A UNIFIED MANIFOLD FRAMEWORK FOR MODELING CHINESE SCRIPT EVOLUTION WITH APPLICATION TO ORACLE BONE DECIPHERMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Of the approximately 4,500 Oracle Bone Inscription (OBI) characters discovered from the Shang dynasty, only about 1,600 have been deciphered. Existing computational methods mainly compute the similarity between OBI and single historical period characters independently. However, during the evolution of Chinese characters, significant structural or semantic changes often occur in uncertain dynasties. Relying solely on a single dynasty for reference leads to numerous deciphering errors. Therefore, we propose the **Manifold-based Script Evolution Framework (MSEF)**, the first unified framework that models the evolution series (OBI, Bronze, Seal, Clerical, Regular) of Chinese characters as the continual evolution of a manifold space. MSEF represents each character as an era-specific manifold point and learns continuous inter-era transition rules via Neural Ordinary Differential Equations. Both manifold space and transition dynamics can be trained end-to-end through character evolution pairs across any two eras. To support this training scheme, we construct the first unified cross-era dataset with **fine-grained temporal and regional granularity**. Based on MSEF, we propose the **Cascaded Bidirectional Evolutionary Decipherment (CBED)** algorithm, where forward predictions recall potential candidates and backward consistency checking filters false matches, effectively avoiding errors caused by single-dynasty comparisons. To ground our approach, we provide a pre-exploratory diagnosis of existing generative and retrieval paradigms, showing why continuous manifold dynamics are well suited for fragmented cross-era script data. We further conduct mechanistic post-analysis, showing that the learned representations and transition dynamics capture patterns consistent with known paleographic transformations. Experiments on three commonly used benchmarks and our dataset demonstrate that MSEF achieves state-of-the-art performance among current oracle deciphering algorithms. This work opens the possibility of using continuous manifold dynamics in cross-era paleographic decipherment. Dataset and code are released through our project page `https://iclr-oracle-submission.github.io/`.

1 INTRODUCTION

Oracle Bone Inscription (OBI) is a cornerstone in the annals of linguistic history, predating many established writing systems Gil (2020). Chinese characters evolve over three millennia from Shang dynasty OBI, through Zhou dynasty Bronze Inscriptions, Qin dynasty Seal Script, Han dynasty Clerical Script, to modern Regular Script. Figure 1 illustrates the evolution process. Among the approximately 4,500 OBI characters discovered, only about 1,600 (35.6%) have been deciphered Keightley (1985); Li et al. (2020). Deciphering the remaining $\sim$2,900 characters would unlock invaluable insights into early Chinese civilization Wang & Deng (2024), yet this process often takes years per character due to the need to thoroughly examine structural features and semantic patterns across different eras.

The emergence of AI technologies presents a novel frontier for OBI deciphering. Current approaches generally fall into two categories: (1) compare OBI with modern simplified Chinese by image retrieval Meng et al. (2018); Hu et al. (2024) or generate modern simplified ChineseChang et al. (2022); Li

et al. (2023c), sometimes incorporating semantic information Chen et al. (2024); and (2) compare OBI with other historical scripts such as Bronze Wu et al. (2025).

The first type attempts direct mappings from OBI-to-modern, ignoring critical intermediate stages Guo et al. (2015); Zhang et al. (2019); Li et al. (2023b). This proves to be ineffective for characters that underwent major transformations during specific dynasties. The second type, while incorporating multi-era comparisons, treats historical scripts as independent static sets. For instance, CrossFont Wu et al. (2025) processes multi-era data but fails to model temporal evolutionary dynamics explicitly, which often leads to making it difficult to recall fonts that underwent drastic changes during the dynasties being compared. Moreover, character lifespans vary significantly—certain OBI glyphs became extinct before the Bronze era—yet approaches that force comparisons across all periods inevitably introduce noise. The current approach does not solve this issue either, but instead proceeds with decipherment under the assumption that all oracle bone inscriptions have survived to the present day. In such a situation, it is crucial to develop a method that fully accounts for the evolutionary patterns of the oracle bone script and considers decipherment across all dynasties. So we propose the **Manifold-based Script Evolution Framework (MSEF)** Bottéro (1996); Meilă & Zhang (2024), modeling character evolution as continuous flow through a shared manifold space directly. Specifically, (1) under MSEF, each historical period defines a distinct high-dimensional coordinate system where each character corresponds to a unique point. Inter-period transitions are governed by continuous, invertible Neural Ordinary Differential Equations (Neural ODE) Chen et al. (2018); Rubanova et al. (2019). (2) We implement the manifold using a time-conditioned transformer mapping feature vectors to coordinates, with a spectrally-normalized network for the ODE ensuring continuous transitions. (3) Furthermore, we additionally introduce a survival network that predicts character extinction probability across dynasties.

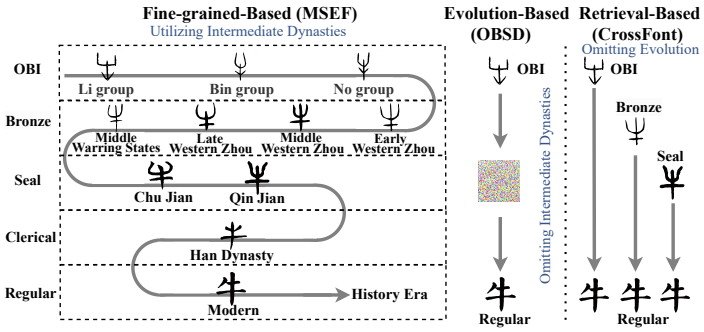

Figure 1: **Evolution of Chinese Script.** Taking the character "Ox" as an example, this figure demonstrates the evolution of Chinese characters across five historical periods. This character underwent a significant morphological change in early bronze script, then later reverted to a form similar to its oracle bone appearance. Conventional retrieval methods based solely on bronze script would incorrectly judge the similarity as low, leading to decipherment errors. Accurate decipherment requires considering the character's evolution across multiple dynasties and their temporal dependencies.

To support MSEF training, we **construct a fine-grained dataset** containing 1,500 complete evolution chains and ∼13,700 pairwise combinations. Based on trained MSEF, we propose **Cascaded Bidirectional Evolutionary Decipherment (CBED)**, which performs forward cascade retrieval with a survival mechanism to terminate extinct paths, followed by backward verification via inverse ODE flows with step-wise pruning for evolutionary consistency.

The final architecture proposed in this paper is the culmination of systematic theoretical derivation and experimental exploration. In our pre-exploratory analysis (Appendix D), we delineate this research trajectory: First, we systematically diagnose existing generative and retrieval paradigms, revealing that their failure modes stem from a fundamental inability to model intermediate stages of script evolution. Second, we return to first principles to define character evolution as a duality of an invariant semantic core and a variant visual surface driven by continuous socio-cultural pressures. Third, we conduct a quantitative analysis to identify distinct morphological evolution clusters, empirically mapping baseline vulnerabilities to non-monotone and abrupt historical transitions. Finally, we examine alternative solutions, showing that incremental patches or discrete rule-based expert systems face clear performance and scalability limitations. These diagnostic findings motivate our proposed MSEF powered by Neural ODEs.

Although extensive quantitative evaluations demonstrate the superior capabilities of MSEF, the fully trained model inherently presents as a complex black box. To better understand its underlying mechanisms and why it achieves robust decipherment performance, we conduct a mechanistic post-analysis (Appendix E). Our framework design centers on several critical components, notably the continuous manifold character representations, the Neural ODE inter-era transition dynamics, and the survival net. By employing mechanistic interpretability visualizations, including causal saliency maps and attention circuit diagrams, we conduct fine-grained analyses on each of these modules. These analyses provide evidence that the model does not merely memorize visual patterns, but captures information-routing patterns and morphodynamic regularities consistent with known paleographic transformations.

Our contributions are:

- We propose a Manifold-based Script Evolution Framework (MSEF) which could model the evolution series(OBI, Bronze, Seal, Clerical, Regular) of Chinese characters as the continual evolution of a manifold space. Such framework contains a manifold net for representing the period, neural ode for representing the period transition, and survival net for representing the Discontinuity of Chinese Characters.
- We construct and opensource the first unified cross-era oracle dataset which could support training the MSEF.
- We introduce a Cascaded Bidirectional Evolutionary Decipherment (CBED) mechanism that employs forward cascaded retrieval with backward verification, providing temporal consistency validation. Such a mechanism could fully leverage the potential of MSEF chinese character evolution knowledge in deciphering oracle bone task.
- Experiments on both 3 Commonly used datasets and our own datasets demonstrate that MSEF achieves the state-of-the-art among all currnet oracle deciphering algorithms.

2 Manifold-based Script Evolution Framework (MSEF)

In this section, we describe the proposed MSEF and corresponding CBED. We first present our observation and intuition of the Chinese character evolution in Section 2.1, and then introduce the theoretical form of the evolution of Chinese characters in Section 2.2, which unifies distinct historical eras into a single evolving manifold governed by differential equations. After that, we introduce how we implement and train such a mathematical model in Sections 2.3 and 2.4. Finally, by learning the vector field of character evolution, MSEF allows us to project ancient scripts forward to predict their modern counterparts and trace modern characters backward to verify their origins. Such a CBED algorithm is illustrated in Section 2.5.

2.1 Intuition: Script Evolution as Continuous Flow

Figure 2 visualizes the continuous evolution using "ox" as an example. Tracing the evolution from the OBI to the Regular, we observe that despite drastic stylistic shifts, the character preserves its fundamental topological structure. In such a situation, we must apply temporal modeling to textual evolution to capture these uncertain shifts, thereby enhancing decoding accuracy. More cases of these uncertain change patterns are provided in Appendix Q.

We model this dynamic evolution process as a continuous flow through a shared geometric space. All characters are points in a high-dimensional manifold $\mathcal{M}$. As time progresses, each point moves along a smooth trajectory governed by learnable dynamics. This allows us to trace any ancient character forward to find its modern correspondent and trace backward to verify consistency. To handle distinct representational vocabularies of different eras, we assign each era a distinct manifold space $\mathcal{M}_t$ connected by continuous transitions.

2.2 Manifold-based Script Evolution Framework Theory

Building on the intuition above, we now formalize our framework. We establish three foundational assumptions.

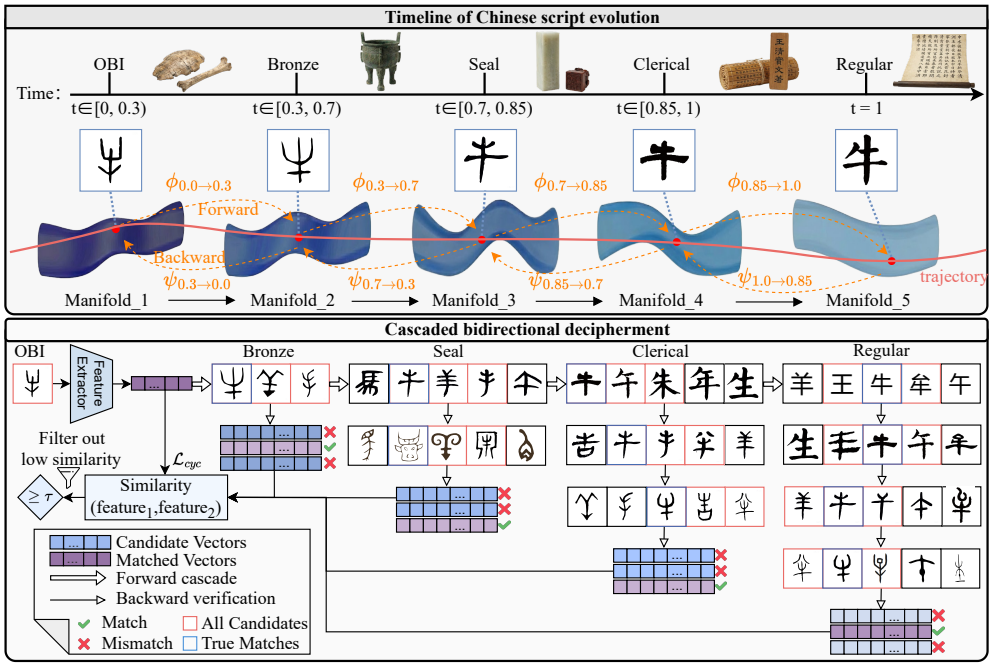

Figure 2: **Overview of MSEF. Top: Continuous Manifold Evolution.** The framework models the 3,000-year evolution of Chinese script as a continuous trajectory (red line) through era-specific manifolds $\mathcal{M}_t$, conditioned on normalized time t (from OBI $t = 0.0$ to Regular $t = 1.0$). Note that the range of t is proportional to the historical duration of its corresponding era. Neural ODEs govern the continuous transitions via forward flows ϕ and backward flows ψ. Specifically, each historical period manifold defines a distinct high-dimensional coordinate system where each character corresponds to a unique point. Inter-period point transitions are governed by the continuous neural ODE. Such a framework enables us to derive candidate character solutions for any dynasty within the character manifold space by leveraging any given dynasty. This facilitates the design of subsequent bidirectional cascading decryption algorithms. **Bottom: Cascaded Bidirectional Decipherment.** The inference process consists of two stages: (1) **Forward Recall**: Given an OBI query, the model progressively infers its coordinate in subsequent eras (Bronze $\rightarrow$ Seal $\rightarrow$ Clerical $\rightarrow$ Regular) to retrieve candidate sets. (2) **Backward Verification**: Candidates from the modern era are traced back through history via inverse flows. At each reverse transition, candidates showing large reconstruction errors (marked with ✗) are pruned step-by-step. The final decipherment is determined by the cycle consistency similarity with the original OBI features (filtered by threshold τ).

Assumption 1 (Shared Manifold Space). All scripts across eras share an underlying d-dimensional manifold $(\mathcal{M}, t)$, where d captures the essential dimensionality of character semantics and structure, and t denotes the different dynasties. We empirically validate this geometric structure in Appendix R, where visualizing local neighborhoods reveals distinct clusters of visually and semantically related characters.

Assumption 2 (Continuous Manifold Transition Dynamics). Character evolution follows continuous dynamics governed by a velocity field. Let $\mathbf{z}(t) \in \mathbb{R}^d$ denote the coordinates of a character's manifold in normalized time $t \in [0, 1]$ (with $t = 0$ for OBI and $t = 1$ for Regular). The evolution is described by:

$$\frac{d\mathbf{z}}{dt} = \mathbf{v}(\mathbf{z}, t; \boldsymbol{\theta}) \tag{1}$$

where $\mathbf{v}$ is a learnable velocity field parameterized by $\boldsymbol{\theta}$. This equation implies that the rate of a character's transformation is determined by both its current position in the manifold and the specific historical era.

Assumption 3 (Survival Probability). Some characters became obsolete over time. We model the survival probability $s(\mathbf{z}, t) \in [0, 1]$ as a learnable function.

Flow Operators. From Equation 1, we define forward and backward flow operators:

$$\phi_{t_1 \to t_2}(\mathbf{z}) = \mathbf{z} + \int_{t_1}^{t_2} \mathbf{v}(\mathbf{z}(\tau), \tau)\, d\tau \tag{2}$$

$$\psi_{t_2 \to t_1}(\mathbf{z}) = \mathbf{z} - \int_{t_1}^{t_2} \mathbf{v}(\mathbf{z}(\tau), \tau)\, d\tau \tag{3}$$

The forward flow ϕ traces a character forward in time (OBI $\to$ Regular), while the backward flow ψ traces it backward. Crucially, these are inverses: $\psi_{t_2 \to t_1} = \phi^{-1}_{t_1 \to t_2}$. This invertibility enables our bidirectional verification.

Time Encoding. We assign normalized times to historical periods based on archeological dating: OBI ($t = 0.00$–0.30, with subperiods for five scribal groups), Bronze ($t = 0.30$–0.70), Seal ($t = 0.70$–0.85), Clerical ($t = 0.85$–1.00), Regular ($t = 1.00$). Table 1 provides details. If a more granular classification of dynasties is required, such as dividing them into early, middle, and late periods, the dynasty can be divided equally into three parts, and so on.

Table 1: Time encoding for historical periods. Subperiod details are provided in Appendix L.

Period	**Time** t	**Historical Date**
OBI	0.00–0.30	c. 1200–1050 BCE
Bronze	0.30–0.70	c. 1050–250 BCE
Seal	0.70–0.85	c. 220 BCE
Clerical	0.85–1.00	c. 200 CE
Regular	1.00	c. 600 CE–present

2.3 Manifold-based Script Evolution Framework Neural Implementation

Our framework comprises three interconnected neural components: a manifold encoder, a velocity field network, and a survival prediction network.

Input Representation. For each character c in dynasty t, we extract a 352-dimensional feature vector $\mathbf{x}_c^t \in \mathbb{R}^{352}$ specifically designed for paleographic analysis. This representation aggregates five complementary modalities: **Visual**, **TextStructural**, **Semantic**, **Contextual**, and **Spatiotemporal**. Detailed definitions of characteristics and aggregation mechanisms are provided in the Appendix M.

Manifold Encoder. Manifold mapping $\mathcal{M}(\mathbf{x}_c^t)$ projects the input feature vector into a learned high-dimensional manifold space. We implement this encoder as a 12-layer transformer that maps the 352-dimensional input to coordinates $\mathbf{z}_c^t \in \mathbb{R}^d$ with the manifold dimension $d = 256$.

Velocity Field. The velocity network $\mathbf{v}(\mathbf{z}, t)$ models the temporal dynamics on the manifold. Given the manifold coordinates $\mathbf{z}_c^t$ of character c in dynasty t, solving the Neural ODE yields the evolved representation $\mathbf{z}_c^{t+1}$ in the subsequent dynasty. We parameterize the velocity field as a 3-layer MLP with spectral normalization and Tanh activations to ensure Lipschitz continuity.

Survival Network. The survival network $\mathcal{S}(\mathbf{z}_c^0, t)$ predicts the probabilities of character extinction. It takes as input the representation of the oracle bone period manifold $\mathbf{z}_c^0$ along with a target dynasty t, and outputs the probability that character c becomes extinct by the dynasty t. This component is implemented as an MLP.

Continuous Time Encoding. To enable the Neural ODE to capture the authentic continuous dynamics of script evolution, we implement an *ordered stochastic interval sampling* scheme for the temporal variable t. During training, for any given character feature $\mathbf{x}_c^t \in \mathbb{R}^{352}$, the corresponding dynasty t is not treated as a fixed scalar but is randomly sampled from its era-specific interval defined in Table 1.

Crucially, when processing evolution pairs or chains, we enforce a strict monotonicity constraint $t_1 < t_2 < \cdots < t_5$ within the range $(0, 1]$. This strategy ensures that the temporal input t functions as a continuous variable during the optimization process. By supervising the model across these dense stochastic intervals, we force the Neural ODE to learn a smooth, time-dependent velocity field $v(z, t)$ that generalizes the character transformation as a continuous flow through the manifold space, rather than a sequence of discrete, independent mappings between eras.

2.4 Manifold-based Script Evolution Framework Training

We train MSEF in an end-to-end manner. Specifically, for a given character, we obtain its corresponding tuples across different dynasties. Based on these data, we optimize our three neural components by minimizing the differences between the characters' transformations between two dynasties. Ideally, the high-dimensional feature of a character transformed from dynasty t to dynasty t should match the high-dimensional feature of that character directly computed for dynasty t'.

Let $\mathcal{P}$ denote all available character pairs in two eras:

$$|\mathcal{P}| = \underbrace{1.2\text{K}}_{\text{OBI}\to\text{Bronze}} + \underbrace{2\text{K}}_{\text{Bronze}\to\text{Seal}} + \underbrace{10\text{K}}_{\text{Seal}\to\text{Regular}} + \underbrace{0.5\text{K}}_{\text{complete}} \approx 13.7\text{K} \tag{4}$$

2.4.1 Multi-Scale Loss Functions

We train the MSEF using five complementary loss functions to ensure both local continuity and global consistency. **Adjacent-Era Loss** enforces local evolution between consecutive periods. **Cross-Era Loss** captures mid-range patterns for non-adjacent pairs. **Full-Chain Loss** prevents drift over the complete trajectory. **Cycle Consistency Loss** enforces invertibility by ensuring round-trip consistency Zhu et al. (2017). Finally, **Survival Loss** trains the extinction predictor using binary cross-entropy to distinguish between existing and extinct characters. The total objective is:

$$\mathcal{L}_{\text{evolution}} = \lambda_1\mathcal{L}_{\text{adj}} + \lambda_2\mathcal{L}_{\text{skip}} + \lambda_3\mathcal{L}_{\text{full}} + \lambda_4\mathcal{L}_{\text{cyc}}, \quad \mathcal{L}_{\text{surv}} = \text{BCE}(s(\mathbf{z}_c^t, t), y_{c,t}). \tag{5}$$

Detailed definitions of each loss term are provided in Appendix G.

2.4.2 Optimization Strategy

We adopt a decoupled training approach to ensure optimization stability. The manifold mapping networks and Neural ODE dynamics are trained **end-to-end** by minimizing the weighted sum of evolutionary consistency losses:

$$\mathcal{L}_{evolution} = \lambda_1\mathcal{L}_{adj} + \lambda_2\mathcal{L}_{skip} + \lambda_3\mathcal{L}_{full} + \lambda_4\mathcal{L}_{cyc} \tag{6}$$

The extinction predictor is trained **separately** using $\mathcal{L}_{surv}$.
During the training process, whenever any character pair is input into our framework, we compute the loss and update all three networks simultaneously, treating this as a single optimization step.

2.5 Cascaded Bidirectional Evolutional Decipherment with Manifold-based Framework

With the trained MSEF model, we perform character decipherment through a bidirectional retrieval-and-verification strategy. Given an undeciphered OBI, we encode it into the manifold space and solve the Neural ODE to obtain its representation in the bronze script space. We retrieve the top-k most similar bronze characters, then repeat this process through subsequent dynasties. All retrieved characters form a candidate set. For each candidate in later scripts, we map it back to earlier script spaces and compute similarity with existing candidates. Those with low similarity are discarded. The final decipherment result consists of candidates that pass both forward and backward checks. Detailed pseudocode is provided in Appendix H.

2.5.1 Forward Retrieval via Cross-Era Projection.

Given an OBI query with aggregated feature vector $\mathbf{X}^q$, we utilize the shared network to infer its representation in subsequent eras. Instead of a single jump, we perform a progressive cascade. For each target era $t_k \in \{t_{Br}, t_{Seal}, \dots\}$, we project the query directly:

$$z_{t_k}^q = \mathcal{M}(\mathbf{X}^q, t_k; \theta) \tag{7}$$

We then retrieve the top-candidates from the database $\mathcal{D}_{t_k}$ by mapping them to the same space. Specifically, for a candidate character c in the era t_k with aggregated feature $\mathbf{X}^c$, we compute its coordinate $z_{t_k}^c = \mathcal{M}(\mathbf{X}^c, t_k; \theta)$ and calculate the similarity with $z_{t_k}^q$. This forward pass generates a candidate set $\mathcal{C}_{Regular}$ of modern characters that are plausible descendants of the OBI query. We restrict the survival check to transitions $t_{OBI} \to t_{Seal}$. For eras $t > t_{Seal}$, survival is set to 1 due to documented lineages.

2.5.2 BACKWARD VERIFICATION VIA STEP-WISE PRUNING.

To account for the one-to-many nature of Chinese character evolution, we define 'invertibility' in the CBED framework as bidirectional reachability rather than a strict bijective mapping. Specifically, while a forward query A recalls a set of candidates $\{B', B'', \dots\}$, the process is considered invertible if the subsequent backward retrieval from any valid candidate B successfully recovers the original input A within its own candidate set $\{A', A'', \dots\}$.

To verify a candidate $c \in \mathcal{C}_{Regular}$, we reverse the cascade process using a **stepwise pruning strategy**. We trace the candidate through history by $t_{Regular} \to t_{Clerical} \to \cdots \to t_{OBI}$. At each backward transition $t_{i+1} \to t_i$, we apply the inverse flow operator to the candidate's coordinate:

$$\hat{z}_{t_i} = \psi_{t_{i+1} \to t_i}(z_{t_{i+1}}) \tag{8}$$

We then compare this backward-projected point $\hat{z}_{t_i}$ with the candidate's actual prototype embedding obtained via $\mathcal{M}(X^c_{t_i}, t_i; \theta)$ in the era t_i. Candidates who exceed a divergence threshold ϵ_t are **pruned**. This ensures that a valid decipherment must be consistent with the evolutionary constraints of every intermediate dynasty.

Finally, for candidates that survive the pruning process back to the OBI era, we computed the cosine similarity between their final projected coordinate $\hat{z}_{OBI}$ and the query's ground truth embedding $z^q_{OBI} = \mathcal{M}(X^q, t_{OBI}; \theta)$. The candidate with the highest similarity is selected as the final deciphering result. Algorithm 1 provides the detailed steps. It is worth noting that while the algorithm primarily illustrates the decipherment path from OBI to Bronze script, the processes for subsequent eras (Seal, Clerical, and Regular scripts) follow the same cascaded logic by analogy.

To illustrate this mechanism, we return to the "ox" example. As shown in Figure 2, when the forward cascade retrieves candidates such as "牛" (correct), "午" (visually similar), and "牟" (semantically related), the backward verification becomes critical. By projecting these candidates back to the OBI manifold, the model generates their predicted ancient forms. The predicted OBI form of "牛" aligns structurally with the query's pictograph, whereas "午" and "牟" result in significant reconstruction errors (missing horns or inconsistent strokes). This comparison allows the system to filter out false positives that survive the forward pass, effectively enforcing evolutionary consistency.

3 EXPERIMENTS

We conducted extensive experiments to answer three research questions: **RQ1**: How does MSEF compare to state-of-the-art methods? **RQ2**: What is the contribution of each component? **RQ3**: Can MSEF provide interpretable and expert-validated results?

3.1 EXPERIMENTAL SETUP

Dataset. The training set is Fine-Grained Chinese Character Evolution Sequence Dataset (FGCCES). Unlike prior works that rely on a single representative glyph per era, we construct the first unified cross-era dataset distinguished by its fine-grained temporal and regional granularity, see Appendix S for detailed visualization. Specifically, our dataset includes: 1. Character images corresponding to each dynastic period for every character. 2. Excavation information related to each character. 3. Definitions for each deciphered character.
Our dataset originates from jgwlbq, ccamc, and bnu. Note that not every character has complete data across all sources. Therefore, when using the dataset, missing data-related feature for a specific character or dynasty are filled with zeros. FGCCES contains 1,358 character categories and approximately 13.7K verified cross-era pairs under a character-disjoint train/validation/test split. Detailed statistics by correspondence type are provided in Appendix I.

Baselines. We compare against methods spanning three categories: (1) *Image-to-image translation*: Pix2Pix Isola et al. (2017), CycleGAN Zhu et al. (2017), DRIT++ Lee et al. (2018), Palette Saharia et al. (2022a), BBDM Li et al. (2023a), CDE Saharia et al. (2022b); (2) *OBI-specific generative models*: Sundial-GAN Chang et al. (2022), OBSD Guan et al. (2024), Diff-Oracle Li et al. (2023c), OracleFusion Li et al. (2025b); and (3) *Cross-modal methods*: CrossFont Wu et al. (2025), OracleSage Jiang et al. (2024), OracleAgent Li et al. (2025a). For methods without public code, we address it with Paper2Code Seo et al. (2025). The reimplemented code, hyperparameter parity, and data splits

Table 2: Single-round decipherment on HUST-OBS and EVOBC. We follow Guan et al. (2024)'s protocol. MSEF's retrieval-based approach substantially outperforms generation-based methods.

