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 the evaluated 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 an important source for studying early Chinese writing and Shang history (Keightley, 1985). 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 Chinese (Chang 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).

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

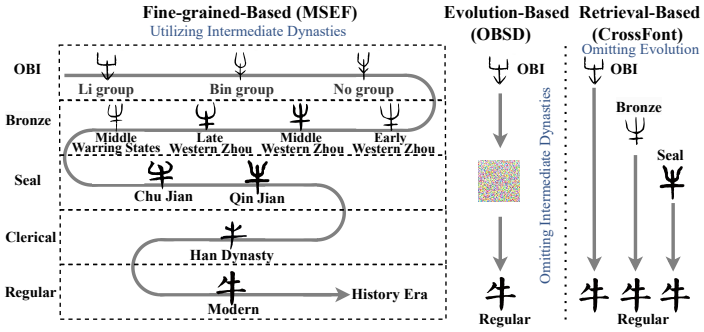


Figure 1: **Evolution of Chinese script.** Taking the character "Ox" as an example, this figure demonstrates the evolution of Chinese characters across five historical periods. This character underwent a significant morphological change in early bronze script, then later reverted to a form similar to 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.

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 makes it difficult to retrieve glyphs that underwent drastic changes during the periods 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. Existing approaches do not solve this issue either, but instead proceed 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 an era-conditioned representation in a shared high-dimensional coordinate space, where each observed glyph corresponds to a 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. To support MSEF training, we **construct a fine-grained dataset** containing complete and partial evolution chains and their available pairwise correspondences. 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 C), we delineate this research trajectory: First, we systematically diagnose existing generative and retrieval paradigms, revealing failure modes associated with not explicitly modeling intermediate stages of script evolution. Second, we return to first principles to model the labeled evolution chains as a duality of a shared correspondence identity and a variant visual surface across historical periods. 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, examining the performance of incremental patches and the design of discrete rule-based alternatives. 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 D). 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.

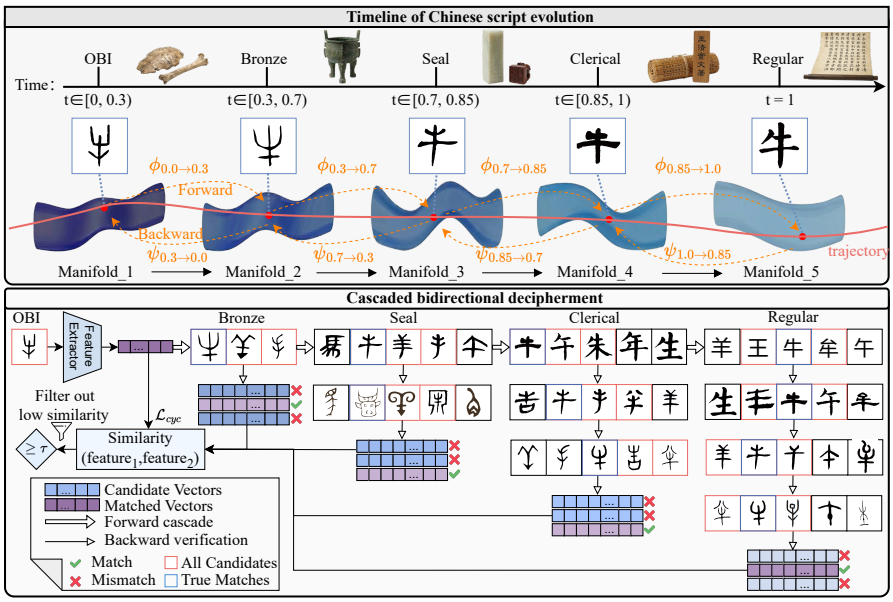


Figure 2: **Overview of MSEF. Top: Continuous Manifold Evolution.** The framework models the 3,000-year evolution of Chinese script as a continuous trajectory (red line) through era-specific manifolds $\mathcal{M}_t$, conditioned on normalized time t (from OBI $t = 0.0$ to Regular $t = 1.0$). The intervals of t define an ordered model-time coordinate; their lengths are not proportional to calendar-year durations. Neural ODEs govern the continuous transitions via forward flows ϕ and backward flows ψ. Specifically, each historical period manifold is represented in a shared high-dimensional coordinate space, where each observed glyph corresponds to a 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 backward compatibility with the original OBI encoding.

These analyses characterize information-routing patterns and morphodynamic regularities consistent with known paleographic transformations; they do not independently exclude memorization. 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 a framework contains a manifold net for representing the period, Neural ODE for representing inter-period transitions, and survival net for representing character survival across periods.
- We construct and open-source 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's knowledge of Chinese character evolution in the oracle bone decipherment task.
- Experiments on three commonly used benchmarks and our own dataset demonstrate that MSEF achieves state-of-the-art performance among the evaluated 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 retains recognizable structural components. 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 P. 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. **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 examine this geometric structure in Appendix Q, 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 OBI times in $[0, 0.30)$ 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.** Let $\mathbf{z}_a(u)$ solve Equation 1 with $\mathbf{z}_a(t_1) = \mathbf{a}$, and let $\mathbf{z}_b(u)$ solve it backward with $\mathbf{z}_b(t_2) = \mathbf{b}$. We define:

$$\phi_{t_1 \to t_2}(\mathbf{a}) = \mathbf{a} + \int_{t_1}^{t_2} \mathbf{v}(\mathbf{z}_a(u), u)\, du, \tag{2}$$

$$\psi_{t_2 \to t_1}(\mathbf{b}) = \mathbf{b} - \int_{t_1}^{t_2} \mathbf{v}(\mathbf{z}_b(u), u)\, du. \tag{3}$$

The forward flow ϕ traces a character forward in time, while the backward flow ψ traces it backward. Under the regularity assumptions in Appendix I, the exact flows are inverses: $\psi_{t_2 \to t_1} = \phi^{-1}_{t_1 \to t_2}$. This latent-space invertibility enables bidirectional verification. **Time Encoding.** We assign ordered model-time intervals to historical periods: OBI ($t = 0.00$–0.30, with fine-grained scribal-group metadata), 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. Fine-grained assignments must follow the record-to-time convention in Appendix K, rather than treating every scribal-group label as an equally spaced chronological period.

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

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**, **Structural**, **Semantic**, **Contextual**, and **Spatiotemporal**. Detailed definitions of characteristics and aggregation mechanisms are provided in the Appendix L. **Manifold Encoder.** Manifold mapping $\mathcal{M}(\mathbf{x}_c^t, t)$ 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 predicted representation $\widetilde{\mathbf{z}}_c^{t_j} = \phi_{t \to t_j}(\mathbf{z}_c^t)$ at a later observation time t_j. 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. We write its extinction output as $e(\mathbf{z}_c^0, t) = \mathcal{S}(\mathbf{z}_c^0, t)$ and the corresponding survival probability as $s = 1 - e$. **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 $[0, 1]$, with Regular script fixed at $t = 1$. 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 train 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 the available supervised cross-era pairs:

$$\mathcal{P} = \mathcal{P}_{\text{adj}} \cup \mathcal{P}_{\text{skip}}. \tag{4}$$

Complete chains are recorded separately, rather than added to a pair count. The counting units are specified in Appendix H.

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** measures numerical round-trip consistency (Zhu et al., 2017); exact invertibility follows from the flow assumptions. Finally, **Survival Loss** trains the extinction predictor using binary cross-entropy to distinguish between existing and extinct characters. The total objective is:

$$\mathcal{L}_{\text{evolution}} = \lambda_1 \mathcal{L}_{\text{adj}} + \lambda_2 \mathcal{L}_{\text{skip}} + \lambda_3 \mathcal{L}_{\text{full}} + \lambda_4 \mathcal{L}_{\text{cyc}}, \quad \mathcal{L}_{\text{surv}} = \text{BCE}(s(\mathbf{z}_c^0, t), y_{c,t}). \tag{5}$$

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

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** with $\mathcal{L}_{\text{evolution}}$ in Equation 5. The extinction predictor has a separate objective $\mathcal{L}_{\text{surv}}$. Only examples providing an OBI representation and a specified survival label contribute to the survival objective; an arbitrary cross-era pair does not supply such supervision.

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 G.

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. We first encode the query at its assigned OBI time t_0: $z_{t_0}^q = \mathcal{M}(\mathbf{X}^q, t_0; \theta)$. For each later checkpoint t_k, we solve the learned dynamics:

$$z_{t_k}^q = \phi_{t_{k-1} \to t_k}(z_{t_{k-1}}^q) \tag{6}$$

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 $t > t_{Seal}$, no additional survival-based rejection is applied; cumulative survival is not reset to one.

2.5.2 Backward Verification via Step-wise Pruning.

To account for the one-to-many nature of Chinese character correspondences, CBED uses set-valued candidate retrieval and bidirectional reachability. This discrete retrieval relation need not be bijective. It is distinct from the invertibility of the latent ODE flow under the assumptions in Appendix I. 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:

$$\hat{z}_{t_i} = \psi_{t_{i+1} \to t_i}(\hat{z}_{t_{i+1}}) \tag{7}$$

We then compare this backward-projected point $\hat{z}_{t_i}$ with the candidate's actual prototype embedding obtained via $\mathcal{M}(X_{t_i}^c, t_i; \theta)$ in the era t_i. Candidate paths that exceed a divergence threshold ϵ_t are **pruned**. Checks use available intermediate prototypes; missing glyph observations are skipped, not replaced by zero vectors. Finally, for candidates that survive the pruning process back to the OBI era, we compute the cosine similarity between their final projected coordinate $\hat{z}_{OBI}$ and the query's observed embedding $z_{OBI}^q = \mathcal{M}(X^q, t_{OBI}; \theta)$. The candidate with the highest valid-path cosine

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**

similarity is selected as the final deciphering result, following the aggregation rule in Appendix G. Algorithm 1 provides the detailed steps. The algorithm covers all five script periods and returns an abstention when no candidate path passes verification. 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 predicts their ancient latent representations. The backward representation of "牛" is consistent with the query's encoding, whereas "午" and "牟" result in significant reconstruction errors in the latent comparison. 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 R 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 under a character-disjoint train/validation/test split. Counting units and dataset organization are described in Appendix H.

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., 2019), 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 strictly follow the original content of the paper. See appendix O.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

Table 3: Results on PictOBI-20k. MSEF (140M params) outperforms the displayed LMM baselines, 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	—	26.31	25.52	26.23
Gemini 2.5 Pro	—	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**

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**

& Hutter, 2017) with learning rate 10^{-4}, batch size 256, for 100 epochs (∼18 hours on 2 NVIDIA H100 GPUs). We report mean±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 O.1 **Comparison with LMMs on PictOBI-20k.** Table 3 compares MSEF with the seven displayed LMM baselines on PictOBI-20k's 4-choice visual decipherment task (Chen et al., 2025). 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 benchmark provides a direct comparison with contemporary OBI methods. The exploratory Joint-Consistent CrossFont result also reaches 72.5% R@1 (Appendix C.1.2); a method advantage over that comparator requires matched evaluation protocols. Appendix O.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 of 19.7 percentage points, confirming fragmented correspondences are vital for low-resource OBI. **(2) Manifold Modeling:** Replacing Neural ODE

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	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^{*}

Table 6: Expert evaluation on 100 challenging cases.

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

with a static manifold degrades performance by 10.3 percentage points, supporting the time-indexed modeling configuration in this comparison. **(3) Inference Mechanism:** Removing cascaded retrieval and bidirectional verification leads to drops of 8.5 and 6.2 percentage points, validating the effectiveness of our "recall-check" strategy. Appendix O.11 provides a micro-level ablation demonstrating how backward verification effectively prunes false positives. **Effect of Intermediate Eras.** Adding intermediate eras consistently improves performance: 2 eras (OBI→Regular) achieves 61.2%, while adding Bronze (+4.5 points), Seal (+3.8 points), and Clerical (+3.0 points) progressively reaches 72.5%. These configurations vary both era coverage and available supervision; the results do not isolate their individual effects. Full results are provided in Appendix O.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 the interpretation of confidence scores are discussed in Appendices O.6 and O.7.

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.

AI USE STATEMENT

We used Paper2Code (Seo et al., 2025) to assist with reimplementing baseline methods for which public implementations were unavailable, as described in Section 3. We also used ChatGPT for manuscript consistency checking, feedback on the methodological and experimental descriptions, and assistance with revising LaTeX text, mathematical statements, and proofs. AI-generated suggestions do not constitute empirical evidence or independent validation of mathematical claims. The authors take responsibility for the final content of the paper, including all AI-assisted text, code, mathematical claims, figures, and other artifacts.

ETHICS STATEMENT

This work supports expert assessment of ancient-script correspondences. Model outputs are candidate hypotheses rather than definitive readings. Model rejection is not evidence of historical extinction. Source provenance, third-party asset treatment, and responsible use are discussed in Appendices T and U.

REPRODUCIBILITY STATEMENT

The model is described in Section 2.2, the losses in Appendix F, and the inference procedure in Algorithm 1. Appendices L, K, and M provide feature, time, and available implementation details. Dataset organization and theoretical assumptions are discussed in Appendices H and I.

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. Review of *The Origin and Early Development of the Chinese Writing System*, by William G. Boltz. *Journal of the American Oriental Society*, 116(3):574–577, 1996. doi:10.2307/605196.

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 T. Q. 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.

Zijian Chen, Wenjie Hua, Jinhao Li, Lirong Deng, Fan Du, Tingzhu Chen, and Guangtao Zhai. PictOBI-20k: Unveiling large multimodal models in visual decipherment for pictographic oracle bone characters. *arXiv preprint arXiv:2509.05773*, 2025.

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.

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*. University 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, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. DRIT++: Diverse image-to-image translation via disentangled representations. *arXiv preprint arXiv:1905.01270*, 2019.

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 T. Q. 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.

