Enabling Stroke-Level Structural Analysis of Hieroglyphic Scripts without Language-Specific Priors

Anonymous ACL submission

Abstract

Hieroglyphs, as logographic writing systems, encode rich semantic and cultural information within their internal structural composition. Yet, current advanced Large Language Models (LLMs) and Multimodal LLMs (MLLMs) usually remain structurally blind to this information. LLMs process characters as textual tokens, while MLLMs additionally view them as raw pixel grids. Both fall short to model the underlying logic of character strokes. Furthermore, existing structural analysis methods are often script-specific and labor-intensive. In this paper, we propose **Hiero**glyphic **S**troke **A**nalyzer (HieroSA), a novel and generalizable framework that enables MLLMs to automatically derive stroke-level structures from character bitmaps without handcrafted data. It transforms modern logographic and ancient hieroglyphs character images into explicit, interpretable line-segment representations in a normalized coordinate space, allowing for cross-lingual generalization. Extensive experiments demonstrate that HieroSA effectively captures character-internal structures and semantics, bypassing the need for language-specific priors. Experimental results highlight the potential of our work as a graphematics analysis tool for a deeper understanding of hieroglyphic scripts.

1 Introduction

Hieroglyphs are among the earliest known writing systems in human history, originating as visually motivated symbols that depict objects, actions, or abstract concepts through recognizable forms (Boltz, 1986; Davies, 1990; Henderson, 1997). As a foundational medium of written communication, hieroglyphic scripts preserve rich information about early human cognition, cultural practices, and systems of knowledge representation (Allen, 2000; Woods and Woods, 2011; Zhang and Mao, 2017; Wong, 2018; Qi et al., 2025). Moreover, some modern writing systems continue to exhibit hieroglyphic properties, such as Chinese characters (Zhao, 2005; Liu et al., 2019) and Japanese Kanji (Heisig, 2008), whose logographic characters possess intrinsic linguistic and cultural significance (Joyce, 2011; Fan, 2014).

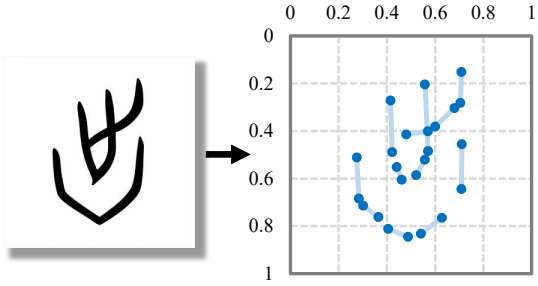

Figure 1: Given a bitmap image of a hieroglyphic character (for example, the Oracle Bone Script character on the left), our proposed HieroSA infers stroke-level structure and converts the image into explicit line-segment representations in a normalized coordinate space (right).

Unlike phonetic scripts, hieroglyphs integrate imagery and language, in which logographic forms function as active carriers of meaning rather than neutral phonetic placeholders. In such writing systems, semantic information is encoded in the internal structure of characters, particularly in the configuration and combination of strokes. Consequently, systematic variations in these patterns reflect semantic relationships between characters. The structural organization of characters therefore plays a central role in understanding hieroglyphic writing systems, motivating computational approaches that aim to model character-internal structure (Muñoz Sánchez, 2024).

However, capturing stroke-level structure remains challenging, despite the rapid development of large language models (LLMs) and multimodal large language models (MLLMs). LLMs usually encode text into unified symbolic representations (Bouma, 2009; Sennrich et al., 2016; Kudo and Richardson, 2018; Land and Bartolo, 2024), where characters or tokens are mapped to discrete indices and processed as abstract sequences, dis-

carding character-internal structural information. As a complement to textual tokens, MLLMs additionally operate on visual inputs and process characters as images composed of pixels (Dosovitskiy, 2020; Fei et al., 2024; Tschannen et al., 2025; Wu et al., 2025). While such pixel-level representations allow models to incorporate visual appearance, they do not effectively model the structural relationships among strokes or the underlying compositional principles of character formation, limiting the ability of current models to achieve generalizable understanding of hieroglyphic characters.

Other existing methods for structural analysis are largely confined to specific writing systems and often rely on language-dependent assumptions (Cao et al., 2018; Xiong et al., 2021; Chen et al., 2025) Moreover, existing approaches typically rely on substantial amounts of labor-intensive annotation or external linguistic knowledge, such as predefined stroke inventories or decomposition rules (Assael et al., 2022; Sommerschield et al., 2023). As a result, their applicability is limited to well-studied scripts, and extending them to other hieroglyphic or pictographic writing systems requires significant additional effort, which constrains their scalability and generalization. This motivates the need for a more general approach that can infer character structure without depending on script-specific stroke definitions or annotations.

In this work, we propose **Hiero**glyphic **S**troke **A**nalyzer (**HieroSA**), a novel and generalizable approach for automatically deriving stroke-level structure from bitmap images of hieroglyphic characters without annotated training data. Specifically, our method enables MLLMs to transform character images into explicit line-segment representations, yielding a compact and interpretable representation of stroke structure, as shown in Figure 1. It does not rely on manual annotations or prior knowledge of language-specific stroke systems, allowing the method to generalize naturally across diverse hieroglyphic and logographic writing systems. Extensive experiments demonstrate that HieroSA effectively models stroke representations, and presents the potential as a scalable tool for computational graphematics analysis and further exploration of hieroglyphic scripts. Our contributions are threefold:

- We introduce a novel training framework that enables MLLMs to effectively capture and model stroke-level structures in hieroglyphic characters, providing explicit access to character-internal structural information.
- Extensive experimental results verify that our proposed **HieroSA** consistently generalizes across multiple pictographic writing systems, without relying on external linguistic expertise or annotations.
- We show that incorporating stroke representations generated by HieroSA benefits downstream task of glyph recognition, and demonstrates potential as a scalable tool for broader graphematics analysis and further exploration of hieroglyphic scripts.

2 Method

We design an approach based on reinforcement learning (RL) that enables a flexible and expressive representation of strokes as line segments. The supervision signals are derived solely from bitmap images, which frees the training process from reliance on predefined stroke definitions, external linguistic expertise, or additional annotations. We introduce the stroke representation, the corresponding reward design, and the overall training paradigm in Sections 2.1, 2.2, and 2.3, respectively.

2.1 Stroke Representation

As illustrated in Figure 2(a), our method maps a binarized character image, with black strokes treated as the foreground against a white background, to an geometric representation in a normalized coordinate space. A character is modeled as a collection of line segments, each specified by the coordinates of its two endpoints. Formally, the stroke structure is represented as

$$\mathcal{S} = \left\{ \left(\mathbf{p}_s^k, \mathbf{p}_e^k \right) \right\}_{k=1}^{n}, \ \mathbf{p}_s^k, \mathbf{p}_e^k \in \mathbb{R}^2, \quad (1)$$

where n denotes the number of strokes.

This coordinate-based representation allows the model to learn stroke structures directly from bitmap images, without relying on language-specific or rule-based stroke definitions. In particular, curved or complex stroke shapes are not explicitly parameterized; instead, the model learns to approximate such structures through flexible decompositions into multiple line segments that best explain the observed visual evidence.