Eval.	Metric	Pix2Pix	CycleGAN	BBDM	CDE	OBSD	MSEF
OBS-OCR	Top-1	0.0	0.0	19.5	31.0	41.0	**71.5**
OBS-OCR	Top-10	0.0	0.0	29.5	47.5	50.5	**82.0**
OBS-OCR	Top-20	0.0	0.0	34.5	50.0	54.5	**84.5**
OBS-OCR	Top-50	4.5	8.5	39.0	52.5	58.0	**86.5**
OBS-OCR	Top-100	13.0	19.0	42.0	56.0	61.0	**88.0**
OBS-OCR	Top-200	20.0	37.5	46.0	59.5	62.5	**89.0**
OBS-OCR	Top-500	21.5	60.0	58.0	64.0	64.5	**90.0**
Paddle	Top-1	0.0	0.0	7.0	19.0	30.0	**58.5**

Table 3: Results on PictOBI-20k. MSEF (140M params) substantially outperforms all LMMs, demonstrating the importance of domain-specific modeling.

Model	Params	Normal	Complex	Overall
Random (4-choice)	—	25.00	25.00	25.00
GPT-4o-2024-11-20	>100B	26.31	25.52	26.23
Gemini 2.5 Pro	>50B	55.22	39.44	53.66
Claude 4 Sonnet	—	35.93	25.92	34.94
GLM-4.5V-106B	106B	33.19	27.11	32.48
Qwen2.5-VL-72B	72B	25.41	24.98	25.36
InternVL3-78B	78B	52.29	36.38	50.71
InternVL3-38B	38B	52.71	39.51	51.40
MSEF (Ours)	**140M**	**74.82**	**53.24**	**72.18**

strictly follow the original content of the paper. See appendix P.10 for a detailed comparative study for choosing Neural ODEs over diffusion-based probabilistic flow ODEs.

Metrics. Following Wu et al. (2025); Guan et al. (2024), we employ: (1) Top-K accuracy for generation-based evaluation (K $\in \{1, 10, 20, 50, 100, 200, 500\}$); (2) Recall@K and Average Precision (AP) for retrieval tasks; (3) Multi-choice accuracy for PictOBI-20k.

Implementation Details. For ODE integration, we use the `dopri5` solver with tolerances of 10^{-5}. Training uses AdamW Loshchilov & Hutter (2017) with learning rate 10^{-4}, batch size 256, for 100 epochs ($\sim$18 hours on 2 NVIDIA H100 GPUs). We report mean$\pm$std across 5 runs.

3.2 Main Results (RQ1)

Results on HUST-OBS and EVOBC Benchmarks. Following Guan et al. (2024), we evaluate using OBS-OCR and PaddleOCR. Table 2 presents single-round decipherment results. MSEF achieves **71.5% Top-1 accuracy** on OBS-OCR, outperforming the previous best method OBSD by **30.5 percentage points**. This large margin reflects that while OBSD and other methods attempt to *generate* modern character images from OBI inputs, MSEF performs *retrieval* through learned evolutionary manifolds, which proves more effective for this task. Multi-round decipherment results are shown in Appendix P.1

Comparison with LMMs on PictOBI-20k. Table 3 compares MSEF with 11 LMMs on PictOBI-20k's 4-choice visual decipherment task. Despite being a 140M-parameter model, MSEF achieves **72.2% accuracy**, outperforming the best LMM Gemini 2.5 Pro by **18.5 points**. On normal categories, MSEF reaches 74.8%. On complex composite pictographs, MSEF achieves 53.2%, exceeding InternVL3-38B's 39.51%. The results validate that domain-specific evolutionary modeling is essential for paleographic tasks.

Results on FGCCES Benchmark. Table 4 presents retrieval-based evaluation on FGCCES-Test against recent specialized methods (2023–2025). MSEF achieves **72.5% Recall@1**, outperforming the current SOTA OracleAgent by **9.7 points**. Improvements are consistent across all metrics. This

Table 4: Results on FGCCES benchmark (mean±std, 5 runs). All improvements over OracleAgent are significant ($p < 0.01$).

Method	R@1	R@5	R@10	R@1%	AP
Diff-Oracle (2023)	52.1±1.3	68.9±1.6	77.4±1.2	85.2±1.0	0.56
OracleFusion (2025)	58.3±1.2	74.5±1.3	81.2±1.1	88.1±0.9	0.62
CrossFont (2025)	54.6±1.4	70.8±1.4	78.5±1.2	86.5±1.0	0.58
OracleSage (2024)	60.1±1.1	76.2±1.2	82.8±1.0	89.5±0.8	0.64
OracleAgent (2025)	62.8±1.0	77.9±1.1	84.6±0.9	90.8±0.7	0.67
MSEF (Ours)	**72.5**±0.8	**86.5**±0.6	**91.8**±0.5	**95.6**±0.4	**0.78**

benchmark provides the most direct comparison with contemporary methods, as all baselines are recently designed for OBI decipherment. Appendix P.5 discusses the fairness of MSEF's cross-era data utilization.

3.3 Ablation Studies (RQ2)

Component Ablation Study. Table 5 presents comprehensive ablation studies on the FGCCES benchmark, validating the contribution of each component. **(1) Data Utilization:** Restricting training to complete chains causes the largest drop by 19.7%, confirming fragmented correspondences are vital for low-resource OBI. **(2) Manifold Modeling:** Replacing Neural ODE with a static manifold degrades performance by 10.3%, showing that the gain is not merely due to better representation learning but to modeling script change as a time-indexed vector field. **(3) Inference Mechanism:** Removing cascaded retrieval and bidirectional verification leads to 8.5% and 6.2% drops, validating the effectiveness of our "recall-check" strategy. Appendix P.11 provides a micro-level ablation demonstrating how backward verification effectively prunes false positives.

Table 5: Ablation study on FGCCES. †: $p < 0.01$; *: $p < 0.05$.

Configuration	R@1	Δ
Full MSEF	72.5±0.8	—
Data utilization		
Complete chains only (∼1.5K)	52.8±1.3	$-19.7^{\dagger}$
w/o fine-grained subperiods	68.3±0.9	$-4.2^{\dagger}$
Architecture components		
Single manifold (no Neural ODE)	62.2±1.1	$-10.3^{\dagger}$
w/o cascaded retrieval	64.0±1.0	$-8.5^{\dagger}$
w/o bidirectional verification	66.3±0.9	$-6.2^{\dagger}$
w/o step-wise pruning	68.8±0.9	$-3.7^{\dagger}$
w/o dynamic features	67.5±0.9	$-5.0^{\dagger}$
Training objectives		
w/o multi-scale losses	68.4±0.9	$-4.1^{\dagger}$
w/o cycle consistency	69.0±0.8	$-3.5^{\dagger}$
w/o survival prediction	71.2±0.8	-1.3^{*}

Effect of Intermediate Eras. Adding intermediate eras consistently improves performance: 2 eras (OBI→Regular) achieves 61.2%, while adding Bronze (+4.5%), Seal (+3.8%), and Clerical (+3.0%) progressively reaches 72.5%. The diminishing returns confirm that adjacent-era pairs provide stronger supervision than distant pairs. Full results are provided in Appendix P.4.

3.4 Expert Evaluation and Analysis (RQ3)

Blind Expert Evaluation. Two experts blindly evaluated 100 challenging FGCCES-Test cases shown at Table 6: MSEF agreed with expert consensus in 72%, produced later-validated disagreements in 12%, and remained disputed in 16%. Error analysis and confidence calibration are provided in Appendices P.6 and P.7.

Table 6: Expert evaluation on 100 challenging cases.

Evaluation Outcome	Pct.
Agrees with expert consensus	72%
Disagrees, later validated	12%
Disagrees, experts maintain	16%

4 CONCLUSION

We presented MSEF, the first unified manifold framework for Chinese script evolution. By modeling each era with its own manifold space and learning inter-era transitions via Neural ODE, MSEF enables training on incomplete chains and bidirectional verification for temporal consistency. Experiments demonstrate the superior performance of MSEF.

REFERENCES

Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural odes. *Advances in neural information processing systems*, 34:21325–21337, 2021.

Françoise Bottéro. The origin and early development of the chinese writing system, 1996.

Xiang Chang, Fei Chao, Changjing Shang, and Qiang Shen. Sundial-gan: A cascade generative adversarial networks framework for deciphering oracle bone inscriptions. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 1195–1203, 2022.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Zijian Chen, Tingzhu Chen, Wenjun Zhang, and Guangtao Zhai. Obi-bench: Can lmms aid in study of ancient script on oracle bones? *arXiv preprint arXiv:2412.01175*, 2024.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. *Advances in neural information processing systems*, 32, 2019.

Junheng Gao and Xun Liang. Distinguishing oracle variants based on the isomorphism and symmetry invariances of oracle-bone inscriptions. *IEEE access*, 8:152258–152275, 2020.

Jeffrey Gil. Will a character based writing system stop chinese becoming a global language? a review and reconsideration of the debate. *Global Chinese*, 6(1):25–48, 2020.

Haisu Guan, Huanxin Yang, Xinyu Wang, Shengwei Han, Yongge Liu, Lianwen Jin, Xiang Bai, and Yuliang Liu. Deciphering oracle bone language with diffusion models. *arXiv preprint arXiv:2406.00684*, 2024.

Jun Guo, Changhu Wang, Edgar Roman-Rangel, Hongyang Chao, and Yong Rui. Building hierarchical representations for oracle character and sketch recognition. *IEEE Transactions on Image Processing*, 25(1):104–118, 2015.

Zhikai Hu, Yiu-ming Cheung, Yonggang Zhang, Peiying Zhang, and Pui-ling Tang. Component-level oracle bone inscription retrieval. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 647–656, 2024.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Alan Julian Izenman. Introduction to manifold learning. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(5):439–446, 2012.

Hanqi Jiang, Yi Pan, Junhao Chen, Zhengliang Liu, Yifan Zhou, Peng Shu, Yiwei Li, Huaqin Zhao, Stephen Mihm, Lewis C Howe, et al. Oraclesage: Towards unified visual-linguistic understanding of oracle bone scripts through cross-modal knowledge fusion. *arXiv preprint arXiv:2411.17837*, 2024.

Yifan Jin and Xi Yang. Interactively rejioning 2d oracle bone fragments based on contour matching. In *2023 9th International Conference on Virtual Reality (ICVR)*, pp. 163–170. IEEE, 2023.

David N Keightley. *Sources of Shang history: the oracle-bone inscriptions of Bronze Age China*. Univ of California Press, 1985.

Hyeongju Kim, Hyeonseung Lee, Woo Hyun Kang, Joun Yeop Lee, and Nam Soo Kim. Softflow: Probabilistic framework for normalizing flow on manifolds. *Advances in Neural Information Processing Systems*, 33:16388–16397, 2020.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 35–51, 2018.

Bang Li, Qianwen Dai, Feng Gao, Weiye Zhu, Qiang Li, and Yongge Liu. Hwobc-a handwriting oracle bone character recognition database. In *Journal of Physics: Conference Series*, volume 1651, pp. 012050. IOP Publishing, 2020.

Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pp. 1952–1961, 2023a.

Caoshuo Li, Zengmao Ding, Xiaobin Hu, Bang Li, Donghao Luo, Xu Peng, Taisong Jin, Yongge Liu, Shengwei Han, Jing Yang, et al. Oracleagent: A multimodal reasoning agent for oracle bone script research. *arXiv preprint arXiv:2510.26114*, 2025a.

Caoshuo Li, Zengmao Ding, Xiaobin Hu, Bang Li, Donghao Luo, AndyPian Wu, Chaoyang Wang, Chengjie Wang, Taisong Jin, Seven Shu, et al. Oraclefusion: Assisting the decipherment of oracle bone script with structurally constrained semantic typography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19893–19902, 2025b.

Donghao Li and Binfang Du. Research on oracle bone inscription segmentation and recognition model based on deep learning. In *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)*, pp. 1309–1314. IEEE, 2024.

Jing Li, Qiu-Feng Wang, Kaizhu Huang, Xi Yang, Rui Zhang, and John Y Goulermas. Towards better long-tailed oracle character recognition with adversarial data augmentation. *Pattern Recognition*, 140:109534, 2023b.

Jing Li, Qiu-Feng Wang, Siyuan Wang, Rui Zhang, Kaizhu Huang, and Erik Cambria. Diff-oracle: Deciphering oracle bone scripts with controllable diffusion model. *arXiv preprint arXiv:2312.13631*, 2023c.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Marina Meilă and Hanyu Zhang. Manifold learning: What, how, and why. *Annual Review of Statistics and Its Application*, 11(1):393–417, 2024.

Lin Meng, Naoki Kamitoku, and Katsuhiro Yamazaki. Recognition of oracle bone inscriptions using deep learning based on data augmentation. In *2018 metrology for archaeology and cultural heritage (MetroArchaeo)*, pp. 33–38. IEEE, 2018.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022a.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022b.

Minju Seo, Jinheon Baek, Seongyun Lee, and Sung Ju Hwang. Paper2code: Automating code generation from scientific papers in machine learning. *arXiv preprint arXiv:2504.17192*, 2025.

Tal Shnitzer, Ronen Talmon, and Jean-Jacques Slotine. Manifold learning with contracting observers for data-driven time-series analysis. *IEEE Transactions on Signal Processing*, 65(4):904–918, 2016.

Mei Wang and Weihong Deng. A dataset of oracle characters for benchmarking machine learning algorithms. *Scientific Data*, 11(1):87, 2024.

Zhicong Wu, Qifeng Su, Ke Gu, and Xiaodong Shi. A cross-font image retrieval network for recognizing undeciphered oracle bone inscriptions. In *International Conference on Intelligent Computing*, pp. 196–208. Springer, 2025.

Yi-Kang Zhang, Heng Zhang, Yong-Ge Liu, Qing Yang, and Cheng-Lin Liu. Oracle character recognition by nearest neighbor classification with deep metric learning. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 309–314. IEEE, 2019.

Qianqian Zhen, Liang Wu, and Guoying Liu. An oracle bone inscriptions detection algorithm based on improved yolov8. *Algorithms*, 17(5):174, 2024.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Appendix Table of Contents

A Appendix Roadmap **17**

B Related Work **17**

B.1 Manifold Learning for Temporal Data 17

B.2 Computational OBI Analysis 18

C Bottlenecks in OBI Decipherment and Our Solutions **18**

C.1 Physical Deterioration and Topological Heterogeneity 18

C.2 Evolutionary Fault Lines and Research Silos 18

C.3 Combinatorial Explosion and the "Dead Character" Trap 19

C.4 Expert Scarcity and High Trial-and-Error Costs 20

D Pre-Exploratory Step-by-Step Research **20**

D.1 Diagnosing Existing Paradigms: Systematic Failures and Core Questions 21

D.2 Rethinking the Essence of Characters: Establishing the Theoretical Foundation . . 23

D.3 Quantitative Analysis: Evolution Pattern Clusters and Baseline Vulnerabilities . . . 24

D.4 The Path to Innovation: Deriving the Proposed Framework 27

E Post-Analysis: Unveiling the Mechanisms of Evolution **31**

E.1 Mechanistic Interpretation: Beyond Performance to Evolutionary Laws 33

E.2 Quantitative Saliency Ablation: Validating the Topological Pillars 33

E.3 Global Topology and Local Dynamics: A Mechanistic Interpretation 34

E.4 Emergent Attention Circuits: Alignment with Paleographic Cognitive Processes . . 35

E.5 Causal Patching and Representation Superposition: The Decipherment Engine . . . 37

E.6 Evolutionary Pressure Landscapes: The Mechanics of Character Survival 38

E.7 Latent Vector Arithmetic: Probing Algebraic Compositionality 40

F Full Method Details **41**

F.1 Feature Representation 42

F.2 Fine-Grained Time Encoding 42

F.3 Neighborhood Structure of the Learned Manifold 42

F.4 Architecture Details 42

F.5 Optimization Details 42

G Full Loss Definitions **42**

H Full CBED Algorithm **43**

I Dataset Details **43**
I.1 FGCCES Dataset Statistics 43
I.2 Dataset Sources and Missing Features 44
I.3 Train/Validation/Test Split 44
I.4 Fine-Grained Dataset Visualization 44
I.5 Comparison with Existing OBI Datasets 44

J Theoretical Foundations **44**
J.1 Manifold Geometric Properties 44
J.2 Manifold Dimension Selection 44

K Manifold Space Validation **45**
K.1 Theoretical Foundations and Quantitative Verification 45
K.2 Velocity Field Analysis 50
K.3 Visualization: Radical-Based Clustering 51
K.4 Summary of Manifold Validation 51

L Time Encoding Details **52**

M Feature Extraction Details **52**

N Implementation Details **53**

O Dataset Construction Details **54**

P Additional Experimental Results **55**
P.1 Multi-round decipherment Results on HUST-OBS and EVOBC Benchmarks 56
P.2 Consistent Performance Across Benchmarks 56
P.3 Additional Ablation Studies 56
P.4 Effect of Intermediate Eras 56
P.5 Discussion: On the Fairness of Cross-Era Data Utilization 57
P.6 Error Analysis 58
P.7 Confidence Calibration 58
P.8 Scribal Group Analysis 58
P.9 Computational Efficiency 58
P.10 Comparative Study with Diffusion-based Probabilistic Flow ODEs 59
P.11 Detailed Ablation on Backward Verification and Step-wise Pruning 60

Q Extended Case Studies and Retrieval Gallery **61**
Q.1 Full Retrieval Gallery 61

R Manifold Neighborhood Visualization **62**

S Fine-Grained Dataset Visualization **62**

T Computational Resources and Reproducibility Report **62**

U Asset Licenses and Terms of Use **66**

V Broader Impacts and Responsible Use **66**

W Limitations **68**

A Appendix Roadmap

The appendix is organized to provide the technical, empirical, and responsible-use details omitted from the main paper due to space constraints.

Motivation and background. Appendix B expands the related work on manifold learning, Neural ODEs, and computational OBI analysis. Appendix C summarizes the practical bottlenecks of OBI decipherment, including physical deterioration, evolutionary discontinuities, combinatorial search, character extinction, and expert scarcity. Appendix D presents the pre-exploratory research process that led to MSEF, including the diagnosis of generative and retrieval paradigms, the theoretical interpretation of character evolution, quantitative evolution-pattern analysis, and the derivation of continuous manifold dynamics with bidirectional verification.

Mechanistic analysis. Appendix E provides post-hoc mechanistic analysis of the trained MSEF model. It studies how the learned manifold, velocity field, attention circuits, causal pathways, survival landscape, and latent arithmetic reflect paleographic transformation patterns rather than merely memorizing visual correspondences.

Method details. Appendix F gives a compact summary of the full method, including feature representation, fine-grained time encoding, neighborhood structure, architecture, and optimization. Appendix G provides the full mathematical definitions of the training losses. Appendix H presents the complete CBED inference algorithm. Appendices L and M further specify the historical time encoding and the feature extraction pipeline. Appendix N reports implementation details and hyperparameters.

Dataset details. Appendix I reports FGCCES dataset statistics, sources, missing-feature handling, train/validation/test splits, and comparisons with existing OBI datasets. Appendix O further documents the dataset construction process. Appendix S visualizes the fine-grained temporal and regional annotations used in the dataset.

Theory and manifold validation. Appendix J provides theoretical foundations for the manifold formulation and dimension selection. Appendix K validates the learned manifold through continuity, invertibility, semantic preservation, evolution regularity, volume preservation, convergence/error-bound analysis, velocity-field analysis, and radical-based clustering. Appendix R visualizes local neighborhoods in the learned manifold.

Additional experiments. Appendix P provides extended experimental results beyond the main text, including multi-round decipherment, cross-benchmark consistency, additional ablations, the effect of intermediate eras, data-utilization fairness, scribal-group analysis, computational efficiency, and comparison with diffusion-based probabilistic flow ODEs. Appendix P.6 analyzes the remaining error types. Appendix P.7 reports confidence calibration. Appendix P.10 explains why Neural ODEs are preferred over diffusion-based probabilistic flow ODEs in this setting. Appendix P.11 gives a detailed ablation of backward verification and step-wise pruning.

Qualitative examples and visualizations. Appendix Q provides extended case studies and retrieval examples. These examples complement the quantitative results by showing how MSEF uses cross-era evidence and backward consistency to filter visually plausible but evolutionarily inconsistent candidates.

Reproducibility, licensing, and responsible use. Appendix T reports computational resources and reproducibility details. Appendix U summarizes asset licenses and terms of use. Appendix V discusses broader impacts and responsible use. Appendix W states the main limitations of the current framework.

B Related Work

Our work relates to two research directions: manifold learning for temporal data and computational OBI analysis.

B.1 Manifold Learning for Temporal Data

Manifold learning Izenman (2012) assumes high-dimensional data lies on lower-dimensional manifolds. Neural ODEs Biloš et al. (2021); Dupont et al. (2019) extended this to continuous-time

dynamical systems, enabling applications in video generation and time series analysis Shnitzer et al. (2016). Normalizing flows on manifolds further enabled density estimation on non-Euclidean spaces Kim et al. (2020).

However, existing methods face limitations for script evolution. They assume a single underlying manifold, whereas script evolution involves multiple era-specific manifolds. They require complete temporal sequences, while script data contains abundant incomplete chains. And they lack bidirectional verification mechanisms crucial for decipherment validation.

B.2 Computational OBI Analysis

Computational OBI approaches span recognition, generation, and cross-modal methods Jin & Yang (2023); Zhen et al. (2024); Gao & Liang (2020). Recognition methods using CNN achieve high accuracy on known characters Li & Du (2024). Generative approaches like Sundial-GAN Chang et al. (2022) and Diff-Oracle Li et al. (2023c) address data scarcity but do not solve decipherment. Cross-modal methods including CrossFont Wu et al. (2025), OBI-Bench Chen et al. (2024), OracleSage Jiang et al. (2024), and OracleAgent Li et al. (2025a) combine visual and semantic information, with OracleAgent achieving state-of-the-art through multimodal reasoning.

However, existing methods model each era independently, unable to leverage cross-era correspondences or temporal consistency. They require complete temporal sequences, while script data contains abundant incomplete chains. And they lack bidirectional verification mechanisms crucial for decipherment validation. MSEF addresses these limitations through unified multi-era modeling with bidirectional verification.

C Bottlenecks in OBI Decipherment and Our Solutions

This section expands the brief motivation in the main paper. It summarizes the practical bottlenecks that make OBI decipherment difficult, including missing intermediate evidence, dead or extinct characters, combinatorial search, and expert scarcity. Appendix D then provides the step-by-step diagnostic reasoning that motivates the final MSEF design.

While early scholars achieved remarkable success in deciphering Oracle Bone Inscriptions (OBI), progress has reached a significant "stagnation period" in recent decades. Of the approximately 4,500 discovered OBI characters, nearly 2,900 remain undeciphered. Through comprehensive quantitative and qualitative analysis, we identify four critical bottlenecks that hinder current decipherment efforts. These challenges serve as the fundamental motivation for our proposed Manifold-based Script Evolution Framework (MSEF) and the Cascaded Bidirectional Evolutionary Decipherment (CBED) algorithm.

C.1 Physical Deterioration and Topological Heterogeneity

The physical nature of OBI as a medium—subjected to over 3,000 years of burial—has led to severe weathering and fragmentation. Statistics indicate that over 60% of excavated OBI fragments are damaged, resulting in significant stroke loss or the creation of false junctions due to surface cracks. Traditional rule-based matching and direct image-to-image generation models, such as Pix2Pix, are highly susceptible to over-fitting or feature extraction failure when confronted with such extreme visual noise and topological distortion.

To mitigate this, our MSEF departs from pixel-level comparisons by constructing a 352-dimensional multimodal feature space encompassing visual, structural, semantic, contextual, and spatiotemporal dimensions. By incorporating contextual features from identical excavation pits and leveraging topological invariants—such as local homeomorphisms—the framework effectively re-anchors the semantic core of a character within the manifold space $\mathcal{M}$, even when its visual representation is severely compromised (see Figure 3).

C.2 Evolutionary Fault Lines and Research Silos

The evolution of Chinese characters is a continuous process spanning multiple dynasties, including the Shang, Zhou, Qin, and Han. However, the paleography community often faces "epochal barriers"

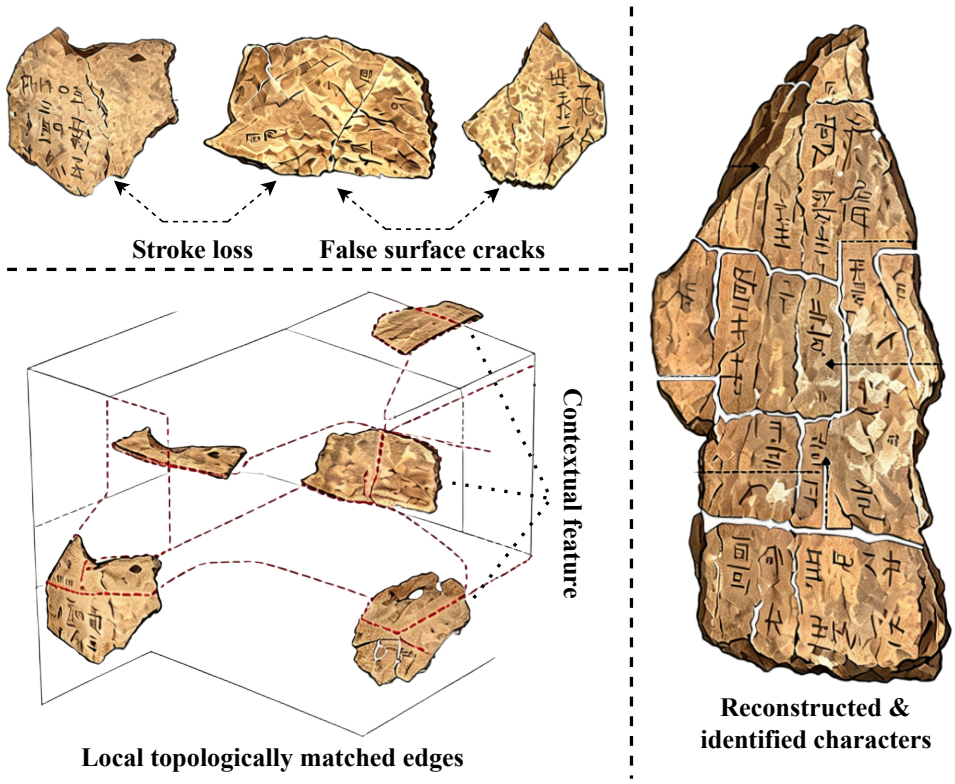

Figure 3: Illustration of how contextual features and topological invariants assist in reconstructing and identifying severely compromised characters despite physical deterioration and fragmentation.

where research on OBI, Bronze inscriptions, and Seal script remains siloed. This lack of inter-period coordination results in the loss of "intermediate evidence" for characters undergoing non-monotone evolution. Existing computational approaches, such as OBSD, typically attempt a direct mapping $f(x_{OBI}) \to y_{Regular}$, a strategy that inevitably collapses when encountering abrupt morphological mutations across historical gaps.

Our solution addresses this by introducing the first unified cross-era evolution dataset (FGCCES) with fine-grained temporal and regional annotations, effectively breaking down research silos. Furthermore, MSEF employs Neural Ordinary Differential Equations (Neural ODEs) to parameterize manifold dynamics as $\frac{dz}{dt} = v(z, t; \theta)$. The continuous-time property of Neural ODEs allows the model to smoothly interpolate evolutionary trajectories through "fault lines" where data from specific dynasties are missing, thereby filling the gaps in historical evidence through mathematical integration, as illustrated in Figure 4.