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 and Continuous-Time Models 17

B.2 Computational OBI Analysis 18

C Pre-Exploratory Step-by-Step Research **18**

C.1 Diagnosing Existing Paradigms 18

C.2 Character Identity and Era-Specific Form 19

C.3 Evolution Pattern Clusters and Baseline Vulnerabilities 19

C.4 From Exploratory Findings to MSEF 21

D Post-Analysis: Representation and Dynamics Diagnostics **23**

D.1 Spatial Attribution Across Eras 23

D.2 Saliency-Based Sensitivity Analysis 24

D.3 Latent Geometry and Trajectories 24

D.4 Attention Patterns and Modality Routing 25

D.5 Activation Interventions and Modality Contributions 25

D.6 Survival-Score Landscapes 26

D.7 Latent Vector Arithmetic 27

E Full Method Details **30**

E.1 Feature Representation 30

E.2 Time Conditioning 30

E.3 Encoding and Evolution 30

E.4 Architecture 30

E.5 Optimization and Inference 30

F Full Loss Definitions **30**

F.1 Evolution Objectives 31

F.2 Survival Supervision 31

F.3 Gradient Paths 31

F.4 Interpretation of Cycle Consistency 32

F.5 Discriminative Representation Requirements 32

G Full CBED Algorithm **32**

G.1 Inputs and Reference Information 32

G.2 Reference Paths and Missing Eras 32

G.3 Path Scores 33

H Dataset Details **34**
H.1 Counting Units . . . 34
H.2 Sources and Missing Features . . . 35
H.3 Train, Validation, and Test Separation . . . 35
H.4 Fine-Grained Visualization . . . 35
H.5 Comparison with Existing Resources . . . 35

I Theoretical Foundations **35**
I.1 Regularity Assumptions . . . 36
I.2 Existence, Uniqueness, and Invertibility . . . 36
I.3 Composition . . . 36
I.4 Manifold Dimension . . . 36

J Manifold Space Validation **36**
J.1 Conditional Properties and Diagnostics . . . 36
J.2 Velocity Field Analysis . . . 39
J.3 Local Neighborhood Illustration . . . 40
J.4 Summary . . . 40

K Time Encoding Details **41**
K.1 Ordered Time Coordinate . . . 41
K.2 Fine-Grained Metadata . . . 41
K.3 Training and Inference Times . . . 41
K.4 Consequences for Interpretation . . . 42

L Feature Extraction Details **42**
L.1 Visual Features: 128 Dimensions . . . 42
L.2 Structural Features: 64 Dimensions . . . 42
L.3 Semantic Features: 64 Dimensions . . . 42
L.4 Context Features: 64 Dimensions . . . 42
L.5 Spatiotemporal Features: 32 Dimensions . . . 42
L.6 Missing Features and Evaluation Availability . . . 42
L.7 Scope of End-to-End Training . . . 43

M Implementation Details **43**
M.1 Architecture . . . 43
M.2 Numerical Integration . . . 43
M.3 Optimization . . . 43
M.4 Parameter Counts and Timing . . . 43

N Dataset Construction Details **44**

N.1 Source Records 44
N.2 Correspondence Verification 44
N.3 Pairs, Chains, and Availability 44
N.4 Comparison with Existing Resources 45

O Additional Experimental Results **45**
O.1 Multi-Round Decipherment 45
O.2 Benchmark-Specific Results 46
O.3 Additional Ablation Studies 46
O.4 Effect of Intermediate Eras 46
O.5 Fairness of Cross-Era Data Utilization 47
O.6 Error Analysis 47
O.7 Compatibility Scores and Calibration Scope 48
O.8 Scribal-Group Analysis 48
O.9 Computational Efficiency 48
O.10 Comparison with a Diffusion-Based Probability-Flow ODE 49
O.11 Backward Verification and Step-Wise Pruning 49
O.12 Scope of Expert Agreement and Significance Claims 49

P Extended Case Studies and Retrieval Gallery **50**
P.1 Full Retrieval Gallery 50

Q Manifold Neighborhood Visualization **51**

R Fine-Grained Dataset Visualization **52**

S Computational Resources and Reproducibility Report **54**
S.1 Main Training Runs 54
S.2 Other Experiments 54
S.3 Inference Accounting 54

T Asset Licenses and Terms of Use **54**

U Broader Impacts and Responsible Use **56**

V Limitations **56**

A Appendix Roadmap

Scope of the supplementary evidence. The appendix separates mathematical consequences of the stated model from experimental summaries. Numerical summaries without a documented sample-level protocol are explicitly identified below as unvalidated observations; they are not additional verified comparisons. In particular, unspecified implementation choices are not replaced with assumed settings, and a retained table does not establish that its configuration matches another table. These distinctions also limit the interpretation of the corresponding main-text summaries. The appendix provides methodological details, supplementary analyses, and documentation supporting the main paper.

Background and motivation. Appendix B discusses related work. Appendix C presents exploratory analyses motivating multi-era modeling and candidate verification. These analyses motivate design choices rather than establish the necessity of a unique architecture.

Representation and dynamics diagnostics. Appendix D examines input sensitivity, attention patterns, latent geometry, survival scores, and vector analogies. These analyses characterize model behavior; they do not independently establish historical causality, equivalence to human cognition, or the absence of memorization.

Method and implementation. Appendix E summarizes the model. Appendix F defines the training objectives. Appendix G specifies CBED inference. Appendices K, L, and M describe time encoding, feature extraction, and implementation settings.

Dataset. Appendices H and N describe dataset organization, counting units, splits, and provenance. Appendix R illustrates the available temporal and scribal-group annotations.

Theory and empirical diagnostics. Appendix I states the regularity assumptions and properties of the latent ODE flow. Appendix J distinguishes conditional mathematical results from finite-sample diagnostics. Appendix Q visualizes local retrieval neighborhoods.

Additional experiments. Appendix O reports supplementary evaluations, ablation analyses, computational measurements, and benchmark-specific protocol requirements. Error analysis and confidence evaluation appear in Appendices O.6 and O.7. Appendices O.10 and O.11 examine alternative dynamics and backward verification.

Examples, reproducibility, and responsible use. Appendix P presents retrieval examples. Appendix S documents computational resources. Appendices T, U, and V describe asset treatment, responsible use, and limitations.

B Related Work

Our work connects continuous-time representation learning with computational analysis of Oracle Bone Inscriptions.

B.1 Manifold Learning and Continuous-Time Models

Manifold learning studies low-dimensional structure in high-dimensional observations (Izenman, 2012). Neural ODEs parameterize the derivative of a latent state and compute transformations through numerical integration (Chen et al., 2018). Latent ODEs explicitly address irregularly sampled observations and variable temporal gaps (Rubanova et al., 2019). Augmented Neural ODEs and neural flows investigate alternative parameterizations and representational properties of continuous-time models (Dupont et al., 2019; Biloš et al., 2021). Related flow-based methods also study probability modeling for data with manifold structure (Kim et al., 2020). Accordingly, incomplete observations and continuous time are not capabilities unique to MSEF. Our focus is their use in cross-era glyph representation learning, together with candidate retrieval and verification against available intermediate-era evidence.

B.2 Computational OBI Analysis

Computational OBI research includes recognition, retrieval, image generation, and multimodal analysis (Jin & Yang, 2023; Zhen et al., 2024; Gao & Liang, 2020). Recognition methods address identification within labeled character inventories (Li & Du, 2024). Generative approaches include Sundial-GAN (Chang et al., 2022), Diff-Oracle (Li et al., 2023c), and OBSD (Guan et al., 2024). Retrieval methods include cross-font comparison with historical reference glyphs (Wu et al., 2025). OBI-Bench, OracleSage, and OracleAgent investigate multimodal understanding and reasoning (Chen et al., 2024; Jiang et al., 2024; Li et al., 2025a). MSEF differs in jointly learning time-conditioned latent representations and inter-era dynamics from observed correspondences, followed by candidate-level backward verification. This distinction does not imply that competing methods cannot be augmented with temporal information. Such augmentations are important comparators, including the sequence-consistency baseline examined in Appendix C.

C Pre-Exploratory Step-by-Step Research

This section presents exploratory analyses motivating multi-era modeling and candidate verification. These analyses concern the specific implementations and datasets examined. They do not establish architectural impossibility results or prove that every component of MSEF is necessary.

C.1 Diagnosing Existing Paradigms

We consider generation-based and retrieval-based approaches. Generation-based methods predict a target-script representation or image. Retrieval-based methods rank glyphs in a reference collection. Their performance depends on the available inputs, reference data, training supervision, and evaluation interface. Table 7 restates selected results from the main paper without combining different models under a single "best paradigm" label.

Table 7: Selected baseline results under their stated benchmark protocols. Values across different benchmarks are not directly comparable.

Method	Evaluation	Reported value (%)
OBSD	HUST-OBS/EVOBC, OBS-OCR Top-1	41.0
Diff-Oracle	FGCCES, R@1	52.1
CrossFont	FGCCES, R@1	54.6

C.1.1 Generation Performance and Evolutionary Complexity

We examine how the reported generation results vary with glyph-distance categories and an abrupt-transition label. A finite difference between observed glyphs is not itself a test of mathematical continuity. The reported accuracy is lower for the high-distance and abrupt-transition groups. This association motivates evaluating intermediate-era evidence, but does not identify a unique cause of failure or rule out alternative generative architectures.

Table 8: Unvalidated generation-accuracy summary by complexity label. The three distance-group counts sum to 1,007; the membership relationship of the 112 abrupt-transition examples is not established. The four rows are not treated as a disjoint partition.

Group	Count	Top-1 (%)	Difference from Low
Low, distance < 0.3	228	67.5	—
Medium, distance 0.3–0.6	445	48.3	-19.2
High, distance > 0.6	334	22.7	-44.8
Abrupt transition	112	14.2	-53.3

C.1.2 Retrieval with Temporal Consistency

We compare independent CrossFont retrieval with an augmentation that combines candidate sets through known cross-era correspondences and cumulative similarity scores. The reported OBI-to-

Table 9: Reported CrossFont retrieval with and without sequence-consistency filtering.

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

Regular result of 72.5 has the same point estimate as MSEF in Table 4. If the evaluation protocols match, these results do not establish an R@1 advantage of MSEF over the augmented CrossFont baseline. The supplied summaries do not establish matching splits, reference graphs, features, candidate vocabularies, or search budgets. The equality of the two point estimates is therefore disclosed without claiming a controlled tie or an advantage over the augmented comparator. An advantage over the unaugmented OracleAgent result does not isolate the contribution of continuous dynamics.

C.1.3 Implications for Model Design

The exploratory results suggest that intermediate-era evidence can be useful. They leave open whether the principal benefit comes from representation learning, additional supervision, candidate-graph access, or the dynamics parameterization. These factors require matched comparisons.

C.2 Character Identity and Era-Specific Form

C.2.1 Shared Correspondence Labels and Variable Glyphs

Verified correspondence labels allow different historical glyph observations to supervise a shared representation. For selected pictographic examples, such as “Ox”, “Sun”, and “Mountain”, this relation can be illustrated using a common referent and varying glyph forms. This is a modeling abstraction for the labeled correspondences, not an assumption that every character preserves an invariant meaning throughout history.

C.2.2 Temporal Regularity as a Modeling Prior

MSEF uses continuous latent dynamics as an inductive bias. Sparse glyph observations do not identify a unique continuous historical trajectory. A smooth latent path can coexist with substantial differences between observed glyph images. The previously reported fraction of transitions within two standard deviations of a distance mean is a distributional summary, not a proof of temporal continuity.

C.3 Evolution Pattern Clusters and Baseline Vulnerabilities

C.3.1 Structural Features

We analyze seven structural descriptors of observed glyphs. These descriptors are used for exploratory analysis and are distinct from the full 352-dimensional model input. Missing glyph observations are not treated as measured zero-valued glyphs in this analysis.

C.3.2 Clustering Procedure

The counts in Table 11 sum to 881. This is a separate reported summary from the 1,007 distance-group examples in Table 8; no common sample population is established by these aggregates. The summaries are not pooled to estimate dataset-wide frequencies. The exploratory pipeline applies

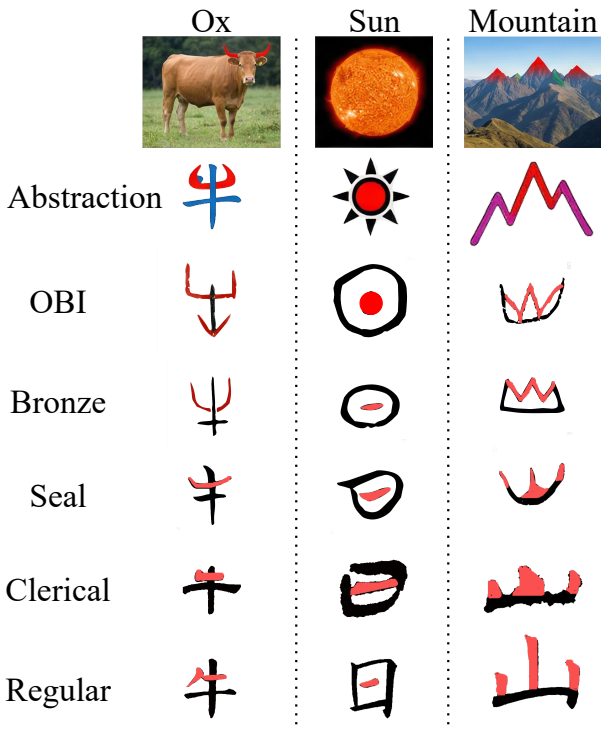


Figure 3: Selected examples of shared correspondence labels and era-specific glyph forms. These examples do not imply universal semantic invariance.

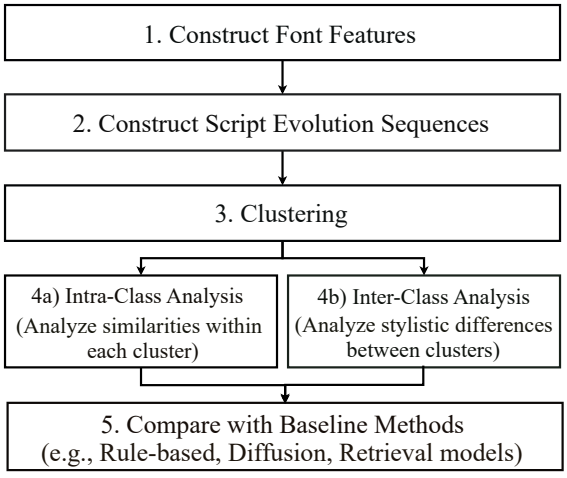


Figure 4: Exploratory workflow: feature construction, cross-era sequence organization, clustering, and baseline error analysis.

per-feature clustering and combines assignments through a co-association matrix and hierarchical clustering. The reported four-cluster solution is summarized below.

Table 10: Structural descriptors used in the exploratory analysis.

Feature	Description
Stroke count	Number of identified strokes
Inflection points	Direction changes in the contour
Endpoint count	Number of open stroke tips
Mean endpoint distance	Mean pairwise endpoint distance
Horizontal/vertical ratio	Relative stroke orientations
Enclosed area ratio	Fraction of enclosed glyph area
Symmetry score	Bilateral symmetry measure

Table 11: Reported exploratory cluster membership counts.

ID	Description	Count
A	Monotone simplification	312
B	Non-monotone change	224
C	Radical reorganization	158
D	Abrupt transformation	187

C.3.3 Cluster-Conditioned Errors

These results describe different error distributions across the reported groups. They do not show that either method fails on every sample outside Cluster A, or prove that the clustering labels identify causal failure mechanisms.