We note that the automatically inferred stroke decompositions may not always align with human-intuitive or conventionally defined stroke segmentations. However, this flexibility enables the model to learn more adaptive and generalizable representations, particularly for scripts where standard stroke

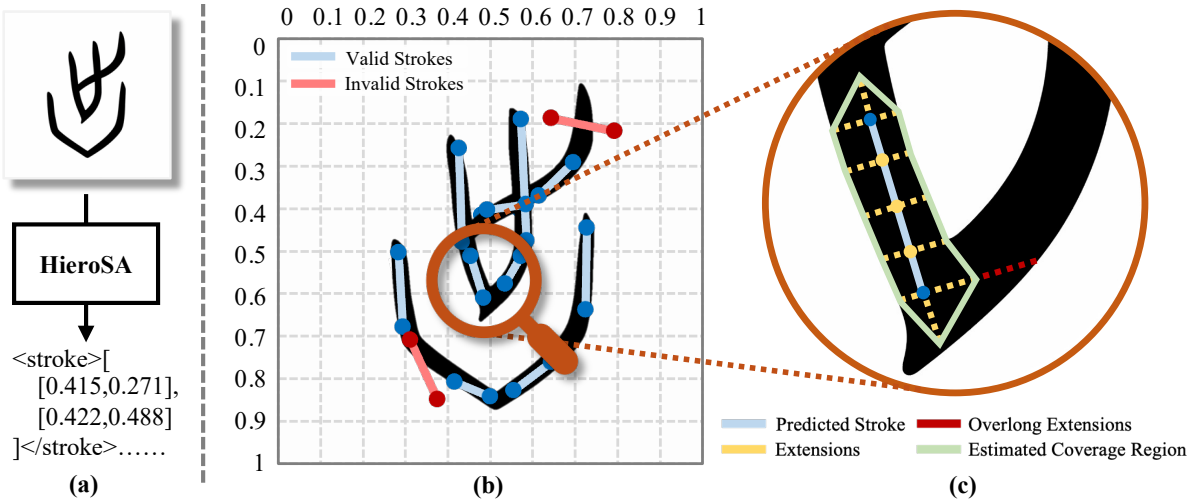

Figure 2: Overview of **HieroSA**. (a) HieroSA takes a binarized character image as input and outputs a structured stroke representation, where strokes are represented as a set of line segments in a normalized coordinate space. (b) Illustration of the training objective: the predicted stroke segments are optimized to maximize their overlap with the black pixels in the binarized character image. (c) Geometric estimation of black-pixel coverage for a single stroke.

definitions are ambiguous or absent. For example, for the oracle bone script character shown in Figure 1, to the best of our knowledge, no standard stroke decomposition is available; nevertheless, our model produces a set of line segments that effectively capture its structural characteristics.

2.2 Reward Design for Stroke Representation

We use bitmap images as the sole source of supervision. Our training objectives involve maximizing the overlap between generated strokes and black pixels in the bitmap image, increasing coverage over black-pixel regions, and decreasing the number of strokes outside black pixels. Our reward computation consists of three steps: identifying valid strokes, estimating coverage for each stroke, and aggregating the reward.

2.2.1 Valid Stroke Identification

Our first step is to identify valid strokes by examining whether each stroke lies within the black-pixel regions. For each predicted stroke defined by its endpoints $(\mathbf{p}_s, \mathbf{p}_e)$, we uniformly sample points along the line segment:

$$\mathbf{p}_i = \frac{i\,\mathbf{p}_s + (m+1-i)\,\mathbf{p}_e}{m+1}, \quad i = 1, \cdots, m, \tag{2}$$

where m is the minimum value such that the distance between neighboring sampled points is smaller than a threshold D. We then check whether all sampled points $\{\mathbf{p}_i\}_{i=1}^{m}$ and endpoints $\mathbf{p}_s$ and $\mathbf{p}_e$ lie within the black-pixel region of the binarized image. If any point falls outside the black-pixel area, the corresponding stroke is marked as invalid. For example, as illustrated in Figure 2(b), the two red strokes are identified as invalid under this criterion. By filtering out such strokes, the model is encouraged to generate stroke segments that closely follow the foreground structure of the character.

2.2.2 Single-Stroke Coverage Estimation

Given a stroke represented by a line segment, we estimate its coverage region by extending the segment along both tangential and normal directions within the black-pixel region, as shown in Figure 2 (c). To avoid over-extension that may intrude into neighboring strokes, excessively long extensions are truncated, yielding a polygonal region that approximates the spatial coverage of the stroke.

Specifically, given the endpoints $(\mathbf{p}_s, \mathbf{p}_e)$ of a stroke and the sampled points $\{\mathbf{p}_i\}_{i=1}^{m}$ obtained in the previous step, we define

$$\mathbf{t} = \frac{\mathbf{p}_e - \mathbf{p}_s}{\|\mathbf{p}_e - \mathbf{p}_s\|} \tag{3}$$

to denote the unit tangent direction of the stroke, and let $\mathbf{n}$ be the corresponding counterclockwise unit normal direction, defined as

$$\mathbf{n} = (-\mathbf{t}_y, \mathbf{t}_x), \tag{4}$$

where $(\mathbf{t}_x, \mathbf{t}_y)$ are the coordinates of $\mathbf{t}$. For each point $\mathbf{p}_j \in \{\mathbf{p}_s, \mathbf{p}_1, \cdots, \mathbf{p}_m, \mathbf{p}_e\}$, we extend rays along both normal directions until reaching the black-white boundary:

$$\begin{cases} d_j^+ = \max\{d > 0 \mid [\mathbf{p}_j, \mathbf{p}_j + d\,\mathbf{n}] \subset \Omega_B\}, \\ d_j^- = \max\{d > 0 \mid [\mathbf{p}_j, \mathbf{p}_j - d\,\mathbf{n}] \subset \Omega_B\}, \end{cases} \tag{5}$$

where Ω_B denotes the set of black pixels in the binarized image.

We first compute the mean normal extension length as follows,

$$\bar{d} = \frac{1}{2(m+2)} \sum_j \left(d_j^+ + d_j^- \right). \tag{6}$$

Then, excessively large extension lengths are regarded as abnormal:

$$\mathcal{A} = \left\{ j \;\middle|\; \max(d_j^+, d_j^-) > \lambda \bar{d} \right\}, \tag{7}$$

where λ is the threshold for identifying abnormal extensions. Using the remaining samples, we compute a refined mean

$$\tilde{d} = \frac{1}{2|\mathcal{I}|} \sum_{j \in \mathcal{I}} \left(d_j^+ + d_j^- \right), \tag{8}$$

where $\mathcal{I} = \{s, 1, \ldots, m, e\} \setminus \mathcal{A}$. Extension lengths are then truncated as

$$\hat{d}_j^{\pm} = \min \left(d_j^{\pm}, \lambda \tilde{d} \right), \tag{9}$$

For the endpoints $\mathbf{p}_s$ and $\mathbf{p}_e$, we further extend along the tangent direction. The tangential extension length is defined as

$$\begin{cases} \ell_s = \dfrac{\hat{d}_s^+ + \hat{d}_s^-}{2}, \\ \ell_e = \dfrac{\hat{d}_e^+ + \hat{d}_e^-}{2}. \end{cases} \tag{10}$$

The endpoints are thus extended to

$$\begin{cases} \mathbf{p}'_s = \mathbf{p}_s - \ell_s \, \mathbf{t}, \\ \mathbf{p}'_e = \mathbf{p}_e + \ell_e \, \mathbf{t}. \end{cases} \tag{11}$$

Finally, the set of offset vertices $\mathcal{Q}$ computes as

$$\mathcal{Q} = \{ q_j^{\pm} \mid q_j^{\pm} = p_j \pm d_j^{\pm} \mathbf{n}, \; j = 1, \cdots, m \}, \tag{12}$$

together with the tangentially extended endpoints $\mathbf{p}'_s$ and $\mathbf{p}'_e$ define a polygonal coverage region $\mathcal{C}$ for the stroke. Examples of the polygonal coverage regions estimated by our method are provided in Appendix A. The region is subsequently used for reward computation in the Section 2.2.3.

