where $\boldsymbol{I}_d$ denotes the rubbing image of a damaged stele, $\mathcal{F}$ is the restoration model, $\boldsymbol{I}_r$ is the restored inscription, and $\boldsymbol{C} = \{(\boldsymbol{c}_i, \boldsymbol{b}_i)\}_{i=1}^{N}$ is the set of illegible characters. Here, $\boldsymbol{c}_i$ and $\boldsymbol{b}_i$ denote the character content and its location (represented by a bounding box), respectively.

3.2 BAGEL: A Unified MLLM

UniHIR is initialized based on BAGEL (Deng et al., 2025), a widely adopted unified MLLM that natively integrates visual understanding and generation. Due to the image input and output of HIR, BAGEL's inherent support for image understanding and image-to-image editing provides a well-suited foundation for this task. Architecturally, BAGEL consists of a SigLIP2 (Tschannen et al., 2025) vision encoder, a FLUX VAE (Black Forest Labs, 2024), and two Transformers initialized from Qwen2.5 (Qwen et al., 2025), organized as a Mixture-of-Transformer-Experts (MoT).

For visual understanding, one Transformer processes text tokens alongside ViT-encoded image tokens to autoregressively generate responses. For visual generation, the other Transformer processes VAE tokens and synthesizes continuous latents via Rectified Flow (Esser et al., 2024). This unified

design allows BAGEL to fluidly alternate between understanding and generation, making it a strong foundation for UniHIR to accomplish the HIR task.

4 Method

UniHIR restores a damaged inscription by three steps: Draft-Guided Localization (DGL), Hierarchical Self-Refinement (HSR), and Appearance Restoration (AR). A schematic of UniHIR is presented in Fig. 2.

4.1 Draft-Guided Localization

Before restoration, models are required to localize the damaged regions within the inscriptions. Yet, steles often contain hundreds of such regions, causing even detection-optimized models like Qwen3-VL (Bai et al., 2025a) to hallucinate misaligned or overlapping boxes. This challenge is amplified for unified MLLMs. To address such dense localization, we propose a novel Draft-Guided Localization (DGL) mechanism, inspired by how humans use intermediate drafts to solve complex problems. As shown in Fig. 2 (Step 1), DGL operates through draft generation and draft-guided reasoning.

(1) Draft Generation. Given a damaged image I_d, UniHIR first synthesizes a damage-location draft $D \in \mathbb{R}^{H\times W\times 3}$, which spatially highlights damaged regions using blue masks. This draft is generated via the flow matching–based generation branch, conditioned on visual features extracted by the VAE encoder. Formally, we sample from the learned distribution:

$$D \sim p_\theta(D \mid I_d) = \int p_\theta(D \mid z) p_\phi(z \mid I_d) dz, \quad (2)$$

where z denotes the VAE latent representation and θ, ϕ are learnable parameters.

(2) Draft-Guided Reasoning. Conditioned on both the original image I_d and the generated draft D, UniHIR performs dense localization and illegible content prediction. Specifically, the model autoregressively generates a sequence of damage descriptors:

$$C = \{(c_i, b_i)\}_{i=1}^{N} = \arg\max_{C'} p_\psi(C' \mid I_d, D), \quad (3)$$

where each tuple (c_i, b_i) contains the predicted character c_i and its bounding box $b_i = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$, serialized with special tokens in conventional inscription reading order (right-to-left, top-to-bottom) (Zhang et al., 2025a).

Discussion. DGL introduces a novel paradigm that departs fundamentally from existing methods. Whereas most unified models use understanding to guide generation, e.g., BAGEL first interprets the image and instruction before producing outputs, we invert this workflow by using generation to facilitate understanding. Experiments (Sec. 5.2) validate the effectiveness of this design, showing that the visual draft provides explicit spatial cues that improve content prediction. Moreover, as illustrated in Fig. 2, unlike BAGEL, which employs ViT and VAE features throughout inference, UniHIR uses both only during draft generation and omits VAE features in the subsequent reasoning stage, thereby reducing inference latency and memory overhead.

4.2 Hierarchical Self-Refinement

During Historical Inscription Restoration, experts typically examine damaged areas repeatedly and validate illegible content against the surrounding context. They may also incorporate historical cues such as the inscription's date, source text, and engraver to reconstruct the illegible information. Motivated by this iterative workflow, we design a Hierarchical Self-Refinement (HSR) mechanism. As illustrated in Fig. 2 (Step 2), HSR leverages autoregressive generation to implement structured self-refinement at two complementary granularities. Given initial predictions $\mathcal{C}^{(0)} = \{(c_i^{(0)}, b_i^{(0)})\}$ and draft $D^{(0)}$ from DGL, HSR performs iterative refinement:

$$\mathcal{C}^{(t+1)}, A^{(t+1)} = f_{\text{refine}}(\mathcal{C}^{(t)}, D^{(t)}, I_d, A^{(t)}) \quad (4)$$

where t indexes iterations, $A^{(t)}$ denotes global attributes, and a new draft $D^{(t+1)}$ is regenerated after each iteration to provide iterative feedback.

(1) Macro-Level Refinement addresses the semantic grounding problem by inferring global attributes $A = \{a_{\text{source}}, a_{\text{date}}, a_{\text{engraver}}\}$, including text source, temporal information, and engraver identity.

(2) Micro-Level Refinement performs fine-grained error correction through three operations:

(1) *Spatial Refinement*: The model cross-references draft $D^{(t)}$ with bounding boxes to identify false negatives (missed detections), false positives (hallucinations), and localization drift, then adds, removes, or adjusts boxes accordingly.

(2) *Semantic Refinement*: We then perform contextual coherence checking by explicitly eliciting the local context. For each character $c_i^{(t)}$, the model

outputs the preceding and following five characters (when available) and uses them, together with $A^{(t)}$ and I_d, to self-assess coherence and revise inconsistent predictions.

(3) *Draft-Mediated Feedback*: The visual drafts are regenerated after each iteration, which serves two purposes: validating spatial–semantic alignment via the generation branch and providing updated visual evidence for the next iteration.