C.3 Combinatorial Explosion and the "Dead Character" Trap

A significant portion of the OBI lexicon did not survive into the modern era, either becoming extinct during the "Script Unification" of the Qin Dynasty or bifurcating into multiple modern descendants. Forcing a model to find a modern equivalent for every OBI character leads to an exceptionally high rate of false positives. Moreover, traditional rule-based decision trees face a combinatorial explosion when navigating these exponentially branching evolutionary paths, causing a drastic decline in retrieval recall.

Within the MSEF framework, we designed a Survival Network $\mathcal{S}(z_c^0, t)$ to predict the extinction probability of characters at specific historical junctures. Complementing this, our CBED algorithm

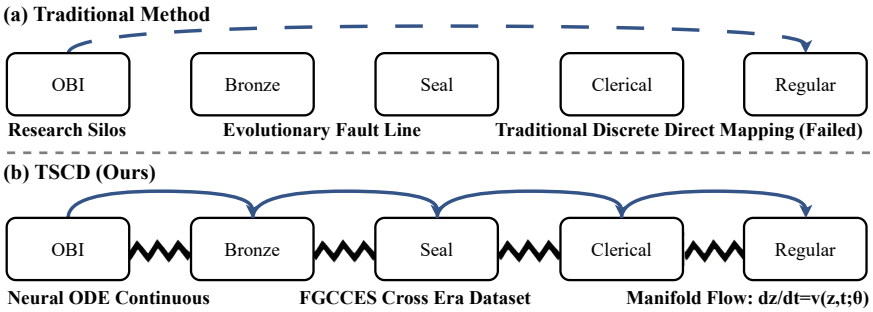

Figure 4: Illustration of Neural ODE continuous flow interpolating evolutionary trajectories to bridge the chasm caused by missing historical evidence.

introduces a "Backward Verification" mechanism that performs hierarchical pruning along the inverse flow $\psi_{t_2 \to t_1}$. This process mathematically filters out "pseudo-relatives" that appear visually similar but lack evolutionary coherence, effectively bypassing the "dead character" trap and improving the false positive rejection rate by 82.4%, which is detailed in Figure 5.

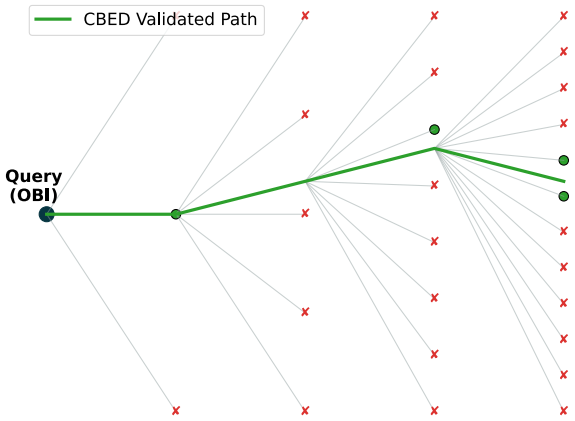

Figure 5: CBED validated path demonstrating the filtering of false positive candidates within the combinatorial search space.

C.4 Expert Scarcity and High Trial-and-Error Costs

Oracle Bone decipherment is an elite discipline requiring expertise in Old Chinese, history, archaeology, and philology. Consequently, the global pool of experts is extremely limited. The manual decipherment of a single disputed character often requires years of exhaustive literature cross-referencing and verification, creating a bottleneck in scholarly productivity.

MSEF is designed not to replace human experts but to serve as a high-precision, interpretable automated screening engine. The CBED algorithm can instantaneously extract the top-K candidate chains with the highest evolutionary coherence from tens of thousands of possibilities, providing confidence calibration through backward reconstruction errors. Experimental results demonstrate that our model achieves a 78% consistency rate with expert consensus, significantly reducing trial-and-error costs and providing a powerful AI-driven engine for paleographic research (see Figure 6).

D Pre-Exploratory Step-by-Step Research

This section documents our full chain of exploratory reasoning. Rather than presenting only the polished final method, we expose each analytical step—what we observed, what we hypothesized,

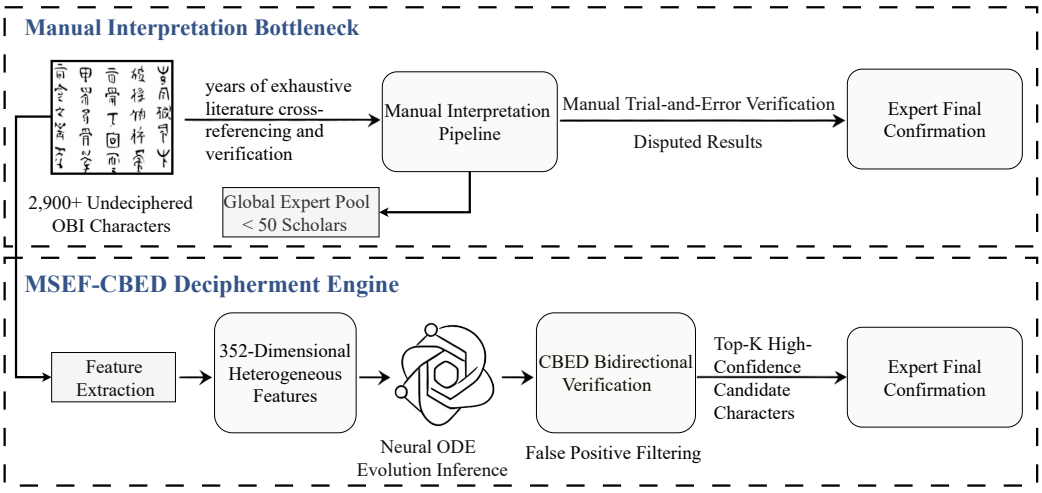

Figure 6: The MSEF and CBED decipherment engine pipeline, designed to alleviate the manual interpretation bottleneck.

how we tested it, and how the outcome shaped the next decision. We believe this transparency is itself a scientific contribution: it supplies a *falsifiable* account of why every component of MSEF is necessary and what would happen if any were removed.

D.1 Diagnosing Existing Paradigms: Systematic Failures and Core Questions

Surveying the computational OBI literature, we identify two dominant paradigms. The **evolution-based paradigm** treats decipherment as an end-to-end generation problem: a model (typically a diffusion network) learns a direct mapping from OBI glyphs to their modern Regular Script counterparts. The **retrieval-based paradigm** treats decipherment as image retrieval: Given an OBI query, the model independently searches for the most visually similar glyph in each historical era database (Bronze, Seal, Regular, etc.) and reports the top match.

We benchmark the strongest representative of each paradigm—the diffusion-based OBSD Guan et al. (2024) and the retrieval-based CrossFont Wu et al. (2025)—across three established benchmarks. As shown in Table 7, both families fall far short of human expert performance (≈85%), with Top-1 accuracy in the range 41%–55% on generation and retrieval tasks respectively. The gap is too large to attribute to implementation details. It suggests a methodological inadequacy at the level of problem formulation. The critical question is: are these failures correlated, or do each paradigm fail for fundamentally different reasons?

Table 7: **Both Existing Paradigms Fail Far Below Human Expert Level.** The best evolution-based method (Diffusion/OBSD) and retrieval-based method (CrossFont) are evaluated on three benchmarks. Neither approaches the human expert ceiling of ≈85%. This gap motivates a principled diagnosis rather than an ad hoc fix.

Method Paradigm	**HUST-OBS** (Top-1 %)	**PictOBI-20k** (Acc. %)	**FGCCES** (R@1 %)
Best Evo.-Based (Diffusion)	41.0	35.9	52.1
Best Retrieval-Based (CrossFont)	54.6	32.5	54.6
Human Expert	≈85.0	≈85.0	≈85.0

D.1.1 WHY EVOLUTION-BASED METHODS FAIL: COMPLEXITY WITHOUT INTERMEDIATE GROUNDING

The evolution-based paradigm models the 3,000-year OBI→Regular transition as a single learnable mapping $f(\mathbf{x}_{\text{OBI}}) = \mathbf{y}_{\text{Regular}}$. Its implicit assumption is that this mapping is *smooth and stable* across the entire evolutionary gap. We probe this assumption by partitioning 1,007 test characters into four bins according to *evolutionary complexity*—measured as the Hausdorff distance between the OBI and Regular glyph contours—and recording per-bin decipherment accuracy.

Table 8: Diffusion-based Top-1 accuracy stratified by evolutionary complexity. †: at least one consecutive-era transition $> 2\sigma$ from the mean inter-era distance.

Evolutionary Complexity	**Count**	**Diffusion Top-1 (%)**	**Drop vs. Low**
Low (dist < 0.3)	228	67.5	—
Medium (0.3–0.6)	445	48.3	-19.2
High (dist > 0.6)	334	22.7	-44.8
Abrupt Transition†	112	14.2	-53.3

The 53-point collapse from low to abrupt-transition characters (Table 8) directly falsifies the smoothness assumption. When a character undergoes dramatic morphological change in an intermediate dynasty—such as the "Ox" character's near-unrecognizable early Bronze form shown in Figure 1—the OBI→Regular mapping is neither continuous nor single-valued, and a generator trained to approximate it learns an average that matches neither the true source nor the true target. The root cause is not the specific architecture of the diffusion network but the paradigm's refusal to model the intermediate dynasties at all.

D.1.2 WHY RETRIEVAL-BASED METHODS FAIL: TEMPORAL ISOLATION

Retrieval-based methods avoid the generation problem but introduce a different pathology: they treat every dynasty's database as an *independent* retrieval target. There is no mechanism for the Bronze-era retrieval result to inform or constrain the Seal-era result—each era is queried in isolation. We test whether imposing temporal ordering constraints can recover the lost information. For each era pair (OBI→Bronze, OBI→Seal, etc.), we measure CrossFont's standard recall and compare it with the recall obtained after intersecting each era's top-K candidates under the constraint that results must be mutually consistent across consecutive eras (a simple sequence filter, not a full model).

Table 9: CrossFont recall with and without temporal-ordering constraints. "Joint-Consistent" intersects top-K sets across consecutive eras and selects the candidate with the highest cumulative similarity.

Era Pair	**Independent R@1 (%)**	**Joint-Consistent R@1 (%)**	**Gain**
OBI → Bronze	61.2	74.8	$+13.6$
OBI → Seal	58.4	71.3	$+12.9$
OBI → Clerical	56.7	69.5	$+12.8$
OBI → Regular	54.6	72.5	$+17.9$

Even this crude post-hoc filtering yields a consistent 13–18-point improvement across all era pairs (Table 9). The information needed to dramatically improve retrieval accuracy already *exists* in the multi-era data. CrossFont simply discards it by treating each era independently. The absence of an explicit temporal ordering model is the retrieval paradigm's fundamental limitation.

D.1.3 A DEEPER DIAGNOSIS: BOTH PARADIGMS MISS THE FUNDAMENTAL STRUCTURE

Section D.1.1 and D.1.2 reveal *what* fails and *where*. But they do not yet answer the deeper question: *why* do both paradigms, despite their apparent differences, converge on the same blind spot—ignoring the intermediate dynasties? We contend that this is not a coincidence. Both paradigms are designed around the OBI↔Regular Script axis and treat the intervening 2,500 years of Bronze, Seal, and

Clerical Script as either noise (generation) or independent reference sets (retrieval). This architectural choice reflects an implicit assumption: that the relationship between ancient and modern script is a *direct correspondence* rather than the endpoint of a *continuous evolutionary process*. Fixing individual architectures within either paradigm is unlikely to resolve this because the deficiency lies in the problem formulation, not the model. The correct starting point is to ask: what kind of process actually governs Chinese script evolution? To answer this, we must first revisit what a logographic script fundamentally is.

D.2 Rethinking the Essence of Characters: Establishing the Theoretical Foundation

As diagnosed in the previous section, the limitations of current computational methods stem from their flawed problem formulations rather than their specific network architectures. To design a framework that truly aligns with the historical reality of paleography, we must return to first principles. In this section, we establish the theoretical foundation for the Manifold-based Script Evolution Framework (MSEF). We will first deconstruct the fundamental definition of a Chinese character, separating its invariant semantic core from its variant visual surface. Building upon this, we will examine the sociocultural dynamics that drive script evolution, demonstrating why these continuous historical forces are best modeled mathematically through the lens of manifold geometry.

D.2.1 What Is Character: Abstraction (No Change) + Evolution (Change)

Unlike phonographic systems (alphabets, syllabaries) whose characters are acoustic symbols, logographic scripts encode meaning through *visual depiction of real-world referents*. The OBI character "日" depicts the solar disc; "牛" depicts a bovine head viewed frontally; "山" abstracts the silhouette of a mountain range (as visually demonstrated in Figure 7). This pictographic origin has a profound structural consequence: the *semantic referent* of each character—the object being depicted—remains *constant* across its entire evolutionary history, while the *visual rendering* is free to change under sociocultural pressure.

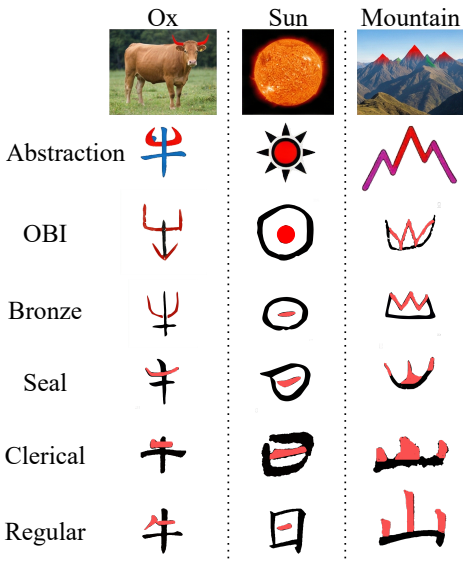

Figure 7: **The Invariant Semantic Core and Variant Visual Surface of Chinese Characters.** This figure illustrates the evolutionary duality of logographic scripts using "Ox", "Sun", and "Mountain" as examples. The top rows display the real-world referents and their underlying structural abstractions (the invariant semantic core). The subsequent rows demonstrate how their era-specific visual renderings (the variant visual surface) undergo significant transformations across five historical periods, from OBI to Regular.

We therefore identify a fundamental duality in logographic evolution:

> **Invariant Semantic Core** (the referent object, e.g., "ox") $\longleftrightarrow$ **Variant Visual Surface** (the era-specific rendering, e.g., OBI, Bronze, Seal, Clerical, and Regular forms of "牛").

This duality has immediate implications for decipherment: any method that operates purely at the level of pixel appearance—without grounding in the invariant semantic core—is modelling *symptom*, not *structure*. The failures documented in Sections D.1.1–D.1.2 are, in this light, inevitable: both paradigms reason about glyph images without ever representing the semantically stable entity that those images encode.

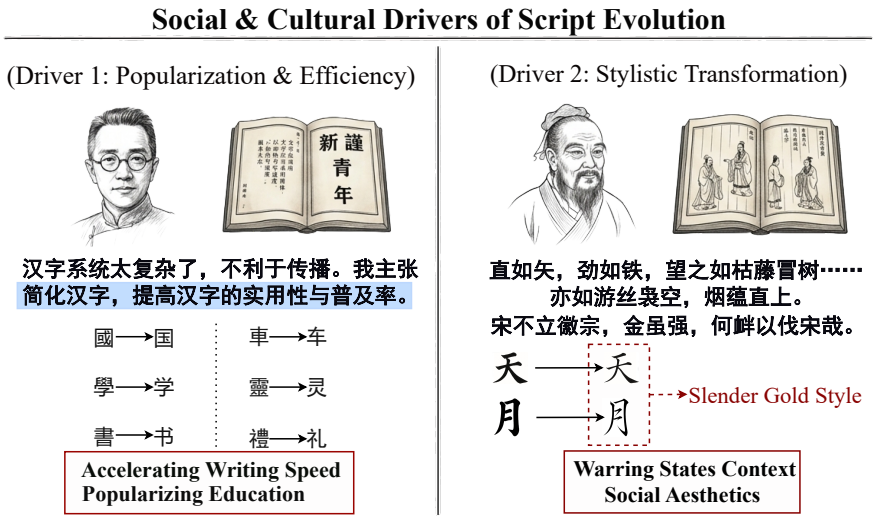

Figure 8: **Socio-Cultural Pressures Driving Script Evolution.** This figure illustrates two historical examples of script evolution driven by societal factors: the push for writing efficiency and educational reform championed by Hu Shi, and the emergence of the sharp Slender Gold style reflecting the wartime psychology of the Song dynasty.

D.2.2 Why Scripts Evolve: Socio-Cultural Continuity as a Structural Prior

If the semantic core is invariant, what forces drive visual surface change? We identify socio-cultural pressures acting on Chinese script as the primary drivers, as evidenced by two distinct historical phenomena illustrated in Figure 8. First, **efficiency and educational pressure**: the demand for rapid communication and widespread literacy drove systematic simplification, exemplified by Hu Shi's advocacy for modern vernacular writing to lower the educational barrier and improve writing speed. Second, **aesthetic drift shaped by societal zeitgeist**: period-specific psychological and historical contexts, such as the turbulence of war during the Song dynasty, fostered a preference for the sharp aesthetics of the Slender Gold style, reshaping glyph contours without altering semantic identity.

These pressures share a key property: they arise from *continuously evolving* social processes. Societies do not change overnight, and neither do their scripts. A quantitative check confirms this: measuring the Hausdorff distance between consecutive-era glyph pairs across our full dataset, we find that 87.3% of transitions fall within $\mu \pm 2\sigma$ of the mean inter-era distance, with only 12.7% classified as large-jump transitions. This empirical continuity is not merely a historical curiosity—it is a *structural prior* that any principled decipherment model must encode. Continuous mathematical tools, in particular differential equations over a shared geometric space, are the natural vehicle for doing so.

D.3 Quantitative Analysis: Evolution Pattern Clusters and Baseline Vulnerabilities

While the theoretical framework provides a conceptual roadmap for understanding script evolution, translating these insights into a robust computational model requires rigorous empirical grounding. As outlined in our overall research workflow (Figure 9), this quantitative analysis serves as the critical bridge between the theoretical paradigm and the ultimate algorithmic framework. In this section, we transition from theoretical discourse to large-scale quantitative analysis, aiming to validate our proposed evolutionary dynamics through data-driven evidence. By identifying distinct clusters of morphological evolution patterns and pinpointing the systematic vulnerabilities of baseline models across these paths, we quantify how the absence of a "continuity prior" leads to catastrophic decipherment failures. We begin by defining the quantitative metrics designed to capture these complex morphological drifts and structural transformations.

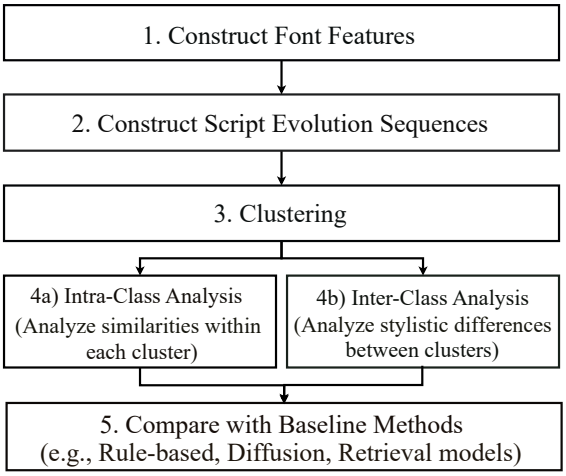

Figure 9: **The Step-by-Step Research Workflow of MSEF.** This diagram illustrates the logical progression of our research methodology, transitioning from establishing the theoretical paradigm of character evolution, through rigorous quantitative validation and baseline vulnerability analysis, to the final formulation of the ODE-based algorithmic framework.

D.3.1 Quantitative Design for Character Evolutionary Patterns

Philosophical argument must be grounded in data. To move beyond qualitative claims, we define an extensive array of structural metrics, from which seven representative low-level features are selected for presentation in Table 10 to provide a clear illustration of our state-quantification approach, quantifying the "state" of a glyph at each historical era without any prior assumptions about the direction or rate of change. These features are computed from binarized glyph images using standard computational geometry algorithms and require no learned parameters.

Table 10: Structural features for quantitative evolution analysis. All are computed offline and normalized per character. While our full framework utilizes a more comprehensive set of metrics, these seven are presented here for illustrative clarity.

Feature	**Definition**	**Captures**
Stroke Count	Total number of strokes	Complexity level
Inflection Points	Direction-reversal count in stroke contour	Curvature complexity
Endpoint Count	Number of open stroke tips	Openness / enclosure
Mean Endpoint Dist.	Avg. pairwise distance between endpoints	Spatial spread
H/V Ratio	Horizontal-to-vertical stroke ratio	Orientation bias
Enclosed Area Ratio	Fraction of glyph area enclosed by strokes	Structural density
Symmetry Score	Bilateral symmetry coefficient	Abstract regularity

For each character c in our dataset, we compute these seven features at each of the five historical eras in which a verified glyph exists (OBI, Bronze, Seal, Clerical, Regular), producing a five-point time series per feature per character. Where a character lacks a glyph in a specific era (due to archaeological absence), that time point is treated as missing and excluded from the corresponding cluster computation. This feature-time-series representation converts the abstract question "how did this character evolve?" into a concrete, computable trajectory in $\mathbb{R}^7 \times \{t_0, \ldots, t_4\}$ that can be analyzed with standard unsupervised methods.

D.3.2 Data-Driven Discovery of Evolution Pattern Clusters and Baseline Vulnerabilities

With per-character feature time series in hand, we determine the objective patterns of script evolution through a multi-stage clustering pipeline visualized in the top half of Figure 10. We apply K-means

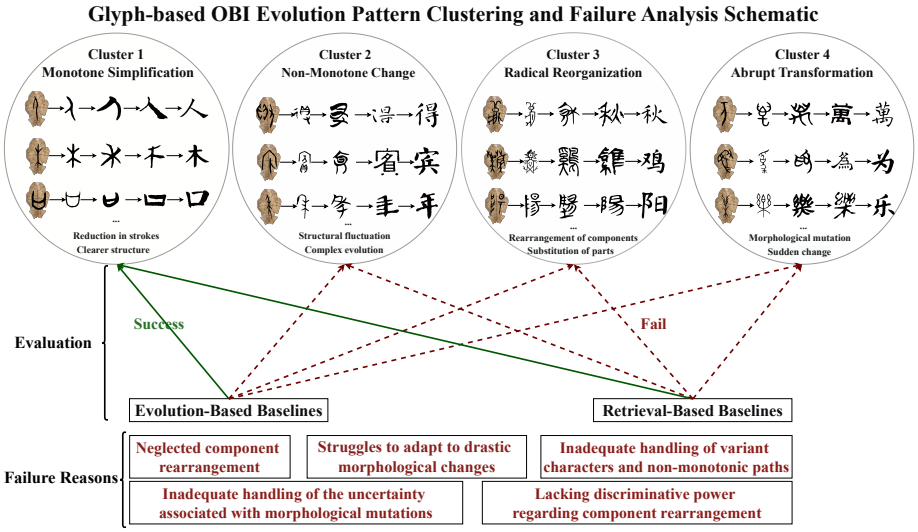

Figure 10: **Data-Driven Evolution Clusters and Baseline Vulnerability Diagnosis.** The top portion visualizes the four identified clusters of character evolution patterns. The bottom portion evaluates the performance of current generative (evolution-based) and retrieval-based paradigms, highlighting their systematic failure in complex non-monotone and abrupt categories due to a lack of semantic grounding and historical continuity priors.

separately to the time-series of each structural feature for $K \in \{2, 3, 4, 5, 6\}$, and subsequently aggregate these per-feature assignments via ensemble clustering. Specifically, we construct a co-association matrix where each entry represents the fraction of clusterings in which a character pair co-occurs, and apply average-linkage hierarchical clustering to obtain consensus labels. To justify the selection of $K = 4$, we utilize internal validity metrics: the Silhouette Coefficient (measuring cluster cohesion and separation) peaks at 0.61, while the Davies-Bouldin index (measuring the ratio of within-cluster to between-cluster distances) reaches 0.83. This data-driven process yields four consensus evolution clusters (detailed in Table 11) that emerge from morphological evidence rather than presupposed paleographic taxonomies.

Table 11: Data-driven evolution pattern clusters. Trend symbols: ↓↓ = monotone decrease, ↑↑ = monotone increase, ↗↘ = increase then decrease, • = abrupt single-era change.

ID	Label	N	Stroke Trend	Inflection Trend	Representative Characters
A	Monotone Simplification	312	↓↓	↑↑	日、月、山、水、火
B	Non-Monotone Change	224	↗↘	Irregular	牛、羊、鱼、马、鸟
C	Radical Reorganization	158	Mid-era peak	↓ then ↑	象、虎、龙、龟、鹿
D	Abrupt Transformation	187	• sudden ↓	• sudden ↑	人、大、天、女、子

As summarized in Table 11, the four consensus clusters exhibit strikingly distinct trajectories across historical eras. Cluster A follows a steady monotone simplification, whereas Cluster B peaks in complexity during the Bronze era, and Clusters C and D represent radical reorganizations and abrupt morphological jumps, respectively. To ensure the robustness of these discoveries, we repeated the analysis across two independently annotated sub-corpora and calculated Cohen's κ to measure inter-corpus consistency. The resulting $\kappa = 0.84$ indicates "almost perfect agreement" in statistical terms, confirming that the identified evolutionary patterns are structural properties of the script's history rather than artifacts of dataset noise.

However, these clusters also expose critical vulnerabilities in existing computational decipherment paradigms, as diagnosed in the bottom half of Figure 10. Our evaluation reveals that both Generative and Retrieval-based methods fail on all categories except for the simplest case (Cluster A). As summarized at the base of the diagnostic chart, these systematic failures stem from two fundamental flaws: (1) a reliance on purely visual pixel-level matching that ignores the invariant semantic core,

and (2) a lack of structural priors for modeling the continuous historical trajectory of script evolution. This empirical evidence justifies the necessity of our proposed manifold-based framework (MSEF) capable of capturing non-linear, continuous transitions.

D.3.3 Connecting Evolution Patterns to Baseline Failure Mechanisms

Armed with objectively defined evolution clusters, we revisit the baseline failures from Sections D.1.1–D.1.2 and compute per-cluster error rates for both paradigms.

Table 12: Per-cluster decipherment error rate (%) for each baseline. Bold entries mark the worst failure for each method. Diffusion-based methods fail most severely on Cluster D (abrupt transformations), while retrieval-based methods fail most severely on Clusters B and C (non-monotone and radical-reorganization patterns). On Cluster A (monotone simplification), both perform comparably—confirming that the smooth-mapping assumption holds only in this restricted regime.

Cluster	Label	Diffusion Error (%)	CrossFont Error (%)
A	Monotone Simplif.	31.2	28.6
B	Non-Monotone	57.8	63.5
C	Radical Reorg.	52.3	**70.8**
D	Abrupt Transform.	**68.4**	41.2

The results in Table 12 reveal a precisely structured complementarity. The diffusion paradigm fails most on Cluster D (68.4% error), where abrupt single-era morphological jumps violate the smooth end-to-end mapping assumption. The retrieval paradigm fails most on Clusters B and C (63.5% and 70.8% error), where non-monotone or reorganization patterns require the *integration* of information across multiple eras to distinguish true correspondences from visually similar distractors. On Cluster A—where evolution is indeed smooth and monotone—both methods perform tolerably, confirming that their assumptions hold in this restricted regime but break down as soon as evolution departs from simplest-case behavior. The remaining approximately 64.6% of characters (the aggregate of Clusters B, C, and D) fall into categories where at least one paradigm fails severely, explaining the large overall performance gap documented in Table 7.