2.2.3 Reward Aggregation

Given the coverage regions $\{\mathcal{C}_k\}_{k=1}^n$ estimated for individual strokes, we aggregate them sequentially to compute the final reward, while penalizing invalid strokes. Let

$$\mathcal{U}_k = \bigcup_{i \in \mathcal{V}_k} \mathcal{C}_i \tag{13}$$

where $\mathcal{V}_k$ denotes the set of previously accepted valid strokes. For the k-th stroke, we examine whether its coverage region $\mathcal{C}_k$ provides sufficient novel contribution. Specifically, the stroke is marked as invalid if

$$\frac{|(\mathcal{C}_k \setminus \mathcal{U}_k) \cap \Omega_B|}{|\Omega_B|} < \tau, \tag{14}$$

where τ is the overlap ratio threshold. This encourages the model to cover new foreground regions with each stroke while penalizing redundant strokes, thereby preventing solutions that repeatedly exploit the same local areas.

All invalid strokes, including those identified in Section 2.2.1 as well as those identified during aggregation, are discarded. Let $\mathcal{V}$ denote the index set of remaining valid strokes, and let

$$\mathcal{C}_{\text{final}} = \bigcup_{k \in \mathcal{V}} \mathcal{C}_k \tag{15}$$

be the final aggregated coverage region. The overall reward is then defined as

$$r_s = \frac{|\mathcal{C}_{\text{final}} \cap \Omega_B|}{|\Omega_B|} \cdot (1 - \alpha \, N_{\text{invalid}}), \tag{16}$$

where N_{invalid} denotes the number of invalid strokes and α is a penalty coefficient. This formulation provides a compact training signal that balances foreground coverage promotion with penalties on invalid strokes.

2.3 Training Paradigm

We adopt Group Relative Policy Optimization (GRPO) as our training algorithm. Prior work has widely applied GRPO in LLM training by combining a rule-based answer reward with a format reward, which yields strong and reliable model performance. Following this established setting, we integrate our stroke-representation reward r_s with a format reward r_f, encouraging the model to generate outputs that conform to the structured format as exemplified in Figure 2(a), thereby facilitating reliable parsing. The final reward is defined as

$$r = r_s + \beta \, r_f, \tag{17}$$

where β is a hyperparameter that balances the contribution of the format reward.

3 Experiments

3.1 Implementations

The training data is curated from two sources: bitmap images generated from SVG glyphs in font files, and publicly available datasets of hieroglyphic character images, without requiring any additional annotations. Our training dataset covers three writing systems: contemporary Chinese characters (hereafter referred to as Chinese), Japanese Kanji, and Oracle Bone Scripts (OBS), an ancient Chinese hieroglyphic writing system.

Model	Chinese (ZH)			Japanese (JA)			Oracle Bone Script (OBS)			AVG		
	RE↑	CO (%)↑	IS (%)↓	RE↑	CO (%)↑	IS (%)↓	RE↑	CO (%)↑	IS (%)↓	RE↑	CO (%)↑	IS (%)↓
GPT-5	0.133	3.6	88.2	0.129	1.5	92.8	0.139	3.8	92.0	0.134	3.0	91.0
Claude Sonnet 4	0.137	11.9	86.9	0.131	10.1	89.9	0.129	6.3	93.4	0.132	9.4	90.1
Qwen3-VL-4B	0.032	0.5	97.9	0.028	0.5	98.0	0.063	0.7	98.5	0.041	0.6	98.1
HieroSA (ZH)	**0.837**	**78.5**	**6.1**	**0.591**	**60.9**	19.7	**0.522**	50.5	**17.6**	**0.650**	**63.3**	**14.5**
HieroSA (JA)	0.756	72.2	10.2	0.584	59.4	**19.1**	0.455	45.0	23.9	0.598	58.9	17.7
HieroSA (OBS)	0.446	64.6	23.1	0.295	52.2	33.9	0.344	**53.3**	25.2	0.362	56.7	27.4

Table 1: Results on the stroke structure parsing task. RE reports the reward computed on the test set using the same reward function as in training. CO measures the percentage of the black pixel region Ω_B that is covered. IS denotes the percentage of invalid strokes relative to the total number of strokes. The evaluated model versions are `gpt-5-2025-08-07` (GPT-5), `claude-sonnet-4-20250514` (Claude Sonnet 4) and `Qwen3-VL-4B-Instruct` (Qwen3-VL-4B). ZH and JA denote Chinese and Japanese.

Model	Chinese (ZH)	Japanese (JA)	OBS
Input Bitmap Image			
HieroSA (ZH)			
HieroSA (JA)			
HieroSA (OBS)			

Table 2: Qualitative visualization of stroke-structure parsing results across scripts and training settings. Rows correspond to models trained on different scripts, and columns correspond to evaluation scripts.

We adopt Qwen3-VL-4B-Instruct as the base model and train a separate model for each script for two epochs. During training, each image is augmented with an overlaid coordinate system to enhance coordinate prediction accuracy and spatial localization capability. Details of data curation are provided in Appendix B.1, with training configurations described in Appendix C. The effectiveness of the overlaid coordinate system is further validated through ablation studies in Appendix E.

3.2 Main Results

We evaluate the stroke-structure parsing performance of our model on Chinese, Japanese, and OBS characters. Given an input character image, the model predicts a set of stroke segments that represent the underlying stroke structure of the character. To assess the stroke parsing quality, we measure the spatial overlap between the polygonal coverage induced by the predicted strokes and the foreground region of the original character image. Results are reported in Table 1. Details of the test set are provided in Appendix B.2.

Training Set	CC	FB	FA	BAR
Chinese	7.466	8.670	0.178	49.827
Japanese	4.389	7.371	0.138	55.566
OBS	2.680	6.154	0.098	63.922

Table 3: Structural statistics of binarized images across training sets in different languages, including the average number of connected components (CC), foreground boundary length of black regions (FB), foreground area (FA), and boundary-to-area ratio (BAR).

Quantitative results. Models trained with HieroSA consistently outperform baselines by a large margin across all scripts. They achieve higher reward (RE), indicating a better alignment with the training objective, while simultaneously yielding higher coverage (CO) and substantially lower invalid-stroke rates (IS). These improvements are evident not only in intra-script experiments, but also in cross-script evaluation settings.

Qualitative results. Qualitative results further illustrate the effectiveness of our approach. As shown in Table 2, the predicted stroke structures align with the core stroke composition of the input characters across scripts. The models capture the main structural skeletons while preserving essential stroke connectivity, yielding visually coherent stroke-structure decompositions. These observations are consistent with the quantitative results.

