SWPC-BNF: Recursive Structure-Constrained Representation of Chinese Characters

Li Weigang and João Paulo Vieira Costa

TransLab, Department of Computer Science, University of Brasilia, 70910-900, Brazil
weigang@unb.br, jpcosta1990@gmail.com
https://cic.unb.br/professores/89-weigang

Abstract. Traditional sequence-based representations struggle to capture the two-dimensional and hierarchical structure of Chinese characters (Hanzi). This paper introduces SWPC-BNF, a recursive structure-constrained framework based on Backus–Naur Form (BNF). By integrating structural composition, pictophonetic encoding, and visual attributes, SWPC-BNF represents Hanzi as explicit structural expressions rather than atomic symbols. To support large-scale evaluation, we construct a manually verified SWPC-BNF database containing 3,866 records covering 13 IDS-aligned structural categories. Experimental results on representative character sets demonstrate clear structural clustering and hierarchical consistency in the induced similarity space. Recursive representation further reduces the number of leaf components by approximately 56%, representation depth by 47%, and matching cost by up to 76% compared with full expansion. Representative recursive families such as 羸”/瀛” and 雚”/罐” illustrate the effectiveness of structural reuse. These results show that SWPC-BNF provides an interpretable, scalable, and computationally efficient alternative to sequence-based representations, supporting structure-aware modeling of Chinese characters.

Keywords: Backus–Naur Form · Chinese Character Recognition · Multimodal Processing · Recursive Representation · SWPC-BNF

1 Introduction

Recent advances in large language models (LLMs) and vision-language models (VLMs), exemplified by ChatGPT, Gemini, DeepSeek, InternVL, and others [1,2,3,4], have significantly accelerated artificial intelligence (AI). These systems are commonly understood as *foundation models* [5], trained on large-scale data and adaptable to diverse downstream tasks.

However, their representations are predominantly based on linear sequences or visual patches, which are not well aligned with the intrinsic properties of logographic writing systems such as Chinese characters (Hanzi). Structurally rich information in Hanzi is often reduced to tokenized sequences or compressed visual features, leading to information loss and limited interpretability. Existing optical character recognition (OCR) and CNN-based approaches further operate largely as black-box systems without explicit structural modeling.

To address this issue, Chinese NLP research has explored various encoding schemes. Early methods such as Four-Corner, Wubi, and Cangjie map characters into symbolic sequences, facilitating computation but flattening structure. More recent approaches attempt to recover structure using Ideographic Description Sequences (IDS) of UNICODE, such as the graph-based radical structure tree representation (GRSTR) [6], or visual feature methods based on OCR and SIFT [7,8]. However, structural information remains implicit, approximate, or computationally expensive.

Recent benchmarks on Chinese LLMs and VLMs, including AlignMMBench [9], CSVQA [10], CVLUE [11], and VisTW [12], indicate that performance limitations stem from representation mismatch rather than data scale alone. Linguistically, models rely on statistical priors rather than intrinsic structure, as reflected in the "paradox of poetic intent" [13]. Visually, recognition degrades when structural complexity exceeds perceptual thresholds [8], often leading to language-driven completion.

These observations reveal a fundamental limitation: existing multimodal models lack structure-aware representations for Chinese characters. Unlike alphabetic scripts, Hanzi are inherently two-dimensional and hierarchically composed, typically recognized through holistic perception ("*Once Learning*") [14]. This motivates explicit structural modeling rather than sequence approximation.

In previous work, we proposed Six-Writings Pictophonetic Coding (SWPC) [15], which integrates structural, phonetic, and semantic information. Although SWPC 1.0 and 2.0 improved Chinese character representation, they remained limited in handling complex characters due to encoding length, structural ambiguity, and restricted compositional expressiveness.

To address these limitations, we introduce **SWPC-BNF**, a structure-constrained framework based on Backus–Naur Form (BNF) [16,17]. SWPC-BNF models Chinese characters as recursively compositional symbolic structures, integrating structural representation, pictophonetic encoding, and visual attributes within a unified formal framework. SWPC-BNF provides a foundation for structure-aware matching and retrieval by exploiting hierarchical attributes such as layout, substructures, and recursive composition. A key feature is its layered recursive design, which enables structural reuse, reduces representation depth, and lowers matching complexity while preserving interpretability.

To support large-scale evaluation, we constructed a manually verified SWPC-BNF database containing 3,866 records, covering the 3,500 most commonly used Chinese characters and all 13 IDS-aligned structural categories. Experimental results demonstrate clear structural clustering, hierarchical consistency, and substantial reductions in representation depth and matching cost.

Overall, SWPC-BNF reframes Hanzi representation as a structure-constrained problem in a multimodal space, providing an interpretable and scalable framework for structure-aware modeling of Chinese characters. This study focuses on representation and structural modeling, leaving full downstream OCR/VLM integration for future work.

2 Related Work

Research on Hanzi processing has evolved from sequence-based encodings to structure-aware representations. Early approaches relied on *sequence-based component encoding*, where characters are decomposed into ordered sequences of discrete components. For example, CalliGAN [9] represents each character as a sequence of component identifiers processed by recurrent neural networks. While effective for integration with sequence models, such methods treat characters as linear sequences, ignoring their two-dimensional and hierarchical structure.

To address this limitation, subsequent work introduced *structure-aware representations*. Radical-based models, such as the Radical Analysis Network (RAN) [18], and SideNet [19], incorporate component relationships and glyph features to improve representation learning. However, structural information is typically learned implicitly within latent embeddings rather than explicitly encoded.