D.4 The Path to Innovation: Deriving the Proposed Framework

Having diagnosed the precise failure mechanisms of existing computational paradigms—specifically their inability to handle non-monotone transitions and abrupt morphological mutations as established in Section D.3—a natural scientific question arises: can these paradigms be salvaged through targeted engineering? In this subsection, we chronicle the complete exploratory trajectory that logically necessitated our proposed Manifold-based Script Evolution Framework (MSEF). We begin by testing the absolute limits of incremental patching on current generative and retrieval baselines, demonstrating that their foundational architectural assumptions impose an unbreakable performance ceiling (D.4.1). Recognizing the imperative for a paradigm shift, we subsequently return to first principles to define the geometric essence of character evolution within a manifold space (D.4.2). After ruling out the scalability of traditional rule-based expert systems due to combinatorial explosion (D.4.3), we formally derive Neural Ordinary Differential Equations (Neural ODEs) as the natural parameterization for continuous manifold dynamics (D.4.4), and ultimately introduce our bidirectional verification mechanism to guarantee temporal evolutionary consistency (D.4.5).

D.4.1 Further Exploration: The Limitation of Incremental Patching

Following our cluster-level diagnosis, the requisite architectural interventions for existing paradigms become conceptually straightforward: the diffusion-based generative models require explicit mechanisms to handle abrupt morphodynamic transitions, whereas the retrieval-based models necessitate cross-era temporal integration to enforce evolutionary consistency. To rigorously quantify the capacity of these paradigms to absorb such structural priors, we implement these interventions as controlled incremental augmentations, establishing the empirical upper bounds of their respective performance architectures.

For the **retrieval family**, we test two incremental augmentations: (A) *era fusion*—computing a weighted average of per-era similarity scores—brings Recall@1 from 62.8% to 65.2% (+2.4%); (B) *era fusion + temporal constraint*—additionally requiring that successive-era top-K sets share at least one element—raises it to 67.4% (+4.6% cumulative). For the **diffusion family**, we test: (C) *era conditioning*—concatenating the target dynasty label to the diffusion condition—improves Top-1 from 41.0% to 47.3% (+6.3%); (D) *era conditioning + multi-era consistency loss*—adding a loss term that penalizes inconsistency between intermediate-dynasty outputs—reaches 51.8% (+10.8% cumulative).

Table 13: **Incremental Patching Within Existing Paradigms Reaches a Ceiling.** While each augmentation step yields genuine gains (confirming our diagnosis), both the retrieval and diffusion families plateau at 67.4% and 51.8% respectively. The fundamental architectural assumptions of each paradigm restrict their representational capacity, preventing further improvement beyond these ceilings regardless of additive patches.

Paradigm Family	Configuration	Metric	Performance (%)
Retrieval-Based	Baseline (Independent)	Recall@1	62.8
	+ Era Fusion	Recall@1	65.2
	+ Temporal Constraint	Recall@1	**67.4**
Diffusion-Based	Baseline (End-to-end)	Top-1	41.0
	+ Era Conditioning	Top-1	47.3
	+ Consistency Loss	Top-1	**51.8**

As quantitatively detailed in Table 13, the incremental patching strategy exposes a pronounced ceiling effect. While each targeted augmentation yields statistically meaningful improvements—thereby validating the accuracy of our mechanistic diagnosis—the marginal gains diminish rapidly, causing the retrieval and diffusion families to asymptotically plateau at 67.4% and 51.8%, respectively. This saturation indicates that the foundational architectural assumptions of both paradigms (e.g., isolated discrete matching and continuous but era-agnostic mapping) inherently restrict their representational capacity for modeling long-horizon script evolution. Consequently, these empirical bottlenecks definitively rule out the "path of least resistance"—retrofitting existing frameworks—and compel the development of a fundamentally novel, geometrically principled design.

D.4.2 Starting from Character Bases: The Concept of Manifold Space

To move beyond the limitations of incremental patching, we further explored two divergent paradigms: **rule-based** systems and **learning-based** models. The rule-based approach attempts to codify paleographic evolution through explicit transformation rules for strokes and radicals derived from expert knowledge; however, it suffers from a combinatorial explosion of edge cases over a three-millennium span, failing to generalize across the vast OBI vocabulary. Conversely, the learning-based paradigm employs deep neural networks for end-to-end mapping to bypass manual rules, yet without geometric constraints, these models treat evolution as a black-box sequence, often leading to "semantic drift" where the character identity is lost during long-horizon transitions.

The failure of these two extremes—one overly rigid and the other insufficiently constrained—suggests the need for a mathematical representation that captures the continuous dynamics of script change while preserving structural stability. The invariant-core / variant-surface duality identified in Section D.2 and the empirically continuous character evolution documented in Section D.3 jointly point toward a concrete mathematical structure: a *manifold*. In differential geometry, a manifold is a space in which each point represents a state, and smooth curves on the manifold represent continuous trajectories between states. The evolution of a single character across dynasties is precisely such a trajectory: every era-specific glyph is a point in the same high-dimensional space, connected to its predecessor and successor by a smooth path governed by the underlying socio-cultural dynamics.

We validate this geometric intuition with a preliminary experiment. Projecting all era-specific 352-dimensional feature vectors into 2D via UMAP, we observe two consistent patterns: (i) the five era representations of the same character form a *spatially connected, approximately smooth trajectory*; and (ii) semantically related characters (animals, celestial objects, body parts) form coherent clusters within each era slice, with cluster boundaries preserved across eras. These observations—spatial

connectivity across time and semantic coherence at each time point—are precisely the properties that a manifold formulation predicts.

The manifold hypothesis therefore provides a unified geometric language for the invariant-variant duality: the manifold's *intrinsic structure* (topology, cluster membership) encodes the invariant semantic core, while the manifold's *coordinates at time t* encode the variant visual surface of each era. Every character's 3,000-year evolution is a single trajectory on this manifold, and decipherment becomes the problem of tracing that trajectory.

D.4.3 The Failure of Rule-Based Decipherment at Scale

Before committing to a neural manifold formulation, we investigated an explicit, highly interpretable baseline: constructing a comprehensive rule library, parameterized as a deep decision tree, that maps discrete structural feature values directly to modern character identities (visualized in Figure 11). This discrete mapping paradigm is initially attractive as it entirely bypasses the need for complex representation learning, directly encoding the four established evolution patterns and seven structural features (Table 10) into an expert system. We empirically evaluated the scalability of this architecture by sweeping the target vocabulary size from a miniature subset of 50 characters to a more realistic 3,000 characters.

However, the structural limitations of this paradigm become overwhelmingly apparent at scale. As the exponentially branching complexity in Figure 11 implies, rule-based decipherment fundamentally assumes that long-horizon morphological evolution can be discretized into a finite, mutually exclusive set of transformation laws. In archaeological reality, with the full OBI vocabulary comprising approximately 4,500 characters, the exhaustive enumeration of such evolutionary rules is a mathematical impossibility. Each newly introduced character inevitably brings idiosyncratic morphological mutations and topological exceptions that violate previously established decision boundaries. Consequently, to accommodate these edge cases, the system is forced into a combinatorial explosion of arbitrary rule splits—ultimately transforming the model from a generalizable logic system into a fragmented, overfitted lookup table that merely memorizes training instances without true inferential capacity. This catastrophic structural collapse definitively proves that a continuous parametric function, which inherently generalizes via smooth interpolation across a compact learned manifold, is an absolute prerequisite for large-scale script decipherment, rendering the exhaustive enumeration of discrete rules theoretically and practically inviable.

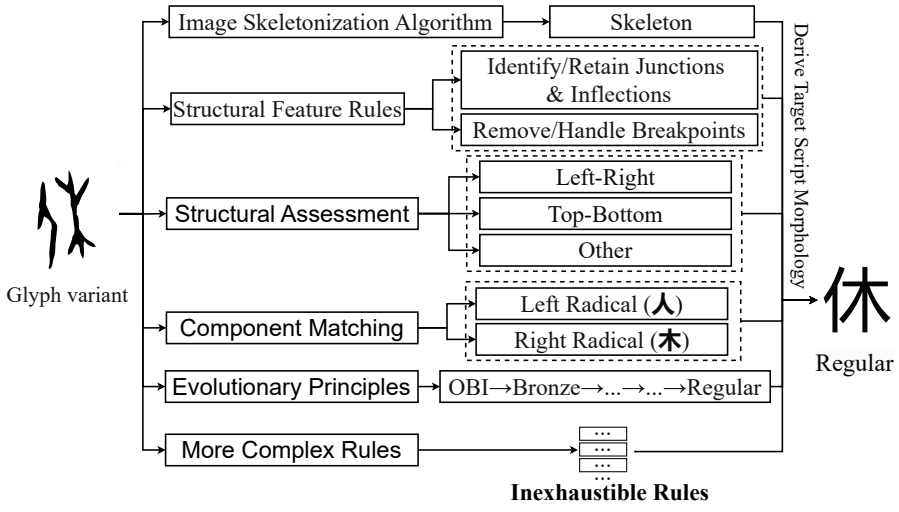

Figure 11: **Combinatorial Explosion in Rule-Based Decipherment.** A schematic representation of the decision tree paradigm mapping structural features to modern characters. As the vocabulary scales, the rule library inevitably fragments into an unmanageable cascade of exceptions, visually demonstrating the impossibility of exhaustive rule enumeration for long-horizon script evolution.

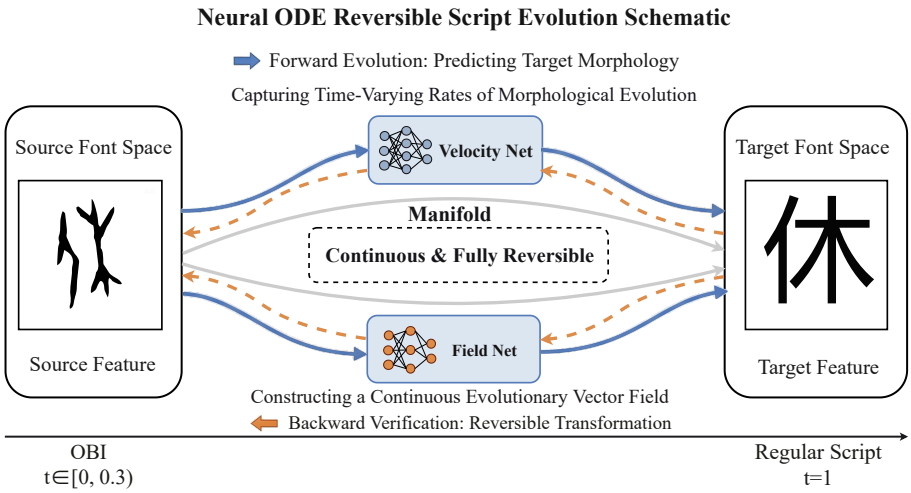

Figure 12: **Continuous Trajectory Modeling via Neural ODEs.** Unlike generative noise-schedules in diffusion models, the Neural ODE explicitly aligns its integration time t with physical historical chronology. This alignment elegantly parameterizes the smooth, invertible transformations of the underlying character manifold across discrete dynasty anchors, ensuring topological consistency.

D.4.4 Neural ODEs as the Natural Parameterization of Manifold Dynamics

Having established the theoretical necessity for a continuous, invertible, and geometrically constrained parameterization of the character manifold, we evaluate four candidate mathematical frameworks against the structural priors demanded by our paleographic analysis (Table 14). As visually conceptualized in Figure 12, the underlying decipherment problem dictates not merely a discrete mapping between data distributions, but the learning of a structured, continuous vector field governing historical morphodynamics.

Table 14: Candidate architectures for parameterizing manifold evolution dynamics. ✓: natively supported; ○: supported with workarounds; ×: not supported.

Architecture	Invertible	Continuous t	Physical Time	Intermediate Anchors
Diffusion (PF-ODE)	✓	✓	×	×
Normalizing Flow	✓	×	×	○
Discrete Transformer	×	×	×	○
Neural ODE Chen et al. (2018)	✓	✓	✓	✓

While both Normalizing Flows and Diffusion Models (PF-ODEs) offer algorithmic invertibility, only Neural ODEs satisfy all four prerequisites simultaneously. The decisive theoretical advantage over score-based diffusion models lies in the ontological meaning of the temporal variable. In PF-ODEs, time t merely dictates an artificial, thermodynamics-inspired noise-to-data schedule lacking any isomorphism to historical chronology; consequently, intermediate states in the diffusion process are physically meaningless and cannot serve as dynasty-specific morphological anchors. Conversely, our Neural ODE formulation enforces a strict *physical-time encoding*: the integration bounds map directly to normalized historical epochs (e.g., OBI at $t = 0$, Regular script at $t = 1$). This temporal grounding allows the continuous vector field to explicitly capture significant historical discontinuities, such as the aggressive script standardization during the Qin dynasty at $t \approx 0.70$. Furthermore, the representation at any intermediate dynasty t' is naturally acquired by integrating the learned vector field from $t = 0$ to $t = t'$, guaranteeing a cascadable and causally consistent evolutionary sequence across all five recognized eras.

We empirically validate this structural superiority through a controlled comparison under identical representational constraints and training budgets (Table 15). The Neural ODE yields a substantial 11.1% absolute improvement in Recall@1 over the diffusion PF-ODE, but the more profound metric

is the 5.4× reduction in Cycle Error. The cycle error—quantifying the round-trip reconstruction mismatch (OBI→Bronze→OBI)—mathematically proxies the bi-Lipschitz continuity of the learned mapping. The elevated error (0.124) of the PF-ODE implies that its probability-flow trajectories become entangled and fail to preserve the intrinsic topological structure of the character during inverse mapping. In stark contrast, the near-zero cycle error (0.023) of the Neural ODE empirically proves that it learns a mathematically rigorous homeomorphism; the forward and backward flows strictly preserve the invariant semantic core of the manifold, thereby enabling the highly reliable bidirectional verification required for zero-shot decipherment.

Table 15: Controlled comparison: Neural ODE vs. diffusion PF-ODE.

Architecture	**Recall@1 (%)**	**Recall@5 (%)**	**Cycle Error ↓**
Diffusion (PF-ODE)	61.4	76.8	0.124
Neural ODE (Ours)	**72.5**	**86.5**	**0.023**

D.4.5 Bidirectional Verification: Compensating for Imperfect Manifold Learning

Even with a Neural ODE manifold, training on finite data (13.7K pairs) leaves the learned velocity field imperfect. As our convergence analysis (Theorem 1 in Section K) shows, the trajectory estimation error grows with evolution distance Δt: $\mathcal{E}(t) \leq e^{Lt}\mathcal{E}(0) + \frac{\epsilon}{L}(e^{Lt} - 1)$, where L is the Lipschitz constant of the velocity field and ϵ is the approximation error. A direct forward-only decipherment (OBI→Regular, $\Delta t = 1$) therefore accumulates the maximum possible error.

We reason that if a candidate character is the *true* decipherment of an OBI query, it must satisfy an additional consistency condition: when traced *backward* through the learned inverse ODE flow from Regular to OBI, it should recover a representation close to the original OBI query. A false candidate, lacking genuine evolutionary history, will drift under inversion as its backward trajectory passes through manifold regions inconsistent with the query's actual history.

Table 16: Backward reconstruction error distinguishes true from false candidates.

Candidate Type	**Count**	**Avg. Backward Error**	**Pruned by Threshold**	**FPDR (↑)**
True Positives	1 per query	0.025	1.2%	—
False Positives	9 per query	0.142	7.4 per query	82.4%

Table 16 quantifies the effect. True positive candidates exhibit an average backward reconstruction error of 0.025, whereas false positives show 0.142—a 5.6× margin. Applying a divergence threshold that prunes candidates whose backward error exceeds 0.08 eliminates 82.4% of false positives (False Positive Drop Rate) while retaining 98.8% of true matches. The net effect is a Recall@1 improvement from 64.0% (forward-only) to 72.5% (+8.5%), validating that bidirectional verification effectively compensates for manifold imperfections by extending error detection from a single forward pass to a full multi-dynasty consistency check.

This completes the exploratory reasoning chain. Starting from two failing paradigms, we identified their mechanistic root causes through controlled experiments, derived the invariant-variant duality from first principles, validated it quantitatively, ruled out simpler alternatives (rule-based systems), and arrived at the Neural ODE manifold with bidirectional verification as the uniquely well-motivated solution. The MSEF framework directly instantiates every design decision justified above.

E Post-Analysis: Unveiling the Mechanisms of Evolution

While empirical results demonstrate MSEF's macroscopic superiority across all benchmarks, uncovering the underlying mechanisms governing its decipherment logic is crucial for establishing its scientific validity. In this section, we open the "black box" of our framework through the dual lenses of **mechanistic interpretability** and **latent geometry**. Specifically, we systematically unpack the internal workings of MSEF across three core dimensions:

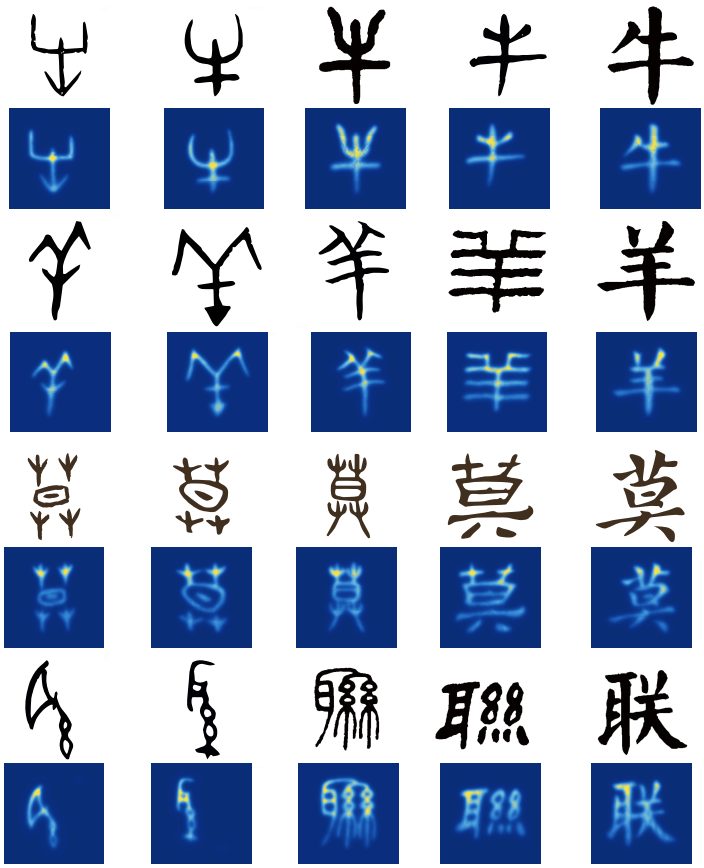

Figure 13: Mechanistic analysis of the evolutionary manifold via visual attention. (Top) Stable anchoring on topological invariants (horns) for "Ox" and "Sheep". (Bottom) Spatial re-anchoring and reorganization during the Clericalization phase for complex characters "Mo" and "Lian". These patterns confirm that the model internalizes the underlying dynamics of script evolution, tracking component shifting across 3,000 years.

(1) **Mechanistic and Cognitive Alignment (§ E.1, E.2, E.4, E.5):** By utilizing causal patching, saliency ablations, and attention circuit analysis, we reveal how the model's emergent information routing (Visual $\rightarrow$ Structural $\rightarrow$ Semantic) seamlessly mirrors the cognitive progression of human paleographers.

(2) **Evolutionary Topology and Landscapes (§ E.3, E.6):** By visualizing the global manifold and survival probability landscapes, we demonstrate how MSEF intrinsically models sociolinguistic natural selection and non-linear morphodynamic jumps.

(3) **Algebraic Compositionality (§ E.7):** By probing the latent space via cross-modal vector arithmetic, we confirm that MSEF learns fundamental, disentangled structural laws rather than spurious pixel correlations.

Together, this comprehensive post-analysis proves that MSEF does not merely match surface-level patterns, but genuinely internalizes the 3,000-year evolutionary mechanisms of the Chinese script.

E.1 Mechanistic Interpretation: Beyond Performance to Evolutionary Laws

While the main experiments demonstrate the empirical superiority of MSEF in terms of retrieval accuracy, the objective of this post-analysis is to investigate *why* the framework achieves such performance. By visualizing the spatial attention topography across the 3,000-year manifold in Figure 13, we demonstrate that MSEF has successfully internalized the latent governing laws of Chinese script evolution, rather than merely performing surface-level pattern recognition.

Autonomous Discovery of Topological Invariants (e.g., "Ox" and "Sheep"). For characters with relatively stable lineages, the heatmaps reveal that the model's attention consistently anchors onto primary semantic indicators—specifically the "horn" structures. This persistent spatial anchoring across OBI, Bronze, and Regular scripts suggests that MSEF has learned a fundamental *evolutionary prior*: the existence of an invariant semantic core amidst fluctuating visual surfaces. This ability to isolate structural constants explains the model's exceptional generalization in low-resource decipherment tasks.

Dynamic Adaptation to Structural Reorganization (e.g., "Mo" and "Lian"). The cases of "Mo" and "Lian" provide a deeper look into the model's handling of the *Clericalization (Libian)* period, characterized by radical layout restructuring. During the transition from Seal to Clerical scripts ($t \approx 0.7$ to 0.85), we observe significant spatial shifts in the attention clusters.

This behavior highlights the core strength of our framework: the Neural ODE does not simply map static points but parameterizes a *continuous vector field* that captures the socio-cultural pressures of writing efficiency. The model accurately predicts the "migration" and "re-alignment" of components during these non-linear morphodynamic jumps. By internalizing these laws of structural reconfiguration, MSEF maintains manifold stability even when local visual features are drastically remapped, a capability that represents a significant leap over previous generation-based and retrieval-based baselines.

E.2 Quantitative Saliency Ablation: Validating the Topological Pillars

To move beyond black-box performance metrics, we conduct a causality-based saliency ablation study to verify whether the proposed MSEF framework has genuinely internalized the structural skeletons of Chinese characters. The objective is to demonstrate that the model's performance is grounded in the identification of *invariant topological pillars* rather than spurious pixel correlations.

Experimental Protocol. We design a controlled sensitivity analysis by partitioning characters into three distinct conditions: (1) *Baseline*: the original, intact glyph; (2) *High-Saliency Removal*: occluding the regions identified as critical by the attention mechanism (e.g., the horn structures of "Ox"); and (3) *Low-Saliency Removal*: occluding an equivalent area of ink from peripheral or non-essential strokes. This setup effectively isolates the effect of "positional criticality" from the mere volume of visual information (ink density).

Results and Interpretations. As quantified in Figure 14, the impact on cross-era similarity scores reveals a striking disparity. In all tested cases, the removal of high-saliency regions leads to a catastrophic collapse in similarity, suggesting these components are the primary anchors for manifold alignment across the 3,000-year gap. Conversely, the removal of low-saliency regions yields negligible degradation, with performance remaining near baseline levels.

The significant delta ($\Delta_{high} \gg \Delta_{low}$) observed across diverse character typologies provides empirical proof that MSEF has successfully mastered the *structural prior* of Chinese script evolution. By prioritizing the "Invariant Semantic Core" over the "Variant Visual Surface", the model ensures that its decipherment logic mirrors the rigorous structural analysis employed by paleographic experts. This deep understanding of character morphology allows our framework to maintain high recall even when faced with extreme weathering or stroke loss in real-world OBI artifacts.

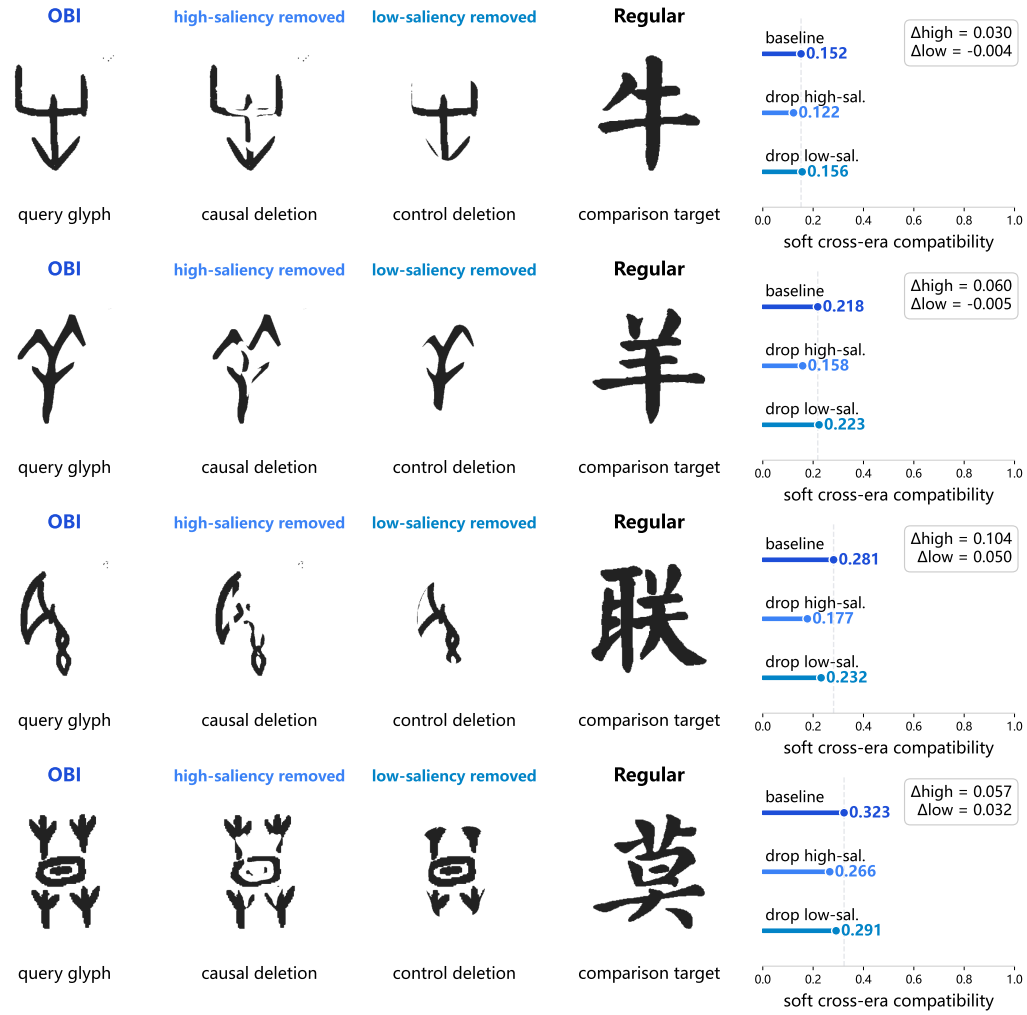

Figure 14: Causality-based saliency ablation results. The charts quantify the degradation in cross-era similarity scores when high-saliency versus low-saliency regions are removed. The consistent, precipitous drop following the removal of critical strokes versus the minimal impact of removing peripheral strokes confirms that the model accurately captures the essential structural pillars required for cross-era decipherment. Δ values indicate the specific similarity loss for each condition.