Performance differences across training languages. We observe noticeable performance differences among models trained on different writing systems. To better understand this behavior, we analyze structural statistics of the corresponding training datasets (Table 3), including the number of eight-connected components, foreground boundary length and area, and the boundary-to-area ratio. From Chinese to Japanese to OBS, the number of connected components, boundary length, and area decrease monotonically, indicating progressively simpler glyph structures. In contrast, the boundary-to-area ratio increases, suggesting more curved and

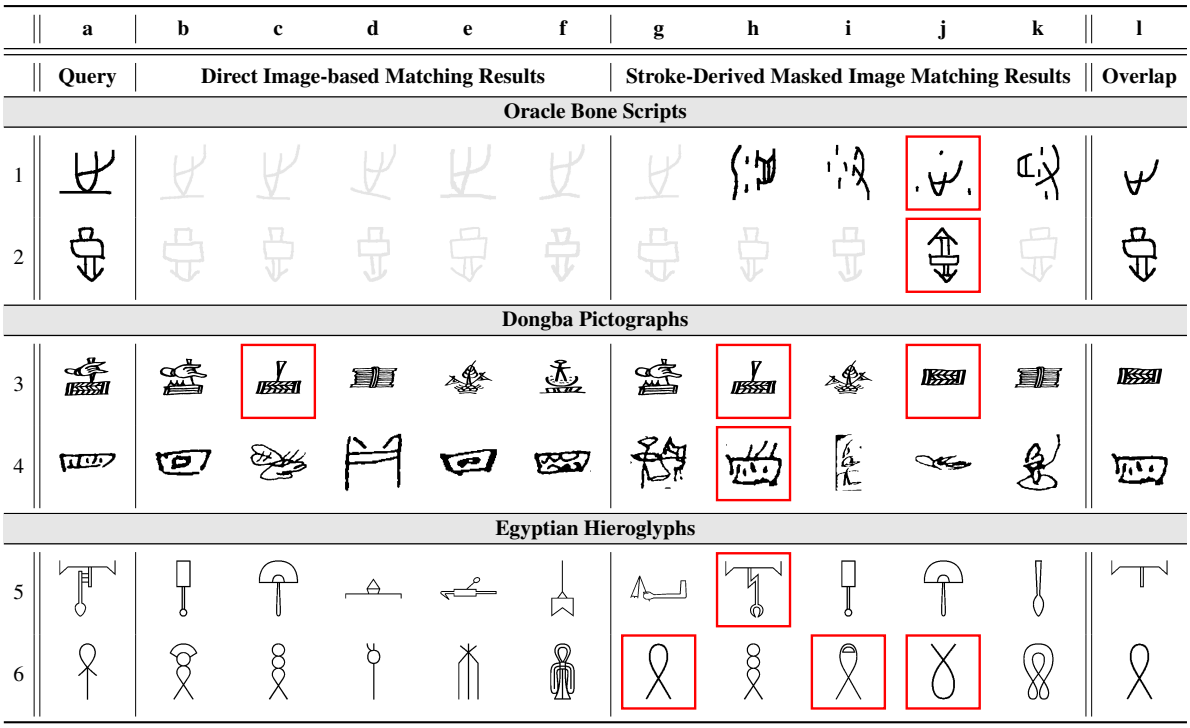

Table 4: Illustrative cases of structure-guided character exploration. We aim to find characters that **differ in identity but share structural similarity with queries**. For each query, we show top-5 results with and without stroke-guided masking; red boxes highlight cases **where the characters are structurally similar to the queries while clearly differing in identity**; gray characters are the identical ones that discourage exploration. The "Overlap" column (***for illustration only***) visualizes shared structures between queries and selected results. See more cases in Appendix H.

less smooth stroke boundaries.

These suggest a plausible reason for the observed performance gap: overly simple structures reduce structural diversity and weaken stroke-structure priors, whereas excessive boundary curvature introduces geometric noise during training. The Chinese characters, with more complex yet cleaner stroke structures, enables models to learn more robust and transferable stroke-structure representations, leading to superior cross-script generalization.

3.3 Results on Downstream Tasks

We further demonstrate the practical utility of HieroSA on downstream tasks, structure-guided hieroglyphic character exploration and optical character recognition, showing that the learned underlying logic of strokes can be effectively leveraged in practical character-related tasks.

3.3.1 Structure-guided Character Exploration

We demonstrate that the stroke-structure representations produced by HieroSA enable the exploration of relationships among hieroglyphic characters by facilitating the matching of structurally similar characters with different identities.

We conduct structure-guided exploration on three writing systems that are generally unfamiliar to contemporary readers: Oracle Bone Script (OBS), Dongba pictographs, an indigenous writing system traditionally used by Naxi priests in southern China, and Egyptian hieroglyphs, an ancient Egyptian writing system (see Appendix F for details). We compare a structure-guided matching strategy using masks derived from predicted strokes with a direct image-based matching baseline. Further experimental details are provided in Appendix G.2. Representative illustrations are presented in Table 4.

Overall, we observe that structure-guided exploration yields a richer set of structurally related characters compared to image-based matching baseline. In addition, many of the identified characters exhibit not only stroke-level structural similarity but also semantic resemblance to the query character, enabling more informative character-level analysis.

OBS character analysis.[1] The OBS characters at positions (1a) and (1j) share a common structural component in the central part of the glyph, which is illustrated in (1l). It depicts human feet, which reflects a pictographic origin related to foot move-

[1]For explanations of OBS characters, we primarily refer to *Yinxu Jiagu Wen Shiyong Zidian* (A Practical Dictionary of Yinxu Oracle Bone Script) by Rusen Ma (2008).

ment or stepping actions. The character at position (1a) conveys the meaning of "arrival" or "reaching a destination", where the foot component directly encodes the act of movement. In contrast, the character at position (1j) denotes a small landform within water (the three surrounding dots), where the same foot-related component can be interpreted as indicating a place of standing or stable footing within the surrounding environment.

Dongba pictographs analysis.[2] A similar pattern is observed for Dongba pictographs. The characters at positions (3a), (3c), (3h), and (3j) exhibit a shared structural pattern centered on the depiction of a book-like object. Characters at positions (3a), (3c), and (3h) represent variant forms of a pictograph denoting the act of writing records, whereas the character at position (3j) denotes the book itself as an object. Despite these semantic distinctions, the shared pictographic structure reveals a common conceptual grounding in writing practices.

Egyptian hieroglyphs analysis.[3] Consistent evidence is also found in Egyptian hieroglyphs. Stroke-guided masking identifies alternative character variants, such as the pair at positions (5a) and (5h), both of which convey meanings related to "night". In addition, the method links multiple variants of the same phonemogram, as illustrated by the correspondence between characters at positions (6a), (6g), and (6i), as well as a closely related phonemogram at position (6j).

Finally, even for characters with undocumented meanings, such as positions (2a) and (4a), stroke-guided masking remains effective in identifying structurally related forms. Our proposed method surfaces plausible structurally related characters. Notably, character (2j) corresponds to a deciphered OBS character meaning "the outer city wall".

Taken together, these qualitative results suggest that *structure-guided exploration provides a principled way to uncover latent structural and semantic relationships across diverse writing systems, extending to those which are under-documented.*

3.3.2 Optical Character Recognition

Optical Character Recognition (OCR) is a common and conventional task that aims to recognize characters from visual inputs. For this task, we examine whether the stroke-structure representations parsed by HieroSA can improve OCR performance.

We fine-tune Qwen3-VL-4B for this task under different settings, spanning different character fonts of both Chinese and Japanese OCR. For unbiased comparison, all characters and fonts used in this experiments are disjoint from those used to train HieroSA. Due to page limitation, we report detailed settings and results in Appendix I. Quantitative results show that fine-tuning with stroke-structure representations substantially improves OCR performance over the zero-shot setting, indicating that the parsed stroke information provides complementary structural cues beyond raw visual features. These findings suggest the potential of stroke-structure representations as an auxiliary signal in OCR.

[2]For explanations of Dongba pictographs, we primarily refer to *Naxi Xiangxing Wenzi Pu* (A Corpus of Naxi Pictographic Script) by Guoyu Fang (2017).

[3]For explanations of Egyptian hieroglyphs, we primarily refer to the Egyptian Hieroglyphs block (U+13000–U+1342F) in the Unicode Explorer, based on Gardiner's sign list; see https://unicode-explorer.com/b/13000.