4.3 Appearance Restoration

After Hierarchical Self-Refinement, we obtain precise bounding boxes for damaged regions along with predictions of their illegible content, denoted as $\mathcal{C} = \{(c_i, b_i)\}_{i=1}^{N}$. While existing methods are limited by patch-wise inscription generation, we perform full page-wise generation, enabling the model to exploit global writing style cues for higher text fidelity. By rendering predicted regions onto the original image, we construct a full-page content image I_g that provides explicit spatial guidance:

$$I_g = I_d \odot (1 - \mathbf{M}) + \mathcal{R}(\mathcal{C}) \odot \mathbf{M}, \quad (5)$$

where I_d is the damaged image, $\mathbf{M}$ is the damage mask. $\mathcal{R}(\mathcal{C})$ renders predicted characters $\mathcal{C}$ at their locations, and $\odot$ denotes dot multiplication.

This reformulates inscription restoration as an image-to-image translation task. We model restoration in VAE latent space via Rectified Flow:

$$\frac{d\mathbf{z}_t}{dt} = \mathbf{v}_\theta(\mathbf{z}_t, t, \mathcal{E}(I_d), \mathcal{E}(I_g)), t \in [0, 1], \quad (6)$$

where $\mathcal{E}$ is the VAE encoder and $\mathbf{v}_\theta$ is the velocity field. The content guidance $\mathcal{E}(I_g)$ leverages global style references from all characters, substantially reducing restoration difficulty while improving stylistic consistency over patch-wise methods with only local style access.

4.4 Training Strategy and HIRBench Dataset

UniHIR is trained via instruction fine-tuning. Yet, its training is challenged by both computational constraints and data scarcity. (1) Inscription images have high resolution (around 2048×1200) and restoration sequences are long (over 5,000 tokens), forbidding the full input–output sequence from being processed within a single training step due to GPU memory constraints. (2) Unified MLLMs typically require large-scale training data, while existing datasets that could serve as a foundation for data construction remain limited in size.

To address these challenges, we propose UHIRFactory, a unified training framework for historical inscription restoration that restructures the training of UniHIR and jointly guides dataset construction. The core idea of UHIRFactory is **step-isolated optimization**, which decomposes model optimization into its own three stages and optimizes each step independently. During training, only the inputs and outputs of the target step are retained in memory. For example, when optimizing Step 3, UHIRFactory stores and back-propagates solely through the Step 3 input–output tensors, avoiding storing the I/O tensors of other steps. This significantly reduces GPU memory consumption, while still optimizing step-specific capabilities.

Building on this design, UHIRFactory further introduces **memory-adaptive data sampling**, which dynamically combines samples from different steps into one batch as long as the overall memory remains within GPU limits. This enables efficient utilization of computational resources under extreme image resolutions and long text sequences.

Importantly, the step-isolated optimization of UHIRFactory necessitates step-aware supervision, which existing datasets do not provide. To fill this gap, we build HIRBench upon the open-source FPHDR dataset (Zhang et al., 2025d), organizing the data to align with UHIRFactory's training scheme. FPHDR contains only 6,543 synthetic and 1,663 real inscription restoration samples, whose scale is far from sufficient for training unified MLLMs. Therefore, we first assign 490 real inscription samples for testing as the HIRBench-Test set, and then generate abundant realistic training data from the remaining samples via data synthesis to form three training subsets: *HIRBench-DGL*, *HIRBench-HSR*, *HIRBench-AR*. Data examples are demonstrated in Fig. 3.

HIRBench-DGL. We overlay damaged regions with blue masks to form damage-location drafts and serialize the coordinates and illegible content into fixed-format text, producing 7,706 draft-text pairs (Fig.3 (a)).

HIRBench-HSR. Since existing MLLMs often produce localization errors (false positives, missed detections, misalignment) that propagate to content prediction, we simulate such errors by randomly manipulating bounding boxes and replacing characters based on FPHDR annotations. This yields 167,478 Draft-Refinement pairs (Fig.3 (b)).

HIRBench-AR. We randomly render standard-font characters with blue backgrounds onto intact

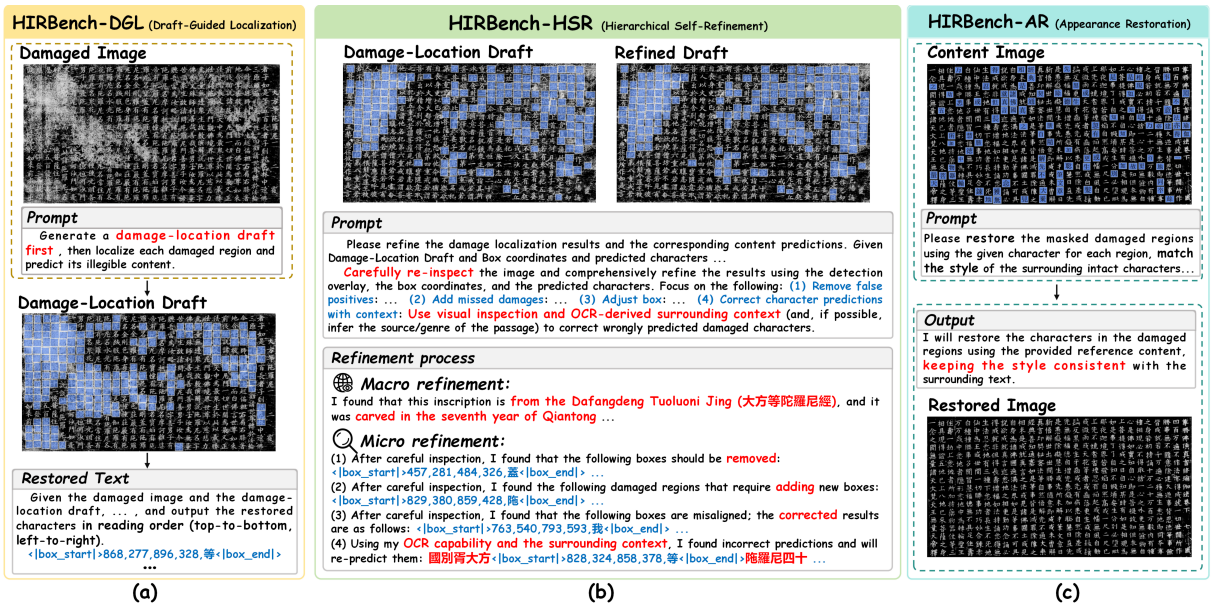

Figure 3: Overview of HIRBench. It contains three subsets: HIRBench-DGL, HIRBench-HSR, and HIRBench-AR.

images to occlude original characters, generating 55,156 content-restored pairs (Fig.3 (c)).