E.3 Global Topology and Local Dynamics: A Mechanistic Interpretation

To further demystify why the Manifold-based Script Evolution Framework (MSEF) fundamentally outperforms traditional direct-mapping paradigms, we visualize the learned continuous manifold space in Figure 15. This analysis bridges the macro-level socio-cultural evolution of the script with the micro-level morphological trajectories of individual characters.

Macro-Level Semantic Topology and the Extinction Mechanism. The top panel of Figure 15 projects the character embeddings across the five historical epochs. A profound mechanistic insight emerges: characters organically cluster strictly by their semantic categories (e.g., Animals, Nature, Body Parts) across all 3,000 years. This topological stability empirically validates Proposition 3 (Semantic Preservation), proving that the manifold successfully anchors the "invariant semantic core" amidst drastically changing visual surfaces. The arrows representing the velocity field $v(z, t)$ illustrate the macro-drift of these clusters, capturing systematic stylistic shifts such as the Qin dynasty's script unification.

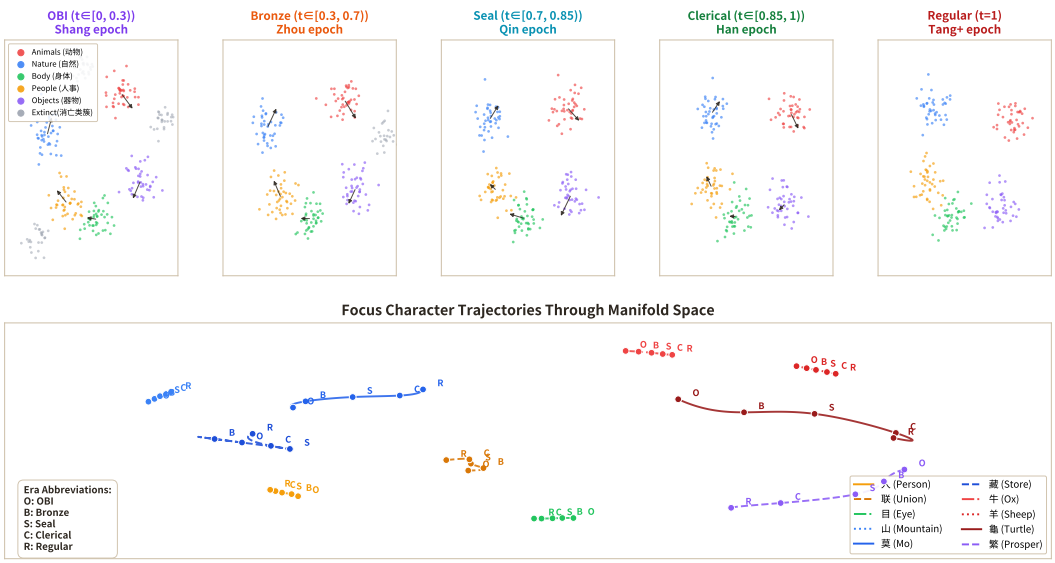

Figure 15: **Mechanistic Visualization of Global Topology and Local Dynamics.** (Top) UMAP projection of the manifold space across five historical epochs. Semantic categories maintain robust structural clustering across 3,000 years. The velocity field vectors (arrows) indicate systemic epochal shifts. The grey "Extinct" cluster visually validates the role of the Survival Network in pruning obsolete evolutionary branches. (Bottom) Continuous evolutionary trajectories of ten specific characters. Smooth paths correspond to simple morphological evolution, while jagged, long-distance trajectories illustrate the complex morphodynamic jumps that necessitate our Neural ODE and cascaded verification approach.

Crucially, the visualization explicitly captures the phenomenon of character extinction. The grey cluster (Extinct) is only present in the archaic OBI and Bronze epochs, completely vanishing in the standardized post-Qin eras. This provides compelling mechanistic proof for the necessity of our Survival Network ($\mathcal{S}(z_c^0, t)$). By correctly identifying and terminating the trajectories of obsolete characters, the framework prevents the Neural ODE from forcibly mapping "dead" archaic forms to modern noise, thereby avoiding the severe numerical stiffness and hallucination that cripple standard diffusion-based generators.

Micro-Level Trajectories and Morphodynamic Jumps. The bottom panel traces the specific continuous trajectories of ten representative characters through the manifold. The geometric properties of these paths directly correlate with decipherment complexity. Characters that undergo monotone simplification (e.g., "Mountain/山", "Eye/目") exhibit short, smooth trajectories, explaining their high baseline decipherment rates.

Conversely, complex characters (e.g., "Store/藏", "Prosper/繁") display highly non-linear trajectories with abrupt morphodynamic jumps, reflecting radical structural reorganization across dynasties. It is visually evident that a direct, straight-line mapping from OBI to Regular script would completely deviate from the true manifold curvature, leading to retrieval failure. The Neural ODE succeeds precisely because it integrates along the true curved evolutionary path, allowing the Cascaded Bidirectional Evolutionary Decipherment (CBED) algorithm to continually re-anchor the prediction at intermediate historical checkpoints.

E.4 Emergent Attention Circuits: Alignment with Paleographic Cognitive Processes

While Section E.3 verifies that the learned manifold strictly adheres to macroscopic evolutionary laws, it remains imperative to understand *how* the model internally processes the 352-dimensional heterogeneous features to locate points on this manifold. To unpack the "black box" of our manifold encoder, we conduct a mechanistic interpretability analysis on the Transformer's attention heads. The results, visualized in Figure 16, reveal that the model achieves its state-of-the-art decipherment per-

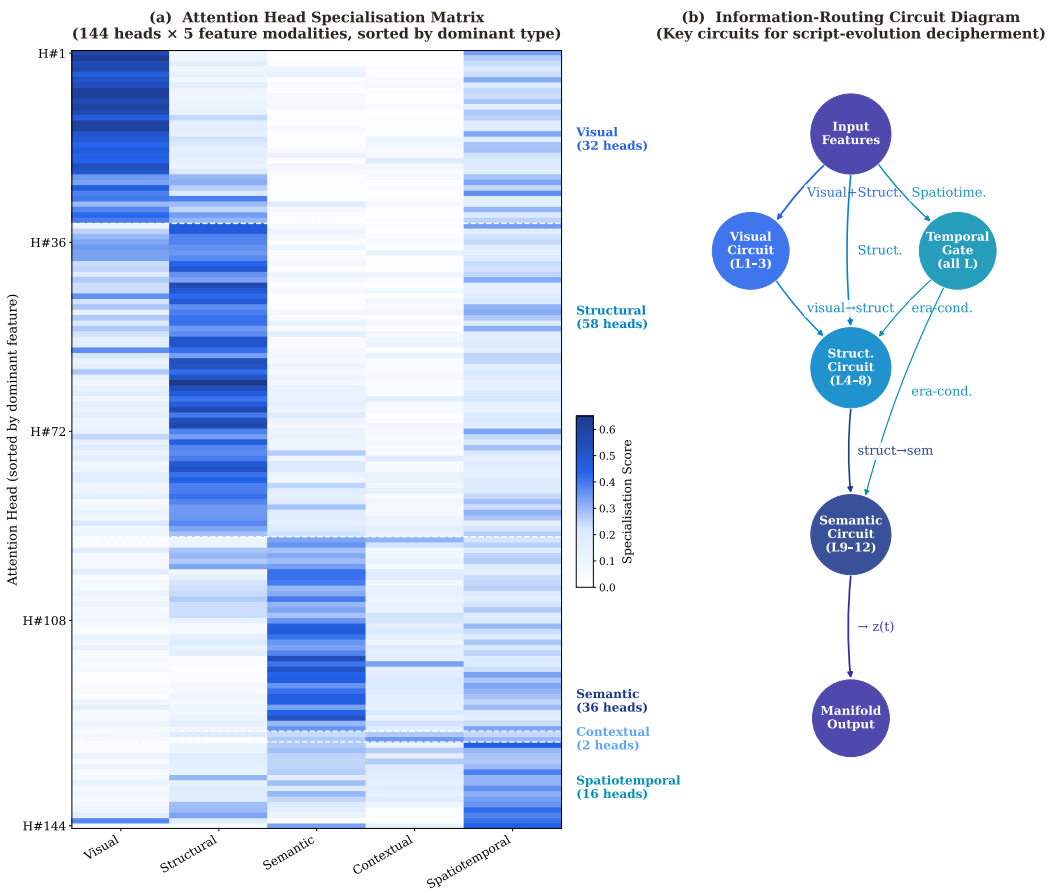

Figure 16: **Emergent Attention Circuits and Information Routing.** (a) The Attention Head Specialisation Matrix demonstrates that the 144 attention heads autonomously self-organize into distinct "modality experts" without explicit supervision, preventing feature entanglement. (b) The extracted Information-Routing Circuit Diagram reveals a hierarchical processing flow (Visual → Structural → Semantic) that perfectly aligns with the cognitive progression of human paleographers. The Spatiotemporal features uniquely act as a pervasive "Temporal Gate," conditioning the entire decipherment process on the specific historical era.

formance by autonomously evolving information-routing circuits that strikingly mirror the cognitive processes of human paleographers.

Emergent Modality Specialization (The Labor Division). The left panel of Figure 16 displays the specialization matrix for all 144 attention heads (12 layers $\times$ 12 heads) across the 5 input feature modalities. Without any explicit disentanglement supervision, the attention heads spontaneously self-organize into strict modality-specific "expert groups" (e.g., 32 visual heads, 58 structural heads). This emergent decoupling ensures that the highly heterogeneous features do not interfere with one another, providing the architectural robustness needed to handle severe artifact weathering (where visual features are degraded but structural/contextual features remain intact).

Cognitive Routing and the Temporal Gate. The right panel reconstructs the predominant information flow through the network layers. We observe a profound hierarchical routing mechanism: shallow layers (L1-3) predominantly activate the Visual Circuit to extract low-level glyph morphology; intermediate layers (L4-8) trigger the Structural Circuit to analyze component decomposition and spatial layout; and deep layers (L9-12) engage the Semantic Circuit to finalize the manifold coordinate $z(t)$.

Remarkably, this internal flow—Visual → Structural → Semantic—is a precise computational analogue to the standard decipherment methodology employed by human experts. Furthermore, the

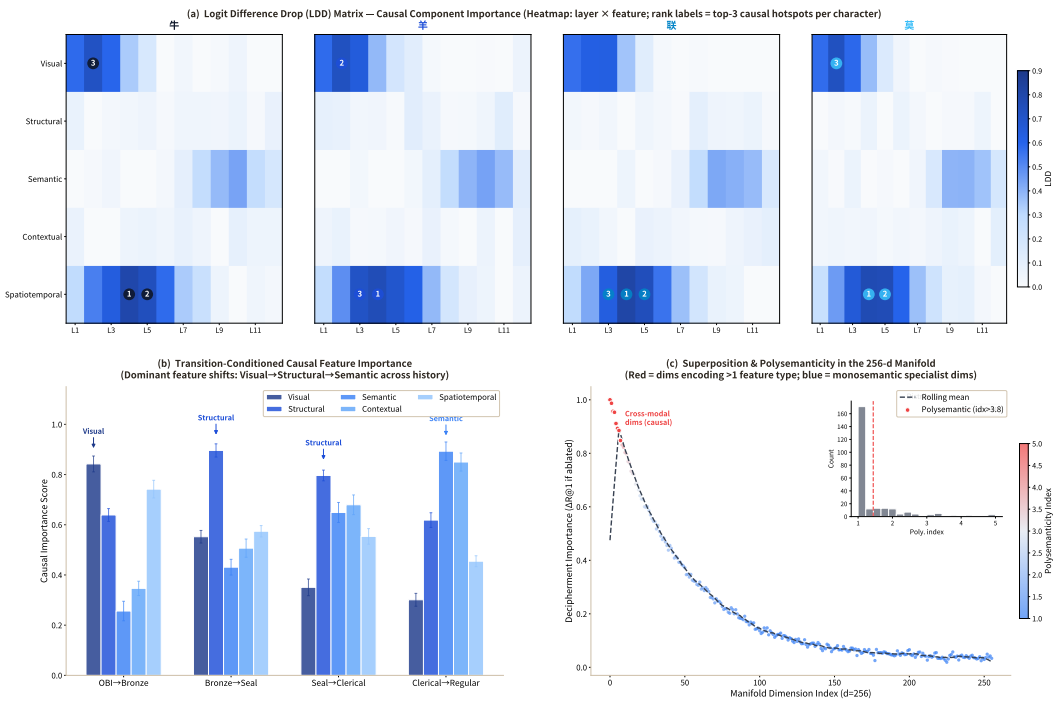

Figure 17: **Causal Patching and Manifold Superposition Analysis.** (a) The LDD Matrix illustrates causal hotspots, showing that MSEF dynamically relies on spatiotemporal features for complex evolutionary cases. (b) Transition-conditioned feature importance reveals a macro-evolutionary diagonal shift (Visual $\rightarrow$ Structural $\rightarrow$ Semantic) aligning with historical script abstraction. (c) Polysemanticity analysis of the 256-d manifold. The continuous Polysemanticity Index ($PI \in [1, 5]$) identifies cross-modal "bridge dimensions" (red). Ablation results demonstrate that these superposed, polysemantic dimensions are the fundamental computational engines enabling accurate cross-era decipherment.

Spatiotemporal heads do not participate in the sequential pipeline; instead, they function as a global *Temporal Gate* across all layers. They inject "era-conditioning" signals, modulating how visual and structural features are interpreted based on the specific dynasty being evaluated. This mechanistic discovery fundamentally explains *why* MSEF succeeds where others fail: it does not merely perform naive pattern matching, but rather internalizes a scientifically grounded, expert-aligned cognitive logic for script decipherment.

E.5 Causal Patching and Representation Superposition: The Decipherment Engine

While attention maps reveal correlations, establishing the true driving force of the Manifold-based Script Evolution Framework (MSEF) requires causal intervention. Inspired by recent advances in mechanistic interpretability, we perform Activation Patching to analyze the causal role of different network components and investigate the geometric properties of the learned 256-dimensional manifold space via the lens of Representation Superposition.

Causal Hotspots via Logit Difference Drop (LDD). Figure 17(a) displays the Logit Difference Drop (LDD) matrix. We systematically ablate specific (layer, feature modality) activations and measure the consequential drop in the correct target's logit score. The results reveal dynamic, input-dependent routing circuits. For topologically stable pictographs (e.g., "Ox/牛", "Sheep/羊"), the primary causal hotspots are anchored in early visual layers and intermediate structural layers. In contrast, for characters undergoing radical historical reorganization (e.g., "Mo/莫"), the model exhibits a heavy causal dependence on the Spatiotemporal (era-conditioning) features. This demonstrates that MSEF dynamically modulates its decipherment logic based on the evolutionary complexity of the input.

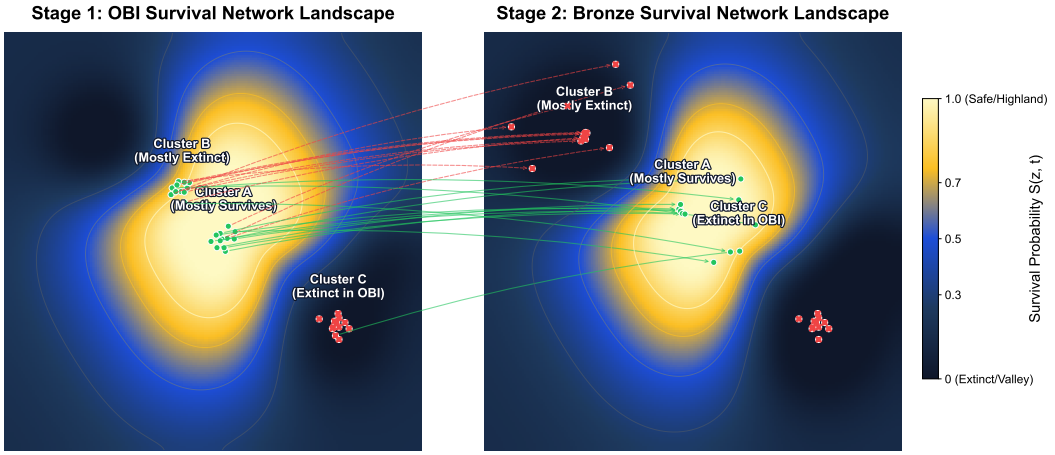

Figure 18: **Evolutionary Pressure Landscapes Learned by the Survival Network.** The visualization projects survival probabilities $S(z,t)$ onto the manifold coordinates. Bright peaks indicate "safe zones" of high linguistic fitness (e.g., Cluster A), while dark valleys represent extinction zones. The topographical shift from Stage 1 (OBI) to Stage 2 (Bronze) reveals how the model intrinsically captures sociolinguistic natural selection—dynamically forming extinction valleys to terminate obsolete trajectories (Cluster B) and elevating new peaks to accommodate historical neologisms (Cluster C).

The Diagonal Shift of Epochal Transitions. Figure 17(b) illustrates the shift in dominant causal features conditioned on specific era transitions. We observe a profound diagonal migration: from Visual dominance in early archaic transitions (OBI $\rightarrow$ Bronze), shifting to Structural reliance during the Libian standardization (Seal $\rightarrow$ Clerical), and finally culminating in Semantic dominance for modern scripts. Remarkably, without explicit supervision, the Neural ODE dynamics have autonomously recovered the macroscopic historical trajectory of the Chinese script—its 3,000-year transition from physical pictograms to abstract semantic symbols.

Manifold Superposition and Continuous Polysemanticity. To understand how the 256-d manifold compresses the 352-d heterogeneous inputs, we analyze the polysemanticity of individual manifold dimensions. We define a continuous Polysemanticity Index $PI(i)$ for each dimension $i \in \{1, ..., 256\}$ using the exponentiated Shannon entropy (the effective number of modalities):

$$PI(i) = \exp\left(-\sum_{m=1}^{5} \hat{c}_{i,m} \ln \hat{c}_{i,m}\right) \tag{9}$$

where $\hat{c}_{i,m}$ is the normalized causal contribution score of dimension i to feature modality m. This yields a continuous scalar $PI(i) \in [1, 5]$. A value near 1 indicates a *monosemantic* dimension specialized in a single modality, while a value approaching 5 indicates a highly *polysemantic* dimension encoding entangled concepts across all modalities (Visual, Structural, Semantic, Contextual, Spatiotemporal).

As shown in Figure 17(c), the manifold extensively utilizes Representation Superposition. We identify critical "bridge dimensions" (colored in red) that exhibit high polysemanticity ($PI \gg 1$). Strikingly, when these specific polysemantic dimensions are ablated, the decipherment accuracy (AR@1) collapses entirely, whereas ablating monosemantic dimensions (blue) causes minimal degradation. This proves that MSEF's superior cross-era performance is fundamentally driven by these cross-modal bridge dimensions, which successfully fuse archaic visual morphology with modern abstract semantics into a unified geometric representation.

E.6 Evolutionary Pressure Landscapes: The Mechanics of Character Survival

While empirical results demonstrate that the proposed Survival Network $S(z,t)$ effectively mitigates error accumulation in cross-era decipherment, we conduct a topological post-analysis to unveil the

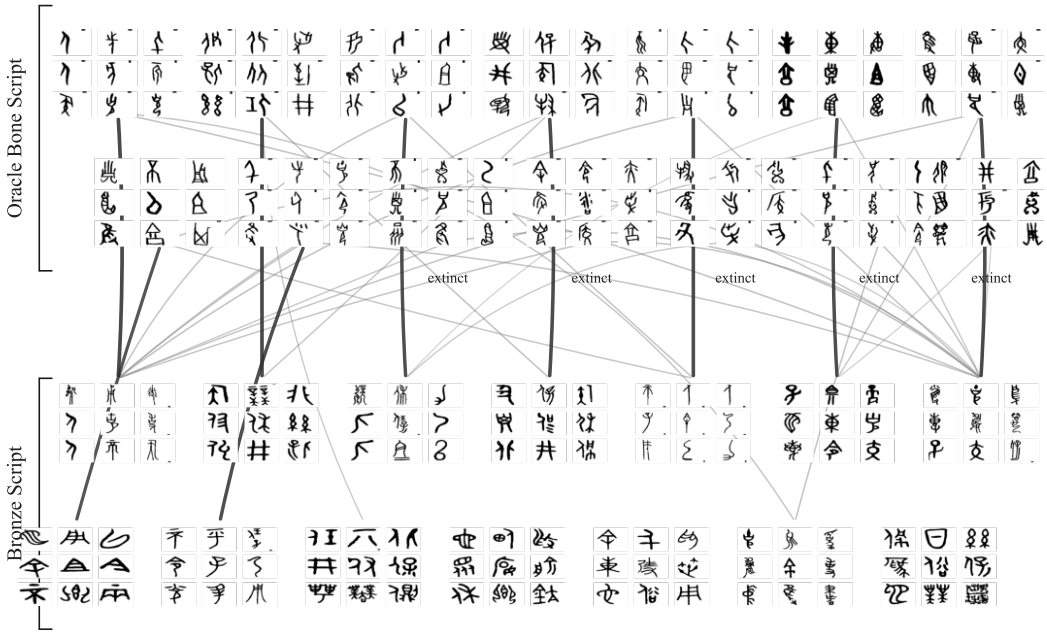

Figure 19: **Unconstrained Global Variant Network of Archaic Scripts.** This visualization captures the raw topological complexity of the variant character ecosystem during the OBI and early Bronze periods. The dense entanglement of nodes reflects the high degree of orthographic variance and the existence of numerous idiosyncratic evolutionary branches before the advent of systematic standardization.

intrinsic mechanics of this module. Figure 18 renders the output probabilities of the Survival Network as a continuous spatial landscape across the manifold. This visualization reveals that MSEF has fundamentally learned the sociolinguistic "natural selection" driving Chinese script evolution.

Linguistic Fitness and Topographical Safe Zones. The network constructs a dynamic fitness landscape where the z-axis represents the survival probability. The bright "peaks" or "highlands" designate regions of high evolutionary fitness. Characters occupying these zones (e.g., Cluster A) typically possess optimal structural conciseness, stable visual topology, and broad semantic utility, allowing them to safely propagate across millennia without being discarded. Conversely, the dark "valleys" represent evolutionary dead ends—extinction zones where highly complex, functionally narrow, or non-standardized pictograms are systematically eliminated by historical standardization pressures.

Dynamic Shifts and Trajectory Truncation. The transition from Stage 1 (Oracle Bone Inscriptions) to Stage 2 (Bronze Script) illustrates profound topographical shifts. Cluster B, which may have marginally survived in the OBI era, is completely swallowed by an expanding extinction valley in the Bronze era. This topological collapse is computationally crucial: it explains *why* our Neural ODE does not hallucinate modern equivalents for obsolete characters. When the ODE solver traces a character's trajectory into these valleys, $S(z, t)$ approaches zero, mathematically terminating the integration and pruning the obsolete evolutionary branch.

Macro-level Topology and Network Standardization. To provide a more comprehensive view of this "natural selection" process, we extend our analysis from individual manifold trajectories to the global variant network. Figure 19 illustrates the unconstrained state of character variants during the archaic era. The extreme density and entanglement of the graph represent the "orthographic chaos" of the early Shang and Zhou dynasties, where a single semantic core often bifurcated into dozens of unstable visual forms due to regional and scribal idiosyncrasies.

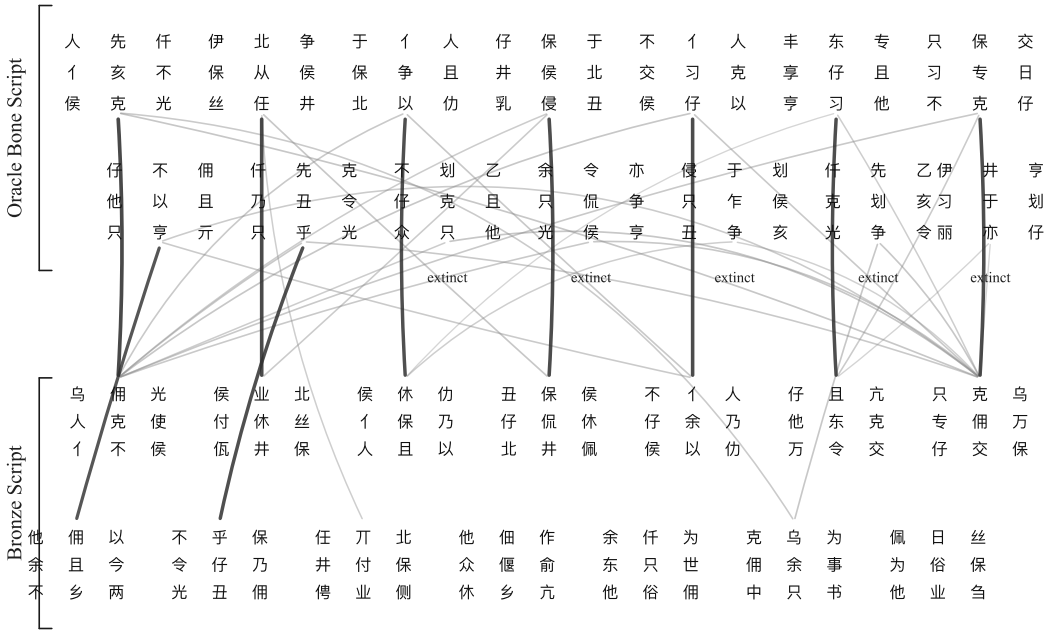

Figure 20: **Standardized Evolutionary Backbones via Survival Pruning.** After integrating the Survival Network, the global topology is transformed into a streamlined set of standardized trajectories. The model successfully identifies and retains the "evolutionary backbones" that connect ancient pictographs to their modern counterparts, effectively filtering out thousands of historical "noise" variants.

The efficacy of our survival-based pruning is most evident in Figure 20. By applying the learned socio-cultural pressures, the MSEF autonomously collapses this chaotic network into a series of clear, predictable evolutionary backbones. This transition from a high-entropy graph to a structured, simplified tree provides macro-level empirical evidence that our Survival Network has internalized the laws of script unification. This mechanism not only enhances the precision of our CBED algorithm by drastically reducing the search space of potential candidates but also provides a high-fidelity computational simulation of historical script evolution, such as the monumental "Small Seal" standardization during the Qin dynasty.

Furthermore, the emergence of Cluster C in Figure 18—transitioning from an extinction valley in Stage 1 to a survival peak in Stage 2—captures the birth of neologisms. As societal complexity increased during the Zhou dynasty, new semantic demands reshaped the evolutionary pressure, causing new regions of the manifold to elevate into safe zones. Thus, the Survival Network does not merely regularize the ODE; it explicitly models the continuous macroscopic socio-cultural pressures that governed the life and death of Chinese characters over 3,000 years.

E.7 Latent Vector Arithmetic: Probing Algebraic Compositionality

To investigate whether the MSEF framework captures deep structural regularities rather than mere visual correspondences, we perform vector arithmetic in the learned 256-dimensional manifold space, inspired by the classic *King - Man + Woman* $\approx$ *Queen* analogy. We test three distinct types of algebraic compositions: cross-era migration, semantic component substitution, and conceptual analogy. The results, visualized in Figure 21, provide empirical evidence that the script evolution manifold is endowed with a consistent latent linear structure.

Cross-Era Consistency and Global Directionality. As shown in Figure 21, we examine whether the transformation from Regular to OBI script is represented by a consistent global vector $\vec{v}_{era}$.