α	CO (%)↑	IS (%)↓	CS(%)↑	TS
0.00	41.7	64.6	1.8	26.6
0.01	21.7	46.0	4.8	4.8
0.10	22.0	45.2	4.8	4.8
0.50	2.7	75.7	1.4	2.0

Table 5: Results under different values of the penalty coefficient α in the reward function. Experiments are conducted on the Chinese training set for 100 training steps. CO measures black pixel region coverage, CS indicates the average coverage per stroke, IS denotes the percentage of invalid strokes, and TS indicates the total number of generated strokes. The definitions of CO and IS are consistent with those used in Table 1.

4 Analyses

4.1 Effect of Penalty for Invalid Strokes

We analyze the effect of the penalty coefficient α in the reward aggregation described in Section 2.2.3, with results summarized in Table 5.

When $\alpha = 0$, no penalty is applied to invalid strokes, resulting in excessive stroke generation, low average coverage per stroke, and a high proportion of invalid strokes. Although overall coverage remains relatively high, the generated characters lack coherent and structurally valid shapes, indicating the necessity of the penalty term.

As α increases, both the total number of strokes and the proportion of invalid strokes decrease, demonstrating that the penalty effectively suppresses invalid stroke generation. However, overly large values of α (e.g., $\alpha = 0.5$) impose excessive constraints, leading to reduced coverage and degraded performance.

A moderate penalty coefficient provides a good

Number of Points	RE↑	CO (%)↑	IS (%)↓
2	0.311	22.0	45.2
3	0.237	13.8	51.2
4	0.230	13.6	52.1

Table 6: Effect of the number of points used to represent a single stroke. Experiments are conducted on the Chinese training set for 100 training steps. Metric definitions follow those used in Table 1.

balance between suppressing invalid strokes and preserving sufficient coverage. In our experiments, $\alpha = 0.1$ yields reasonable stroke counts, low invalid-stroke ratios, and sufficient coverage, and is therefore adopted.

4.2 Effect of Number of Stroke Points

We study the effect of stroke representation granularity on model performance. In the main experiments, each stroke is represented by two points corresponding to its start and end locations. We vary the number of points per stroke to evaluate whether a finer representation is beneficial.

Results in Table 6 show that using two points consistently achieves the best performance in terms of reward and coverage, while yielding fewer invalid strokes. Increasing the number of points degrades performance across all metrics.

Qualitative inspection indicates that *representing a stroke with more points does not improve geometric modeling*, but instead increases the difficulty of learning stable and valid stroke structures. Therefore, we adopt the two-point representation and model complex characters through the composition of multiple strokes rather than increasing the complexity of individual strokes.

5 Related Work

5.1 Computational Analysis of Hieroglyphs

Hieroglyphs, such as OBS and Dongba, encode meaning primarily through structural composition, distinguishing them from purely phonetic characters (see Appendix F for details). Early computational approaches often treated them as discrete text via standardized codes (Morioka, 2015; Lu, 2019), while more recent work shifts towards visual-centric methods that often formulate the problem as image classification to address pixel-level patterns via Convolutional Neural Networks (CNNs) (Liu et al., 2020, 2024; Zhou et al., 2025).

The emphasis on structures naturally aligns hieroglyph analysis with sketch learning, which models abstract line drawings and skeleton patterns (Ha and Eck, 2017; Singer et al., 2020; Aksan et al., 2020). In both domains, semantic context is primarily conveyed by compositional layouts and positioning of strokes, driving methods to prioritize line-based input over texture or appearance (Yu et al., 2017; Jia et al., 2020). These motivate us to further explore stroke-structure representations for hieroglyph and logograph analysis.

5.2 Stroke-based Language Modeling

Stroke-based modeling captures internal character structures in hieroglyphic and logographic writing systems to provide structural and compositional inductive biases for language modeling (Nguyen et al., 2020). Previous work encodes characters as stroke sequences or sub-character embeddings via neural encoders (Cao et al., 2018; Xiong et al., 2021), demonstrating advantages over purely symbolic representations, and proving effective for downstream tasks such as Named Entity Recognition (NER) (Cao et al., 2018; Yi et al., 2023), character recognition (Chen et al., 2023; Huang et al., 2025), and Neural Machine Translation (NMT) (Wang et al., 2022). Building on these, recent work further integrates stroke-related visual features with MLLMs for ancient hieroglyph recognition (Chen et al., 2025).

However, these methods largely depend on handcrafted stroke annotations and predefined stroke labels (Assael et al., 2022; Sommerschield et al., 2023), limiting scalability and generalization across diverse writing systems. Contrastly, our **HieroSA** is annotation-free, scalable, and generalizable across languages.

6 Conclusion

In this paper, we propose **HieroSA**, a novel and generalizable framework that enables MLLMs to autonomously derive stroke-level structures from character bitmaps without manual annotations. We conduct extensive experiments on modern logographic and ancient hieroglyphic writing systems and demonstrate that HieroSA robustly generalizes across diverse writing systems without linguistic priors. It also achieves consistent advancements in downstream tasks such as OCR and structure-based character exploration. Results and case study highlight the potential of HieroSA as a scalable tool for the computational linguistic analysis and decipherment of hieroglyphic scripts.

Limitations

While our framework demonstrates strong performance across multiple hieroglyphic scripts, several limitations remain. First, the quality of the learned stroke-level representations can be affected by variability in structural complexity and geometric noise in the training data. Developing more effective data curation and structure-aware denoising strategies could further improve robustness. Second, due to computational constraints, our experiments are conducted with moderate-scale models and datasets. Training larger models with more diverse data may further enhance representation quality and generalization. Third, our current framework adopts a single-model setting; incorporating ensemble-based strategies could help stabilize predictions and improve overall performance. Finally, while our method enables structure-guided character exploration, the identification of the most informative exploration results still relies on a limited degree of manual inspection. Future work could investigate automatic filtering or ranking mechanisms to streamline exploration and improve usability.

References

Emre Aksan, Thomas Deselaers, Andrea Tagliasacchi, and Otmar Hilliges. 2020. Cose: Compositional stroke embeddings. *Advances in Neural Information Processing Systems*, 33:10041–10052.

James P Allen. 2000. *Middle Egyptian: An introduction to the language and culture of hieroglyphs*. Cambridge University Press.

T George Allen. 1951. Egyptian grammar, being an introduction to the study of hieroglyphs.

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

Xiaojun Bi, Shuo Li, Junyao Xing, Ziyue Wang, Fuwen Luo, Weizheng Qiao, Lu Han, Ziwei Sun, Peng Li, and Yang Liu. 2025. DongbaMIE: A multimodal information extraction dataset for evaluating semantic understanding of dongba pictograms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

William G Boltz. 1986. Early chinese writing. *World Archaeology*, 17(3):420–436.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. 2018. cw2vec: Learning chinese word embeddings with stroke n-gram information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yingfa Chen, Chenlong Hu, Cong Feng, Chenyang Song, Shi Yu, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. Multi-modal multi-granularity tokenizer for chu bamboo slips. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6201–6211, Abu Dhabi, UAE. Association for Computational Linguistics.

Zongze Chen, Wenxia Yang, and Xin Li. 2023. Stroke-based autoencoders: Self-supervised learners for efficient zero-shot chinese character recognition. *Applied Sciences*, 13(3):1750.

W Vivian Davies. 1990. Egyptian hieroglyphs. *JT Hooker et al., Reading the Past*, pages 75–135.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Jiageng Fan. 2014. *A Study of Characters in Chinese and Japanese, including Semantic Shift*. Ph.D. thesis, University of Canterbury.

Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. *Advances in neural information processing systems*, 37:57207–57239.