C.4 From Exploratory Findings to MSEF

C.4.1 Incremental Augmentations

The results show gains for the tested augmentations. They do not exclude stronger augmentations or alternative architectures. In particular, the 67.4 result cannot be presented as a general retrieval ceiling in view of Table 9.

C.4.2 A Shared Latent Representation

A shared latent representation offers a common space for comparing observations across eras. Time conditioning represents era dependence, while the learned flow models transitions between latent states. Two-dimensional projections can illustrate local structure, but do not prove the existence of a particular data manifold or preservation of high-dimensional topology.

C.4.3 Rule-Based Alternatives

Rule-based models provide an interpretable alternative to learned dynamics. Their scalability and accuracy depend on the chosen rule representation, learning procedure, and vocabulary.

Table 12: Reported error percentages conditional on the exploratory cluster assignments.

Cluster	Description	Generation error	CrossFont error
A	Monotone simplification	31.2	28.6
B	Non-monotone change	57.8	63.5
C	Radical reorganization	52.3	70.8
D	Abrupt transformation	68.4	41.2

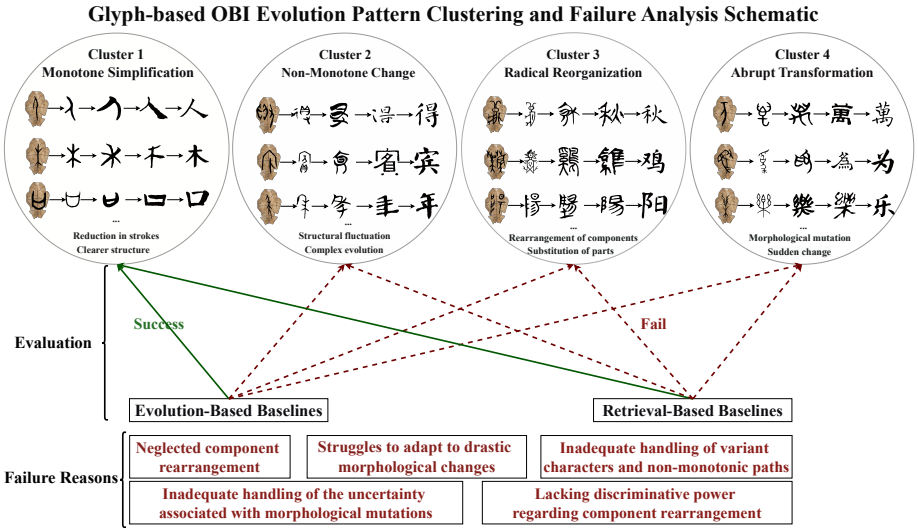


Figure 5: Illustration of the reported exploratory cluster labels. The diagram is not an independent test of baseline failure mechanisms.

Table 13: Reported results for particular baseline augmentations. These values are not architectural performance ceilings.

Family	Configuration	Metric	Value
Retrieval	Baseline	R@1	62.8
Retrieval	+ Era fusion	R@1	65.2
Retrieval	+ Temporal constraint	R@1	67.4
Generation	Baseline	Top-1	41.0
Generation	+ Era conditioning	Top-1	47.3
Generation	+ Consistency loss	Top-1	51.8

C.4.4 Continuous-Time Dynamics

Neural ODEs parameterize a latent velocity field and allow evaluation at selected integration times (Chen et al., 2018). MSEF associates this time coordinate with an ordered script-period convention. The coordinate is not a linear calendar-year scale. Time-conditioned discrete models, normalizing flows, and diffusion-based constructions remain relevant alternatives. The comparison in Appendix O.10 concerns a particular implemented baseline rather than all members of these model families.

C.4.5 Backward Candidate Verification

Backward verification starts from an observed candidate representation and compares its predicted earlier states with available reference observations and the query. This differs from integrating a query forward and immediately reversing the same numerical trajectory. Appendix O.11 reports the corresponding ablation. In Table 5, removing bidirectional verification changes R@1 from 72.5 to 66.3. The 64.0 row removes cascaded retrieval and represents a different configuration. Overall, these analyses motivate explicit multi-era modeling and verification, but do not establish Neural ODEs as the uniquely valid solution.

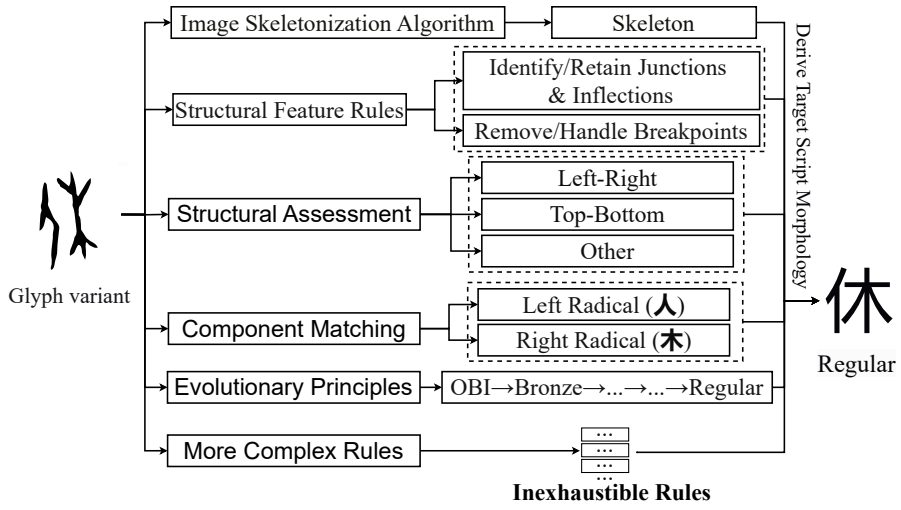


Figure 6: Schematic rule-based matching pipeline. A schematic branching structure is not a proof that finite rule enumeration is mathematically impossible.

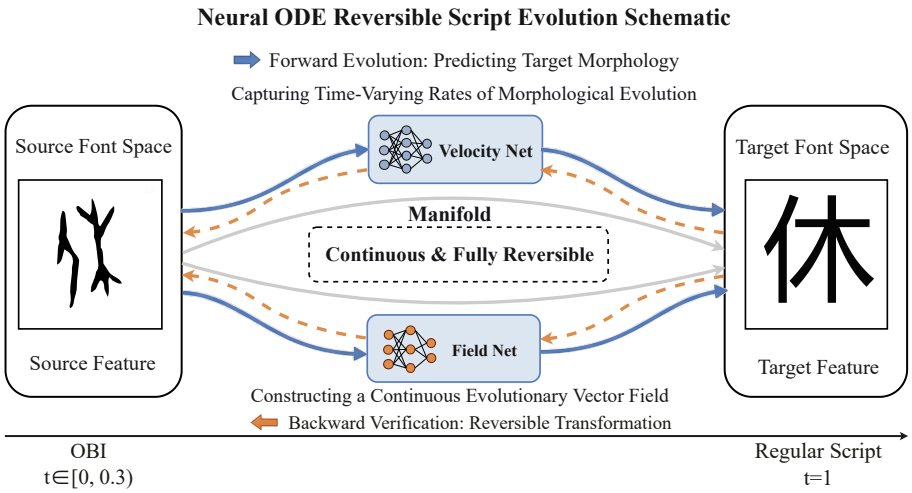


Figure 7: Continuous-time latent modeling under the time convention in Appendix K. The diagram does not establish recovery of a unique physical historical trajectory.

D Post-Analysis: Representation and Dynamics Diagnostics

This section examines the behavior of the learned representations, dynamics, and survival scores. The analyses are diagnostic: they do not establish equivalence to human cognition, historical causality, or the absence of memorization.

D.1 Spatial Attribution Across Eras

Figure 8 presents spatial attribution examples for selected characters. Corresponding highlighted structures across eras can suggest sensitivity to recurring glyph components. They are not, by themselves, evidence of a learned invariant semantic core. Because the encoder consumes precomputed features, pixel-level attribution requires an explicit mapping between image perturbations and the resulting features.

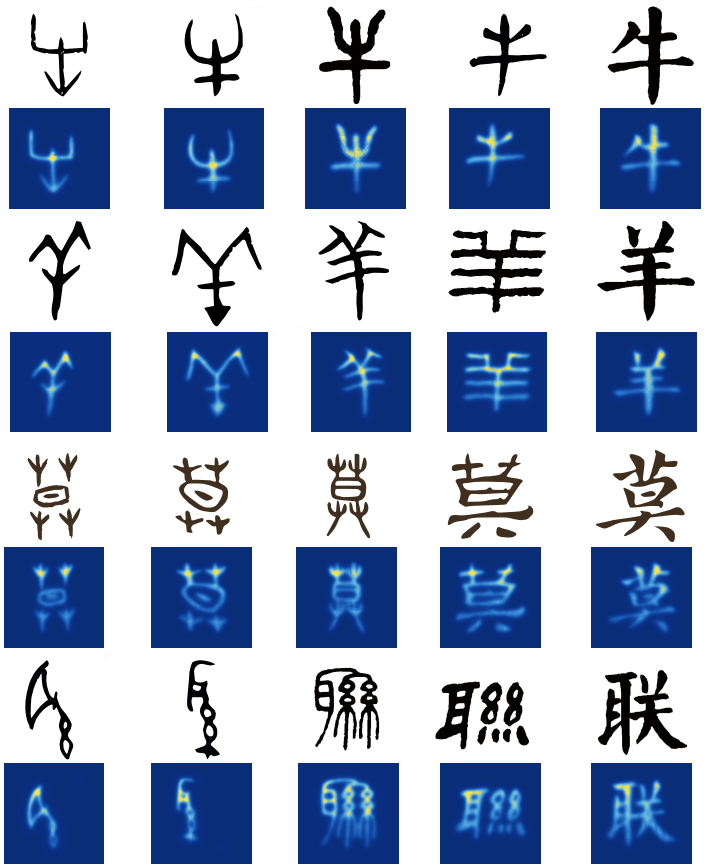


Figure 8: Spatial attribution examples across script periods. Highlighted regions describe the selected attribution procedure, not direct measurements of historical evolutionary laws.

D.2 Saliency-Based Sensitivity Analysis

We compare intact glyphs with high-saliency and low-saliency occlusions. Matched occlusion area controls one aspect of the perturbation, but does not control every possible change in glyph structure. For a declared compatibility score r, define

$$\Delta_{\text{high}} = r(x) - r(x_{\text{high}}), \qquad \Delta_{\text{low}} = r(x) - r(x_{\text{low}}).$$

The displayed examples provide a local sensitivity analysis. Population-level conclusions require aggregate results over a specified sample and appropriate perturbation controls.

D.3 Latent Geometry and Trajectories

Figure 10 illustrates low-dimensional projections of latent representations and selected trajectories. Distances and apparent curvature in a projection need not equal those in the original latent space. A regular ODE trajectory is continuous. A jagged plotted polyline can reflect sparse sampling, projection, or large but continuous changes; it is not evidence of an actual discontinuity in the latent ODE solution. Similarly, removing points from a displayed graph does not establish historical extinction or improved numerical conditioning.

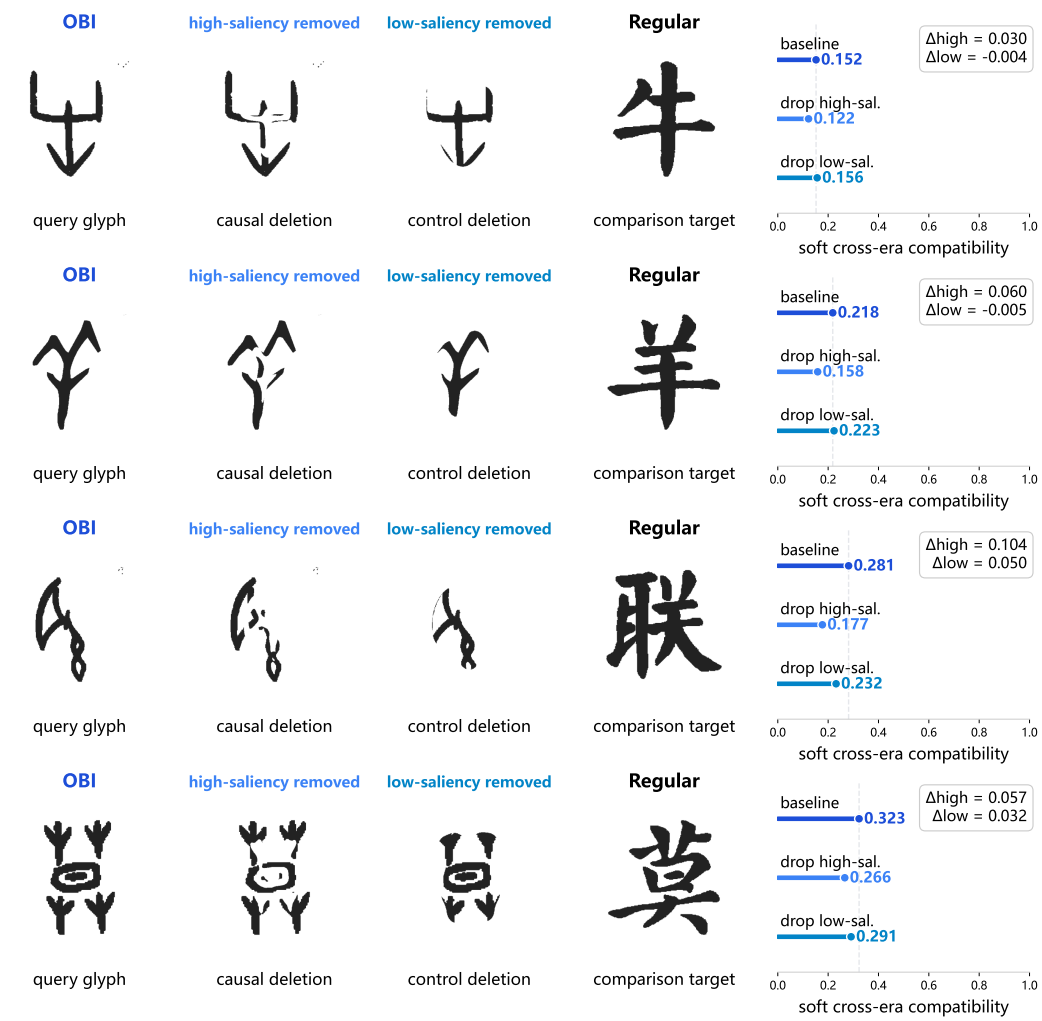


Figure 9: Compatibility-score changes under high-saliency and low-saliency occlusions for the displayed examples. Effects vary across examples and should not be described as uniformly catastrophic or negligible.