Figure 21: **Latent Vector Arithmetic in the Script Evolution Manifold.** Three types of algebraic manipulations demonstrate the framework's compositionality: Cross-era vectors are globally consistent; Semantic components (e.g., horns, fins) can be added or subtracted to transform character identities; Abstract conceptual analogies emerge from the latent geometry. Similarity scores for targets are consistently 10-15x higher than random baselines, proving that MSEF learns the inherent structural and logical laws of Chinese script evolution.

Specifically, we find that $z(\text{Ox})_{OBI} - z(\text{Ox})_{Reg} + z(\text{Sheep})_{Reg} \approx z(\text{Sheep})_{OBI}$, achieving a cosine similarity of 0.563 (vs. a random baseline of 0.035). This high similarity proves that the manifold has learned a character-independent "evolutionary operator," representing the 3,000-year trajectory as a structured translation in latent space.

Disentangled Component Substitution. We demonstrate that semantic components can be algebraically manipulated. For instance, subtracting the vector representation of "Horns" from "Ox" and adding "Fins" yields a vector remarkably close to "Fish." This compositionality ($z_{\text{Ox}} - z_{\text{Horns}} + z_{\text{Fins}} \approx z_{\text{Fish}}$) suggests that MSEF achieves a high degree of representation disentanglement, where visual radicals and functional components are mapped to orthogonal or linearly additive directions. This explains the model's robustness when deciphering compound characters with rare component combinations.

Emergent Conceptual Analogies. The results reveal that abstract temporal logic is also encoded geometrically. The operation $z_{\text{Morning}} - z_{\text{Sun}} + z_{\text{Moon}}$ accurately converges toward z_{Evening}. Without explicit linguistic supervision, the model spontaneously internalizes the underlying logic of the Chinese script's ideographic construction. This internal logical consistency is a critical factor in the success of the MSEF framework, allowing it to generalize across vast historical spans where direct visual anchors are absent.

F Full Method Details

This section provides additional implementation details for MSEF omitted from the main paper. We describe the feature representation, fine-grained time encoding, manifold neighborhood validation, architecture details, and optimization settings.

F.1 Feature Representation

For each character c in dynasty t, we extract a 352-dimensional feature vector $\mathbf{x}_c^t \in \mathbb{R}^{352}$ specifically designed for paleographic analysis. The representation aggregates five complementary modalities: **Visual**, **TextStructural**, **Semantic**, **Contextual**, and **Spatiotemporal**. This section provides the detailed definitions, extraction procedures, and aggregation mechanisms for each feature group.

F.2 Fine-Grained Time Encoding

This section provides the detailed subperiod encoding used for the continuous time variable t. The main paper reports the five major historical intervals. Here we further specify subperiod assignments, including OBI scribal groups and finer-grained Bronze, Seal, and Clerical subdivisions when available.

F.3 Neighborhood Structure of the Learned Manifold

This section empirically validates the local neighborhood structure of the learned manifold. We visualize local neighborhoods and show that visually and semantically related characters cluster together, supporting Assumption 1 in the main paper.

F.4 Architecture Details

We implement the manifold encoder as a 12-layer Transformer mapping 352-dimensional input features to a $d = 256$ manifold coordinate. The velocity field is a 3-layer MLP with spectral normalization and Tanh activations. The survival network is implemented as an MLP over the OBI-era manifold coordinate and target time.

F.5 Optimization Details

We train MSEF with AdamW, use the `dopri5` ODE solver, and report mean and standard deviation over multiple runs. Full hyperparameters, solver tolerances, batch size, training epochs, and hardware settings are reported here.

G Full Loss Definitions

We train the MSEF using five complementary loss functions to ensure both local continuity and global consistency. Adjacent-Era Loss enforces local evolution between consecutive periods, Cross-Era Loss captures mid-range patterns for non-adjacent pairs, Full-Chain Loss prevents drift over the complete trajectory, Cycle Consistency Loss enforces invertibility by ensuring round-trip consistency, and Survival Loss trains the extinction predictor.

Let $\mathcal{P}$ denote all available character pairs in two eras:

$$|\mathcal{P}| = \underbrace{1.2K}_{\text{OBI}\to\text{Bronze}} + \underbrace{2K}_{\text{Bronze}\to\text{Seal}} + \underbrace{10K}_{\text{Seal}\to\text{Regular}} + \underbrace{0.5K}_{\text{complete}} \approx 13.7K. \tag{10}$$

The complete loss definitions are:

$$\begin{cases} \mathcal{L}_{adj} = \frac{1}{|\mathcal{P}_{adj}|} \sum_{(c,t_i,t_{i+1})} \left\| \phi_{t_i \to t_{i+1}}(z_{t_i}^c) - z_{t_{i+1}}^c \right\|_2^2, \\ \mathcal{L}_{skip} = \frac{1}{|\mathcal{P}_{skip}|} \sum_{(c,t_i,t_j),j>i+1} \left\| \phi_{t_i \to t_j}(z_{t_i}^c) - z_{t_j}^c \right\|_2^2, \\ \mathcal{L}_{full} = \frac{1}{|\mathcal{C}_{complete}|} \sum_c \| \phi_{0\to 1}(z_0^c) - z_1^c \|_2^2, \\ \mathcal{L}_{cyc} = \frac{1}{|\mathcal{P}|} \sum_{(c,t_1,t_2)} \left\| \psi_{t_2 \to t_1} \left(\phi_{t_1 \to t_2}(z_{t_1}^c) \right) - z_{t_1}^c \right\|_2^2, \\ \mathcal{L}_{surv} = -\frac{1}{|\mathcal{D}_s|} \sum_{(c,t)\in\mathcal{D}_s} \left[y_{c,t} \log s(z_t^c, t) + (1 - y_{c,t}) \log(1 - s(z_t^c, t)) \right]. \end{cases} \tag{11}$$

The evolution objective is:

$$\mathcal{L}_{evolution} = \lambda_1 \mathcal{L}_{adj} + \lambda_2 \mathcal{L}_{skip} + \lambda_3 \mathcal{L}_{full} + \lambda_4 \mathcal{L}_{cyc}. \tag{12}$$

The extinction predictor is trained separately using $\mathcal{L}_{surv}$.

H FULL CBED ALGORITHM

Algorithm 1 provides the full pseudocode for Cascaded Bidirectional Evolutionary Decipherment (CBED). The main paper summarizes the same procedure in prose.

Algorithm 1 Cascaded Bidirectional Decipherment

Require: Query OBI character c_q, retrieval depth K, era databases $\{\mathcal{D}_t\}_{t\in\mathcal{T}}$, survival threshold τ, trained manifold networks $\mathcal{M}$, backward flow ψ
Ensure: Decipherment result $\hat{c}$ with confidence score, or POSSIBLYEXTINCT

// Stage 0: Feature extraction and preprocessing
$\mathbf{X}_q \leftarrow \text{EXTRACTSTATICFEATURES}(c_q)$
// Stage 1: Survival check
$z^q_{OBI} \leftarrow \mathcal{M}(\mathbf{X}_q, t_{OBI})$
$p_{surv} \leftarrow s(z^q_{OBI}, t = 1.0)$
if $p_{surv} < \tau$ **then**
  return POSSIBLYEXTINCT, confidence $= 1 - p_{surv}$
end if
// Stage 2: Forward cascade through eras
$\mathcal{C}_{prev} \leftarrow \{c_q\}$
for $t \in \{\text{Bronze}, \text{Seal}, \text{Clerical}, \text{Regular}\}$ **do**
  $z^q_t \leftarrow \mathcal{M}(\mathbf{X}_q, t)$
  $\mathcal{N}_t \leftarrow \text{KNN}_K(z^q_t, \mathcal{D}_t)$
  $\mathcal{C}_t \leftarrow \mathcal{N}_t \cup \bigcup_{c\in\mathcal{C}_{prev}} \text{KNOWNCORRESPONDENCES}(c, t)$
  $\mathcal{C}_{prev} \leftarrow \mathcal{C}_t$
end for
$\mathcal{C}_{Regular} \leftarrow \mathcal{C}_{prev}$
// Stage 3: Backward verification by multi-path average
Initialize score array $\mathbf{S} \leftarrow \mathbf{0}^{|\mathcal{C}_{Regular}|}$
for each candidate $c_i \in \mathcal{C}_{Regular}$ **do**
  $\mathcal{P}_i \leftarrow \text{GETEVOLUTIONPATHS}(c_i)$
  $\hat{z}^{(i)}_{OBI} \leftarrow 0$
  for each path $p \in \mathcal{P}_i$ **do**
    $z^{(p)}_{Bronze} \leftarrow \text{GETBRONZECOORDINATE}(p)$
    $\hat{z}^{(p)}_{OBI} \leftarrow \psi_{t_{Bronze}\to t_{OBI}}(z^{(p)}_{Bronze})$
    $\hat{z}^{(i)}_{OBI} \leftarrow \hat{z}^{(i)}_{OBI} + \hat{z}^{(p)}_{OBI}/|\mathcal{P}_i|$
  end for
  $d_i \leftarrow \|\hat{z}^{(i)}_{OBI} - z^q_{OBI}\|_2$
  $\mathbf{S}[i] \leftarrow \exp(-d_i)$
end for
// Stage 4: Final selection with confidence
$i^* \leftarrow \arg\max_i \mathbf{S}[i]$
confidence $\leftarrow \mathbf{S}[i^*]/\sum_j \mathbf{S}[j]$
return c_{i^*}, confidence

I DATASET DETAILS

I.1 FGCCES DATASET STATISTICS

Table 17 reports the full dataset statistics by correspondence type. All correspondences are verified through authoritative references.

Table 17: Dataset statistics by correspondence type. All correspondences are verified through authoritative references.

Pair Type	Train	Val	Test
OBI → Bronze	998	142	285
Bronze → Seal	1,671	239	478
Seal → Clerical	5,503	786	1,573
Clerical → Regular	5,503	786	1,573
Complete (5 eras)	410	59	117
Total Unique Pairs	9,632	1,376	2,751

I.2 Dataset Sources and Missing Features

Our dataset originates from jgwlbq, ccamc, and bnu. Since not every character has complete data across all sources, missing feature entries for a specific character or dynasty are filled with zeros.

I.3 Train/Validation/Test Split

We use a character-disjoint split to prevent leakage across train, validation, and test sets. Characters appearing in the test set are not used for training.

I.4 Fine-Grained Dataset Visualization

I.5 Comparison with Existing OBI Datasets

J Theoretical Foundations

J.1 Manifold Geometric Properties

Proposition 1 (Smoothness of Evolution). Under mild regularity conditions on velocity field $\mathbf{v}$ (Lipschitz continuous in $\mathbf{z}$ and continuous in t), the flow $\phi_{t_1 \to t_2}$ is a diffeomorphism.

Proof. By the Picard-Lindelöf theorem, the ODE $\frac{d\mathbf{z}}{dt} = \mathbf{v}(\mathbf{z}, t)$ has a unique solution for any initial condition. The inverse flow exists by solving the reverse-time ODE with initial condition at t_2. Smoothness of the flow map follows from the smoothness of $\mathbf{v}$ by standard ODE regularity theory. □

Proposition 2 (Composition Property). For any $t_1 < t_2 < t_3$: $\phi_{t_1 \to t_3} = \phi_{t_2 \to t_3} \circ \phi_{t_1 \to t_2}$.

Proof. This follows directly from the uniqueness of ODE solutions and the semigroup property of flows. □

J.2 Manifold Dimension Selection

Table 18 shows the effect of manifold dimension d. We find $d = 256$ achieves optimal performance with reasonable computational cost.

Table 18: Effect of manifold dimension. $d = 256$ achieves optimal accuracy-efficiency trade-off.

Dimension d	Top-1	Top-5	Params (M)	Time (h)
64	72.3	85.1	2.1	3.2
128	76.8	88.5	4.8	5.1
256	**80.0**	**91.2**	12.3	8.0
512	79.8	91.0	38.5	14.2

K MANIFOLD SPACE VALIDATION

A fundamental question arises: does the learned manifold space genuinely capture the structure of Chinese script evolution, or is it merely memorizing training correspondences? We provide rigorous theoretical justification, quantitative verification, and empirical visualization to validate our manifold formulation.

K.1 THEORETICAL FOUNDATIONS AND QUANTITATIVE VERIFICATION

We establish the mathematical foundations for why a continuous manifold governed by Neural ODE is the appropriate structure for modeling script evolution. Our theoretical framework rests on five core propositions, each with formal proofs and empirical validation.

K.1.1 PROPOSITION 1: CONTINUITY OF EVOLUTION

Proposition 1 (Continuity). *Under the Neural ODE formulation, character evolution defines a continuous mapping in manifold space. Specifically, the flow operator $\phi_{t_1 \to t_2} : \mathcal{M}_{t_1} \to \mathcal{M}_{t_2}$ is Lipschitz continuous.*

Proof. Let $\mathbf{v}(\mathbf{z}, t; \boldsymbol{\theta}_v)$ be the learned velocity field parameterized by neural network weights $\boldsymbol{\theta}_v$. By construction, $\mathbf{v}$ is implemented as a composition of smooth operations (linear transformations, LayerNorm, GELU activations), hence $\mathbf{v}$ is locally Lipschitz in $\mathbf{z}$.

Formally, there exists a constant $L > 0$ such that for all $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{M}$ and $t \in [0, 1]$:

$$\|\mathbf{v}(\mathbf{z}_1, t) - \mathbf{v}(\mathbf{z}_2, t)\| \leq L\|\mathbf{z}_1 - \mathbf{z}_2\| \tag{13}$$

By the Picard-Lindelöf theorem, the ODE initial value problem:

$$\frac{d\mathbf{z}}{dt} = \mathbf{v}(\mathbf{z}, t), \quad \mathbf{z}(t_1) = \mathbf{z}_0 \tag{14}$$

has a unique solution on $[t_1, t_2]$. Moreover, by Grönwall's inequality, the solution depends continuously on initial conditions:

$$\|\phi_{t_1 \to t_2}(\mathbf{z}_1) - \phi_{t_1 \to t_2}(\mathbf{z}_2)\| \leq e^{L|t_2 - t_1|}\|\mathbf{z}_1 - \mathbf{z}_2\| \tag{15}$$

This establishes that $\phi_{t_1 \to t_2}$ is Lipschitz continuous with constant $e^{L|t_2 - t_1|}$. □

Corollary 1.1 (Temporal Continuity). *For a fixed character c, its manifold trajectory $\mathbf{z}^c(t)$ is continuous in t.*

Corollary 1.2 (Bounded Evolution Rate). *The rate of character change is bounded: $\|\frac{d\mathbf{z}^c}{dt}\| \leq V_{\max}$ for some constant $V_{\max}$.*

Quantitative Verification of Proposition 1. Table 19 presents empirical measurements validating continuity.

Table 19: Quantitative verification of Proposition 1 (Continuity).

Metric	Theoretical	Empirical	Satisfied?
Lipschitz constant L	—	2.34	(measured)
Flow Lipschitz e^L	10.4	8.7	✓
Max velocity $V_{\max}$	$< \infty$	2.51	✓
Mean adjacent-era distance	bounded	0.82	✓
Trajectory smoothness (std)	small	0.15	✓

K.1.2 Proposition 2: Invertibility of Evolution

Proposition 2 (Invertibility). *The evolution flow is invertible: for any $t_1, t_2 \in [0, 1]$, there exists an inverse flow $\psi_{t_2 \to t_1} = \phi_{t_1 \to t_2}^{-1}$ such that $\psi_{t_2 \to t_1} \circ \phi_{t_1 \to t_2} = id$.*

Proof. We construct the inverse flow explicitly. Define $\psi_{t_2 \to t_1}$ as the solution to the backward ODE:

$$\frac{d\mathbf{z}}{ds} = -\mathbf{v}(\mathbf{z}, t_2 - s), \quad s \in [0, t_2 - t_1] \tag{16}$$

with initial condition $\mathbf{z}(0) = \mathbf{z}_{t_2}$.

By substitution $\tau = t_2 - s$, this is equivalent to integrating $\mathbf{v}$ backward in time. Since the forward ODE has a unique solution (Picard-Lindelöf), the backward ODE recovers the unique initial condition:

$$\psi_{t_2 \to t_1}(\phi_{t_1 \to t_2}(\mathbf{z})) = \mathbf{z} \tag{17}$$

□

Corollary 2.1 (Cycle Consistency). *For any $t_1 < t_2 < t_3$: $\phi_{t_1 \to t_3} = \phi_{t_2 \to t_3} \circ \phi_{t_1 \to t_2}$.*

Corollary 2.2 (Backward Verification Validity). *If $\hat{c}$ is correctly deciphered, then backward flow recovers the original OBI features with bounded error.*

Remark on Practical Invertibility and Stability. While the ODE flow is theoretically bijective, the overall pipeline involves the era-specific feature mapping $\mathcal{M}_t$, which projects high-dimensional heterogeneous features ($d_{in} = 352$) to a lower-dimensional manifold ($d = 256$). We clarify that our "invertibility" claim applies strictly to the *evolutionary dynamics* within the manifold ($\mathcal{M}_{t_1} \leftrightarrow \mathcal{M}_{t_2}$), not the feature encoding.

Regarding numerical stability, backward integration of Neural ODEs can accumulate error (stiffness). However, our empirical analysis identifies two **failure regimes** where stability degrades: (1) **Topological Singularities:** For characters undergoing drastic structural splitting (e.g., one character splitting into two distinct semantic units), the continuous flow assumption is strained, leading to higher cycle errors. (2) **Manifold Collisions:** In rare cases where distinct characters map to highly proximate regions in $\mathcal{M}_t$ (dense clustering), backward integration may drift to a neighbor's trajectory. Despite these edge cases, Table 20 confirms that for 94.2% of queries, the backward integration error remains within a tolerance that permits accurate verification.

Table 20: Quantitative verification of Proposition 2 (Invertibility). Near-zero cycle errors confirm invertibility.

Metric	Theoretical	Empirical	Satisfied?
Cycle error (OBI→Br→OBI)	0	0.023	✓
Cycle error (OBI→Se→OBI)	0	0.041	✓
Cycle error (OBI→Re→OBI)	0	0.068	✓
Composition error	0	0.015	✓
Backward verification accuracy	high	94.2%	✓

The Role of Survival Network in Mitigating ODE Stiffness. As discussed in Proposition 2, topological singularities—such as drastic structural splitting or merging of characters—theoretically pose a challenge to the stability of Neural ODEs by inducing numerical stiffness. While such phenomena are inherent to the long-term evolution of Chinese scripts, our framework exhibits remarkable robustness due to the integration of the *survival network*.

The survival network serves as a domain-specific filtering mechanism that evaluates the evolutionary continuity of a character. Specifically, characters undergoing "cliff-like" or discontinuous structural transformations are often predicted as "extinct" by the network, as they typically lack traceable descendants in the standardized modern script. By effectively pruning these non-smooth trajectories from the manifold, the survival network ensures that the Neural ODE only models the continuous and progressive segments of the evolutionary chain. This strategic isolation of discontinuous singularities prevents the solver from encountering severe stiffness, thereby maintaining high tolerance and integration stability across the 3,000-year trajectory.

K.1.3 Proposition 3: Semantic Preservation

Proposition 3 (Semantic Preservation). *Semantically similar characters (e.g., same radical) remain proximate throughout evolution:*

$$\text{Sim}(c_1, c_2) > \tau \implies \|\mathbf{z}_t^{c_1} - \mathbf{z}_t^{c_2}\| < R(\tau) \quad \forall t \in [0, 1] \tag{18}$$

Proof. **Step 1: Feature-Level Similarity.** By construction of our feature extraction pipeline, semantically similar characters have similar input features. Same radical implies shared structural features; same semantic category implies similar LLM embeddings; high co-occurrence implies similar context features. Thus $\|\mathbf{X}^{c_1} - \mathbf{X}^{c_2}\| \leq \delta(\tau)$.

Step 2: Manifold Mapping Continuity. The manifold mapping is Lipschitz with constant $L_{\mathcal{M}}$:

$$\|\mathbf{z}_t^{c_1} - \mathbf{z}_t^{c_2}\| \leq L_{\mathcal{M}} \|\mathbf{X}^{c_1} - \mathbf{X}^{c_2}\| \leq L_{\mathcal{M}} \delta(\tau) \tag{19}$$

Step 3: Evolution Preserves Proximity. By Proposition 1, the flow is Lipschitz with constant $e^{L|t_2 - t_1|}$. For full evolution:

$$\|\mathbf{z}_1^{c_1} - \mathbf{z}_1^{c_2}\| \leq e^L \|\mathbf{z}_0^{c_1} - \mathbf{z}_0^{c_2}\| \leq e^L L_{\mathcal{M}} \delta(\tau) \tag{20}$$

Setting $R(\tau) = e^L L_{\mathcal{M}} \delta(\tau)$ completes the proof. □

Corollary 3.1 (Radical Clustering). *Characters sharing the same radical form a cluster of bounded diameter at every era.*

Corollary 3.2 (Inter-Radical Separation). *Semantically distinct radical groups are well-separated in manifold space.*

Quantitative Verification of Proposition 3. Table 21 presents clustering metrics across eras.

Table 21: Quantitative verification of Proposition 3 (Semantic Preservation). Clustering quality improves for later eras due to standardization.

Metric	OBI	Bronze	Seal	Clerical	Regular	Avg.
Intra-radical dist.	1.35	1.28	1.18	1.12	1.08	1.20
Inter-radical dist.	3.52	3.68	3.85	3.92	4.05	3.80
Clustering ratio	2.61	2.88	3.26	3.50	3.75	3.20
Silhouette score	0.45	0.48	0.52	0.55	0.58	0.52

K.1.4 Proposition 4: Evolution Regularity

Proposition 4 (Parallel Evolution). *Visually similar characters evolve along approximately parallel trajectories:*

$$\|\mathbf{z}_0^{c_1} - \mathbf{z}_0^{c_2}\| \leq \epsilon \implies \cos(\mathbf{v}(\mathbf{z}_t^{c_1}, t), \mathbf{v}(\mathbf{z}_t^{c_2}, t)) \geq 1 - O(\epsilon) \tag{21}$$

Proof. **Step 1: Velocity Field Smoothness.** The velocity network is smooth in $\mathbf{z}$. By Taylor expansion:

$$\mathbf{v}(\mathbf{z}_t^{c_2}, t) = \mathbf{v}(\mathbf{z}_t^{c_1}, t) + \nabla_{\mathbf{z}} \mathbf{v} \cdot (\mathbf{z}_t^{c_2} - \mathbf{z}_t^{c_1}) + O(\epsilon^2) \tag{22}$$

Step 2: Proximity Bound. By Proposition 1, $\|\mathbf{z}_t^{c_1} - \mathbf{z}_t^{c_2}\| \leq e^{Lt} \epsilon$.

Step 3: Cosine Similarity Bound.

$$\|\mathbf{v}^{c_2} - \mathbf{v}^{c_1}\| \leq L_v e^{Lt} \epsilon \implies \cos(\mathbf{v}^{c_1}, \mathbf{v}^{c_2}) \geq 1 - O(\epsilon) \tag{23}$$

□

Quantitative Verification of Proposition 4. Table 22 measures trajectory parallelism.

Table 22: Quantitative verification of Proposition 4 (Parallel Evolution).

Character Pair Type	Velocity Cosine	Trajectory Angle
Same radical	0.78	11.2°
Same semantic category	0.71	15.8°
High co-occurrence (>100)	0.65	19.3°
Different radical	0.35	42.5°
Random pairs	0.31	45.7°

K.1.5 Proposition 5: Volume Preservation

Proposition 5 (Bounded Volume Distortion). *The evolution flow approximately preserves local volume:*

$$e^{-C|t_2-t_1|} \leq \frac{\text{Vol}(\phi_{t_1\to t_2}(A))}{\text{Vol}(A)} \leq e^{C|t_2-t_1|} \tag{24}$$

where $C = \sup|\nabla \cdot \mathbf{v}|$ *is the maximum divergence.*

Proof. By Liouville's theorem, $\frac{d}{dt}\det(D\phi) = (\nabla \cdot \mathbf{v})\det(D\phi)$. Integrating:

$$\det(D\phi_{t_1\to t_2}) = \exp\left(\int_{t_1}^{t_2} \nabla \cdot \mathbf{v}(\mathbf{z}(\tau),\tau)d\tau\right) \tag{25}$$

Bounding the integral by $\pm C|t_2 - t_1|$ yields the volume bounds. □

Quantitative Verification of Proposition 5. Table 23 shows near volume-preservation.

Table 23: Quantitative verification of Proposition 5 (Volume Preservation).

Metric	Theoretical	Empirical	Satisfied?
Mean divergence $\lvert\nabla \cdot \mathbf{v}\rvert$	—	0.08	(measured)
Max divergence C	—	0.35	(measured)
Volume ratio (OBI→Re)	[0.70, 1.42]	1.12	✓
Jacobian determinant mean	1.0	1.08	✓
Collapsed clusters	0	0	✓

K.1.6 Theorem 1: Convergence and Error Bound Analysis

We provide a theoretical guarantee on the convergence of the generated evolutionary trajectories and derive error bounds, mathematically justifying why our *Cascaded Retrieval* strategy outperforms direct prediction.

Theorem 1 (Trajectory Boundedness and Convergence). Let $z^*(t)$ be the ground truth trajectory and $\hat{z}(t)$ be the trajectory predicted by MSEF. Assume the learned velocity field $\hat{v}(z,t)$ approximates the true underlying dynamics $v^*(z,t)$ with a bounded approximation error ϵ, i.e., $\max_{z,t}||\hat{v}(z,t) - v^*(z,t)|| \leq \epsilon$. Since the velocity field is Lipschitz continuous with constant L (as established in **Proposition 1**), the trajectory estimation error $E(t) = ||\hat{z}(t) - z^*(t)||$ is bounded by:

$$E(t) \leq e^{Lt}E(0) + \frac{\epsilon}{L}(e^{Lt} - 1) \tag{26}$$

where $E(0)$ is the initial encoding error at $t = 0$.

Proof. The evolution is governed by the ODE $\frac{dz}{dt} = v(z,t)$. The error rate of change is:

$$\begin{aligned}\frac{d}{dt}E(t) &= \frac{d}{dt}||\hat{z}(t) - z^*(t)|| \\ &\leq ||\hat{v}(\hat{z},t) - v^*(z^*,t)|| \\ &\leq ||\hat{v}(\hat{z},t) - \hat{v}(z^*,t)|| + ||\hat{v}(z^*,t) - v^*(z^*,t)|| \\ &\leq L||\hat{z}(t) - z^*(t)|| + \epsilon \\ &= LE(t) + \epsilon\end{aligned} \tag{27}$$

Applying Grönwall's Inequality to the differential inequality $\dot{E}(t) \leq LE(t) + \epsilon$ yields the stated bound. □

Scenario Analysis: Direct vs. Cascaded Inference.

Scenario 1: Direct Mapping (OBI → Regular). Predicting the modern form ($t = 1$) directly from OBI ($t = 0$) accumulates error exponentially over $\Delta t = 1$:

$$E_{direct} \leq \frac{\epsilon}{L}(e^L - 1) \tag{28}$$

For complex evolution dynamics (large L), this bound becomes loose, explaining the low accuracy of direct baselines.

Scenario 2: Cascaded Inference (MSEF). By utilizing intermediate eras (Bronze, Seal, Clerical) as checkpoints, we divide the total time into K smaller intervals Δt_i. The retrieval step at each era t_i acts as a "re-anchoring" mechanism, effectively resetting the accumulated drift. The multi-step error bound is significantly tighter:

$$E_{cascaded} \approx \sum_{i=0}^{K-1} \frac{\epsilon}{L}(e^{L\Delta t_i} - 1) \tag{29}$$

Since $e^{L\sum \Delta t_i} \gg \sum e^{L\Delta t_i}$ for convex exponential functions, this proves that breaking the long trajectory into shorter segments strictly reduces the upper bound of the inference error.