Haisu Guan, Huanxin Yang, Xinyu Wang, Shengwei Han, Yongge Liu, Lianwen Jin, Xiang Bai, and Yuliang Liu. 2024. Deciphering oracle bone language with diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15554–15567.

David Ha and Douglas Eck. 2017. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*.

James W Heisig. 2008. *Remembering the Kanji: A systematic guide to reading the japanese characters*, volume 2. University of Hawaii Press.

John S Henderson. 1997. *The world of the ancient Maya*. Cornell University Press.

Yiheng Huang, Shuang She, Zewei Wei, Jianmin Lin, Ming Yang, and Wenyin Liu. 2025. StrokeNet: Unveiling How to Learn Fine-Grained Interactions in Online Handwritten Stroke Classification. In *International Conference on Document Analysis and Recognition*, pages 200–217. Springer.

Qi Jia, Xin Fan, Meiyu Yu, Yuqing Liu, Dingrong Wang, and Longin Jan Latecki. 2020. Coupling deep textural and shape features for sketch recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 421–429.

Terry Joyce. 2011. The significance of the morphographic principle for the classification of writing systems. *Written Language & Literacy*, 14(1):58–81.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Sander Land and Max Bartolo. 2024. Fishing for magikarp: Automatically detecting under-trained tokens in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11631–11646.

Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. 2019. Ancient–modern chinese translation with a new large training dataset. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–13.

Mengting Liu, Guoying Liu, Yongge Liu, and Qingju Jiao. 2020. Oracle bone inscriptions recognition based on deep convolutional neural network. *Journal of image and graphics*, 8(4):114–119.

Xiaojuan Liu, Xu Han, Shanxiong Chen, Weijia Dai, and Qiuyue Ruan. 2024. Ancient yi script handwriting sample repository. *Scientific Data*, 11(1):1183.

Qin Lu. 2019. Computers and chinese writing systems. In *The Routledge handbook of Chinese applied linguistics*, pages 461–482. Routledge.

Tomohiko Morioka. 2015. Multiple-policy character annotation based on chise. *Journal of the Japanese Association for Digital Humanities*, 1(1):86–106.

Ricardo Muñoz Sánchez. 2024. When hieroglyphs meet technology: A linguistic journey through Ancient Egypt using natural language processing. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 156–169, Torino, Italia. ELRA and ICCL.

Minh Nguyen, Gia H. Ngo, and Nancy F. Chen. 2020. Hierarchical character embeddings: Learning phonological and semantic representations in languages of logographic origin using recursive neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:461–473.

Hengnian Qi, Hao Yang, Zhaojiang Wang, Jiabin Ye, Qiuyi Xin, Chu Zhang, and Qing Lang. 2025. AncientGlyphNet: an advanced deep learning framework for detecting ancient Chinese characters in complex scene. *Artificial Intelligence Review*, 58(3):88.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1715–1725.

Johannes Singer, Katja Seeliger, and Martin N Hebart. 2020. The representation of object drawings and sketches in deep convolutional neural networks. In *NeurIPS 2020 Workshop SVRHM*.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando De Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Pengjie Wang, Kaile Zhang, Xinyu Wang, Shengwei Han, Yongge Liu, Jinpeng Wan, Haisu Guan, Zhebin Kuang, Lianwen Jin, Xiang Bai, and 1 others. 2024. An open dataset for oracle bone script recognition and decipherment. *arXiv preprint arXiv:2401.15365*.

Zhijun Wang, Xuebo Liu, and Min Zhang. 2022. Breaking the representation bottleneck of Chinese characters: Neural machine translation with stroke sequence modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6473–6484, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adelina Wei-Kwan Wong. 2018. Ancient chinese hieroglyphs: Archetypes of transformation. *Jung Journal*, 12(3):54–74.

Mary B Woods and Michael Woods. 2011. *Ancient communication technology: From hieroglyphics to scrolls*. Twenty-First Century Books.

Xiaofeng Wu, Karl Stratos, and Wei Xu. 2025. The impact of visual information in chinese characters: Evaluating large models' ability to recognize and utilize radicals. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 331–350.

Zongyang Xiong, Ke Qin, Haobo Yang, and Guangchun Luo. 2021. Learning chinese word representation better by cascade morphological n-gram. *Neural Computing and Applications*, 33(8):3757–3768.

Fuquan Yang. 2012. *Living Hieroglyphics and Dongba Culture*. Yunnan Education Press.

Fen Yi, Hong Liu, You Wang, Sheng Wu, Cheng Sun, Peng Feng, and Jin Zhang. 2023. Medical named entity recognition fusing part-of-speech and stroke features. *Applied Sciences*, 13(15):8913.

Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. 2017. Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision*, 122(3):411–425.

Hong-Mei Zhang and Yuan-Ming Mao. 2017. An overview of ancient chinese character inscriptions of national minorities. *Journal of Chinese Writing Systems*, 1(1):19–28.

Shouhui Zhao. 2005. Chinese character modernisation in the digital era: A historical perspective. *Current Issues in Language Planning*, 6(3):315–378.

Shu Zhou, Xin Wang, Jingwen Qiu, Wenru Bu, and Hao Wang. 2025. Oraclenet: enhancing oracle bone script recognition with adaptive deformation and texture-structure decoupling. *npj Heritage Science*, 13(1):273.

A Examples of Stroke Coverage Estimation

Table 7 presents examples of polygonal coverage regions estimated for individual strokes. Across Chinese, Japanese, and Oracle Bone Script characters, the estimated regions accurately reflect the spatial extent of individual strokes. The results indicate that the proposed method remains effective even in the presence of complex stroke geometries, such as curved strokes and stroke intersections.

B Data Curation

Our datasets consist solely of bitmap images generated from SVG glyphs in font files, as well as publicly available datasets of hieroglyphic character images. No additional annotations are used. In this section, we describe the construction of the training set and the test set.

B.1 Training Set Curation

Our training dataset includes Chinese characters, Japanese Kanji, and OBS. For Chinese, we select 2,000 characters. For each character, bitmap images are rendered from SVG glyphs of six Chinese font families: `SimHei`, `KaiTi`, `SimLi`, `Microsoft YaHei`, `SimYou`, and `SimSun`, resulting in a total of 12,000 images. For Japanese Kanji, we follow the same procedure using six Japanese font families: `M PLUS 1p`, `Zen Maru Gothic`, `Klee One`, `Zen Kurenaido`, `Noto Sans Japanese`, and `Noto Serif Japanese`. For OBS, we collect character images from the undeciphered subset of the HUST-OBC dataset (Wang et al., 2024).

B.2 Test Set Curation

Our test set is curated following the same procedure as the training set, with all font families and character identities disjoint from those used for training. For Chinese, test images are rendered from SVG glyphs of two font families, `JinNianYeYaoJiaYouYa` and `Source Han Sans`, with 500 characters sampled from each font, resulting in a total of 1,000 test images. For Japanese Kanji, we follow the same setting, using `As Winter Comes` and `BIZ UDPGothic`, with 500 characters per font and 1,000 images in total. For OBS, the test set consists of 1,000 images collected from the deciphered subset of the HUST-OBC dataset (Wang et al., 2024).