5 Experiments

5.1 Evaluation Metrics

Restored Content Accuracy. We evaluate damaged localization using precision, recall, and F1 score at an IoU threshold of 0.5. Damaged content prediction is measured by Top-1 accuracy. For appearance restoration, since pixel-level ground truth on real data is unavailable, we assess restoration quality via character recognition accuracy. Specifically, we train a text-line OCR using AHCDB (Xu et al., 2019), MTHv2 (Ma et al., 2020), HisDoc1B (Shi et al., 2025b), MegaHan97K (Zhang et al., 2025c), and M^{5}HisDoc (Shi et al., 2023) to recognize the restored data, and adopt the commonly used Accurate Rate (Zhang et al., 2025e, 2026) to measure the **OCR Accuracy**, which is formulated as:

$$AccurateRate = (N_t - D_e - S_e - I_e)/N_t, \quad (7)$$

where N_t is the total number of characters in annotations, while D_e, S_e, and I_e denote deletion, substitution, and insertion errors, respectively.

Generation Quality. We then introduce Style-LPIPS to measure font-style similarity between restorations and ground truth (GT). To isolate style evaluation from content prediction errors, we treat all models as image-to-image translators: we provide GT-rendered inscription images as input (e.g., in Step 3 of UniHIR, Fig. 2, we replace the predicted content with blue-masked inscriptions rendered using GT bounding boxes and text). This eliminates confounding effects from mislocalized or incorrectly predicted characters. We then crop all repaired and legible characters according to annotations, resize them to 48×48, and compute LPIPS (Zhang et al., 2018) between each restored-GT character pair, averaging across all pairs for the final Style-LPIPS score. To further assess style consistency, we conduct a user study with 16 participants. Participants are asked to rate the font-style similarity between the restored and original regions on a 0–5 scale (5 = completely consistent, 0 = completely inconsistent).

5.2 Comparison with Existing Methods

We compare UniHIR with nine existing methods, including **specialized pipeline model** (AutoHDR (Zhang et al., 2025d)), **understanding models** (Qwen3-VL (Bai et al., 2025a), Qwen2.5-VL (Bai et al., 2025b), and InternVL3.5 (Wang et al., 2025)), **generation models** (FLUX.1 Kontext (Labs et al., 2025) and Qwen-Image (Wu et al., 2025a)), and **unified MLLMs** (Lumina-DiMOO (Xin et al., 2025), BLIP3o-NEXT (Chen et al., 2025b), and BAGEL (Deng et al., 2025)). We fine-tune all baselines for a fair comparison, except AutoHDR. Specifically, understanding models are fine-tuned on HIRBench-DGL for damage localization and illegible content prediction, and evaluated only on these two tasks. Generation models are

# Row	Method	F1 score ↑	Precision ↑	Recall ↑	Prediction Accuracy ↑	OCR Accuracy ↑		
						Light	Medium	Severe
	Specialized Pipeline Model							
1	*AutoHDR* (Zhang et al., 2025d)	**94.10**	**97.00**	91.40	95.15	91.92	90.21	85.00
	Understanding Model (MLLM)							
2	Qwen2.5-VL-7B (Bai et al., 2025b)	39.27	37.26	41.51	38.88	-	-	-
3	Qwen3-VL-8B (Bai et al., 2025a)	75.50	71.81	79.60	83.80	-	-	-
4	InternVL3.5-8B (Wang et al., 2025)	9.45	13.36	7.31	11.95	-	-	-
	Generation Model							
5	FLUX.1 Kontext (Labs et al., 2025)	50.36	49.17	51.61	-	-	-	-
6	Qwen-Image (Wu et al., 2025a)	69.85	67.22	72.69	-	-	-	-
	Unified MLLM							
7	Lumina-DiMOO (Xin et al., 2025)	7.35	6.41	8.06	0.62	1.14	1.34	1.09
8	BLIP3o-NEXT (Chen et al., 2025b)	6.90	4.38	16.34	0.48	0.33	0.54	0.70
9	BAGEL (Deng et al., 2025)	67.60	85.63	55.84	64.24	85.00	76.54	57.81
11	**UniHIR (Ours)**	91.42	90.41	**92.45**	**96.38**	**92.06**	**90.46**	**86.65**

Table 1: Comparison with existing methods. **Bold** and underline indicate the best and second-best results, respectively. "Prediction Accuracy" measures the accuracy of damaged-content prediction, and "OCR Accuracy" measures OCR recognition accuracy. Light, medium, and severe denote inscriptions with mild, moderate, and severe damage.

fine-tuned on HIRBench-DGL to generate damage-location drafts, and are evaluated only for damage localization by extracting the blue boxes from the generated drafts. Unified MLLMs are fine-tuned on HIRBench-DGL (without DGL and HSR) and HIRBench-AR.

Main Results. The quantitative results are reported in Tab. 1. UniHIR achieves SOTA performance on both illegible content prediction (Prediction Accuracy) and appearance restoration (OCR Accuracy). Compared with the strongest unified MLLM baseline (BAGEL), it improves prediction accuracy by 32.14% (from 64.24 to 96.38) and boosts OCR accuracy by up to 28.84% under severe damage (from 57.81 to 86.65). Moreover, UniHIR outperforms all MLLMs and unified MLLMs in damage localization, and achieves performance comparable to AutoHDR, which relies on a powerful specialized DINO detector (Zhang et al., 2023). Additionally, it is evident from Tab. 1 that advanced generation models (e.g., Qwen-Image) surpass most advanced understanding models (e.g., Qwen2.5-VL-7B) on the damage localization (a dense detection task), except for Qwen3-VL-8B, which is explicitly optimized for object detection. This is exactly what inspired us to propose the Draft-Guided Localization mechanism. Overall, these results demonstrate UniHIR's remarkable capability in historical inscription restoration.

Generation Quality. As described in Sec. 5.1, we use GT-constructed inputs and evaluate only the appearance restoration step (Step3). This removes the influence of localization and content-prediction errors, enabling a more objective comparison of character generation quality and style consistency

Method	OCR Accuracy ↑			Style LPIPS ↓	User Study ↑
	Light	Medium	Severe		
Original Image	85.05	75.57	55.68	-	-
AutoHDR	93.73	94.18	94.63	0.2055	3.02
Flux.1 Kontext	82.94	81.53	68.11	0.2072	3.01
Qwen-Image	91.55	91.48	91.27	0.2059	3.76
Lumina-DiMOO	1.21	1.44	1.17	0.5229	0.18
BLIP3o-NEXT	0.32	0.55	0.72	0.2631	0.05
BAGEL	91.89	92.23	92.42	0.1975	3.73
UniHIR (Ours)	**94.11**	**94.57**	**95.08**	**0.1968**	**4.03**

Table 2: Generation-quality evaluation using GT-rendered inputs to isolate restoration/style quality from OCR Accuracy, Style-LPIPS, and user study.

across methods. As shown in Tab. 2, UniHIR achieves the best performance. Compared with the original damaged images, it improves OCR accuracy by 9.06%, 19.00%, and 39.40% under light, medium, and severe damage, respectively. Moreover, UniHIR attains the lowest Style-LPIPS and the highest user-study score, indicating stronger restoration capability with more consistent style and more accurate characters.