D.4 Attention Patterns and Modality Routing

The displayed graphic contains 144 head entries. With the 12-layer encoder in the main text, this would correspond to 12 heads per layer only if all layers contribute equally to that graphic. This conditional arithmetic is not a checkpoint-level architecture specification. The 256-dimensional output does not determine the internal Transformer width or head count. Attention diagnostics must correspond to the encoder used for the reported results. For L layers and H heads per layer, the number of heads is LH.

Attention weights alone do not establish causal importance or disentanglement. These interpretations require interventions and controls beyond inspection of the attention matrix.

D.5 Activation Interventions and Modality Contributions

Zeroing an activation and replacing it with an activation from another input are different interventions. The supplied analysis does not identify the intervention sufficiently to attribute an effect to either operation. The following equations distinguish possible reporting quantities; they do not establish

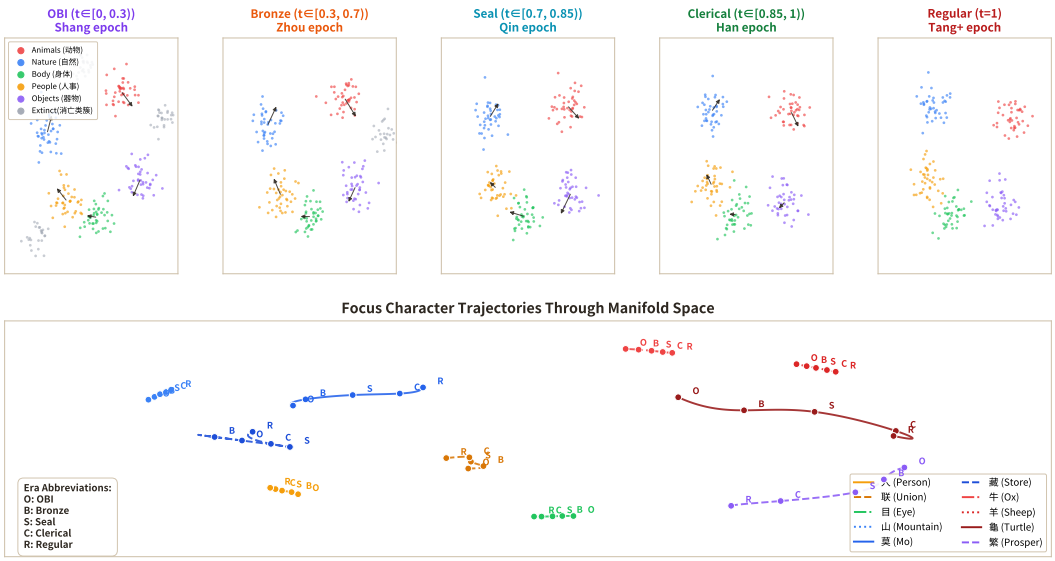


Figure 10: Visualization of latent representations and selected trajectories. Projected cluster structure is an exploratory diagnostic rather than a proof of topological preservation or historical causality.

which quantity was measured. For retrieval, a target-versus-distractor margin can be defined as

$$m(q) = r(c^+, q) - r(c^-, q),$$

where c^+ is a valid target and c^- is a specified distractor. An intervention effect is then

$$\Delta m = m_{\text{original}} - m_{\text{intervened}}.$$

This definition distinguishes a retrieval-score margin from a classification logit difference. The earlier activation-intervention graphic is omitted from the quantitative analysis because its labels (classification-logit drop, causal importance, and polysemanticity) are not tied to an identified estimator in the available specification. In particular, its plotted values are not relabeled as measurements of Δm or $PI(i)$ merely by changing a caption. For nonnegative modality contributions $a_{i,m}$, with $\sum_m a_{i,m} > 0$, define

$$\widehat{a}_{i,m} = \frac{a_{i,m}}{\sum_{m'=1}^{5} a_{i,m'}}, \qquad PI(i) = \exp\left(-\sum_{m=1}^{5} \widehat{a}_{i,m} \log \widehat{a}_{i,m}\right),$$

using $0 \log 0 = 0$. Then $PI(i) \in [1, 5]$ measures the effective number of contributing modalities under this construction. This index alone does not prove semantic superposition, orthogonality, or a causal role for particular latent dimensions.

D.6 Survival-Score Landscapes

Survival-score visualizations describe the output of a predictive module. They are not direct measurements of linguistic fitness or historical selection pressures.

Survival and extinction notation. Throughout the formulation, $e_\omega = \mathcal{S}$ denotes the extinction output and $s_\omega = 1 - e_\omega$ denotes survival. The earlier landscape graphic used S on an axis labeled survival; its numeric output convention is not established by the available plotting description. That graphic is omitted rather than silently complementing its values or assigning it a different output. A small score does not mathematically terminate the ODE. Rejection occurs only through an explicit inference rule. If the output is interpreted as cumulative survival for the same initial lineage, it must be non-increasing with target time. A transition from a low score to a high score does not establish the birth of a new character.

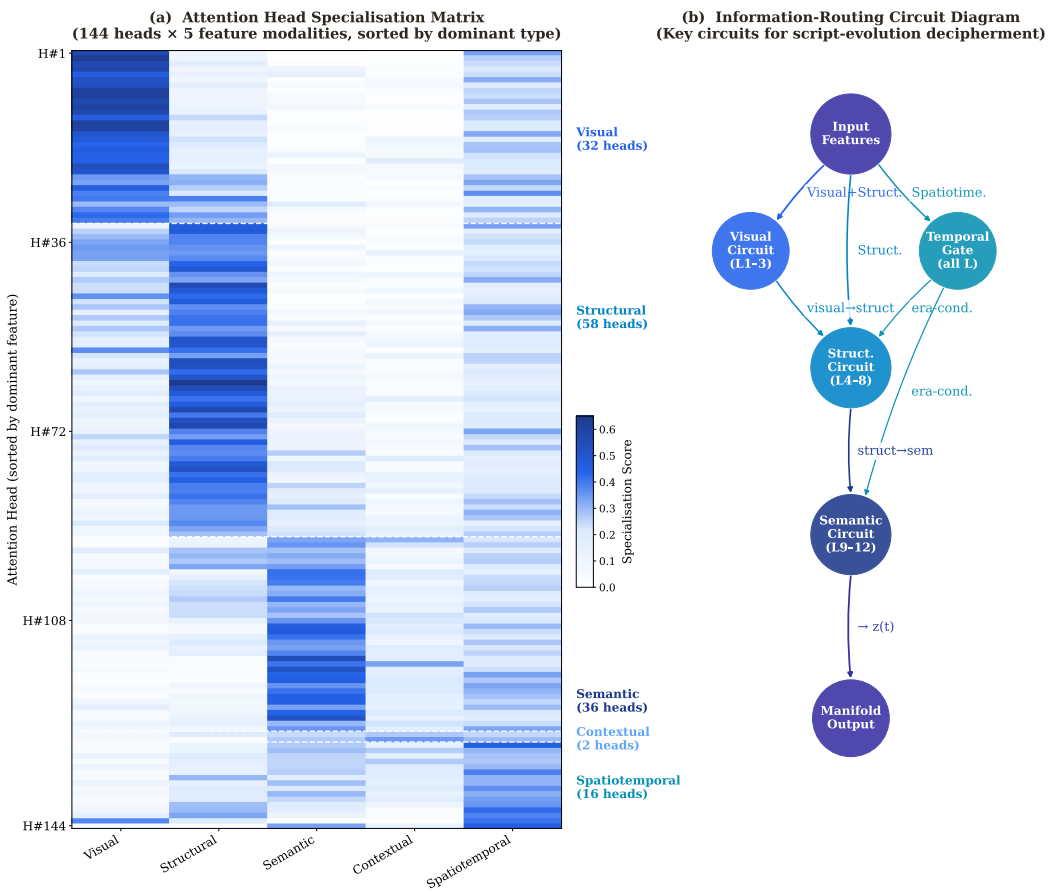


Figure 11: Retained, unvalidated attention-pattern visualization. The graphic shows 144 entries, but its correspondence to the reported checkpoint is not independently established here. Head preferences are descriptive, not causal or cognitive evidence.

D.7 Latent Vector Arithmetic

Vector analogies provide exploratory tests of local representation structure. For example, a cross-era analogy can be written as

$$a = z^{\text{Ox}}_{\text{OBI}} - z^{\text{Ox}}_{\text{Regular}} + z^{\text{Sheep}}_{\text{Regular}}.$$

The existing example reports a cosine similarity of 0.563 between a and $z^{\text{Sheep}}_{\text{OBI}}$, compared with a random-reference value of 0.035. These values characterize that example only. Ratios to near-zero random cosine values are not used as the principal effect measure. A broader evaluation should specify the analogy set, candidate vocabulary, rank-based metric, and random controls. Because semantic embeddings are included among the inputs, these results must not be described as emerging without linguistic information.

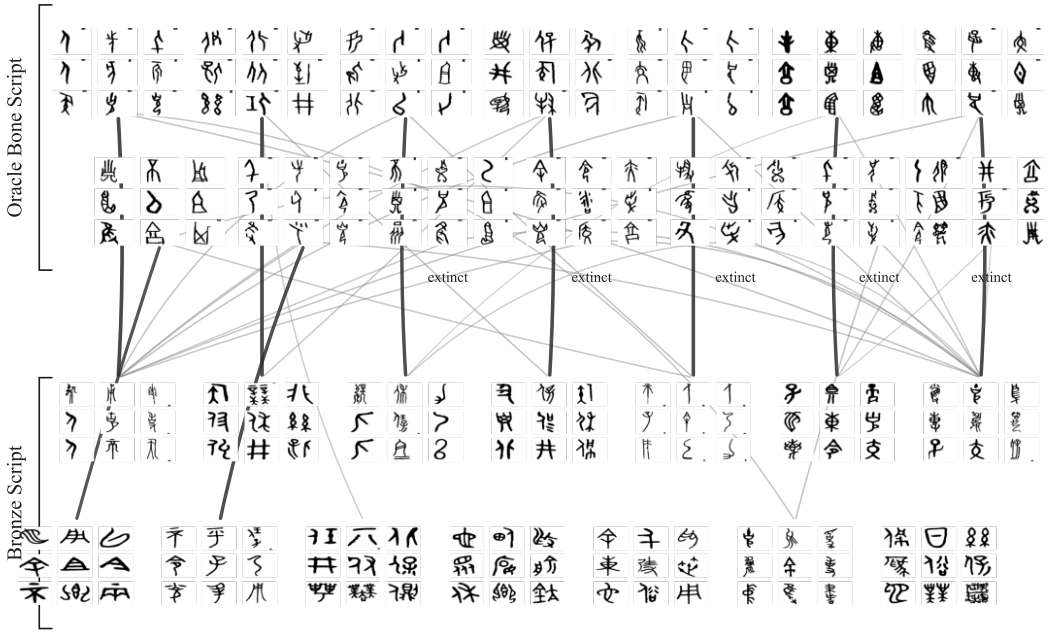


Figure 12: Reference variant network before the illustrated filtering procedure.

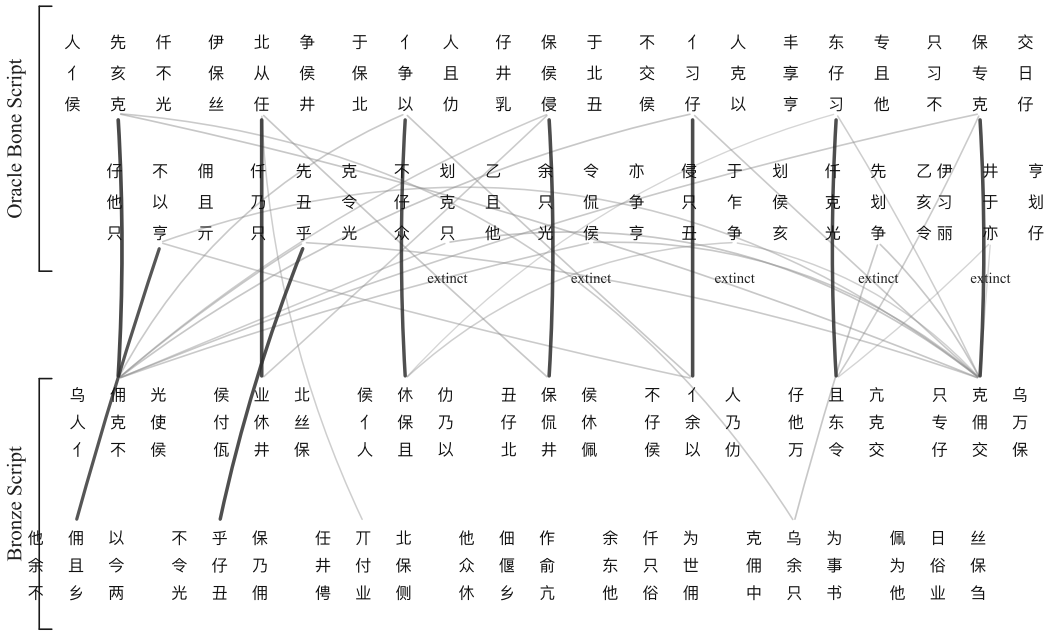


Figure 13: Reference network after the illustrated filtering procedure. Reduced graph size does not establish the historical invalidity of removed variants.



Figure 14: Selected latent vector analogies. Individual examples do not establish a globally character-independent transformation operator or disentangled semantic components.

E Full Method Details

MSEF consists of a time-conditioned encoder, a latent velocity field, and a survival-score module. CBED uses their outputs for candidate retrieval and backward verification.

E.1 Feature Representation

For an observed glyph x assigned time t, the input feature vector is

$$X = [f_{\text{vis}}; f_{\text{str}}; f_{\text{sem}}; f_{\text{ctx}}; f_{\text{st}}] \in \mathbb{R}^{352},$$

with dimensions 128, 64, 64, 64, and 32, respectively. Extraction and missing-feature handling are described in Appendix L.

E.2 Time Conditioning

The model uses the ordered time convention in Appendix K. Script type, scribal group, and archaeological date range are distinct metadata fields. A missing date is not an additional historical era.

E.3 Encoding and Evolution

Write

$$E_\theta(X, t) := \mathcal{M}(X, t; \theta).$$

An observed glyph is encoded at its assigned observation time:

$$z_t = E_\theta(X_t, t).$$

Its predicted state at another time u is

$$\widetilde{z}_u = \phi_{t\to u}(z_t).$$

In general,

$$E_\theta(X_t, u) \neq \phi_{t\to u}(E_\theta(X_t, t)).$$

Changing the encoder's time argument is therefore not interchangeable with ODE integration. Neighborhood visualizations are provided in Appendix Q.