Examples of Estimated Coverages of Strokes

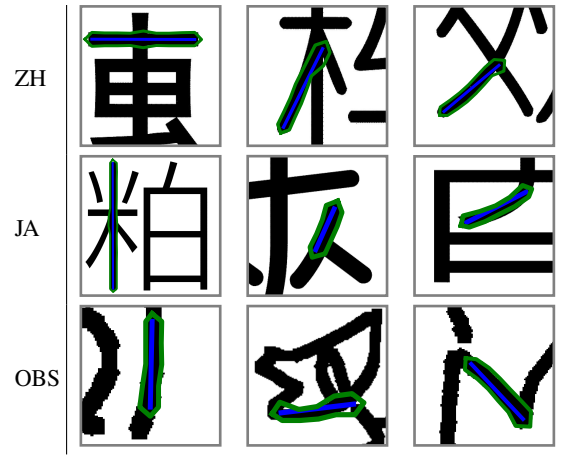

Table 7: Examples of estimated polygonal coverage regions for individual strokes across Chinese (ZH), Japanese (JA), and Oracle Bone Script (OBS) characters. For visualization clarity, the original images are cropped to highlight the target regions. Blue curves indicate the strokes, while green polygons denote the corresponding estimated coverage regions.

β	$r_s \uparrow$	CO (%)↑	IS (%)↓
0.0625	0.172	19.3	34.3
0.1250	0.186	22.0	45.2
0.2500	0.166	19.4	42.6
0.5000	0.158	17.8	47.1
1.0000	0.130	14.8	60.6

Table 8: Effect of the format reward coefficient β. Experiments are conducted on the Chinese training set for 100 training steps. r_s denotes the stroke representation reward defined in Section 2.2.3. Other metrics follow the definitions used in Table 1.

C Training Details

We train models with a batch size of 32, a rollout size of 8, and a learning rate of 1×10^{-6}. The training hyperparameters are fixed to $D = 0.05$, $\lambda = 1.3$, $\alpha = 0.1$, and $\beta = 0.125$. All models are trained on each training set for approximately 22 hours using 8 $\times$ NVIDIA A800 (80GB) GPUs.

Ablation studies for α and β are presented in Section 4.1 and Appendix D, respectively. Ablations for D and λ are not conducted due to the lack of directly comparable quantitative metrics; however, qualitative examples are provided in Appendix A.

D Effect of Format Reward Coefficient

We analyze the effect of the format reward coefficient β in GRPO training, as described in Section 2.3. Results under different values of β are reported in Table 8. As shown in the table, both

β	RE↑	CO (%)↑	IS (%)↓
w/o Coordinate System	0.277	19.8	50.9
w/ Coordinate System	0.311	22.0	45.2

Table 9: Effect of the overlaid coordinate system. Experiments are conducted on the Chinese training set for 100 training steps. Metric definitions follow those used in Table 1.

overly small and overly large values of β lead to degraded training performance. In practice, we select $\beta = 0.125$, which achieves the highest stroke representation reward r_s.

E Effect of Overlaid Coordinate System

We analyze the effect of incorporating the overlaid coordinate system, as described in Section 3.1. The results are reported in Table 9. As shown in the table, introducing the coordinate system consistently improves performance across all evaluated metrics.

F Early Hieroglyphic Writing Systems

Early hieroglyphic writing systems vary widely in form and function across historical and cultural contexts. Compared to modern writing systems, these scripts preserve rich pictographic elements, making visual structure central to their interpretation. In this work, we focus on Oracle Bone Script, Dongba script, and Egyptian Hieroglyphs—three representative writing systems originating from ancient Chinese and Egyptian civilizations (see Table 10 for a comparison).

F.1 Oracle Bone Script

Oracle Bone Script (OBS), dating back to the Shang Dynasty, is one of the earliest known writing systems in ancient China. As noted by Guan et al. (2024), OBS is characterized by a distinctive visual and topological structure, with symbols that closely reflect pictorial representations of objects, actions, and concepts. The script exhibits a high degree of visual variability and structural complexity, where meaning is often conveyed through the spatial arrangement and configuration of strokes. These properties make visual structure a fundamental component in the interpretation of OBS glyphs.

F.2 Dongba Script

The Dongba script, used by the Naxi people in Yunnan Province, China, is widely regarded as the only living pictographic writing system in the world (Yang, 2012). Dongba manuscripts are organized as spatially structured visual compositions rather than strictly sequential inscriptions, with glyphs arranged across a two-dimensional plane to convey meaning. Interpretation therefore depends on the relative spatial placement and visual relationships among symbols, rather than on a fixed reading order. A defining property of Dongba is its visual compositionality, whereby semantic distinctions are expressed through graphic variation at the stroke level. Glyph meaning can be modified by adding, removing, or altering visual components, allowing related concepts to be represented through systematic pictorial transformations.

F.3 Egyptian Hieroglyphs

Egyptian hieroglyphs served as the formal writing system of ancient Egypt and constitute a complex logophonetic system (Allen, 1951), rather than a collection of simple drawings. Although hieroglyphs are organized under a standardized classification scheme known as Gardiner's Sign List, this system primarily serves as a descriptive taxonomy of glyph forms. The interpretation of signs and inscriptions depends not only on categorical membership, but also on visual context, including shape, orientation, and spatial arrangement of glyphs.

G Details for Structure-guided Character Exploration Experiment

G.1 Data Curation

We conduct character exploration experiments on Oracle Bone Script (OBS), Dongba script, and Egyptian hieroglyphs. For OBS, we use all images in the HUST-OBC dataset (Wang et al., 2024) sourced from HWOBC, which provides glyph images with high visual clarity. For the Dongba script, we render glyph images from the `BabelStone Naxi`[4] font and additionally include 100 images manually cropped from DongbaMIE (Bi et al., 2025). DongbaMIE mainly provides paragraph-level images, from which we manually select and crop clear single-glyph instances, excluding samples that contain multiple overlapping glyphs. For Egyptian hieroglyphs, we render character images using the `Noto Sans Egyptian Hieroglyphs`[5] font.

[4] https://www.babelstone.co.uk/Fonts/Naxi.html
[5] https://fonts.google.com/noto/specimen/Noto+Sans+Egyptian+Hieroglyphs

Script	Origin	Region	Type	Status
Oracle Bone Script	∼1600 BC	China (Shang)	Pictograph	Extinct
Dongba Script	∼7th Century	China (Yunnan)	Pictograph	**Living**
Egyptian Hieroglyphs	∼3200 BC	Egypt	Logophonetic	Extinct

Table 10: Comparison of the early hieroglyphic writing systems used in our experiments.

G.2 Experimental Details

Our character exploration process consists of two stages: generating perturbed query images by masking strokes and matching characters based on the cosine similarity of CLIP (Radford et al., 2021) image embeddings.

Each query is defined by an image I and a set of valid strokes $\mathcal{S} = \{(\mathbf{p}_s^k, \mathbf{p}_e^k)\}_{k=1}^n$, where $\mathbf{p}_s^k, \mathbf{p}_e^k \in \mathbb{R}^2$ denote the start and end points of the k-th stroke, generated by HieroSA (Chinese).

We first compute the midpoint of each stroke as

$$\mathbf{p}_m^k = \frac{\mathbf{p}_s^k + \mathbf{p}_e^k}{2}. \tag{18}$$

To generate a stochastic masking pattern, we sample a masking center $\mathbf{c}$ uniformly from $\{\mathbf{p}_m^k\}_{k=1}^n$. For each stroke, we compute its Euclidean distance to the center,

$$d_k = \left\| \mathbf{p}_m^k - \mathbf{c} \right\|_2, \tag{19}$$

and define a distance-decayed weight

$$w_k = \exp\left(-\frac{d_k}{\tau}\right), \tag{20}$$

where $\tau > 0$ is a temperature parameter controlling the spatial locality of masking. We set $\tau = 0.4$ throughout our experiments. To normalize the overall masking strength, we rescale the weights by their mean,

$$\tilde{w}_k = w_k / \left(\frac{1}{n}\sum_{k=1}^{n} w_k\right), \tag{21}$$

and obtain the stroke-wise discard probability

$$p_k = \mathrm{clip}(\rho\, \tilde{w}_k,\, 0,\, 1), \tag{22}$$

with base discard rate $\rho \in (0, 1)$. We set $\rho = 0.5$ throughout our experiments. Each stroke is then independently masked according to $z_k \sim \mathrm{Bernoulli}(p_k)$, where $z_k = 1$ indicates that the stroke $(\mathbf{p}_s^k, \mathbf{p}_e^k)$ is masked. We mask the corresponding polygonal coverage regions of the selected strokes to obtain a perturbed image $I(\mathbf{z})$.