As presented in Fig. 4, we further conduct a qualitative analysis. Lumina-DiMOO and BLIP3o-NEXT can barely generate intact and legible characters: the former may be limited by the representation bottleneck of VQ-based discrete image codes, which hinders fine-stroke recovery, while the latter may suffer from insufficient character-generation data during pretraining, leading to weak character modeling. BAGEL and Flux.1 Kontext can generate characters but often produce blurry shapes. In contrast, Qwen-Image and AutoHDR perform better overall but still exhibit clear issues. Qwen-Image tends to produce relatively uniform, standard-looking font styles, with occasional character-shape errors; it also often generates overly clean backgrounds, causing background

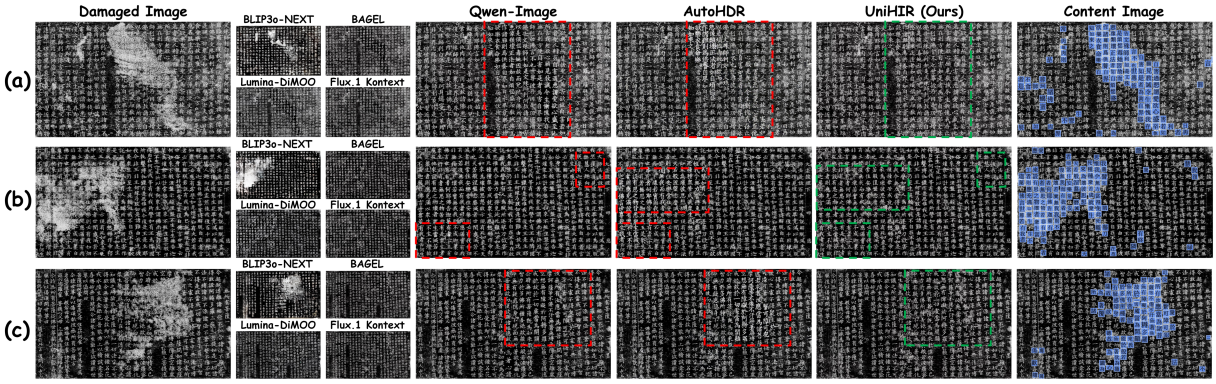

Figure 4: Qualitative comparison. We visualize the results of some evaluated methods. Red highlights regions with inconsistent background or inaccurate characters, while green denotes areas with satisfactory restoration quality.

DGL	HSR data (train-only)	HSR (inference)		F1	Pred. Acc
		Micro	Macro		
✗	✗	✗	✗	67.60	64.24
✓	✗	✗	✗	77.44	85.53
✓	✓	✗	✗	82.86	91.22
✓	✓	✓	✗	90.95	95.95
✓	✓	✓	✓	91.42	96.38

Table 3: Ablation study of Draft-Guided Localization (DGL) and Hierarchical Self-Refinement (HSR). "HSR data (train-only)" denotes training with additional Draft–Refinement paired data, while no refinement is performed at inference time unless Micro/Macro is enabled.

Training Strategy	Mode	F1 ↑	R ↑	P ↑
Understanding-only	Und.	61.02	47.87	84.13
Generation-only	Gen.	63.56	56.27	73.02
Hybrid (no DGL)	Und.	63.39	48.53	91.36
	Gen.	68.89	59.52	81.76
Hybrid (DGL)	Und.	71.77	69.76	73.00
	Gen.	71.25	69.57	73.02

Table 4: Synergy ablation between understanding and generation under different training strategies.

inconsistencies and deviating from the goal of "restoring the old as the old" (Fig. 4 (a)(b)). In comparison, AutoHDR's restored characters are often stylistically inconsistent with the intact ones (Fig. 4 (a)(b)(c)). Overall, UniHIR is the most stable in both character clarity and style consistency, delivering superior restoration quality.

Discussion. As shown in Table 1, UniHIR achieves a lower F1 score than AutoHDR but obtains higher OCR accuracy. This is because the two metrics evaluate different aspects of the task: F1 measures the accuracy of damaged-region localization, while OCR accuracy reflects the correctness of the restored characters. In practice, more precise semantic reconstruction of damaged content can lead to more faithful character recovery, even if the localization is less accurate. In addition, UniHIR performs page-level restoration, producing more visually consistent results that further benefit downstream OCR performance.

5.3 Ablation Study

Effectiveness of Draft-Guided Localization and Hierarchical Self-Refinement mechanism. We further analyze Draft-Guided Localization (DGL) and Hierarchical Self-Refinement (HSR) in Tab. 3. DGL alone improves the F1 score from 67.6 to 77.4, and raises prediction accuracy from 64.24 to 85.53. This indicates that the damage-location draft serves as an effective spatial prior, guiding the model toward damaged regions and making downstream predictions more stable. Moreover, training with HSR data further boosts results to 82.86/91.22 (F1/Pred. Acc) even without inference-time refinement, indicating that the model benefits from the diversity of the HSR training data, which improves robustness to varied degradation patterns. Finally, enabling micro-refinement at inference time boosts performance to 90.95 / 95.95 (F1/Pred. Acc), and adding macro-refinement further improves it to 91.42 / 96.38. These gains indicate that HSR effectively detects and corrects erroneous damage-region and illegible-content predictions, while the additional benefit from macro-refinement highlights the value of global attribute for resolving remaining errors.

Is there synergy between generation and understanding? We conduct a 20k-step ablation on BAGEL under the HIRBench-DGL setting. We design four training modes around BAGEL's understanding and generation branches: (1) Understanding-only: given a damaged image, directly predict the damage-location coordinates; (2) Generation-only: given a damaged image, directly generate the damaged-location draft; (3) Hybrid (no DGL): BAGEL's original training and in-

ference scheme, where the understanding branch predicts coordinates first and then conditions the generation branch on these coordinates; (4) Hybrid (DGL): training with DGL. As shown in Tab. 4, generation and understanding exhibit clear synergy in our task. Compared with understanding-only and generation-only training, joint training without DGL already improves both branches, with a more notable F1 gain on the generation branch (Tab. 4 (Rows 2-5)). Introducing DGL further yields the best F1 for both branches, again confirming that an aligned damage-location draft provides shared spatial cues for downstream prediction. Additional ablations (e.g., the effect of the number of refinement iterations) are provided in the Appendix D.