E.4 Architecture

The main-text configuration uses a 12-layer Transformer encoder with a 256-dimensional output. The velocity field is a 3-layer MLP with spectral normalization and Tanh activations. The survival module is an MLP conditioned on the initial OBI representation and target time. The available architecture settings are listed in Appendix M.

E.5 Optimization and Inference

The encoder and velocity field are optimized using the evolution objectives in Appendix F. The survival objective has separate labels; its gradient routing is an implementation detail not determined by the objective alone. CBED first obtains candidate glyphs from later eras, then compares backward-predicted states with available reference prototypes and the query. Algorithm 1 specifies the proposed consistent inference procedure.

F Full Loss Definitions

Indexing convention. In pair sums, (c, i, j) denotes a particular observed correspondence record, including its glyph variants, rather than one record per identity and era pair. In the complete-chain sum below, $c \in \mathcal{C}_{\text{complete}}$ indexes a chain record; several chain records can share a character identity. This avoids conflating glyph paths with unique characters. Let $E_\theta(X, t) := \mathcal{M}(X, t; \theta)$. For an observed glyph of character c at time t_i, write

$$z_i^c = E_\theta(X_i^c, t_i).$$

Let $\mathcal{P}_{\text{adj}}$ and $\mathcal{P}_{\text{skip}}$ contain verified training pairs from adjacent and non-adjacent script periods, respectively. Let

$$\mathcal{P} = \mathcal{P}_{\text{adj}} \cup \mathcal{P}_{\text{skip}}.$$

A complete chain is a separate record in $\mathcal{C}_{\text{complete}}$, not an additional unit to be added to a pair count.

F.1 Evolution Objectives

For nonempty supervision sets, define

$$\mathcal{L}_{\text{adj}} = \frac{1}{|\mathcal{P}_{\text{adj}}|} \sum_{(c,i,j)\in\mathcal{P}_{\text{adj}}} \left\| \phi_{t_i \to t_j}(z_i^c) - z_j^c \right\|_2^2, \tag{8}$$

$$\mathcal{L}_{\text{skip}} = \frac{1}{|\mathcal{P}_{\text{skip}}|} \sum_{(c,i,j)\in\mathcal{P}_{\text{skip}}} \left\| \phi_{t_i \to t_j}(z_i^c) - z_j^c \right\|_2^2, \tag{9}$$

$$\mathcal{L}_{\text{full}} = \frac{1}{|\mathcal{C}_{\text{complete}}|} \sum_{c\in\mathcal{C}_{\text{complete}}} \left\| \phi_{t_0^c \to t_4^c}(z_0^c) - z_4^c \right\|_2^2, \tag{10}$$

$$\mathcal{L}_{\text{cyc}} = \frac{1}{|\mathcal{P}|} \sum_{(c,i,j)\in\mathcal{P}} \left\| \psi_{t_j \to t_i}\left(\phi_{t_i \to t_j}(z_i^c)\right) - z_i^c \right\|_2^2. \tag{11}$$

The endpoints in $\mathcal{L}_{\text{full}}$ are the assigned observation times of the chain. They are not automatically zero and one when record-specific times are used.

Meaning of the full-chain term. $\mathcal{L}_{\text{full}}$ is an *endpoint-alignment loss on complete chains*. It compares the first and last observations, not every intermediate observation. If the same endpoint pair is also included in $\mathcal{P}_{\text{skip}}$, the term reweights that endpoint supervision; it is not an independent intermediate-path constraint. Intermediate observations enter through the observed pair losses and candidate verification.

The evolution objective is

$$\mathcal{L}_{\text{evolution}} = \lambda_{\text{adj}}\mathcal{L}_{\text{adj}} + \lambda_{\text{skip}}\mathcal{L}_{\text{skip}} + \lambda_{\text{full}}\mathcal{L}_{\text{full}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}.$$

When a minibatch contains no examples for a term, that term contributes zero rather than dividing by an empty-set size.

F.2 Survival Supervision

We use $s_\omega(z_0^c, t)$ to denote the target-era survival score, with label $y_{c,t} = 1$ for supported survival and $y_{c,t} = 0$ for a documented negative outcome under the annotation protocol. If the implemented network outputs extinction probability e_ω, then

$$s_\omega = 1 - e_\omega.$$

For labeled examples $\mathcal{D}_s$,

$$\mathcal{L}_{\text{surv}} = -\frac{1}{|\mathcal{D}_s|} \sum_{(c,t,y)\in\mathcal{D}_s} \left[y \log s_\omega(z_0^c, t) + (1-y)\log(1 - s_\omega(z_0^c, t)) \right]. \tag{12}$$

Absence of an excavated or catalogued descendant is not automatically a confirmed negative label. Unknown outcomes require separate treatment.

F.3 Gradient Paths

Let θ, η, and ω parameterize the encoder, velocity field, and survival module, respectively. The evolution objective supplies gradients with respect to θ and η, while the survival objective supplies a gradient with respect to ω. If the OBI encoding remains connected to the encoder, the chain rule additionally gives

$$\nabla_\theta \mathcal{L}_{\text{surv}} = (D_\theta E_\theta)^\top \nabla_{z_0} \mathcal{L}_{\text{surv}}.$$

With a detached encoding this contribution is zero. These are different implementations, not consequences of using separate loss names. The available configuration does not identify which gradient route produced the reported checkpoints, so neither route is asserted as a verified implementation here. Examples without an observed OBI input do not supply z_0 for this loss.

F.4 Interpretation of Cycle Consistency

For the same velocity field and exact integration,

$$\psi_{t_j \to t_i} \circ \phi_{t_i \to t_j} = \mathrm{id}.$$

Thus the round-trip loss primarily measures numerical self-consistency. It does not independently establish correct cross-era correspondence. Candidate verification instead compares

$$\psi_{t_j \to t_i}(z_j^c)$$

with the observed query encoding. The candidate representation is an independently observed target state rather than the query's own forward prediction.

F.5 Discriminative Representation Requirements

The alignment objectives alone admit a degenerate solution:

$$E_\theta(X, t) \equiv a, \qquad v_\eta(z, t) \equiv 0.$$

All four evolution losses are then zero, although the representation cannot distinguish characters. This is a counterexample to a guarantee of discriminative learning from these objectives alone; it does not prove that the reported training runs actually converged to that solution. Freezing an offline feature extractor alone does not prevent a subsequent trainable encoder from becoming constant. The manuscript does not document an implemented anti-collapse mechanism sufficient to exclude this solution. Consequently, no such guarantee is claimed, and no unrun contrastive, classification, triplet, or variance loss is added to explain existing results.

G Full CBED Algorithm

Output scope. CBED returns Regular-script candidate identities or abstains. A low survival score rejects the entire query in this algorithm; it does not return an early-era reading for a character without a Regular-script descendant. Abstention records model rejection, not a confirmed historical extinction. This section specifies CBED using ODE-based forward prediction and backward checks at available reference observations.

G.1 Inputs and Reference Information

Let

$$t_0 < t_1 < t_2 < t_3 < t_4$$

be the declared inference checkpoints for OBI, Bronze, Seal, Clerical, and Regular scripts. For clarity, the pseudocode uses one checkpoint per era. Finer-grained observations add assigned times to the ordered sequence. A comparison uses the same assigned time for the prototype and prediction; a prototype recorded at a different time needs an explicit transport operation. This algorithmic convention does not supply the missing numeric subperiod-to-time mapping. Let $\mathcal{D}_j$ contain reference observations assigned to checkpoint t_j, with embeddings

$$z_j^c = E_\theta(X_j^c, t_j).$$

The reference graph $\mathcal{G}_{\mathrm{ref}}$ contains only correspondences permitted by the evaluation protocol. It excludes held-out query identity labels and query-to-descendant answer edges. For previously retrieved observations A, $\mathrm{Reach}(A, t_j; \mathcal{G}_{\mathrm{ref}})$ returns reachable reference observations at t_j through permitted time-ordered edges.

G.2 Reference Paths and Missing Eras

A reference path p for candidate c identifies a Regular-script prototype and any available intermediate-era prototypes associated with it. An unavailable intermediate observation is represented as missing, not as a zero vector. When no intermediate reference observations exist, a path containing only the Regular prototype is allowed. Its verification is endpoint-only and must not be described as independently verified at every era. Let $\mathcal{P}(c)$ denote the permitted paths for candidate c.

G.3 PATH SCORES

For a path that passes the available checks, let $\widehat{z}_0^p$ be its backward-predicted OBI state. Define

$$r_p(q) = \frac{\langle \widehat{z}_0^p, z_0^q \rangle}{\max\{\|\widehat{z}_0^p\|_2 \|z_0^q\|_2, \varepsilon_{\cos}\}}, \qquad \varepsilon_{\cos} > 0.$$

The candidate score is

$$r(c, q) = \max_{p \in \mathcal{P}_{\mathrm{ok}}(c)} r_p(q).$$

A candidate with no valid path is excluded. This aggregation scores paths before combining their evidence; it does not average incompatible latent coordinates into a potentially misleading intermediate point.

Survival checks. The algorithm applies survival checks only through the Seal checkpoint. It does not reset cumulative survival to one after that checkpoint. Low survival score and failure of candidate verification are recorded as different rejection reasons.

No implicit re-anchoring. Forward query states are propagated by the ODE. Candidate retrieval expands a reference set; it does not reset the query state to the nearest reference prototype. By the flow-composition property, splitting exact integration into more segments does not change the final query state. Additional information in the cascade comes from intermediate reference retrieval, permitted graph expansion, and observed-prototype checks, not segmentation of the same integral alone.

Scores and evaluation. Compatibility scores are not calibrated probabilities. The score formula alone does not specify tie-breaking; the current summary also does not specify the verification thresholds or path-enumeration limits. These omissions limit exact reproduction of rankings. Retrieval metrics are computed after deduplicating candidate identities. For closed-set evaluation, an abstention counts as an incorrect prediction rather than removing the query from the denominator.

Algorithm 1 CBED with observed-checkpoint verification

Require: Query q; checkpoints $\{t_j\}_{j=0}^{4}$; reference databases $\{\mathcal{D}_j\}$; permitted graph $\mathcal{G}_{\mathrm{ref}}$; encoder E; flows ϕ, ψ; retrieval depth K; survival threshold τ_s; verification thresholds $\{\epsilon_j\}_{j=0}^{3}$
Ensure: Ranked candidates with compatibility scores, or abstention with a recorded reason

$X^q \leftarrow$ ExtractPermittedFeatures(q)
$z_0^q \leftarrow E(X^q, t_0)$
$\widetilde{z}_0^q \leftarrow z_0^q$
$A \leftarrow \varnothing$
for $j = 1, 2, 3, 4$ **do**
  if $j \leq 2$ and $s(z_0^q, t_j) < \tau_s$ **then**
    return ABSTAIN: low survival score
  end if
  $\widetilde{z}_j^q \leftarrow \phi_{t_{j-1} \to t_j}(\widetilde{z}_{j-1}^q)$
  $N_j \leftarrow \mathrm{TopK}(\widetilde{z}_j^q, \mathcal{D}_j, K)$
  $C_j \leftarrow N_j \cup \mathrm{Reach}(A, t_j; \mathcal{G}_{\mathrm{ref}})$
  $A \leftarrow A \cup C_j$
end for
$R \leftarrow \varnothing$
for each distinct candidate identity c in C_4 **do**
  $V_c \leftarrow \varnothing$
  for each permitted reference path $p \in \mathcal{P}(c)$ **do**
    $\widehat{z} \leftarrow z_4^p$
    valid $\leftarrow$ true
    for $j = 3, 2, 1, 0$ **do**
      if valid **then**
        $\widehat{z} \leftarrow \psi_{t_{j+1} \to t_j}(\widehat{z})$
        $u \leftarrow$ MISSING
        if $j = 0$ **then**
          $u \leftarrow z_0^q$
        else
          if prototype z_j^p is observed **then**
            $u \leftarrow z_j^p$
          end if
        end if
        if $u \neq$ MISSING and $\|\widehat{z} - u\|_2 > \epsilon_j$ **then**
          valid $\leftarrow$ false
        end if
      end if
    end for
    if valid **then**
      $V_c \leftarrow V_c \cup \{r_p(q)\}$
    end if
  end for
  if $V_c \neq \varnothing$ **then**
    $R \leftarrow R \cup \{(c, \max V_c)\}$
  end if
end for
if $R = \varnothing$ **then**
  return ABSTAIN: no consistent candidate
end if
return R sorted by decreasing compatibility score

H Dataset Details

H.1 Counting Units

The main text reports 1,358 character categories. The available summaries do not establish final train/validation/test identity counts, per-era image counts, distinct pair counts, complete-chain counts, or survival-label counts. Accordingly, Table 14 defines units but is not presented as a completed numerical dataset manifest. Neither the 586 nor the 1,500 historical chain figure is adopted as a verified replacement count. We distinguish four units: a character identity, an observed glyph image,

a cross-era correspondence pair, and an evolutionary chain. A correspondence pair identifies two observed glyph records and their script periods. A complete chain identifies linked observations across all five periods. Multiple glyph variants can produce several pairs or paths for the same character identity.

Table 14: Definitions of counting units used for FGCCES.

Record type	Definition
Character identity	Distinct correspondence label
Glyph observation	A recorded image in a script period
Correspondence pair	Two linked observed glyph records
Complete chain	Linked observations across all five periods

H.2 Sources and Missing Features

FGCCES combines records from jgwlbq, CCAMC, and the BNU character resources. Unavailable feature entries are zero-filled under the feature convention in Appendix L. A missing feature, a missing glyph observation, an unknown correspondence, and a confirmed negative lineage outcome are different states. They must not be collapsed into the same training label.

H.3 Train, Validation, and Test Separation

The intended evaluation uses character-disjoint training, validation, and test sets. Under this split definition, variants belonging to one held-out identity stay on the same identity side of the training/evaluation boundary. The text specifies the intended protocol; it does not certify a completed artifact-overlap or cross-benchmark leakage audit.

Query-side information. The intended simulated-undeciphered protocol excludes query answer identities, answer definitions, and answer-dependent annotations from feature construction. Actual per-benchmark masking records are not supplied in the available summary.

Candidate-side information. Public reference glyphs and permitted later-era correspondences are distinct from the held-out query's answer information. Matching a dataset name does not establish that competing systems had the same candidate-side information.

Cross-benchmark overlap. Shared source collections can overlap in images, glyph variants, identities, and artifact context. Character-disjoint splitting alone does not rule out these channels, including indirect answer information in neighboring same-identity variants. No completed cross-source overlap audit is claimed here.