For a candidate pool of character images $\mathcal{D} = \{J_m\}_{m=1}^M$, which consists of all characters from the same writing system in our experiment, we compute cosine similarity between masked query and each candidate using CLIP image embeddings[6],

$$s_m^t = \mathrm{sim}\big(I(\mathbf{z}^t), J_m\big), \tag{23}$$

where t indexes the masking trial. We repeat the masking-and-scoring process for $T = 3$ independent trials and aggregate the similarities as

$$\bar{s}_m = \max\left(\{s_m^t\}_{t=1}^T\right). \tag{24}$$

The final matched results are obtained by ranking candidates by $\bar{s}_m$ and selecting the top-5 matches.

Our baseline ranks the candidate pool $\mathcal{D} = \{J_m\}_{m=1}^M$ by the CLIP cosine similarity $\mathrm{sim}(I, J_m)$ and selects the top-5 candidates. Compared to the baseline, our proposed stochastic masking procedure introduces localized, structure-aware perturbations at the stroke level. As a result, the matching process is biased toward identifying characters that differ in identity while sharing salient structural properties with the query, as observed in our experimental results.

H More Cases of Structure-Guided Character Exploration

Additional examples of structure-guided character exploration are presented in Table 11.

I Details for Optical Character Recognition Experiment

In this experiment, we examine whether the stroke-structure representations parsed by HieroSA can improve OCR performance.

We curate datasets using the same font families as those in the Chinese and Japanese test sets. We select 1,000 characters rendered from `JinNianYeYaoJiaYouYa` (Chinese) and `As Winter Comes` (Japanese) for OCR training. For evaluation, we use the same 1,000 character identities rendered from `Source Han Sans`

[6] We employ the version `clip-vit-large-patch14-336`.

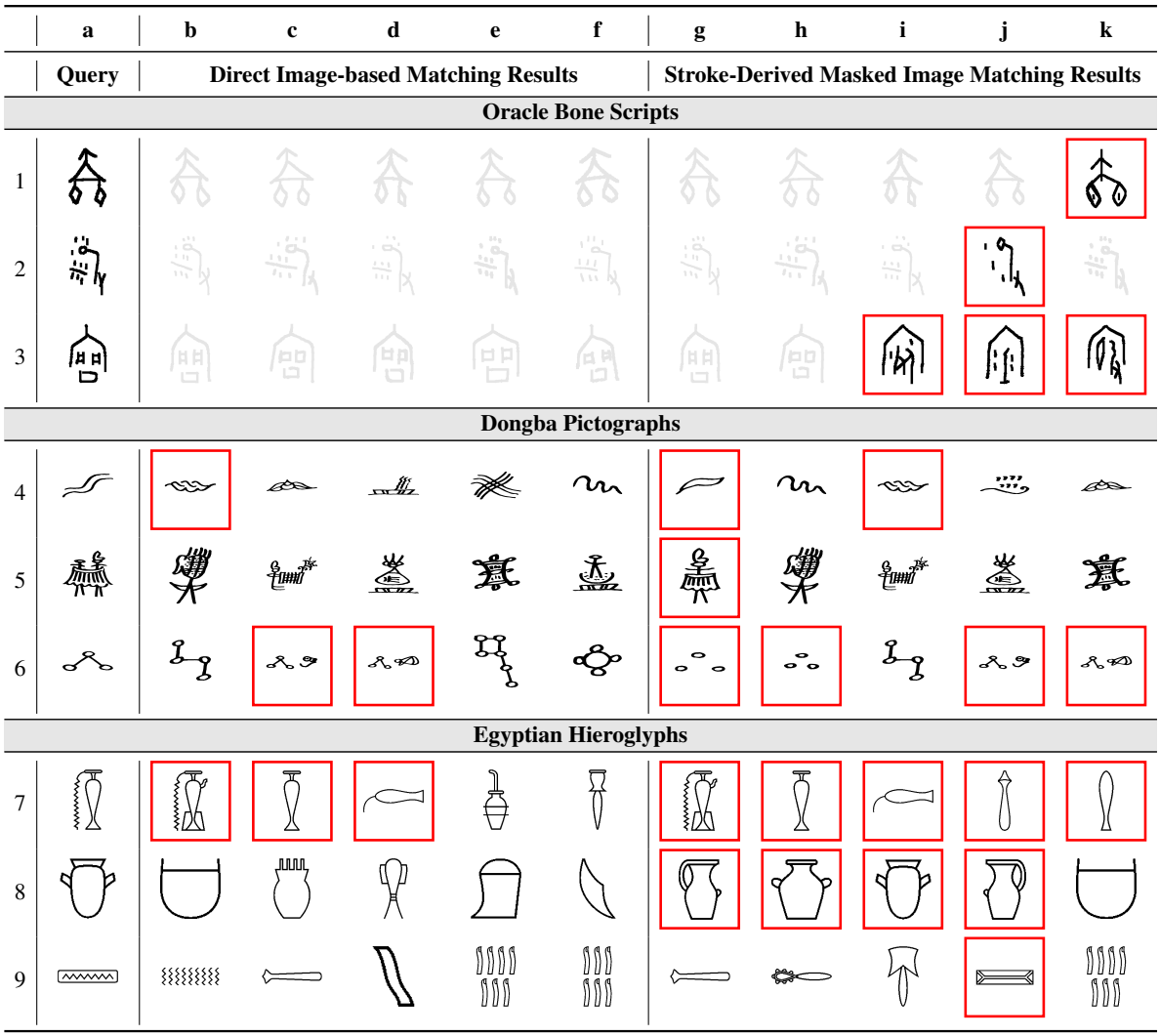

Table 11: More cases of structure-guided character exploration.

OCR Setting	Chinese	Japanese	AVG
Zero-shot	66.4	63.2	64.8
Trained w/o Stroke Rep	88.9	75.8	82.4
Trained w/ Stroke Rep	**89.7**	**77.3**	**83.5**

Table 12: OCR accuracy under different training settings. Stroke representations are generated by HieroSA (Chinese) and are consistently applied to both the OCR training and test sets.

(Chinese) and `BIZ UDPGothic` (Japanese). We ensure that the characters used for OCR are disjoint from those used for HieroSA training. The OCR task uses paired image-character supervision, and the model outputs the corresponding character sequence in text form.

We fine-tune Qwen3-VL-4B for the OCR task under different settings, with or without additional stroke representations. All OCR models are trained for two epochs under the same configuration for fair comparison.

Quantitative results are reported in Table 12. As shown in the table, fine-tuning substantially improves OCR performance over the zero-shot setting. Furthermore, incorporating stroke-structure representations consistently leads to additional performance gains on both Chinese and Japanese datasets, indicating that the parsed stroke information provides complementary structural cues beyond raw visual features.

J LLM Usage Statement

Throughout the preparation of this manuscript, LLMs are used exclusively for spelling and grammatical error checking. They are not employed for any other purposes, including the generation of research ideas or the validation of the proposed methods or experimental results.