6 Conclusion

In this paper, we propose UniHIR, a novel unified MLLM for end-to-end Historical Inscription Restoration. UniHIR introduces Draft-Guided Localization and Hierarchical Self-Refinement to enable iterative reasoning and self-correction, supporting reliable damage localization, illegible-content prediction, and full page-level appearance restoration with improved style consistency. To enable efficient training under high-resolution inputs and long sequences, we further design UHIRFactory, a memory-efficient step-wise training framework, and construct HIRBench with step-aware supervision. Experiments demonstrate that UniHIR significantly improves restoration accuracy and visual rendering quality, and also validate the effectiveness of unified MLLMs over traditional pipeline-based approaches. We hope this work not only provides meaningful support for the preservation of cultural heritage but also opens up a new avenue for automated HIR.

Acknowledgements

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

Limitations

UniHIR currently restores illegible content by rendering standard glyphs onto the original damaged images. As a result, once a character is classified as damaged, it may be fully regenerated even when some original strokes remain, rather than completing only the missing parts. For Historical Inscription Restoration, a more conservative strategy—preserving existing strokes and repairing only the missing ones—may better align with the goal of minimal alteration. We leave stroke-preserving, fine-grained restoration (e.g., with stroke-level constraints and local editing) to future work.

References

Yannis Assael, Thea Sommerschield, Alison Cooley, Brendan Shillingford, John Pavlopoulos, Priyanka Suresh, Bailey Herms, Justin Grayston, Benjamin Maynard, Nicholas Dietrich, and 1 others. 2025. Contextualizing ancient texts with generative neural networks. *Nature*, 645(8079):141–147.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

David Bamman and Patrick J Burns. 2020. Latin bert: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*.

Black Forest Labs. 2024. Flux. Https://github.com/black-forest-labs/flux.

Jiahuan Cao, Yang Liu, Peirong Zhang, Yongxin Shi, Kai Ding, and Lianwen Jin. 2025. Tonggu-vl: Advancing visual-language understanding in chinese classical studies through parameter sensitivity-guided instruction tuning. In *Proceedings of the ACM International Conference on Multimedia (MM)*, page 11111–11120.

Fengjiao Chen, Minhao Jing, Weitao Lu, Yan Feng, Xiaoyu Li, and Xuezhi Cao. 2025a. Unihetero: Could generation enhance understanding for vision-language-model at large data scale? *arXiv preprint arXiv:2512.23512*.

Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, Tianyi Zhou, Junnan Li,

Silvio Savarese, Caiming Xiong, and Ran Xu. 2025b. Blip3o-next: Next frontier of native image generation. *Preprint*, arXiv:2510.15857.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*.

Wei-Sheng Du. 1999. "restoration as old" and "restoration as new" in the repair of ancient books. *Journal of Beijing Library*, (04):99–102.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*.

Lijie Fan, Luming Tang, Siyang Qin, and et al. 2025. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*.

Ethan Fetaya, Yonatan Lifshitz, Elad Aaron, and Shai Gordin. 2020. Restoration of fragmentary babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences*, 117(37):22743–22751.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024. Seed-X: Multimodal Models with Unified Multi-Granularity Comprehension and Generation. *arXiv preprint arXiv:2404.14396*.

Kaixin Han, Weitao You, Huanghuang Deng, Lingyun Sun, Jinyu Song, Zijin Hu, and Heyang Yi. 2024. Lant: finding experts for digital calligraphy character restoration. *Multimedia Tools and Applications*, pages 1–24.

Xin He, Longhui Wei, Jianbo Ouyang, Lingxi Xie, and Qi Tian. 2025. EMMA: Efficient Multimodal Understanding, Generation, and Editing with a Unified Architecture. *arXiv preprint arXiv:2512.04810*.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, and 2 others. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *Preprint*, arXiv:2506.15742.

Koren Lazar, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. Filling the gaps in ancient akkadian texts: a masked language modelling approach. *arXiv preprint arXiv:2109.04513*.

Bin Lin, Zongjian Li, Xinhua Cheng, and et al. 2025. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*.

Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, and Yongpan Wang. 2020. Joint layout analysis, character detection and recognition for historical document digitization. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 31–36. IEEE.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, and et al. 2025. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *CVPR*, pages 7739–7751.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, and et al. 2025. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*.

Katerina Papavassileiou, Dimitrios I Kosmopoulos, and Gareth Owens. 2023. A generative model for the mycenaean linear b script and its application in infilling text from ancient tablets. *ACM Journal on Computing and Cultural Heritage*, 16(3):1–25.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Yang Shi, Yuhao Dong, Yue Ding, and et al. 2025a. Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. *arXiv preprint arXiv:2509.24897*.

Yongxin Shi, Chongyu Liu, Dezhi Peng, Cheng Jian, Jiarong Huang, and Lianwen Jin. 2023. M5HisDoc: A Large-scale Multi-style Chinese Historical Document Analysis Benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 78483–78495.

Yongxin Shi, Dezhi Peng, Yuyi Zhang, Jiahuan Cao, and Lianwen Jin. 2025b. A large-scale dataset for chinese historical document recognition and analysis. *Scientific Data*, 12(1):169.

Quan Sun, Qiying Yu, Yufeng Cui, and et al. 2024a. Emu: Generative pretraining in multimodality. In *ICLR*.

Yutao Sun, Hangbo Bao, Wenhui Wang, Zhiliang Peng, Li Dong, Shaohan Huang, Jianyong Wang, and Furu Wei. 2024b. Multimodal latent language modeling with next-token diffusion. *arXiv preprint arXiv:2412.08635*.

Chameleon Team. 2024. Chameleon: Mixed-Modal Early-Fusion Foundation Models. *arXiv preprint arXiv:2405.09818*.

Shengbang Tong, David Fan, Jiachen Li, and et al. 2025. MetaMorph: Multimodal Understanding and Generation via Instruction Tuning. In *ICCV*, pages 17001–17012.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *Preprint*, arXiv:2502.14786.