H.4 Fine-Grained Visualization

Appendix R illustrates the retained scribal-group and subperiod metadata.

H.5 Comparison with Existing Resources

Appendix N distinguishes FGCCES correspondence records from recognition and multimodal benchmark resources.

I Theoretical Foundations

The following results concern the ODE in the Euclidean latent coordinates used by the model. They do not prove that the observed glyph data form a smooth manifold or that the feature encoder is invertible.

I.1 Regularity Assumptions

Let

$$v : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$$

be continuous in time and continuously differentiable in the latent state, with jointly continuous state Jacobian. Assume that v is uniformly globally L-Lipschitz in the state and that

$$\sup_{t\in[0,1]} \|v(0,t)\| < \infty.$$

These are sufficient assumptions for the flow properties below. A finite-sample estimate of a Lipschitz ratio is not a certificate that these assumptions hold globally.

I.2 Existence, Uniqueness, and Invertibility

Proposition: regular latent flow. Under the stated assumptions, the initial-value problem

$$\dot{z}(t) = v(z(t),t), \qquad z(s) = z_s$$

has a unique solution throughout $[0,1]$. The flow $\phi_{s\to t}$ is a C^1 diffeomorphism, with

$$\phi_{s\to t}^{-1} = \phi_{t\to s}.$$

Proof. Uniform Lipschitz continuity gives local existence and uniqueness. The bound

$$\|v(z,t)\| \le L\|z\| + \sup_u \|v(0,u)\|$$

prevents finite-time escape on the bounded time interval. Continuous differentiability of the vector field gives differentiable dependence on the initial condition. Solving the same ODE backward from time t recovers the unique initial state at s. □

I.3 Composition

For any $s,t,u \in [0,1]$,

$$\phi_{s\to u} = \phi_{t\to u} \circ \phi_{s\to t}.$$

Proof. Both sides solve the same initial-value problem from time s, so uniqueness implies equality. □

A deterministic latent flow therefore does not split a single initial point into multiple endpoints. Multiple glyph candidates arise from the observation and retrieval relation, not from non-uniqueness of the ODE solution.

I.4 Manifold Dimension

The main configuration uses $d = 256$. A dimension comparison must keep the dataset, evaluation protocol, encoder family, and parameter-counting convention explicit.

J Manifold Space Validation

This section separates conditional mathematical properties from empirical diagnostics of a particular trained model.

J.1 Conditional Properties and Diagnostics

J.1.1 Continuity and Sensitivity

Under Appendix I's assumptions, Grönwall's inequality gives

$$\|\phi_{s\to t}(z_1) - \phi_{s\to t}(z_2)\| \le e^{L|t-s|}\|z_1 - z_2\|.$$

A velocity bound additionally requires a bounded region of interest or a separate bound on the network output. It does not follow from local Lipschitz continuity alone.

Table 15: Reported finite-sample sensitivity diagnostics. Sampled maxima are not certified global bounds.

Reported quantity	Value
Estimated velocity-field Lipschitz ratio	2.34
Estimated flow Lipschitz ratio	8.7
Maximum sampled velocity norm	2.51
Mean adjacent-era distance	0.82
Reported trajectory smoothness statistic	0.15

J.1.2 Numerical Round-Trip Consistency

For exact integration of the same field,

$$\psi_{t\to s} \circ \phi_{s\to t} = \mathrm{id}.$$

Numerical implementations can have nonzero round-trip error. A small round-trip error does not establish correct historical correspondence. Conversely, a difficult glyph or a rejected candidate does not establish ODE stiffness. The survival module does not provide a mathematical guarantee against stiffness. Such a claim would require solver diagnostics, not merely lower retrieval error after rejection.

Table 16: Reported numerical consistency measurements. They are distinct from candidate-identification accuracy.

Metric	Value
OBI $\to$ Bronze $\to$ OBI error	0.023
OBI $\to$ Seal $\to$ OBI error	0.041
OBI $\to$ Regular $\to$ OBI error	0.068
Composition error	0.015

J.1.3 Conditional Proximity

Suppose the initial encoder is L_E-Lipschitz and $\|X_1 - X_2\| \le \delta$. Then

$$\|\phi_{s\to t}(E(X_1, s)) - \phi_{s\to t}(E(X_2, s))\| \le e^{L|t-s|} L_E \delta.$$

This follows by applying the encoder bound and then the flow bound. Shared semantic labels do not automatically imply the feature-distance premise. An upper distance bound also does not prove separation between different semantic groups.

Table 17: Reported radical-conditioned geometry. These are descriptive clustering measurements, not proofs of semantic preservation.

Metric	OBI	Bronze	Seal	Clerical	Regular
Intra-radical distance	1.35	1.28	1.18	1.12	1.08
Inter-radical distance	3.52	3.68	3.85	3.92	4.05
Distance ratio	2.61	2.88	3.26	3.50	3.75
Silhouette score	0.45	0.48	0.52	0.55	0.58

J.1.4 Velocity Direction

Let

$$u = v(z_1(t), t), \qquad w = v(z_2(t), t).$$

Assume

$$\min(\|u\|, \|w\|) \geq m > 0, \qquad \|z_1(s) - z_2(s)\| \leq \delta.$$

For $t \geq s$,

$$1 - \cos(u, w) \leq \frac{L^2 e^{2L(t-s)} \delta^2}{2m^2}.$$

Indeed,

$$2\|u\|\|w\|(1 - \cos(u, w)) = \|u - w\|^2 - (\|u\| - \|w\|)^2 \leq \|u - w\|^2,$$

and the flow and velocity Lipschitz bounds bound $\|u - w\|$. Without a nonzero speed lower bound, nearby states need not have aligned velocities.

Table 18: Unvalidated velocity-cosine and trajectory-angle summaries. The two columns are distinct reported quantities; the angle estimator is unspecified and is not inferred by applying arccos to a mean cosine.

Pair type	Velocity cosine	Trajectory angle
Same radical	0.78	11.2°
Same semantic category	0.71	15.8°
High co-occurrence	0.65	19.3°
Different radical	0.35	42.5°
Random pairs	0.31	45.7°

J.1.5 Bounded Volume Distortion

If

$$|\nabla \cdot v(z, t)| \leq C$$

on the trajectories of a measurable set A with positive finite volume, then

$$e^{-C|t-s|} \leq \frac{\mathrm{Vol}(\phi_{s \to t}(A))}{\mathrm{Vol}(A)} \leq e^{C|t-s|}.$$

Proof. The Jacobian determinant satisfies

$$\det D\phi_{s \to t}(z) = \exp\left(\int_s^t \nabla \cdot v(z(u), u)\, du \right).$$

Bounding the integral and applying change of variables gives the volume bounds. □

This result bounds volume distortion. It does not imply volume preservation or an expected Jacobian determinant of one.

Table 19: Reported volume-related diagnostics. A sampled maximum divergence is not a certified supremum over the state space.

Reported quantity	Value
Mean absolute divergence	0.08
Maximum sampled absolute divergence	0.35
Estimated OBI-to-Regular volume ratio	1.12
Mean Jacobian determinant	1.08

J.1.6 TRAJECTORY PERTURBATION BOUND

Let $\widehat{z}$ and z^* solve the learned and a reference dynamics, respectively. Assume the learned field is L-Lipschitz and

$$\|\widehat{v}(z,u) - v^*(z,u)\| \le \epsilon$$

on the relevant domain. For $t \ge s$, let

$$E(u) = \|\widehat{z}(u) - z^*(u)\|,$$

and $L > 0$,

$$E(t) \le e^{L(t-s)}E(s) + \frac{\epsilon}{L}\left(e^{L(t-s)} - 1\right).$$

For $L = 0$,

$$E(t) \le E(s) + \epsilon(t-s).$$

Proof. Almost everywhere,

$$\dot{E}(u) \le LE(u) + \epsilon.$$

The integrated inequality gives the result. At zero error, the same argument can be formulated using the upper right Dini derivative. □

This is a conditional perturbation bound. It does not prove optimizer convergence, statistical consistency, or recovery of a known true historical vector field.

Intermediate corrections. Segmenting integration does not change the exact flow. If an additional correction operator satisfies

$$E_i^+ \le \rho_i E_i^- + \delta_i,$$

then

$$E_i^- \le e^{L\Delta t_i}E_{i-1}^+ + \frac{\epsilon}{L}(e^{L\Delta t_i} - 1),$$
$$E_i^+ \le \rho_i\left[e^{L\Delta t_i}E_{i-1}^+ + \frac{\epsilon}{L}(e^{L\Delta t_i} - 1)\right] + \delta_i.$$

For $L = 0$, the factor $\epsilon(e^{L\Delta t_i} - 1)/L$ is replaced by its limit $\epsilon\Delta t_i$. An improvement requires conditions on the correction operator. Nearest-neighbor retrieval does not automatically satisfy such conditions. The candidate expansion in Algorithm 1 does not itself reset the query state.

J.1.7 DENSITY EVOLUTION AS BACKGROUND

For a smooth probability density transported by the latent flow, the instantaneous change-of-variables relation is

$$\frac{d}{dt}\log p(z(t),t) = -\nabla \cdot v(z(t),t),$$

as used in continuous normalizing flows (Chen et al., 2018). Subject to the required integrability conditions, the differential entropy satisfies

$$\frac{dH(t)}{dt} = \mathbb{E}_{p_t}[\nabla \cdot v(z,t)].$$

These identities are background properties of transported continuous densities. They do not implement the survival module, prove historical standardization, or imply that distinct points merge under a bijective finite-time flow. Mean absolute divergence does not determine the sign of entropy change.

J.2 VELOCITY FIELD ANALYSIS

J.2.1 MAGNITUDE BY ERA

These values depend on latent scaling and the selected time coordinate. They are not changes per calendar year. Under a differentiable, strictly increasing reparameterization $u = g(t)$ with $g'(t) > 0$,

$$\frac{dz}{du} = \frac{v(z,t)}{g'(t)}.$$

Consequently, comparisons of historical change rates require an explicitly justified time scale.

Table 20: Reported latent velocity magnitudes under the experimental coordinate system.

Transition	Mean norm	Standard deviation
OBI → Bronze	1.85	0.42
Bronze → Seal	1.42	0.31
Seal → Clerical	1.12	0.25
Clerical → Regular	0.45	0.12

J.2.2 Direction by Era

Table 21 retains the reported direction-summary series. The estimator behind "direction statistic" and the rule behind "clusters for 80%" are not specified. These entries are unvalidated observations, not a measured entropy, a certified dispersion statistic, or evidence of historical standardization. No missing estimator is inferred from their numeric range.

Table 21: Reported direction-dispersion diagnostics.

Era	Direction statistic	Clusters for 80%
OBI	0.42	5
Bronze	0.31	4
Seal	0.25	3
Clerical	0.18	2
Regular	0.12	2

J.2.3 Character-Type Analysis

Table 22: Reported velocity norms by character-type annotation.

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

Character-type labels describe the analysis groups. Their association with velocity does not establish a causal mechanism.

J.3 Local Neighborhood Illustration

Figure 15 illustrates a local neighborhood around an Ox anchor and a dissimilar Wood reference. It is a schematic, not the empirical radical-group scatter plot suggested by an earlier caption. Its displayed proximity therefore carries no quantitative distance or clustering guarantee.

J.4 Summary

The conditional results characterize the latent ODE under explicit assumptions. The empirical measurements describe a finite set of representations and trajectories. Neither low-dimensional visualization nor small numerical round-trip error establishes historical correctness, global semantic preservation, or absence of memorization.

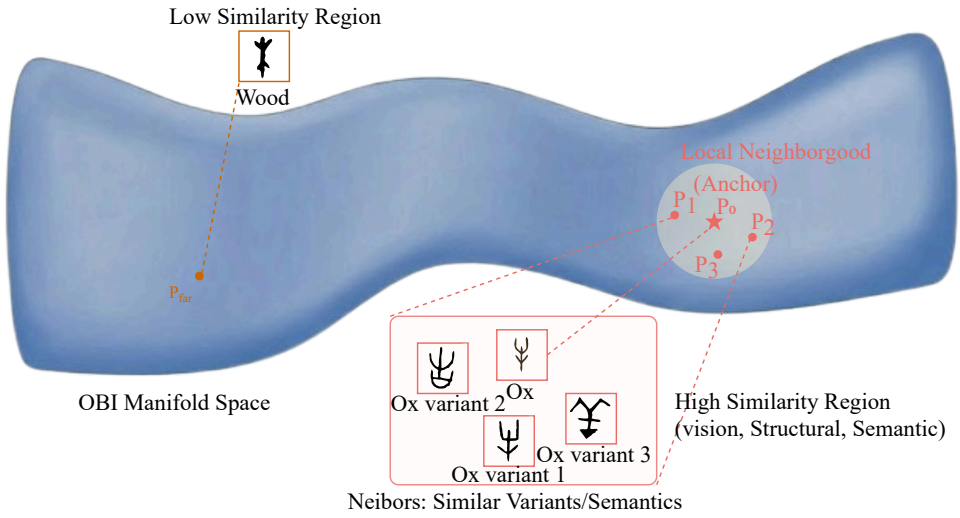


Figure 15: Schematic local neighborhood around an OBI "Ox" anchor. Nearby points illustrate similar variants or semantic associations; the distant "Wood" example illustrates a dissimilar reference. The surface is an illustration, not a measured radical-group scatter plot.

K Time Encoding Details

The following specifies the ordered era intervals and their interpretation. A record-level numeric time map, within-interval sampling distribution, and fixed inference anchors are not present in the available configuration. Thus the era intervals do not constitute a complete executable fine-grained time assignment. No uniform sampling rule or equally spaced scribal-group chronology is assumed.

K.1 Ordered Time Coordinate

We use the ordered script-period convention declared in Table 1:

$$\begin{aligned} I_{\text{OBI}} &= [0, 0.30), \\ I_{\text{Bronze}} &= [0.30, 0.70), \\ I_{\text{Seal}} &= [0.70, 0.85), \\ I_{\text{Clerical}} &= [0.85, 1), \\ t_{\text{Regular}} &= 1. \end{aligned}$$

These intervals are model coordinates rather than durations proportional to calendar years. Their ordering is a modeling convention and does not assert that script categories occupy non-overlapping archaeological date ranges.

K.2 Fine-Grained Metadata

Scribal group, regional provenance, script category, and archaeological date range are recorded separately. A scribal-group label is not automatically an ordered chronological bin. The label *None* denotes an unclassified sample, not an additional historical period.