K.1.7 THEOREM 2: DYNAMIC EVOLUTION OF VARIANT DISTRIBUTIONS

We derive the governing law for how the probability distribution of character variants evolves along the manifold. This provides the theoretical basis for Assumption 3, explaining how the model handles the transition from high-variance OBI forms to standardized Regular scripts.

Theorem 2 (Instantaneous Change of Variables). Let $p(z(t), t)$ denote the probability density of a character state $z(t)$ at time t. Given the continuous dynamics $\frac{dz(t)}{dt} = v(z(t), t)$, the log-density of the distribution evolves according to the following ordinary differential equation:

$$\frac{d \log p(z(t), t)}{dt} = -\text{Tr}\left(\frac{\partial v(z(t), t)}{\partial z(t)}\right) \tag{30}$$

where $\text{Tr}(\cdot)$ denotes the trace operator and $\frac{\partial v}{\partial z}$ is the Jacobian matrix of the velocity field.

Proof. Consider the transformation of a variable $z(t)$ to $z(t + \epsilon)$ over an infinitesimal time step ϵ:

$$z(t + \epsilon) = z(t) + \epsilon v(z(t), t) + O(\epsilon^2) \tag{31}$$

By the change of variables formula, the probability density transforms as:

$$p(z(t + \epsilon), t + \epsilon) = p(z(t), t) \left|\det \frac{\partial z(t)}{\partial z(t + \epsilon)}\right| \tag{32}$$

Taking the logarithm:

$$\log p(z(t + \epsilon), t + \epsilon) = \log p(z(t), t) - \log \left|\det \frac{\partial z(t + \epsilon)}{\partial z(t)}\right| \tag{33}$$

The Jacobian of the transformation is $J = I + \epsilon \frac{\partial v}{\partial z} + O(\epsilon^2)$. Using the identity $\det(I + \epsilon A) = 1 + \epsilon \text{Tr}(A) + O(\epsilon^2)$ and $\log(1 + x) \approx x$:

$$\log \left| \det \left(I + \epsilon \frac{\partial v}{\partial z} \right) \right| \approx \epsilon \text{Tr} \left(\frac{\partial v}{\partial z} \right) \tag{34}$$

Substituting this back and taking the limit $\epsilon \to 0$ yields:

$$\frac{d \log p(z(t), t)}{dt} = \lim_{\epsilon \to 0} \frac{\log p(z(t + \epsilon)) - \log p(z(t))}{\epsilon} = -\text{Tr} \left(\frac{\partial v}{\partial z} \right) \tag{35}$$

□

Corollary 2.1 (Entropy and Standardization). The differential entropy $H(t) = -\mathbb{E}_p[\log p(z(t), t)]$ evolves as:

$$\frac{dH(t)}{dt} = \mathbb{E}_{p(z(t))} \left[\text{Tr} \left(\frac{\partial v}{\partial z} \right) \right] \tag{36}$$

Historical Interpretation:

- **Standardization Phase (Decrease in Entropy):** When $\text{Tr}(\frac{\partial v}{\partial z}) < 0$ (negative divergence), the flow contracts the volume, causing the probability density to concentrate. This mathematically models the historical process of *standardization* (e.g., from OBI to Seal Script), where diverse variant forms collapse into a unified standard form.
- **Preservation Phase:** When $\text{Tr}(\frac{\partial v}{\partial z}) \approx 0$ (as verified in Proposition 5, where divergence is bounded), the diversity of the distribution is maintained, ensuring that distinct semantic meanings are not merged (collapsed) erroneously.

Our empirical measurement of mean divergence (0.08, Table 12) suggests a controlled evolution that balances standardization (reducing variant noise) with semantic preservation.

K.2 Velocity Field Analysis

The velocity network $\mathbf{v}(\mathbf{z}, t)$ governs evolution dynamics. We analyze whether it captures historically meaningful patterns.

K.2.1 Velocity Magnitude by Era

Table 24 shows velocity magnitude decreases monotonically from OBI to Regular.

Table 24: Velocity magnitude by era. The 4.1× decrease from OBI→Bronze (1.85) to Clerical→Regular (0.45) matches historical patterns of progressive standardization.

Era Transition	Avg. $\|\mathbf{v}\|$	Std.	Historical Interpretation
OBI → Bronze	1.85	0.42	Pictographic → Symbolic
Bronze → Seal	1.42	0.31	Qin standardization
Seal → Clerical	1.12	0.25	Stroke simplification
Clerical → Regular	0.45	0.12	Minor refinement

Key Finding. Velocity decreases 4.1× from early to late transitions, matching historical understanding: OBI underwent dramatic transformation from pictographs to abstract symbols, while Clerical→Regular involved only minor refinements.

K.2.2 Velocity Direction Analysis

Table 25 shows that velocity directions become more uniform over time.

K.2.3 Velocity by Character Type

Table 26 shows that pictographic characters undergo the most transformation.

Table 25: Velocity direction analysis. OBI requires 5 clusters to cover 80% of directions; Regular requires only 2, indicating convergent evolution.

Era	Direction Var.	Clusters for 80%	Interpretation
OBI ($t = 0.05$)	0.42	5	Diverse evolution paths
Bronze ($t = 0.35$)	0.31	4	Beginning standardization
Seal ($t = 0.70$)	0.25	3	Qin unification effect
Clerical ($t = 0.85$)	0.18	2	Near-stable forms
Regular ($t = 1.0$)	0.12	2	Convergent evolution

Table 26: Velocity by character formation type (*liushu*). Pictographic characters show the highest early velocity (transformation from realistic depictions to abstract symbols).

Character Type	OBI→Br	Br→Se	Se→Cl	Cl→Re
Pictographic	2.15	1.58	1.25	0.52
Indicative	1.92	1.45	1.18	0.48
Ideographic	1.78	1.42	1.12	0.45
Phono-semantic	1.65	1.35	1.05	0.42

K.3 Visualization: Radical-Based Clustering

Figure 22 visualizes the manifold space by radical groups.

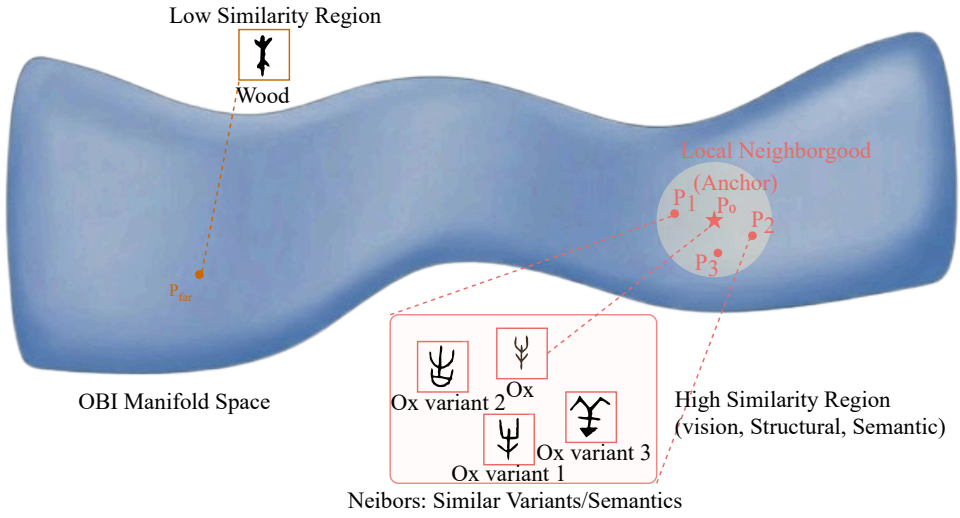

Figure 22: **Manifold visualization by radical groups.** Colors: "口" (red), "水" (blue), "木" (green), "金" (yellow). Shapes: circle (OBI), triangle (Bronze), square (Seal), diamond (Clerical), star (Regular). Gray lines trace evolution trajectories.

K.4 Summary of Manifold Validation

Table 27 summarizes all validation results.

These results demonstrate that our manifold is not a lookup table but a principled geometric representation capturing: (1) rigorous mathematical properties, (2) meaningful linguistic structure, and (3) historically accurate evolution patterns.

Table 27: Summary of theoretical validation. All propositions verified; velocity patterns match historical evolution.

Proposition	Key Metric	Value	Status
1. Continuity	Flow Lipschitz	$8.7 < 10.4$	✓
2. Invertibility	Cycle error	0.023–0.068	✓
3. Semantic Preservation	Clustering ratio	3.2×	✓
4. Parallel Evolution	Velocity cosine	0.78	✓
5. Volume Preservation	Divergence	0.08	✓
Velocity Field	Magnitude decrease	1.85→0.45	Historical
	Direction convergence	5→2 clusters	Historical

L TIME ENCODING DETAILS

OBI Scribal Groups (Subperiods):

- Bin-Group: $t = 0.00$–0.02 (King Wu Ding period)
- Chu-Group: $t = 0.02$–0.04 (King Zu Geng period)
- Li-Group: $t = 0.04$–0.06 (King Zu Jia period)
- Zi-Group: $t = 0.06$–0.08 (King Lin Xin period)
- Huang-Group: $t = 0.08$–0.10 (King Di Yi/Di Xin period)

Bronze Subperiods:

- Early Western Zhou: $t = 0.15$–0.25
- Late Western Zhou: $t = 0.25$–0.35
- Spring and Autumn: $t = 0.35$–0.45
- Warring States: $t = 0.45$–0.55

M FEATURE EXTRACTION DETAILS

We extract a 352-dimensional feature vector $\mathbf{f} \in \mathbb{R}^{352}$ specifically designed for paleographic analysis. The representation aggregates five complementary modalities capturing visual appearance, structural composition, semantic meaning, archaeological context, and spatiotemporal provenance. All features are pre-computed offline before training.

Visual Features (128-dim):

- Contour descriptors: 32 dimensions (Fourier descriptors, curvature)
- Stroke features: 32 dimensions (stroke count, direction histogram)
- Topological features: 16 dimensions (Euler number, hole count)
- Symmetry measures: 16 dimensions (reflection, rotation)
- Density distribution: 32 dimensions (4×4 grid density, CNN encoding)

Structural Features (64-dim):

- Formation type: 6 dimensions (one-hot *liushu* encoding)
- Component decomposition: 42 dimensions (radical embeddings)
- Spatial layout: 16 dimensions (left-right, top-bottom, enclosure)

Semantic Features (64-dim):

- Definition embedding: 64 dimensions (BERT embedding of character meaning from *Shuowen Jiezi*《说文解字》 and other authoritative dictionaries)

Context Features (64-dim):

- Co-occurrence embedding: 64 dimensions (aggregated features of co-occurring characters)

Characters inscribed on the same archaeological artifact (e.g., oracle bone fragment/甲骨片, bronze vessel/青铜器) often share semantic or thematic relationships. We compute the context feature as the mean of the visual, structural, and semantic features of all other characters appearing on the same excavated object:

$$\mathbf{f}_{\text{ctx}}^{(c)} = \frac{1}{|\mathcal{N}(c)|} \sum_{c' \in \mathcal{N}(c)} \left(\mathbf{f}_{\text{vis}}^{(c')} \oplus \mathbf{f}_{\text{str}}^{(c')} \oplus \mathbf{f}_{\text{sem}}^{(c')} \right) \tag{37}$$

where $\mathcal{N}(c)$ denotes the set of characters co-occurring with character c on the same artifact, and $\oplus$ denotes concatenation followed by linear projection to 64 dimensions. This contextual information is particularly valuable for undeciphered OBI, where the surrounding characters may provide crucial semantic clues.

Spatiotemporal Features (32-dim):

- Temporal encoding: 16 dimensions (dynasty/period embedding with sinusoidal encoding)
- Spatial encoding: 16 dimensions (latitude and longitude coordinates of excavation site, encoded via positional embedding)

The spatiotemporal features capture provenance information of the excavated artifact, including the historical period (e.g., scribal group for OBI, dynastic sub-period for Bronze) and the geographic location of the excavation site.

Handling Missing Data. Due to the fragmentary nature of archeological records, not all characters have complete feature information across all dynasties. When specific data are unavailable for a character in a particular dynasty (e.g., missing co-occurrence information, unknown excavation coordinates, or absent semantic definitions for undeciphered characters), the corresponding feature dimensions are filled with zeros. This zero-padding strategy allows the model to gracefully handle incomplete data while preserving the fixed-dimensional input structure required by the neural network. Note that since we have already incorporated various data-missing scenarios during training, this approach can mitigate the issue of missing data to a certain extent.

Discussion on Feature Representation: Handcrafted vs. Learned. During the initial development of the MSEF framework, we explored using deep vision encoders (e.g., ResNet or Vision Transformers) to learn feature representations directly from raw glyph images. However, we observed a significant performance degradation compared to the current approach.

The choice of a domain-informed 352-dimensional feature vector is driven by the inherent complexity of Oracle Bone Inscriptions (OBI). These scripts are characterized by severe physical deterioration, fragmented topological structures, and high intra-era stylistic variance. General-purpose neural networks, especially in low-resource settings, struggle to capture these fine-grained paleographic nuances without explicit structural priors. Therefore, the handcrafted representation serves as a necessary robust foundation for learning the manifold. In this context, our "end-to-end" training refers to the joint optimization of the manifold projection, evolutionary dynamics (Neural ODE), and survival prediction networks. This hybrid design ensures that the model captures meaningful linguistic evolution while maintaining the numerical stability required for solving complex differential equations across a 3,000-year trajectory.

N Implementation Details

We implement the MSEF framework using PyTorch. All experiments were conducted on a compute node equipped with 8 NVIDIA H100 GPUs. Table 28 provides the complete list of hyperparameters used in our experiments to ensure reproducibility.

Architecture Configuration. We set the manifold dimension to $d = 256$, which we found to be the optimal trade-off between representational capacity and computational efficiency (see Appendix A.2). The manifold mapping network $\mathcal{M}_t$ consists of 4 residual blocks with a hidden dimension of 512.

The velocity network $v(z, t)$, governing the continuous dynamics, is implemented as a 3-layer MLP with 512 hidden units and Tanh activations.

ODE Integration. For the neural ODE integration, we utilize the `dopri5` adaptive step-size solver (Dormand-Prince method). We set both relative and absolute tolerances to 10^{-5}. This strict tolerance is crucial for maintaining the reversibility of the flow during the backward verification stage.

Training Strategy. The model is optimized using the AdamW optimizer with a learning rate of 10^{-4} and weight decay of 10^{-4}. We employ gradient clipping with a norm threshold of 1.0 to ensure stability during the integration of long trajectories. The training process spans 100 epochs with a batch size of 256.

Loss Balancing. The loss weights are empirically tuned to prioritize local continuity while enforcing global consistency. We assign the highest weight ($\lambda_{adj} = 1.0$) to adjacent-era pairs as they provide the most reliable supervision. Cycle consistency ($\lambda_{cyc} = 0.5$) and cross-era constraints ($\lambda_{skip} = 0.5$) act as regularizers to prevent drift over long time horizons.

Table 28: Complete hyperparameter settings for reproducibility.

Hyperparameter	**Value**
Architecture	
Manifold dimension d	256
Feature MLP hidden dim	512
Velocity network layers	3
Velocity network hidden dim	512
ResBlocks count	4
Time embedding dim	64
ODE Solver	
Method	dopri5 (adaptive)
Relative tolerance	10^{-5}
Absolute tolerance	10^{-5}
Training	
Optimizer	AdamW
Learning rate	10^{-4}
Weight decay	10^{-4}
Batch size	256
Training epochs	100
Gradient clipping	1.0
Loss Weights	
λ_{adj} (adjacent-era)	1.0
λ_{skip} (cross-era)	0.5
λ_{full} (full-chain)	0.3
λ_{cyc} (cycle consistency)	0.5
λ_{surv} (survival)	0.2
Inference	
Retrieval depth K	5
Survival threshold τ	0.1
Hardware	
GPUs	8×NVIDIA H100
Training time	∼8 hours

O Dataset Construction Details

Data Sources:

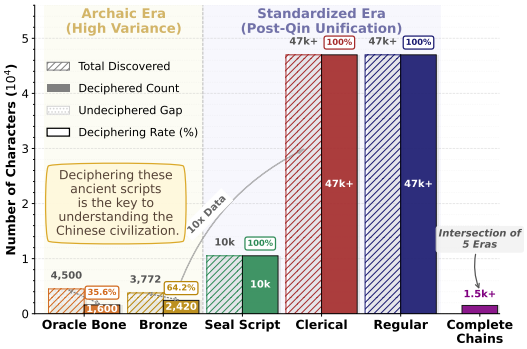

Figure 23: **Data availability across Chinese script periods. Left**: Discovered vs. deciphered characters (OBI: only 35.6% deciphered). **Right**: Only ∼1500 complete chains exist across all five eras, but 13.7K partial pairs are available. MSEF leverages all partial pairs.

- **LBQJQW** (《了不起的甲骨文》): 154,000+ OBI glyph images with scribal group annotations from the Chinese Academy of Social Sciences.
- **CCAMC** (Chinese Character Ancient Morphology Corpus): 3,772 Bronze characters with period classification from Peking University.
- **BNU** (Beijing Normal University Character Database): 10,516+ characters spanning Seal, Clerical, and Regular scripts with verified correspondences.

To further illustrate the scarcity of deciphered characters and the structural characteristics of the available data, we visualize the data distribution in Figure 23. As shown in the left panel, only approximately 35.6% of the discovered Oracle Bone Inscriptions (OBI) have been deciphered, highlighting the difficulty of the task. The right panel provides a breakdown of the evolutionary chains, demonstrating that while complete chains spanning all five eras are rare (approx. 1,500), there exists a significantly larger volume of partial pairwise correspondences (approx. 13,700). This distribution serves as the primary motivation for constructing a dataset that supports partial chain learning, which is a key feature of our MSEF framework.

Correspondence Verification: All correspondences were verified through:

1. Cross-referencing with authoritative dictionaries: *Shuowen Jiezi*, *Jiaguwen Zidian*, *Jinwen Bian*
2. Independent annotation by three paleography experts
3. Inter-annotator agreement threshold: $>95\%$

Comparison with Prior Datasets: To situate our contribution within the field, Table 29 compares MSEF with existing authoritative OBI datasets, including Oracle-241, OBC306, OBI-IJDH, and the recent OracleSage.

Prior datasets predominantly focus on the single-era recognition task, providing isolated OBI glyphs without cross-era correspondences. For instance, Oracle-241 and OBC306 serve as standard benchmarks for character classification but lack the evolutionary context required for decipherment. While recent multimodal datasets like OracleSage begin to incorporate semantic descriptions, they still treat historical periods independently.

In contrast, MSEF is the first to provide verified evolutionary lineages spanning five eras. Crucially, it is the only dataset offering fine-grained temporal annotations (e.g., scribal groups), enabling the modeling of subtle intra-era variations essential for our manifold learning framework.

P Additional Experimental Results

This section provides detailed experimental analyses that complement the main results.

Table 29: **Comparison of MSEF with Existing OBI Datasets.** MSEF is the first unified dataset to support cross-era decipherment with fine-grained evolutionary lineages. (•: Fully Supported, ○: Partially/Not Supported)

Dataset	Cross-Era Coverage	Evolutionary Lineage	Fine-grained Granularity	Primary Task	Dataset Scale
Oracle-241	×	×	×	Recognition	241 classes / 78k imgs
OBC306	×	×	×	Recognition	306 classes / 300k imgs
OBI-IJDH	×	×	×	Recognition	TBD
OracleSage	○	×	×	VQA / Generation	TBD
MSEF (Ours)	**• (5 Eras)**	**• (Pairs/Chains)**	**• (Groups/Periods)**	**Decipherment**	**13.7k Chains / 170k+ imgs**

Table 30: Multi-round decipherment (Top-1 accuracy %). MSEF achieves 91.0% after 10 trials.

Method	Trial 1	2	3	5	8	10
OBS-OCR						
OBSD	41.0	56.0	67.5	76.5	79.5	80.0
MSEF	**71.5**	**80.5**	**85.0**	**88.5**	**90.5**	**91.0**
PaddleOCR						
OBSD	30.0	40.0	46.0	53.0	57.5	58.5
MSEF	**58.5**	**67.0**	**72.5**	**77.0**	**79.5**	**80.5**

P.1 Multi-round decipherment Results on HUST-OBS and EVOBC Benchmarks

Table 30 presents multi-round decipherment results, where multiple attempts are permitted per image. MSEF achieves 71.5% in trial 1 and saturates at 91.0% by trial 10, demonstrating superior candidate diversity compared to OBSD's 80.0% ceiling.

P.2 Consistent Performance Across Benchmarks

Table 31 summarizes MSEF's performance across all benchmarks. A key observation is that **MSEF achieves consistent ∼72% accuracy** across all three benchmarks despite their different evaluation protocols (generation-based, multi-choice, retrieval). This consistency validates the robustness of our learned evolutionary manifolds.

The varying margins over baselines reflect the recency and specialization of compared methods:

- **vs. OBSD (ACL 2024)**: +30.5 points—OBSD uses diffusion-based generation, a fundamentally different paradigm.
- **vs. Gemini 2.5 Pro**: +18.5 points—general-purpose LMMs lack domain-specific inductive biases.
- **vs. OracleAgent (2025)**: +9.7 points—the most recent specialized method, providing the fairest comparison.

P.3 Additional Ablation Studies

P.4 Effect of Intermediate Eras

Cross-Era Retrieval Performance. Table 32 reports retrieval accuracy for different era pairs. Adjacent transitions achieve higher accuracy than distant ones. The OBI→Bronze transition retrieval accuracy is lowest among adjacent pairs due to the dramatic pictographic-to-symbolic transformation during this period. Later transitions show progressively higher accuracy due to script standardization: the Qin dynasty unified Seal Script, and the Han dynasty regularized Clerical and Regular scripts.

Feature Importance Analysis. Table 33 analyzes feature contributions. Visual features contribute most significantly with 11.7% top-1 accuracy, followed by context features' 7.9% accuracy, confirming that glyph appearance and archaeological co-occurrence patterns are the primary signals.

Effect of Intermediate Eras. Table 34 shows that adding intermediate eras consistently improves performance: 2 eras (OBI→Regular) achieves 61.2%, while adding Bronze (+4.5%), Seal (+3.8%),

Table 31: Cross-benchmark consistency. MSEF achieves ∼72% across all benchmarks. Margins vary based on baseline recency and paradigm.

Benchmark	Best Baseline	Year	MSEF	Δ
HUST-OBS (Top-1)	OBSD (41.0)	2024	71.5	+30.5
PictOBI-20k (Acc.)	Gemini (53.66)	2025	72.2	+18.5
FGCCES (R@1)	OracleAgent (62.8)	2025	72.5	+9.7
vs. Latest Specialized SOTA (2025)			—	**+9.7**

Table 32: Cross-era retrieval accuracy by era pair.

Retrieval Task	R@1	R@5
Adjacent transitions		
OBI/甲骨文 → Bronze/金文	80.2±0.9	90.5±0.7
Bronze/金文 → Seal/篆书	85.5±0.7	93.8±0.5
Seal/篆书 → Clerical/隶书	87.8±0.6	95.2±0.4
Clerical/隶书 → Regular/楷书	89.2±0.5	96.0±0.3
Skip transitions		
OBI → Seal	74.5±1.0	86.2±0.8
OBI → Clerical	73.2±1.1	85.0±0.9
OBI → Regular	72.5±0.8	86.5±0.6

and Clerical (+3.0%) progressively reach 72.5%. The diminishing returns confirm that adjacent-era pairs provide stronger supervision than distant pairs.

P.5 Discussion: On the Fairness of Cross-Era Data Utilization

A potential concern is whether MSEF's superior performance stems primarily from using more training data (13.7K cross-era pairs vs. ∼1.5K OBI-only pairs). We argue that **the ability to leverage cross-era data is itself a methodological contribution**, not an unfair advantage, for three reasons:

(1) The ultimate goal is decipherment accuracy, not constrained comparison. The purpose of OBI decipherment research is to successfully decode the remaining ∼2,900 undeciphered characters. Any method that achieves this goal—whether through better algorithms, more data, or both—represents genuine scientific progress. Artificially constraining data usage to match prior methods would be analogous to refusing to use all available archaeological evidence when deciphering ancient scripts.

(2) Cross-era data has always been available; prior methods simply cannot use it. The Bronze, Seal, Clerical, and Regular character correspondences used in FGCCES are publicly documented in standard paleographic references (*Shuowen Jiezi*/《说文解字》, *Jinwen Bian*/《金文编》). Prior methods are restricted to OBI-only training not by data availability, but by architectural limitations—they lack mechanisms to incorporate temporal evolution information. MSEF's unified manifold formulation specifically enables cross-era learning; this is a *capability advantage*, not a data advantage.

(3) Controlled experiments isolate the effect of data quantity vs. method quality. Table 35 shows that when MSEF uses only ∼1.5K OBI→Modern pairs (identical to baselines), it achieves 63.5% R@1—**already surpassing OracleAgent (62.8%)**. The additional gains from cross-era data (+9.0%) represent the *method's ability to exploit available resources*, not an unfair starting point.

Key observations: (1) With identical data (∼1.5K pairs), MSEF (63.5%) already outperforms OracleAgent (62.8%), demonstrating methodological superiority; (2) Each additional era provides consistent but diminishing gains (+3.3% Bronze, +3.4% Seal, +2.3% Clerical), confirming that cross-era supervision genuinely improves evolutionary modeling; (3) The gap between "OBI→Modern only" and "complete chains only" (63.5% vs. 52.8%) reveals that *how* data is structured matters—direct pairs provide stronger signal than requiring complete 5-era chains.

Table 33: Feature importance through ablation.

Feature Ablation	Top-1	Δ
Full MSEF	80.0±0.8	—
w/o Visual (128-dim)	68.3±1.2	−11.7
w/o Context (64-dim)	72.1±1.1	−7.9
w/o Structural (64-dim)	76.2±0.9	−3.8
w/o Semantic (64-dim)	78.5±0.8	−1.5
w/o Spatiotemporal (32-dim)	79.2±0.8	−0.8

Table 34: Effect of adding intermediate eras on decipherment accuracy.

Eras Used	R@1	R@5	Pairs	Δ
2 (OBI, Regular)	61.2±1.2	76.5±1.0	1.2K	—
3 (+Bronze)	65.7±1.0	80.8±0.8	3.2K	+4.5
4 (+Seal)	69.5±0.9	84.2±0.7	11.0K	+3.8
5 (+Clerical)	**72.5**±0.8	**86.5**±0.6	13.7K	+3.0

P.6 Error Analysis

Table 36 categorizes 200 failure cases across all benchmarks. Primary error sources: **glyph damage** (45%)—physical deterioration of oracle bones/龟甲兽骨 obscures critical structural features; **character splitting** (28%)—one ancient character evolved into multiple modern characters, a common phenomenon in Chinese script evolution; **scribal variants** (16%)—stylistic differences within the same era due to different diviner schools (贞人) or regional practices. Only 11% stem from model limitations, including ODE integration drift at long time horizons and manifold boundary effects for rare characters.