Guo-Qiang Wang. 2021. Minimal Intervention Principle for Ancient Book Conservation in China :Techniques and Application Strategy. *Library Forum*, 41(07):141–148.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, and et al. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, and 20 others. 2025a. Qwen-image technical report.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, and et al. 2025b. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *CVPR*, pages 12966–12977.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. 2025c. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, and et al. 2024. VILA-U: A Unified Foundation Model Integrating Visual Understanding and Generation. *arXiv preprint arXiv:2409.04429*.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. 2025. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, and 1 others. 2025. Lumina-dimoo: An omni diffusion large language model for multimodal generation and understanding. *arXiv preprint arXiv:2510.06308*.

Yue Xu, Fei Yin, Da-Han Wang, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. 2019. Casia-ahcdb: A large-scale chinese ancient handwritten characters database. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 793–798. IEEE.

Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Zhendong Wang, Hao Liu, Bin Lin, Hao Li, Xue Xu, and 1 others. 2025. Unified multimodal model as auto-encoder. *arXiv preprint arXiv:2509.09666*.

Zhenhua Yang, Dezhi Peng, Yongxin Shi, Yuyi Zhang, Chongyu Liu, and Lianwen Jin. 2025. Predicting the Original Appearance of Damaged Historical Documents. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. 2023. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations(ICLR)*.

Peirong Zhang, Haowei Xu, Jiaxin Zhang, Xuhan Zheng, Guitao Xu, Yuyi Zhang, Junle Liu, Zhenhua Yang, Wei Zhou, and Lianwen Jin. 2026. OCRGen-Bench: A Comprehensive Benchmark for Evaluating OCR Generative Capabilities.

Peirong Zhang, Jiaxin Zhang, Jiahuan Cao, Hongliang Li, and Lianwen Jin. 2025a. Smaller But Better: Unifying Layout Generation with Smaller Large Language Models. *International Journal of Computer Vision (IJCV)*, 133:3891–3917.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. 2025b. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *Preprint*, arXiv:2505.02567.

Yuyi Zhang, Yongxin Shi, Peirong Zhang, Yixin Zhao, Zhenhua Yang, and Lianwen Jin. 2025c. Mega-Han97K: A large-scale dataset for mega-category Chinese character recognition with over 97K categories. *Pattern Recognition*, 167:111757.

Yuyi Zhang, Peirong Zhang, Zhenhua Yang, Pengyu Yan, Yongxin Shi, Pengwei Liu, Fengjun Guo, and Lianwen Jin. 2025d. Reviving cultural heritage: A novel approach for comprehensive historical document restoration. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Yuyi Zhang, Yuanzhi Zhu, Dezhi Peng, Peirong Zhang, Zhenhua Yang, Zhibo Yang, Cong Yao, and Lianwen Jin. 2025e. Hiercode: A lightweight hierarchical codebook for zero-shot Chinese text recognition. *Pattern Recognition*, 158:110963.

Shipeng Zhu, Ang Chen, Na Nie, Pengfei Fang, Min-Ling Zhang, and Hui Xue. 2026. Epiagent: An agent-centric system for ancient inscription restoration. *Preprint*, arXiv:2604.09367.

Shipeng Zhu, Hui Xue, Na Nie, Chenjie Zhu, Haiyue Liu, and Pengfei Fang. 2024. Reproducing the past: A dataset for benchmarking inscription restoration. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 7714–7723.

A Implementation Details

UniHIR is trained in four successive stages, as summarized in Table 5: Incremental Pretraining, Supervised Fine-Tuning (SFT), and two Resolution Enhancement stages. All experiments were conducted on 10 NVIDIA A100 GPUs.

In the incremental pretraining stage, we train the model on CCS358K (Cao et al., 2025) and HisDoc1B (Shi et al., 2025b). We then perform SFT on the HIRBench dataset. In this stage, the training sampling ratio for DGL, HSR, and AR is set to 2:2:1. In the resolution enhancement stage, the training sampling ratio for DGL, HSR, and AR is set to 3:3:4.

Across all stages, we use the AdamW optimizer with $(\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 1.0 \times 10^{-15})$, set the weight decay to 0.0, and apply gradient norm clipping at 1.0 to stabilize training. We keep a constant learning rate of 2.0×10^{-5} throughout all stages, and maintain an exponential moving average (EMA) of model parameters with a decay ratio of 0.9999. The training objective includes a text-generation loss and an image-generation loss, weighted by λ_{CE} and λ_{MSE}. As shown in Table 5, we set CE:MSE to $2:1$ for Incremental Pretraining and to $1:1$ for SFT and both Resolution Enhancement stages.

We train each stage for a fixed number of steps (80K / 30K / 55K / 44K) without learning-rate warm-up. The maximum sequence length per sample is set to 14K, 18K, 18K, and 28K for the four stages, respectively. Due to BAGEL's 1024 long-side limit for image generation, in the resolution enhancement stage, we increase the generation resolution. The generation resolution is $(512, 1024)$ in Incremental Pretraining and SFT, then increased to $(512, 1536)$ in Resolution Enhancement (1) and further to $(512, 2048)$ in Resolution Enhancement (2), while the understanding resolution is fixed to $(224, 980)$ across all stages. The diffusion timestep shift is set to 4.0 for all stages.

B Data Synthesis Details

HIRBench is constructed based on FPHDR (Zhang et al., 2025d) annotations, which provide fine-grained labels for both intact and damaged characters. The dataset is synthesized through rule-based transformations, without relying on additional generative models or external large language models.

In UniHIR, the initial drafts generated in Step 1 often contain systematic errors, such as incorrectly labeling intact characters as damaged or misclassifying damaged characters as intact. To enable the model to iteratively correct such errors through refinement, we explicitly model and simulate these common error patterns when constructing the training data. Specifically, we consider three types of errors:

(1) **Damage misclassification simulation.** We randomly apply occlusions to intact characters to simulate cases where intact characters are mistakenly identified as damaged. Concretely, square-shaped masks are overlaid on target regions, using a light-colored fill (e.g., white) with visible boundaries and semi-transparency (opacity 160). This introduces occlusion while preserving partial underlying visual structure.

(2) **Localization perturbation.** We simulate detection inaccuracies by perturbing character bounding boxes. The perturbations include random translation and scaling (with translation bounded by 2% of the image size and scaling within 10% of the original box size), structured shifts along horizontal or vertical directions (with offsets ranging from 20% to 40% of the box width or height), as well as detection errors such as box deletion (missed detection), box insertion (false positives), and character misclassification. These perturbations are applied with different probabilities conditioned on the damage level of each character in FPHDR (e.g., uncertain, vague, or deform), resulting in diverse and realistic error patterns.