K.3 Training and Inference Times

During training, the stated interval-sampling procedure samples times within the permitted intervals while preserving the order of supervised correspondences. Sampling additional time values does not create additional historical observations or determine a unique trajectory between observed eras.

K.4 Consequences for Interpretation

Latent velocities depend on the chosen time parameterization. Accordingly, velocity magnitude is interpreted as change per unit model time, not as a measured rate of script change per year.

L Feature Extraction Details

The input is a 352-dimensional concatenation of five feature groups:

$$X = [f_{\text{vis}}; f_{\text{str}}; f_{\text{sem}}; f_{\text{ctx}}; f_{\text{st}}].$$

L.1 Visual Features: 128 Dimensions

The visual representation contains contour descriptors (32 dimensions), stroke descriptors (32), topological descriptors (16), symmetry descriptors (16), and density/appearance descriptors (32).

L.2 Structural Features: 64 Dimensions

The structural representation contains character-formation encoding (6 dimensions), component/radical representation (42), and spatial-layout representation (16). Image-observable structural annotations are distinguished from annotations that require a known modern reading.

L.3 Semantic Features: 64 Dimensions

The semantic block encodes dictionary definitions using a BERT-based representation followed by the dimensionality reduction used in the implementation. In the intended undeciphered-query protocol, the answer definition is unavailable and its semantic block is zero-filled. Providing that definition to a labeled benchmark query would instead evaluate a different information setting; the benchmark labels alone do not establish that the masking was actually executed.

L.4 Context Features: 64 Dimensions

Let $\mathcal{N}(c)$ be the other observed characters on the same archaeological artifact. Define

$$h^{c'} = [f_{\text{vis}}^{c'}; f_{\text{str}}^{c'}; f_{\text{sem}}^{c'}] \in \mathbb{R}^{256}.$$

Making the projection explicit,

$$f_{\text{ctx}}^{c} = \begin{cases} W_{\text{ctx}} \dfrac{1}{|\mathcal{N}(c)|} \displaystyle\sum_{c' \in \mathcal{N}(c)} h^{c'}, & |\mathcal{N}(c)| > 0, \\ 0 \in \mathbb{R}^{64}, & |\mathcal{N}(c)| = 0, \end{cases}$$

where

$$W_{\text{ctx}} \in \mathbb{R}^{64 \times 256}.$$

L.5 Spatiotemporal Features: 32 Dimensions

The spatiotemporal block contains a 16-dimensional temporal/provenance encoding and a 16-dimensional spatial encoding. The temporal information describing the observed artifact is distinct from the target time supplied to the dynamics. Transporting a latent state does not change the query artifact's actual provenance.

L.6 Missing Features and Evaluation Availability

Unavailable feature entries are zero-filled. This preserves input dimensionality but does not imply that missingness has no effect on prediction. Table 23 gives the intended information boundary. It is a protocol specification, not a retrospective verification of every benchmark run. Candidate-side reference information is distinct from query-side answer information.

Table 23: Intended information-access rules; execution is not certified by this table.

Field	Query side	Training/reference side
Glyph and provenance	Observed image and artifact metadata	Observed reference records
Answer identity	Excluded from model input	Training identities only; candidate labels are reference outputs
Dictionary definition	Query answer definition excluded; unavailable block zero-filled	Definitions of permitted reference identities
Structural annotation	Image-observable components only; no answer-dependent labels	Permitted reference annotations
Artifact context	Available neighbors; no hidden query answer or answer-equivalent variant	Separate from proof of artifact-disjoint splitting
Correspondence graph	No held-out query-to-answer edges	Only permitted reference correspondences

L.7 Scope of End-to-End Training

The offline feature pipeline is distinct from the trainable encoder and dynamics. Here, end-to-end optimization refers to the trainable components explicitly connected by the objectives in Appendix F. We do not infer the necessity of handcrafted features from unreported comparisons with deep image encoders. Any such comparison requires matched data, training budgets, and evaluation conditions.

M Implementation Details

Scope of the configuration. The settings below preserve the documented configuration. The Transformer hidden width, heads per layer, feed-forward width, feature-to-token construction, precise time injection, checkpoint selection, seed list, and per-checkpoint pruning thresholds are not specified. The 140M headline and the 144-head graphic cannot be reconstructed from the 352-dimensional input and 256-dimensional output alone. No undocumented encoder is introduced to reconcile the parameter count. The configuration below follows the architecture and training setup declared in the main text.

M.1 Architecture

The encoder is a 12-layer Transformer producing a 256-dimensional latent representation. The velocity network is a 3-layer MLP with spectral normalization and Tanh activations. The survival module is conditioned on the initial OBI representation and target time.

M.2 Numerical Integration

Forward and backward integrations use the same learned velocity field. Solver tolerances control numerical error estimates; they do not guarantee exact reversibility or correct character correspondence.

M.3 Optimization

The survival supervision objective is defined in Appendix F. A separate survival objective does not imply that every cross-era pair supplies a survival label.

M.4 Parameter Counts and Timing

Parameter counts distinguish trainable model parameters from frozen feature extractors. Timing distinguishes offline preprocessing, index construction, model inference, and candidate verification. The main text reports approximately 18 hours on two NVIDIA H100 GPUs for a training run. Appendix S gives the corresponding accounting.

Table 24: Architecture and optimization configuration.

Setting	Value
Input dimension	352
Latent dimension	256
Encoder layers	12 Transformer layers
Velocity network	3-layer MLP
Velocity activation	Tanh
Velocity normalization	Spectral normalization
ODE solver	`dopri5`
Relative tolerance	10^{-5}
Absolute tolerance	10^{-5}
Optimizer	AdamW
Learning rate	10^{-4}
Weight decay	10^{-4}
Batch size	256
Training epochs	100
Gradient clipping norm	1.0
Independent runs	5
λ_{adj}	1.0
λ_{skip}	0.5
λ_{full}	0.3
λ_{cyc}	0.5
Per-checkpoint retrieval depth	5
Survival-score threshold	0.1

N Dataset Construction Details

Relation to prior cross-era data. EVOBC already provides a character-evolution resource spanning six historical stages.[1] FGCCES is positioned here by its fine-grained metadata and organization of partial correspondences, not by a claim that no earlier cross-era dataset exists. The available documentation does not establish priority for the broader category of unified cross-era datasets.

N.1 Source Records

FGCCES combines OBI records from jgwlbq, Bronze-script records from CCAMC, and later-script records from the BNU character resources. Each released correspondence should be traceable to its source glyph records and supporting documentation.

N.2 Correspondence Verification

The construction process uses dictionary cross-references and expert assessment of candidate correspondences. The dictionaries cited in the source manuscript include *Shuowen Jiezi*, *Jiaguwen Zidian*, and *Jinwen Bian*. Disputed correspondences should remain distinguishable from agreed correspondences. Variant-level paths should retain the provenance of each edge.

N.3 Pairs, Chains, and Availability

The data support training from partial correspondences rather than only complete five-era chains. Counts use the units defined in Appendix H. The earlier aggregate availability chart is omitted because its source-level character totals and 1.5K+ complete-chain bar are not reconciled with a final FGCCES manifest. Table 14 defines the four units; it does not contain numerical counts against

[1] Haisu Guan et al., *An open dataset for the evolution of oracle bone characters: EVOBC*, 2024, arXiv:2401.12467v2. `https://arxiv.org/abs/2401.12467`.

which that chart could be verified. Partial correspondences remain a supported data organization without assigning an unverified aggregate size.

N.4 Comparison with Existing Resources

FGCCES emphasizes cross-era correspondence records and associated provenance. Recognition datasets and multimodal benchmarks serve related but different evaluation purposes. We do not equate the MSEF model with the FGCCES dataset, or equate a pair count with a complete-chain count.

Table 25: Roles of selected existing resources and FGCCES. This table compares task focus, not dataset scale.

Resource	Role in this comparison
Oracle-241	Character recognition
OBC306	Character recognition
OBI-IJDH	OBI image resource
OracleSage	Multimodal OBI resource
EVOBC	Multi-era character evolution resource
FGCCES	Cross-era correspondence records

O Additional Experimental Results

Interpretation of retained summaries. Tables explicitly labeled *unvalidated* preserve supplied aggregate observations without asserting that their protocols have been reconstructed. They are not used to establish a matched method advantage, a calibrated success probability, or a common-hardware speedup. No absent log, adapter, seed, or sample-level record is inferred from the displayed numbers. This section supplements the main experiments. Benchmark-specific inputs, candidate collections, and aggregation rules are distinguished explicitly.

O.1 Multi-Round Decipherment

Repeated identical inference from a deterministic model does not produce additional candidates. A multi-round evaluation therefore requires a specified source of variation or a rule for selecting different ranked candidates. For attempts $\widehat{y}_q^{(1)}, \ldots, \widehat{y}_q^{(n)}$ and valid-answer set Y_q, a cumulative hit rate is

$$H_n = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}\left[\exists r \le n : \widehat{y}_q^{(r)} \in Y_q\right].$$

This is not the same as single-attempt Top-1 accuracy, OCR Top-K, or recall at rank n. The definition of H_n does not establish that the retained score series in Table 26 was computed using that definition. Repeated deterministic inference alone cannot produce additional attempts.

Table 26: Unvalidated multi-round score summaries (percent). Attempt generation, feedback, and duplicate handling are not identified by the supplied records; these values are not established cumulative hit rates or evidence of independent candidate diversity.

Method	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

O.2 Benchmark-Specific Results

FGCCES ranks character/glyph candidates. PictOBI-20k instead asks a model to choose a real-object image corresponding to an OBI image (Chen et al., 2025). The available MSEF specification does not identify an encoder of those object images or a glyph-to-option adapter. Likewise, it does not identify whether the OCR interface consumes retrieved glyph images or directly scores character identities. OCR Top-K and the retrieval model's own Top-K are different rankings. Thus the native retrieval algorithm alone does not establish comparability of the PictOBI and OCR scores.

Aggregation of PictOBI subgroups. For disjoint Normal and Complex groups on a common test population, micro-averaging gives $A = wA_{\text{Normal}} + (1 - w)A_{\text{Complex}}$, with a common w across models. The displayed Gemini and Claude summaries imply $w \simeq 0.9011$, whereas the MSEF summary implies $w \simeq 0.8777$. The supplied subgroup denominators and aggregation rule do not resolve this difference. The reported 72.18 is not replaced by a newly inferred accuracy, and is not treated here as a verified common-population micro-average.

Baseline-column provenance. The supplied review identifies an attribution issue in the Pix2Pix column of Table 2: its Top-50/100/200/500 values (4.5, 13.0, 20.0, 21.5) correspond to a DRIT++ series in the cited source. Without a reproduction log identifying that column, the appendix does not interpret it as a verified independent Pix2Pix reproduction or use it to draw a method-specific conclusion. This issue is separate from the MSEF output adapter.

Table 27: Main-text score summaries under different prediction interfaces. Differences are arithmetic only: task adapters and information matching are not established, and the PictOBI aggregate has an unresolved subgroup-weighting convention.

Evaluation	Comparator	MSEF	Difference
OBS-OCR Top-1	OBSD: 41.0	71.5	30.5
PictOBI accuracy	Gemini: 53.66	72.18	18.52
FGCCES R@1	OracleAgent: 62.8	72.5	9.7

The metrics differ in their prediction spaces and evaluation interfaces. We do not infer robustness merely because the reported values are numerically similar.

O.3 Additional Ablation Studies

Table 28: Reported retrieval results by era pair.

Task	R@1	R@5
OBI $\rightarrow$ Bronze	80.2 ± 0.9	90.5 ± 0.7
Bronze $\rightarrow$ Seal	85.5 ± 0.7	93.8 ± 0.5
Seal $\rightarrow$ Clerical	87.8 ± 0.6	95.2 ± 0.4
Clerical $\rightarrow$ Regular	89.2 ± 0.5	96.0 ± 0.3
OBI $\rightarrow$ Seal	74.5 ± 1.0	86.2 ± 0.8
OBI $\rightarrow$ Clerical	73.2 ± 1.1	85.0 ± 0.9
OBI $\rightarrow$ Regular	72.5 ± 0.8	86.5 ± 0.6

Differences between tasks can also reflect candidate-set size, query composition, and available reference coverage. They are not attributed solely to historical standardization.

O.4 Effect of Intermediate Eras

The reported results increase as eras are added. Because supervision and reference coverage also change, this experiment does not isolate the effect of the number of integration checkpoints.

Table 29: Unvalidated era-augmentation summary A. This identifier denotes a report series, not a verified run configuration. Unreconciled pair counts are omitted.

Eras	R@1	R@5
OBI, Regular	61.2 ± 1.2	76.5 ± 1.0
+ Bronze	65.7 ± 1.0	80.8 ± 0.8
+ Seal	69.5 ± 0.9	84.2 ± 0.7
+ Clerical	72.5 ± 0.8	86.5 ± 0.6

O.5 Fairness of Cross-Era Data Utilization

Summaries A and B report different intermediate scores: 61.2 versus 63.5 for the two-era case, 65.7 versus 66.8 after Bronze, and 69.5 versus 70.2 after Seal. Their run identifiers, split correspondence, and reference-access differences are not established. The series are therefore neither pooled nor described as replications, and the absent configuration difference is not invented. Using additional cross-era evidence is a useful capability, but it does not eliminate the need for matched comparisons. Training examples, feature availability, candidate vocabularies, and inference-time reference graphs are separate resources. The 0.7-point difference between 63.5 and 62.8 does not alone establish

Table 30: Unvalidated restricted-supervision summary B. Configuration equivalence with summary A is not established; no matched-information advantage is inferred. Unreconciled pair/chain counts are omitted.

Reported configuration	R@1
OracleSage	60.1
OracleAgent	62.8
MSEF, OBI–Modern only	63.5 ± 1.1
MSEF, complete chains only	52.8 ± 1.3
MSEF, + Bronze	66.8 ± 1.0
MSEF, + Seal	70.2 ± 0.9
MSEF, full	72.5 ± 0.8

a statistically reliable methodological advantage. Uncertainty and a specified comparison test are required. Joint-Consistent CrossFont from Table 9 must also be considered: its reported OBI-to-Regular point estimate equals 72.5.

O.6 Error Analysis

The original analysis assigns 200 failure cases to the categories below. These categories describe observed failure contexts, not mutually exclusive causal explanations of model error. Damage and

Table 31: Reported coding of 200 failure cases.

Assigned category	Share (%)
Glyph damage	45
Character splitting or correspondence ambiguity	28
Scribal variants	16
Other/model-related diagnostics	11

variation are also conditions a decipherment model is expected to handle. The table therefore does not justify the claim that only 11 percent of errors are model limitations.