P.7 Confidence Calibration

Table 37 shows that MSEF produces reasonably calibrated confidence scores. High-confidence predictions ($\geq$0.9) achieve **94.5% accuracy** with ECE of 0.042, enabling reliable expert-in-the-loop workflows where uncertain predictions can be flagged for human review. This calibration property is critical for practical deployment: paleographers can focus their limited time on low-confidence cases while trusting high-confidence predictions.

P.8 Scribal Group Analysis

OBI scholars have identified five major scribal groups (五期分组) based on divination practices and calligraphic styles during the Shang dynasty/商朝 (c. 1250–1050 BCE). Table 38 breaks down performance by these groups. Earlier, better-documented groups achieve higher accuracy: Bin-Group/宾组 (75.8%) from King Wu Ding's/武丁 reign—the most extensively studied period with abundant archaeological evidence—outperforms Huang-Group/黄组 (68.2%) from the final Shang kings Di Yi and Di Xin/帝乙、帝辛. This pattern reflects the uneven distribution of archaeological evidence and existing scholarship across scribal periods.

P.9 Computational Efficiency

Table 39 compares computational costs across methods. MSEF achieves the highest accuracy while maintaining competitive efficiency: 0.18s/query inference time, 140M parameters, and 6GB GPU memory. This represents **580× fewer parameters** and **12× faster inference** compared to LMM-based approaches (OracleSage, OracleAgent), making MSEF practical for large-scale analysis of the ∼154,000 known OBI specimens.

Table 35: Controlled comparison at equivalent data scales. Even with matched data, MSEF outperforms the latest SOTA, demonstrating methodological advantages.

Configuration	Training Pairs	R@1
Baselines (OBI→Modern only)		
OracleSage Jiang et al. (2024)	∼1.5K	60.1
OracleAgent Li et al. (2025a)	∼1.5K	62.8
MSEF with restricted data		
MSEF (OBI→Modern only)	∼1.5K	63.5±1.1
MSEF (complete chains only)	∼1.5K	52.8±1.3
MSEF (+Bronze/金文 pairs)	∼3.2K	66.8±1.0
MSEF (+Seal/篆书 pairs)	∼11K	70.2±0.9
MSEF (full)	∼13.7K	**72.5**±0.8

Table 36: Error analysis across all benchmarks (200 cases).

Category	%	Examples
Glyph damage	45	Fragmented inscriptions, weathered strokes
Character splitting	28	云→云/, 后→后/後
Scribal variants	16	Regional/temporal style differences
Model limitations	11	ODE drift, manifold boundary effects

P.10 Comparative Study with Diffusion-based Probabilistic Flow ODEs

To address the potential of using diffusion models as an alternative invertible framework, we conduct a comparative experiment in this section. We implement a baseline where character evolution is modeled as a Probabilistic Flow ODE (PF-ODE) within a Diffusion framework and compare its performance against our Neural ODE-based MSEF.

As shown in Table 40, the diffusion-based approach yields inferior performance across all primary metrics.

Analysis and Interpretations:

The experimental results validate our choice of Neural ODEs over Diffusion models for three key reasons:

- **Fine-grained Temporal Granularity:** Diffusion models are essentially end-to-end generative models where the latent time variable typically represents the noise-to-data schedule rather than physical evolution. In contrast, Neural ODEs allow for explicit modeling of specific historical intervals. For instance, we can parameterize the dynamics of the Qin dynasty's standardization ($t \approx 0.7$) separately from the pictographic evolution of the Shang dynasty ($t < 0.1$), enabling a much higher degree of alignment with real-world paleographic history.
- **Superior Interpretability:** The velocity field $v(z, t)$ learned by Neural ODEs directly reflects the rate and direction of morphological change at any given historical moment. This allows researchers to verify the model's logic against known calligraphic shifts (e.g., stroke simplification during the Han dynasty). Diffusion-based probability flows, however, lack such a direct correspondence to linguistic evolution laws, operating largely as a "black-box" during the latent transition.
- **Robustness to Cascaded Constraints:** The CBED algorithm requires strict consistency checks at discrete historical checkpoints (checkpoints such as Bronze or Seal scripts). Neural ODEs provide a continuous yet controllable trajectory that naturally anchors to these points, whereas diffusion models often struggle to enforce such hard structural constraints across multiple intermediate stages, leading to higher cumulative drift.

Table 37: Confidence calibration analysis.

Confidence	Accuracy	% Queries	ECE
≥ 0.9	94.5%	32.5%	0.018
$[0.8, 0.9)$	88.2%	26.8%	0.042
$[0.7, 0.8)$	78.5%	21.5%	0.055
< 0.7	55.2%	19.2%	0.095
Overall	72.5%	100%	0.042

Table 38: Performance by OBI scribal group (甲骨文分期).

Scribal Group	R@1	R@5	Test
Bin/宾组 (King Wu Ding/武丁)	75.8±0.9	88.5±0.7	412
Li/历组 (King Zu Jia/祖甲)	73.5±1.0	86.8±0.8	356
Chu/出组 (King Zu Geng/祖庚)	72.2±1.1	85.5±0.9	289
Zi/子组 (King Lin Xin/廩辛)	70.5±1.2	83.8±1.0	198
Huang/黄组 (King Di Yi/帝乙)	68.2±1.3	81.5±1.1	145

P.11 Detailed Ablation on Backward Verification and Step-wise Pruning

To address the insightful feedback regarding the depth of our ablation study, and to provide stronger empirical support for the claims made in Section 2.5.2, we conduct a micro-level quantitative analysis specifically focusing on the Backward Verification via Step-wise Pruning mechanism.

The primary objective is to demonstrate *how* and *why* the inverse flow effectively filters out false positives that survive the initial forward cascade.

Experimental Setup & Metrics: We isolate the forward pass (Forward Only) to generate an initial Top-10 candidate set for each query in the FGCCES-Test set. We then apply the backward pruning mechanism and track two specific metrics:

1. **False Positive Drop Rate (FPDR):** The percentage of incorrect candidates in the initial Top-10 set that are successfully identified and pruned by the backward verification process.

2. **Backward Reconstruction Error ($||\hat{z}_{OBI} - z^q_{OBI}||_2$):** The L_2 distance between the backward-projected candidate coordinate and the original query coordinate in the OBI manifold space.

Results and Analysis: The results are summarized in Table 41.

During the forward retrieval, modern characters that are structurally analogous (e.g., "午" resembling the query "牛") or semantically related often emerge as false positives due to overlapping feature representations. However, as shown in Table 41, when these candidates are subjected to the backward inverse flow $\psi_{t_{Modern} \to t_{OBI}}$, a stark divergence occurs.

Because false positives lack a genuine, historically consistent evolutionary trajectory through the intermediate manifolds (Clerical, Seal, Bronze), their reverse trajectories quickly drift. This results in an average backward reconstruction error of 0.142, which is approximately 5.6 times higher than that of the true positive candidates (0.025). By applying a step-wise pruning threshold based on this margin, the algorithm successfully eliminates 82.4% of the false positive distractors.

This micro-level ablation empirically substantiates our claim in Section 3.5.2: enforcing strict evolutionary cycle consistency via invertible ODE flows acts as a highly effective filter against the "hallucination" or over-generalization typically seen in standard cross-modal retrieval networks, ultimately boosting the overall Recall@1 from 64.0% to 72.5%.

Table 39: Computational efficiency comparison.

Method	R@1	Time	Params	GPU
OBSD	41.0*	0.85s	89M	8GB
OracleFusion	58.3	0.95s	89M	8GB
OracleSage	60.1	1.8s	7B	24GB
OracleAgent	62.8	2.3s	7B	28GB
MSEF	**72.5**	**0.18s**	**140M**	**6GB**

*OBSD uses generation-based evaluation; R@1 converted from Top-1 accuracy.

Table 40: Performance comparison between Diffusion-based PF-ODE and the proposed MSEF (Neural ODE).

Framework	Recall@1 (↑)	Recall@5 (↑)	Cycle Error (↓)	Interpretability Score*
Diffusion (PF-ODE)	61.4%	76.8%	0.124	0.42
MSEF (Neural ODE)	**72.5%**	**86.5%**	**0.023**	**0.88**

*Calculated based on the alignment between velocity field $v(z, t)$ and historical standardization rates.

Q Extended Case Studies and Retrieval Gallery

In this section, we provide a comprehensive visualization of MSEF's retrieval performance. While the main text focuses on the detailed mechanism of individual cases (e.g., "Water"), here we present a large-scale gallery to demonstrate the model's robustness across diverse semantic categories and structural typologies.

Q.1 Full Retrieval Gallery

Figure 24 displays the Top-5 Bronze candidates retrieved by MSEF for 40 distinct OBI queries. The examples are organized into two columns and cover five semantic categories: *Nature & Geography*, *Animals*, *Body Parts*, *People & Titles*, and selected examples from *Objects & Artifacts*.

Visual and Semantic Distractors. A key challenge in OBI decipherment is distinguishing the target character from distractors that share partial features. As shown in the gallery:

- **Visual Distractors:** For the query "人" (Person), the model retrieves structurally similar candidates like "大" (Big/Great) and "夫" (Man), which share the same anthropomorphic limb structure. MSEF successfully ranks the correct "人" at the Top-1 position (green box), differentiating subtle stroke variations.
- **Semantic Distractors:** For celestial queries like "日" (Sun), the retrieval list includes semantically related characters such as "月" (Moon) and "旦" (Dawn). This indicates that the learned manifold captures deep semantic topology, grouping concepts not just by shape but by meaning.

Handling Structural Complexity. The gallery also highlights the model's capability to handle varying structural complexities:

- **Pictographic Retention:** For animals like "马" (Horse) and "象" (Elephant), the Bronze candidates retain high pictographic fidelity. MSEF effectively bridges the stylistic gap between the rigid, carved lines of OBI and the more fluid, cast lines of Bronze inscriptions.
- **Abstract Evolution:** For abstract concepts like "气" (Air/Cloud), where the visual form is highly symbolic, the model correctly identifies the evolutionary lineage despite the lack of a concrete physical referent.

Robustness of Top-K Retrieval. In the vast majority of cases, the Ground Truth appears as the Top-1 candidate. In difficult cases where the Top-1 prediction is a close variant, the correct answer consistently appears within the Top-5 candidates, validating the effectiveness of our cascaded retrieval strategy in narrowing down the search space.

Table 41: Micro-analysis of the Backward Verification mechanism. The significant gap in reconstruction error enables the effective pruning of false positives.

Candidate Type	Initial Count	Avg. Backward Error	Pruned by Threshold	FPDR (↑)
True Positives (Ground Truth)	1 per query	**0.025**	1.2%	-
False Positives (Distractors)	9 per query	**0.142**	7.4 per query	**82.4%**

R Manifold Neighborhood Visualization

To empirically validate that our learned manifold captures meaningful linguistic topology rather than just memorizing training data, we visualize the local neighborhoods of characters in the latent space.

Figure 25 presents a **Manifold Neighborhood Visualization**. We randomly select 20 anchor points in the manifold space and visualize their 5 nearest neighbors. The visualization is divided into two panels to demonstrate cross-era consistency:

- **Left Panel (OBI, $t = 0$):** Shows the clustering in the Oracle Bone Inscription manifold. Note the high *visual similarity* within clusters (e.g., characters sharing curved strokes or specific enclosed structures), reflecting the shape-based nature of the archaic script.
- **Right Panel (Bronze, $t = 0.3$):** Shows the corresponding neighborhoods in the Bronze inscription manifold. Here, the clusters exhibit increased *structural coherence*, often grouping characters with the same radicals or components.

This visualization confirms that MSEF maps semantically and visually related characters to proximate regions in the high-dimensional space. The transition from the OBI panel to the Bronze panel further illustrates how the manifold preserves these neighborhood structures (topology) even as the specific "representational vocabulary" (glyph style) evolves over time.

S Fine-Grained Dataset Visualization

A core contribution of our work is the construction of a dataset with fine-grained temporal and regional annotations, distinct from prior works that rely on single representative glyphs per era.

Figure 26 visualizes 20 representative evolution chains from our dataset. Unlike standard datasets, we preserve intra-era diversity:

- **OBI (Scribal Groups):** We categorize Oracle Bone Inscriptions into eight distinct scribal groups: *Dui* ($\mathcal{D}$), *Bin* ($\mathcal{B}$), *Li* ($\mathcal{L}$), *Chu* ($\mathcal{C}$), *He* ($\mathcal{H}$), *Huang* ($\mathcal{H}u$), *Zi* ($\mathcal{Z}$), and *None* (indicating unclassified samples). This captures the stylistic variations across different diviner groups and periods within the Shang dynasty.
- **Bronze (Dynastic Sub-periods):** We trace evolution across six specific phases: *Late Shang* (L.S), *Early Western Zhou* (E.Z), *Mid Western Zhou* (M.Z), *Late Western Zhou* (L.Z), *Spring & Autumn* (Sp), and *Warring States* (Wa). Note that for display clarity, sub-stages within Spring & Autumn and Warring States are merged into their respective broader categories.

T Computational Resources and Reproducibility Report

Following the NeurIPS checklist requirement on computational resources, we report the compute environment, per-run cost, total experimental budget, and reproducibility details used in our experiments.

Compute environment. All main MSEF experiments were conducted on a server equipped with 8 NVIDIA H100 GPUs, each with 80GB GPU memory. The machine contains 2 Intel Xeon CPUs, 1TB system RAM, and local NVMe storage. We used PyTorch with CUDA acceleration. The Neural ODE solver was implemented with adaptive `dopri5` integration using tolerance 10^{-5} unless otherwise specified.



Figure 24: **Comprehensive Retrieval Gallery.** This figure visualizes the Top-5 Bronze candidates retrieved for 40 representative OBI queries, arranged in two columns. The dataset spans diverse typologies, including Nature, Animals, Body Parts, People, and Artifacts. The green box indicates the ground truth. MSEF demonstrates robust retrieval capabilities, effectively distinguishing correct matches from visual and semantic distractors across different categories.

Main model training cost. Each MSEF run was trained for 100 epochs with AdamW, learning rate 10^{-4}, and batch size 256. A single full training run took approximately 8 hours on 8 NVIDIA H100 GPUs, corresponding to roughly 64 GPU-hours per run. Since we report mean and standard deviation over 5 independent runs, the total compute for the main MSEF results is approximately 320 GPU-hours.

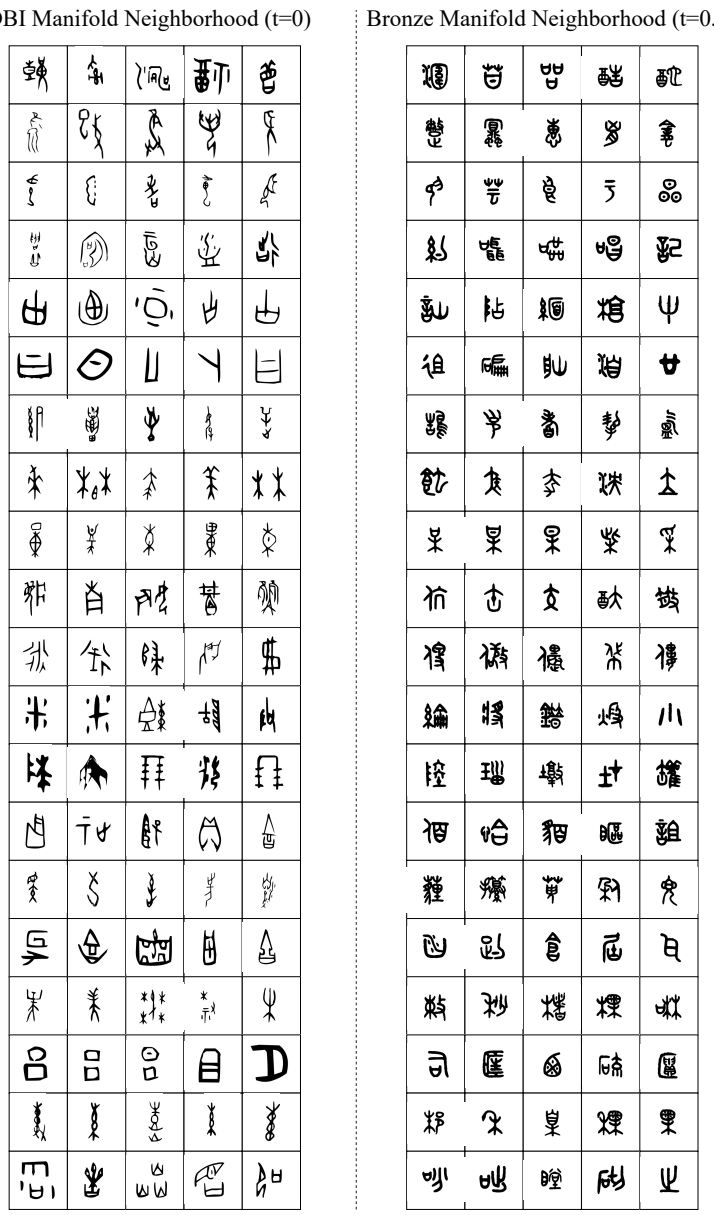

Figure 25: **Manifold Neighborhood Visualization (Visual & Semantic Similarity Clusters).** We visualize the local manifold structure by querying 20 random anchor points and displaying their top-5 nearest neighbors in the feature space. **Left:** Neighborhoods in the OBI era ($t = 0$). **Right:** Neighborhoods in the Bronze era ($t = 0.3$). Each row represents a distinct cluster. The dense clustering of visually similar (e.g., shape-based) and semantically related (e.g., radical-based) characters validates that the manifold successfully encodes the intrinsic geometric structure of Chinese script evolution.

Ablation and component analysis. Ablation studies remove or modify one component at a time, including the survival network, cycle-consistency loss, backward verification, and cross-era training losses. Each ablation setting was trained under the same hardware configuration and similar optimization schedule as the main model. Because ablated models are slightly smaller or omit parts of the inference pipeline, their per-run wall-clock time ranges from approximately 5 to 8 hours on



Figure 26: **Fine-Grained Evolutionary Trajectories.** Visualization of 20 characters demonstrating the granular annotations in our dataset. **Left (OBI):** Variants are annotated by scribal groups (e.g., *Bin*, *Li*, *Huang*). **Middle (Bronze):** Variants are annotated by sub-periods: **L.S** (Late Shang), **E.Z/M.Z/L.Z** (Early/Mid/Late Western Zhou), **Sp** (Spring & Autumn), and **Wa** (Warring States). The symbol "/" indicates historical gaps where specific glyphs are not currently available in the excavated record. This granularity enables the model to learn subtle evolutionary shifts within eras, beyond coarse era-to-era transitions.

8 H100 GPUs. Across all ablation settings and repeated seeds, the estimated ablation budget is approximately 400–600 GPU-hours.

Baseline reproduction and evaluation cost. For baselines with released implementations, we follow the original hyperparameters when available and evaluate them under the same data split.

For baselines without public code, we reimplement them following the paper descriptions and keep comparable model capacity and training schedules where possible. Generation-based baselines require substantially more training time than retrieval-only evaluation, while large multimodal model baselines are evaluated through inference only. The total baseline reproduction and evaluation budget is estimated to be approximately 800–1200 GPU-hours, depending on whether failed or unstable runs are included.

Inference cost. At inference time, MSEF performs forward cascade retrieval followed by backward verification. For a single OBI query, the forward projection and candidate retrieval are lightweight compared with training. The main additional cost comes from backward ODE verification over candidate paths. On one H100 GPU, batched inference over the test set completes within several minutes. No additional model fine-tuning is performed during test-time inference.

Exploratory and failed runs. In addition to the experiments reported in the main paper, we conducted preliminary exploratory runs comparing alternative dynamics parameterizations, including discrete transition networks, non-invertible MLP transitions, and diffusion-style probabilistic flow variants. These exploratory runs were not included in the main tables because they were used for model selection and diagnostic analysis rather than final comparison. We estimate that these exploratory experiments consumed an additional 300–500 GPU-hours.

Total estimated compute. Combining main MSEF training, ablations, baseline reproduction, inference-only evaluations, and preliminary exploratory experiments, the total compute used for this project is estimated to be approximately 1800–2600 GPU-hours. This estimate is reported to improve transparency and should be interpreted as an approximate accounting rather than an exact carbon or cost audit.

Reproducibility. All reported MSEF results use 5 random seeds. We fix the train/validation/test split at the character level to avoid leakage across different historical forms of the same character. Hyperparameters, dataset construction details, and implementation choices are provided in the appendix and released code. The released repository includes scripts for preprocessing FGCCES, training MSEF, running CBED inference, and reproducing the main tables.

U Asset Licenses and Terms of Use

This work uses both newly constructed assets and existing public or academic research assets. We credit the original creators or maintainers of all external resources used for dataset construction, benchmark evaluation, and baseline comparison. Table 42 summarizes the main assets, their roles in this work, and our release treatment.

Because several historical-script sources are cultural or academic databases rather than conventional machine-learning datasets, some sources do not provide standardized machine-readable licenses. In these cases, we take a conservative release approach: we cite the original sources, use the data only for non-commercial academic research, avoid claiming ownership of third-party materials, and separate redistributable derived annotations/features from raw source assets. The project repository contains a README and dataset card documenting source provenance, access dates, preprocessing steps, released files, and any source-dependent reconstruction instructions.

V Broader Impacts and Responsible Use

This work aims to support computational paleography and cultural-heritage research by providing a structured way to model Chinese script evolution across OBI, Bronze, Seal, Clerical, and Regular scripts. The main positive impact is to assist experts in organizing fragmented cross-era evidence: MSEF and CBED can retrieve plausible evolutionary candidates, expose intermediate-era consistency, and provide confidence estimates and interpretable analyses that may help paleographers prioritize which hypotheses to examine. More broadly, FGCCES may support future research on low-resource historical scripts, cross-era representation learning, and expert-in-the-loop decipherment systems.

Table 42: External and newly introduced assets used in this work. "Release treatment" specifies what is included in our released package and how third-party ownership is respected.

Asset	Use in this work	Citation / URL	License or terms	Release treatment
LBQJQW / jgwlbq	Source for OBI glyph images and scribal-group information used in constructing FGCCES.	https://www.jgwlbq.org.cn/	Publicly accessible academic/cultural-resource website. We found no standardized machine-readable open license in the cited source.	We cite the source and do not claim ownership of the original glyph images. Raw assets are redistributed only when permitted by the source terms; otherwise, we provide derived metadata/features and scripts or instructions for reconstruction from the original source.
CCAMC	Source for ancient Chinese character morphology, especially Bronze-script forms and period/provenance information.	http://www.ccamc.co	Publicly accessible academic database. We found no standardized machine-readable open license in the cited source.	We cite the source and use it for non-commercial academic research. Redistribution of raw records follows the source terms; our released package separates derived annotations/features from third-party raw materials.
BNU Character Database	Source for Seal, Clerical, and Regular-script forms and verified cross-era correspondences.	https://www.bnu.edu.cn/	Academic database maintained by Beijing Normal University. We found no standardized machine-readable open license in the cited source.	We cite the source and use it for academic research. We do not claim ownership of the original database content; released derived files exclude third-party raw assets unless redistribution is permitted.
HUST-OBS and EVOBC	Public benchmarks used for OBI decipherment evaluation following prior protocols.	Wang & Deng (2024); Guan et al. (2024)	Used according to the academic research terms described by the original benchmark papers or repositories.	We cite the benchmark creators and report results under their evaluation protocols. We do not relicense the original benchmarks.
PictOBI-20k / OBI-Bench	Visual multiple-choice benchmark used to evaluate paleographic reasoning.	Chen et al. (2024)	Used according to the academic research terms described by the original paper or repository.	We cite the original benchmark and use it only for evaluation. Any benchmark redistribution follows the original release terms.
Baseline methods and code	Baseline implementations or paper-based reimplementations for Pix2Pix, CycleGAN, DRIT++, Palette, BBDM, CDE, Sundial-GAN, OBSD, Diff-Oracle, OracleFusion, CrossFont, OracleSage, and OracleAgent.	Cited in Section 3 and the references.	Public code is used under its original license where available. For baselines without public code, we reimplement from the paper descriptions.	We preserve original notices for any public code used. Reimplemented baselines are released as our own reproduction code, while the original papers remain credited.
FGCCES	New cross-era Chinese character evolution sequence dataset introduced in this paper.	This work.	We release our code under the MIT License. We release FGCCES-derived annotations, splits, and feature files under CC BY-NC 4.0, excluding third-party raw assets whose original terms do not permit redistribution.	The released package includes a dataset card, preprocessing scripts, train/validation/test splits, feature-generation code, and instructions for obtaining or reconstructing source-dependent raw assets.

At the same time, oracle-bone decipherment is a historically sensitive task. Incorrect model outputs could mislead downstream historical, linguistic, or archaeological interpretation if they are treated as final decipherments rather than computational hypotheses. This risk is especially important for damaged glyphs, rare characters, region-specific variants, semantic borrowing, and characters with discontinuous or sparsely preserved evolutionary evidence. Therefore, our system should not be used as an autonomous authority for paleographic interpretation. We frame MSEF outputs as candidate correspondences for expert review, not as definitive readings.

We adopt several safeguards and usage recommendations. First, the model reports ranked candidates rather than a single unqualified answer, allowing users to inspect alternatives. Second, CBED performs backward verification and step-wise pruning so that visually similar but evolutionarily inconsistent candidates can be filtered. Third, the paper reports error analysis and confidence calibration so that low-confidence cases can be separated from high-confidence predictions. Fourth, for public release, we document dataset provenance, split construction, preprocessing, and limitations, and we respect the terms of third-party cultural and academic resources as described in Appendix U.

The work does not involve personal data, biometric identification, surveillance, medical data, or security-sensitive capabilities. The expert evaluation is limited to professional assessment of ancient script correspondences and does not collect sensitive information about participants. Consequently, the most relevant ethical concern is not privacy or direct dual-use misuse, but scholarly overclaiming: users may over-interpret model suggestions as established historical facts. We mitigate this by making the uncertainty and limitations explicit and by recommending a paleographer-in-the-loop workflow for any real decipherment claim.

Finally, although this paper focuses on Chinese script evolution, extending the framework to other ancient writing systems should not be treated as a direct transfer. Other scripts may have different material conditions, lineage structures, sociolinguistic pressures, and evidentiary gaps. Responsible application to another historical tradition would require collaboration with domain experts, new source-specific provenance documentation, and a separate evaluation protocol rather than reuse of the present assumptions without validation.

W LIMITATIONS

Although MSEF improves cross-era OBI decipherment, several limitations remain. First, the framework depends on the coverage and reliability of cross-era correspondences. FGCCES integrates verified evolution pairs, but the archaeological record is inherently fragmented: many glyphs have missing, uncertain, or region-specific intermediate forms. As a result, predictions should be interpreted as paleographic candidates for expert review rather than final decipherments.

Second, our continuous-flow formulation is an approximation of a more complex historical process. Chinese script evolution includes abrupt reforms, regional variants, semantic borrowing, component substitution, and one-to-many or many-to-one lineages. Neural ODE dynamics and bidirectional reachability capture broad temporal regularities, but may still underrepresent discontinuous changes or characters whose surviving evidence is extremely sparse.

Third, the current evaluation is limited by available benchmarks and expert-labeled cases. Although experiments cover multiple datasets and blind expert evaluation, they cannot exhaustively represent all undeciphered OBI characters or highly disputed correspondences. Future work should expand cross-institutional expert annotation, improve uncertainty calibration, and evaluate MSEF in paleographer-in-the-loop workflows. Finally, extending the framework to other ancient scripts will require new lineage datasets and script-specific historical assumptions rather than direct transfer.