(3) **Recognition error simulation.** We simulate recognition errors by randomly replacing characters. The replacement characters are sampled from a predefined character set collected from the dataset, ensuring consistency and plausibility.

C Training Details

C.1 Training Loss

UniHIR is jointly trained for Draft-Guided Localization, Hierarchical Self-Refinement, and appearance restoration with two complementary loss functions:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CE}} \left(-\sum_{i=1}^{N} y_i \log(\hat{y}_i) \right) + \lambda_{\text{MSE}} \left\| \hat{\mathbf{v}} - \mathbf{v} \right\|_2^2 . \tag{8}$$

Here, the CE term is used to optimize text generation, where $\hat{y}_i$ and y_i denote the predicted and target token probabilities. The MSE term is used to optimize image generation in the VAE latent

Hyperparameters	Incremental Pretraining	Supervised Fine-Tuning	Resolution Enhancement (1)	Resolution Enhancement (2)
Learning rate	2.0×10^{-5}	2.0×10^{-5}	2.0×10^{-5}	2.0×10^{-5}
LR scheduler	Constant	Constant	Constant	Constant
Weight decay	0.0	0.0	0.0	0.0
Gradient norm clip	1.0	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9,\ \beta_2 = 0.95,\ \epsilon = 1.0 \times 10^{-15}$)			
Loss weight (CE : MSE)	2 : 1	1 : 1	1 : 1	1 : 1
Warm-up steps	0	0	0	0
Training steps	80K	30K	55K	44K
EMA ratio	0.9999	0.9999	0.9999	0.9999
Max sequence length per samples	14K	18K	18K	28K
Gen resolution (min short side, max long side)	(512, 1024)	(512, 1024)	(512, 1536)	(512, 2048)
Und resolution (min short side, max long side)	(224, 980)	(224, 980)	(224, 980)	(224, 980)
Diffusion timestep shift	4.0	4.0	4.0	4.0

Table 5: Training details.

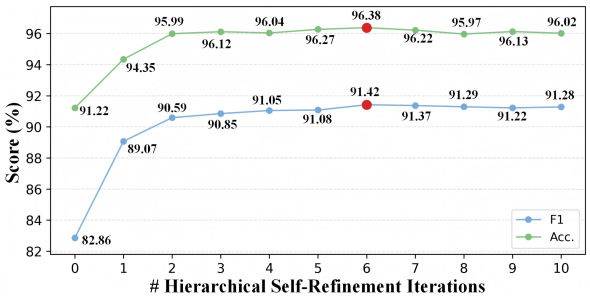

Figure 5: Effect of refinement iterations.

space, where $\hat{\mathbf{v}}$ and $\mathbf{v}$ denote the predicted and target restoration outputs.

C.2 Training Objective of Draft Generation

The draft generator in UniHIR is trained using a flow-matching objective. Specifically, we adopt a mean squared error (MSE) loss to supervise the velocity prediction in the latent space, as defined in Eq. 8. For draft mask supervision, we construct ground-truth masks from annotated bounding boxes of damaged characters in HIRBench-DGL. These masks are rendered as spatial regions and serve as direct supervision signals for draft generation.

C.3 Iterative Refinement Mechanism

UniHIR employs an iterative refinement strategy in which the draft is regenerated at each iteration based on the updated restoration results. This design mitigates error accumulation and enables progressive correction of both localization and content prediction errors.

D Additional Ablation Study

We study how the number of refinement iterations affects performance. As shown in Fig. 5, performance increases sharply in the first two iterations, then the gains slow down and plateau around 3-6 iterations, after which it only fluctuates slightly. We also find that self-refinement rarely terminates by itself: due to the complexity of stele damage, later iterations often get stuck on ambiguous, borderline regions that are arguably fixable or not, leading to repeated deliberation. Balancing accuracy and inference cost, we recommend 3-6 refinement iterations.

E More Visualization Results

E.1 Additional Qualitative Results

As shown in Fig. 6, we provide additional visualizations of restoration results from AutoHDR (Zhang et al., 2025d), Qwen-Image (Wu et al., 2025a), BLIP3o-NEXT (Chen et al., 2025b), BAGEL (Deng et al., 2025), FLUX.1 Kontext (Labs et al., 2025), Lumina-DiMOO (Xin et al., 2025), and UniHIR. The comparisons show that UniHIR produces more accurate character restorations while better preserving stylistic consistency with surrounding text.

E.2 Iterative Refinement Analysis

As shown in Fig. 7, we further visualize the iterative refinement process of UniHIR. From left to right, the results correspond to different iterations, where the model progressively corrects missed detections, false positives, and inaccurate bounding boxes. The red boxes highlight initially incorrect regions, while the green boxes indicate regions

that have been successfully corrected during the refinement process. This demonstrates the effectiveness of our iterative strategy in gradually enhancing restoration quality.

F Inference Time Analysis

We report the inference time of UniHIR to assess the computational cost of iterative self-refinement. The evaluation is conducted on 50 sample pairs using a single NVIDIA A100 GPU. With six refinement iterations, the average total inference time is 267.96 seconds per sample. Specifically, Step 1 (Draft-Guided Localization) takes approximately 56.57 seconds, each refinement iteration requires about 25.01 seconds, and Step 3 (Appearance Restoration) takes around 61.23 seconds.

G OCR-Based Evaluation and Domain Gap Analysis

We use an OCR-based metric to evaluate restoration quality. Although the OCR model is trained on AHCDB, MTHv2, and M5HisDoc, which are not specifically composed of inscription rubbing data, the primary discrepancy lies in the foreground–background contrast (i.e., white background vs. black background). To mitigate this domain gap, we invert the colors of inscription rubbing images prior to recognition, converting them into black text on a white background to better align with the OCR training distribution.

To further assess the impact of the domain gap, we evaluate the OCR model on intact characters from the FPHDR dataset. Out of 651,075 characters, 639,856 are correctly recognized, achieving an accuracy of 98.28%. This result indicates that the OCR model generalizes well to our data, and the domain discrepancy has a minimal effect on evaluation reliability.