O.7 Compatibility Scores and Calibration Scope

No numerical calibration result is established from the available aggregate summaries. The equations below define a possible per-query analysis rather than report a completed calibration experiment. Empty bins are omitted, and all quantities refer to the same evaluation population. Calibration conditional on accepted queries would need its acceptance coverage reported separately. A normalized retrieval score is not automatically a calibrated probability. Let Q be the declared evaluation population, p_q a confidence value, and $a_q \in \{0, 1\}$ the corresponding correctness indicator. For confidence bins B_b, define

$$\mathrm{acc}(B_b) = \frac{1}{|B_b|} \sum_{q \in B_b} a_q,$$

$$\mathrm{conf}(B_b) = \frac{1}{|B_b|} \sum_{q \in B_b} p_q,$$

$$\mathrm{ECE} = \sum_b \frac{|B_b|}{|Q|} \left| \mathrm{acc}(B_b) - \mathrm{conf}(B_b) \right|.$$

For micro-averaging over the same population,

$$\mathrm{Acc}(Q) = \sum_b \frac{|B_b|}{|Q|} \mathrm{acc}(B_b).$$

O.8 Scribal-Group Analysis

Table 32 retains group-specific score summaries. The earlier count column summed to 1,400, but its unit and deduplication convention were not identified. It is omitted rather than relabeled as images, identities, or artifacts without evidence. The five named groups are a subset of the metadata labels, not an exhaustive chronology. Without common candidate sets and sampling records, these differences do not establish historical causes of accuracy.

Table 32: Unvalidated score summaries for five named scribal groups. The count column is omitted because its counting unit is unresolved.

Group	R@1	R@5
Bin	75.8 ± 0.9	88.5 ± 0.7
Li	73.5 ± 1.0	86.8 ± 0.8
Chu	72.2 ± 1.1	85.5 ± 0.9
Zi	70.5 ± 1.2	83.8 ± 1.0
Huang	68.2 ± 1.3	81.5 ± 1.1

O.9 Computational Efficiency

Table 33: Unvalidated computational summaries. Common hardware, batching, precision, timing boundaries, and parameter-counting scope are not established; no cross-method speed or memory advantage is inferred.

Method	Seconds/query	Parameters	GPU memory
OBSD	0.85	89M	8GB
OracleFusion	0.95	89M	8GB
OracleSage	1.8	7B	24GB
OracleAgent	2.3	7B	28GB
MSEF	0.18	140M	6GB

Using the displayed parameter counts, $7\mathrm{B}/140\mathrm{M} = 50$, not 580. This arithmetic does not verify that the parameter-counting scopes are comparable.

O.10 Comparison with a Diffusion-Based Probability-Flow ODE

The value 0.023 in Table 16 denotes an OBI–Bronze–OBI round trip, whereas 0.068 denotes an OBI–Regular–OBI round trip. Neither is substituted into a cross-model comparison without a matched integration path, solver, and state-normalization convention. The baseline schedule, conditioning, and supervision are not fully specified, so the retained retrieval summaries are not a controlled comparison of all probability-flow and Neural ODE models. This comparison concerns the particular probability-flow ODE baseline implemented for the experiment. It does not establish limitations of every diffusion-based or time-conditioned model.

Table 34: Unvalidated retrieval summaries for two dynamics parameterizations. The previous cycle-error column is omitted because its integration paths are not matched.

Model	R@1	R@5
Diffusion PF-ODE	61.4	76.8
MSEF Neural ODE	72.5	86.5

A smaller numerical cycle error is not independent proof of semantic correctness or topological superiority.

O.11 Backward Verification and Step-Wise Pruning

Let C_q be the forward candidate set and Y_q the set of valid answers. Define

$$TP_q = C_q \cap Y_q, \qquad FP_q = C_q \setminus Y_q.$$

For retained candidates C_q^{keep}, the candidate-weighted false-positive removal rate is

$$\mathrm{FPDR} = \frac{\sum_q |FP_q \setminus C_q^{\mathrm{keep}}|}{\sum_q |FP_q|}.$$

The true-positive retention rate is

$$\mathrm{TPR} = \frac{\sum_q |TP_q \cap C_q^{\mathrm{keep}}|}{\sum_q |TP_q|}.$$

These metrics are defined only for nonzero denominators.

Table 35: Unvalidated candidate-filtering diagnostics. The analyzed candidate population and weighting are not established; the percentages are not independently verified estimates of the FPDR and TPR definitions in the text.

Candidate type	Mean backward error	Removal	Retention
Valid targets	0.025	1.2%	98.8%
Distractors	0.142	82.4%	17.6%

In Table 5, removing bidirectional verification changes R@1 from 72.5 to 66.3, a difference of 6.2 percentage points. Removing cascaded retrieval gives 64.0 and is a different ablation. The diagnostic score separation supports testing a verification rule. It does not guarantee rejection of all incorrect candidates or retention of every valid answer.

O.12 Scope of Expert Agreement and Significance Claims

The main text reports outcomes for 100 cases in a 72/12/16 split. Agreement with experts, acceptance after further discussion, and validation by independent historical evidence are distinct outcomes. The available summary does not identify case-level supporting sources, validation dates, or adjudication

records for the 12 cases labeled "later validated". That category is therefore not interpreted here as independently confirmed new decipherments.

The main-text significance symbols are retained source assertions, not tests recomputed in this appendix. Means and standard deviations across five runs alone do not determine a paired test: the pairing, unit of analysis, test statistic, and treatment of multiple comparisons also matter. No additional p-value or significance conclusion is inferred from those aggregates.

P Extended Case Studies and Retrieval Gallery

The main text uses "Ox" to illustrate candidate retrieval and verification. This section presents additional retrieval examples across different glyph structures.

P.1 Full Retrieval Gallery

Figure 16 displays Top-5 Bronze candidates for 40 selected OBI queries.

Visual and semantic distractors. Selected queries have candidates sharing visual components or semantic associations. These examples illustrate the ambiguity of nearest-neighbor retrieval.

Scope of the visualization. The gallery shows reference glyphs retrieved from a database. It does not demonstrate generation of previously unobserved ancient glyph images.

Failure cases and selection. A selected gallery does not establish that the correct answer is always retrieved within the first five positions. Aggregate recall and rejection statistics must be reported separately.



Figure 16: Top-5 Bronze retrieval candidates for 40 selected OBI queries. The green box indicates the annotated target. These examples illustrate retrieval behavior and are not a substitute for evaluation on the complete test population.

Q Manifold Neighborhood Visualization

We inspect local retrieval neighborhoods around 20 selected anchors and display five neighbors per anchor. Visual or semantic similarity among neighbors can be consistent with the representation objective. It does not independently exclude memorization, particularly when semantic information is included among the inputs.

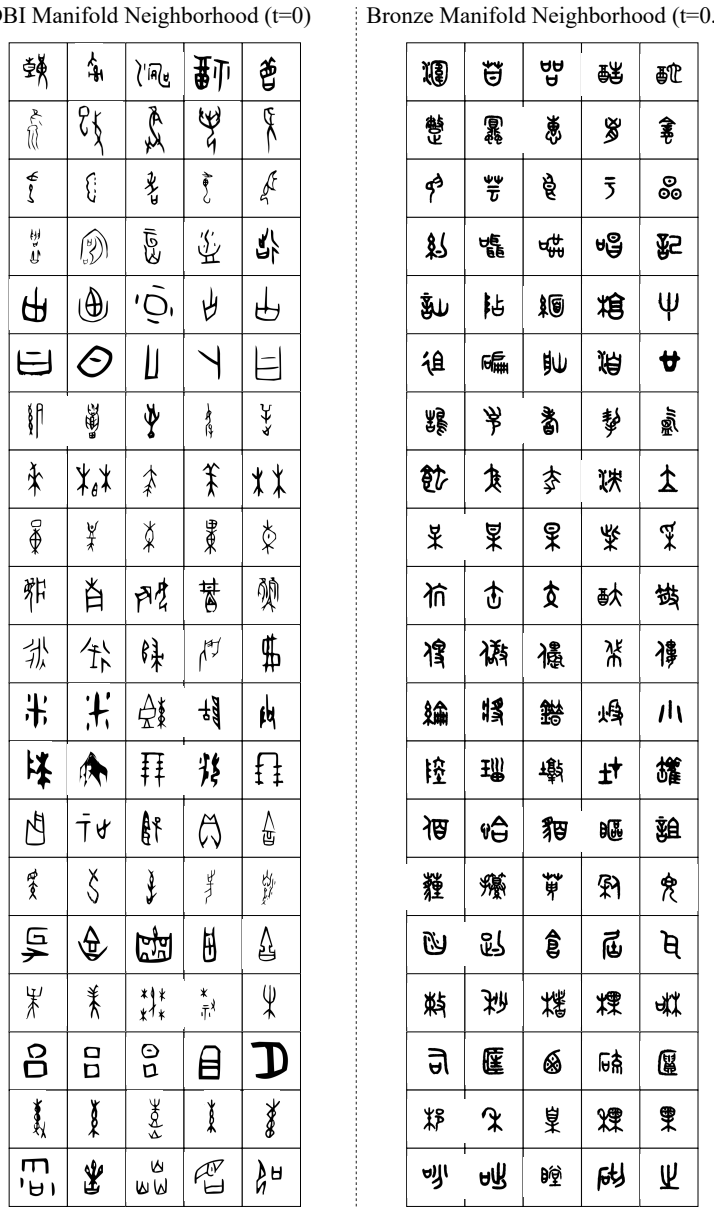


Figure 17: Local neighborhoods in the OBI and Bronze representations. Each row contains an anchor and its retrieved neighbors under the declared distance. The visualization characterizes selected neighborhoods rather than proving preservation of global topology.

R Fine-Grained Dataset Visualization

Figure 18 illustrates the available metadata and glyph observations for 20 selected characters.

OBI labels. The visualization uses seven named scribal-group labels: Dui, Bin, Li, Chu, He, Huang, and Zi. The additional label *None* denotes unclassified observations. It is not an eighth historical period.



Figure 18: Selected glyph observations with fine-grained metadata. OBI variants are annotated by scribal-group labels where available. Bronze variants use the displayed subperiod abbreviations L.S, E.Z, M.Z, L.Z, Sp, and Wa. The symbol "/" denotes an unavailable observation in the displayed collection; it does not establish historical extinction.

Bronze labels. The displayed categories are Late Shang, Early Western Zhou, Mid Western Zhou, Late Western Zhou, Spring and Autumn, and Warring States. Display categories and the actual model-time assignments are documented separately.

S Computational Resources and Reproducibility Report

This section reports computational resources using allocated GPU-hours and distinguishes training, preprocessing, and inference.

S.1 Main Training Runs

The main text reports approximately 18 hours on two NVIDIA H100 GPUs per run. Under that configuration, the allocated compute is approximately

$$2 \times 18 = 36 \quad \text{GPU-hours per run.}$$

Five such independent runs account for approximately

$$5 \times 36 = 180 \quad \text{GPU-hours.}$$

S.2 Other Experiments

Ablations requiring retraining are counted separately from inference-only ablations. Reusing one checkpoint across multiple analyses does not count as multiple training runs. Baseline reproduction, feature extraction, hyperparameter search, and exploratory or failed jobs are separate budget categories.

For jobs $r \in \mathcal{R}$,

$$C_{\text{total}} = \sum_{r \in \mathcal{R}} g_r h_r,$$

where g_r is the allocated GPU count and h_r is the elapsed time in hours. A job is counted only once.

S.3 Inference Accounting

Inference measurements distinguish query feature extraction, reference encoding, index lookup, forward integration, and backward verification. Precomputed reference features should not be presented as zero-cost computation.

T 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 36 summarizes the main assets, their roles in this work, and our release treatment.

Table 36: External and newly introduced assets used in this work. Source terms and release treatment are separate from ownership.

Field	Description
LBQJQW / jgwlbq	
Use	Source for OBI glyph images and scribal-group information used in constructing FGCCES.
Source	`https://www.jgwlbq.org.cn/`
Terms	Publicly accessible academic/cultural-resource website. We found no standardized machine-readable open license in the cited source.
Release treatment	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	
Use	Source for ancient Chinese character morphology, especially Bronze-script forms and period/provenance information.
Source	`http://www.ccamc.co`

Table 36 continued

Field	Description
Terms	Publicly accessible academic database. We found no standardized machine-readable open license in the cited source.
Release treatment	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	
Use	Source for Seal, Clerical, and Regular-script forms and verified cross-era correspondences.
Source	`https://www.bnu.edu.cn/`
Terms	Academic database maintained by Beijing Normal University. We found no standardized machine-readable open license in the cited source.
Release treatment	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	
Use	Public benchmarks used for OBI decipherment evaluation following prior protocols.
Source	(Wang & Deng, 2024; Guan et al., 2024)
Terms	Used according to the academic research terms described by the original benchmark papers or repositories.
Release treatment	We cite the benchmark creators and report results under their evaluation protocols. We do not relicense the original benchmarks.
PictOBI-20k	
Use	Visual multiple-choice benchmark used to evaluate paleographic reasoning.
Source	(Chen et al., 2025)
Terms	Used according to the academic research terms described by the original paper or repository.
Release treatment	We cite the original benchmark and use it only for evaluation. Any benchmark redistribution follows the original release terms.
Baseline methods and code	
Use	Baseline implementations or paper-based reimplementations for Pix2Pix, CycleGAN, DRIT++, Palette, BBDM, CDE, Sundial-GAN, OBSD, Diff-Oracle, OracleFusion, CrossFont, OracleSage, and OracleAgent.
Source	Cited in Section 3 and the references.
Terms	Public code is used under its original license where available. For baselines without public code, we reimplement from the paper descriptions.
Release treatment	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	
Use	New cross-era Chinese character evolution sequence dataset introduced in this paper.
Source	This work.
Terms	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.
Release treatment	The manuscript describes a package containing a dataset card, preprocessing scripts, splits, feature-generation code, and source instructions. This is a stated release scope, not a verified inventory.

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 release description alone does not establish that every listed file is present or accessible. No repository-content or permissions audit is reported in this appendix; public accessibility of a source is not itself a redistribution license.

U 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. 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 discusses error analysis and the limitations of interpreting compatibility scores as calibrated confidence. 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 T. 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.

V Limitations

Specification and evidentiary limits. The stated alignment objectives admit collapse, and the implemented anti-collapse mechanism and survival gradient route are not established by the current specification. Fine-grained time assignments and several architecture and pruning settings remain unspecified. Unvalidated supplementary score summaries do not resolve these implementation omissions or establish leakage-free, information- matched comparisons. The PictOBI and OCR adapters, PictOBI aggregation, independent expert-validation evidence, and final dataset manifest remain unverified; they are not repaired by the formal ODE results. 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.