H Reliability and Human-in-the-Loop Verification

While UniHIR is fine-tuned on historical data, it may still produce plausible yet incorrect restorations in highly ambiguous cases. To enhance reliability in practical applications, we adopt a human-in-the-loop workflow for historical inscription restoration. Specifically, domain experts review and refine the intermediate outputs generated at each stage of the pipeline, ensuring that potential errors are identified and corrected in a timely manner. This process helps guarantee the accuracy and scholarly reliability of the final restoration.

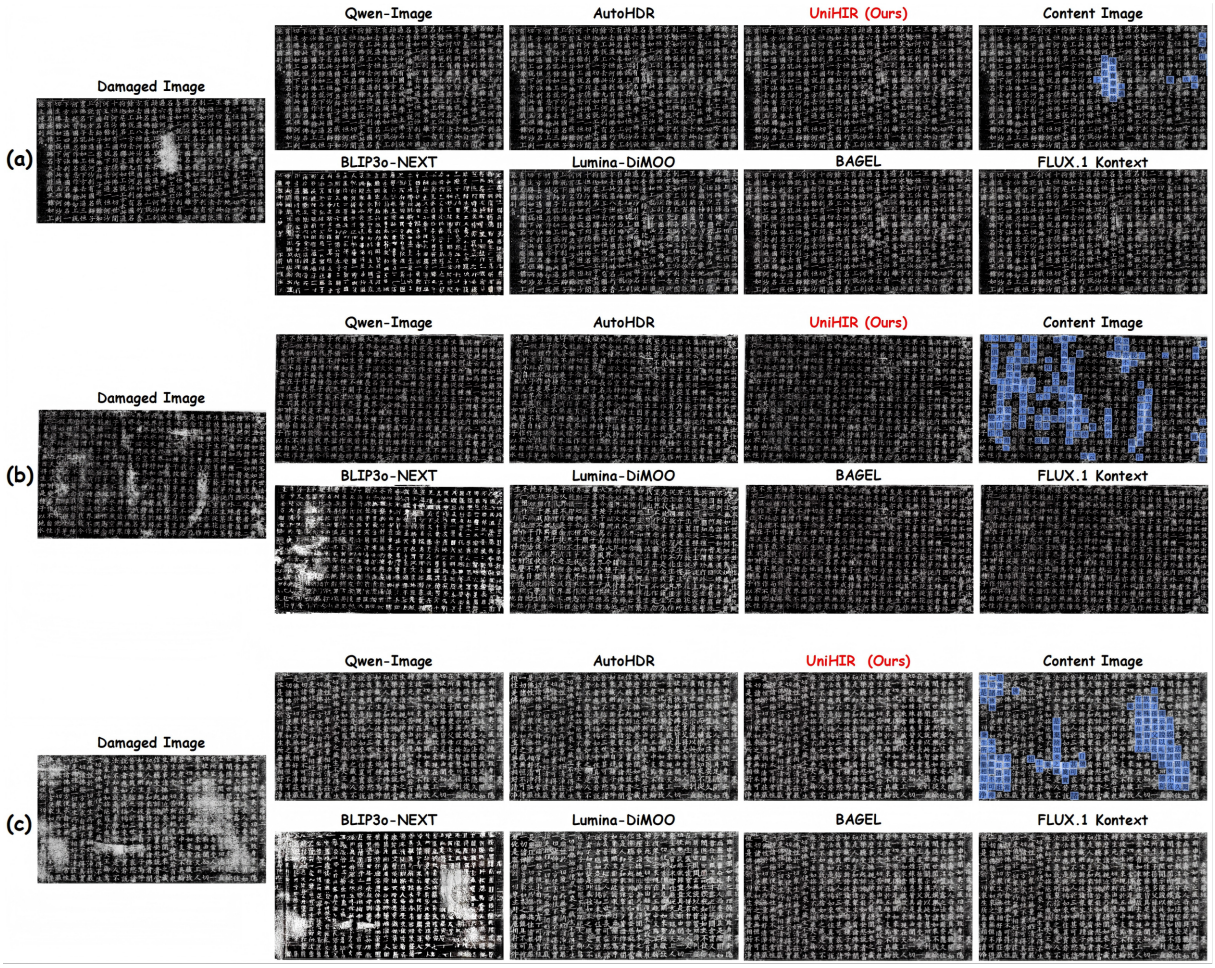

Figure 6: Additional qualitative comparison.

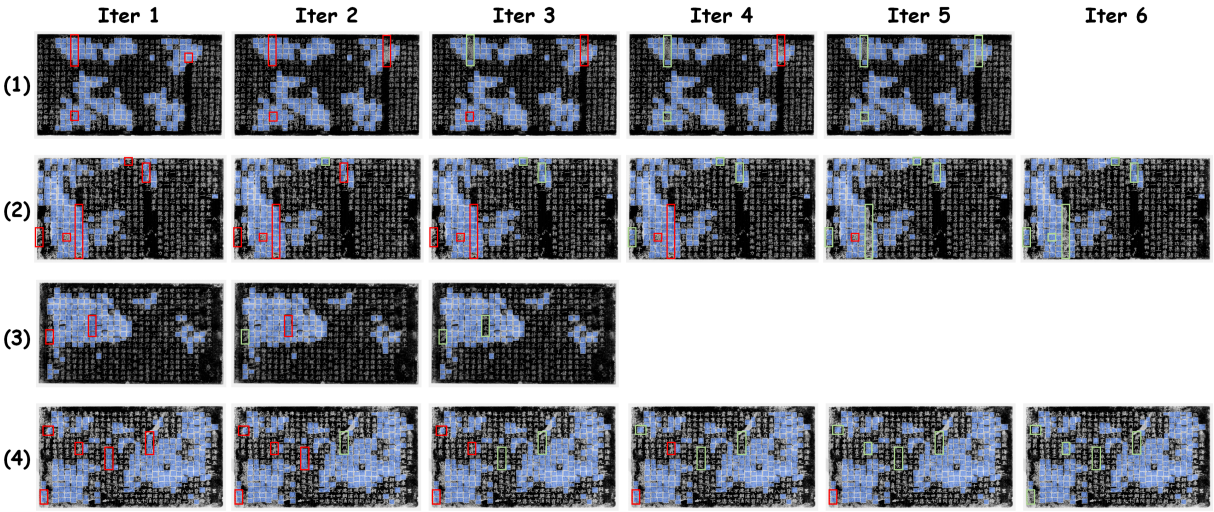

Figure 7: Visualization of the iterative refinement process. From left to right, the results at different iterations ($k = 1$ to $k = 6$) demonstrate the progressive improvement over iterations. Red boxes indicate erroneous regions, while green boxes denote corrected regions.