More recent studies explore *explicit structural modeling*. GRSTR [6] constructs hierarchical trees from IDS and processes them using graph neural networks. While this improves modeling of spatial and hierarchical relations, it remains within a *compositional paradigm*, where characters are processed through traversal of components, which may lead to error propagation and increased computational cost as complexity grows.

From a cognitive perspective, this contrasts with *Once Learning* [14], which characterizes character recognition as a holistic process with parallel structural perception. This suggests that effective modeling may require globally consistent structural representations rather than purely compositional mechanisms.

In addition, most approaches focus on representation learning in continuous spaces, with limited consideration of *structural constraints* and *capacity limits*. Prior work on character length–frequency distributions and visual thresholds [8] shows that violating these constraints can introduce ambiguity rather than improve performance.

In contrast, this work formulates Chinese character processing as a *structure-constrained modeling problem*, integrating (i) holistic structural acquisition, (ii) explicit structural encoding via SWPC, and (iii) statistical and perceptual constraints. Based on this view, we propose SWPC-BNF, a recursive grammar-based framework for interpretable and efficient structural modeling.

3 Technical Background

3.1 Backus–Naur Form (BNF)

Backus–Naur Form (BNF) is a formal notation for describing language syntax through recursive production rules of the form

$$\langle A \rangle ::= \alpha, \tag{1}$$

where $\langle A \rangle$ is a non-terminal symbol and α is a sequence of terminal and/or non-terminal symbols. BNF provides a compact representation of hierarchical structures via recursive definitions. A key property of BNF is its ability to model *compositional hierarchy*: complex structures are formed by recursively combining simpler components, preserving structural relationships [16,17].

For Hanzi modeling, BNF offers three advantages: (i) *structure preservation*, explicitly encoding spatial configurations such as left–right and top–bottom; (ii) *recursive compactness*, enabling nested structures without long linear encodings; and (iii) *interpretability*, as each production rule corresponds to a meaningful structural relation. These properties make BNF well suited for modeling logographic systems. In this work, it serves as the formal foundation of SWPC-BNF.

3.2 Input Method Encoding and Structural Constraints

A central challenge in Hanzi processing is encoding inherently two-dimensional, hierarchical structures into computational representations. Traditional input methods, such as Four-Corner, Wubi, and Cangjie, map characters into symbolic sequences, enabling efficient processing but flattening structural relationships.

As a result, spatial configurations and component hierarchies are only partially preserved, leading to ambiguity and reduced structural discriminability. To address this, SWPC framework integrates structural, phonetic, and semantic information into a unified representation [15]. Chinese character modeling is further constrained by *statistical and perceptual limits*. The Character Length–Character Frequency (CLCF) framework shows stable statistical regularities [8], but increasing encoding granularity often introduces *encoding-induced noise*. Similarly, in the visual domain, excessive stroke complexity relative to pixel resolution leads to structural merging and recognition errors.

These observations indicate that effective modeling must operate under *structural and capacity constraints*. From this perspective, SWPC can be viewed as a threshold-aware encoding strategy, but its lack of explicit recursion and grammar limits its scalability.

3.3 From SWPC 1.0 to SWPC 2.0

SWPC was proposed to encode Chinese characters by integrating morphological, phonetic, and semantic information. In SWPC 1.0, characters are decomposed into semantic radicals and phonetic components and encoded in a fixed-length format using Wubi and Four-Corner features.

Despite incorporating linguistic structure, SWPC 1.0 remains a *linear encoding scheme*, where structural relations are implicit and difficult to scale for complex characters.

SWPC 2.0 extends this approach toward a *structure-aware representation* [20]. It introduces a structural interface layer linking symbolic encoding and composition, and organizes components into a text–image dual-modal matrix, enabling characters to be interpreted as structured combinations.

However, SWPC 2.0 still lacks a formal mechanism for hierarchical composition. Structural relations are not explicitly defined.These limitations motivate a grammar-based formulation.

4 SWPC-BNF: A Structure-Constrained Framework

4.1 Encoding Foundation: Capacity and Primitive Selection

Chinese characters exhibit high structural density while maintaining a relatively limited practical vocabulary. To support large-scale structural modeling, we constructed a manually verified SWPC-BNF database containing 3,866 records, covering the 3,500 most commonly used Chinese characters together with structurally significant components and indivisible character elements.

Capacity constraint. SWPC encoding operates within a **26×26 two-letter matrix**, providing a theoretical capacity of **676 encoding slots**. This design, inherited from Wubi-style encoding systems, ensures compactness and computational efficiency. Consequently, the allocation of Level-1 primitives must prioritize high-frequency and structurally stable units.

Level-1 primitive selection. Level-1 primitives are atomic units that are not further decomposed during encoding. They occupy the 676 slots according to the priority scheme shown in Table 1.

Table 1. Composition of Level-1 primitives (676 encoding slots). Priority 1 (301 radicals) follows the 2022 national standard [21].

Priority	Category	Count	Examples
1	Principal radicals (2022 standard)	201	亻, 口, 木, 氵, 艹
1	Dependent radical variants	100	⻊, ⺮, 衤, ⺗
2	Indivisible high-frequency characters	∼100	之, 乎, 者, 也, 以
3	Indivisible non-character components	∼100	⺧, ⺈, 厂
4	Divisible but high-frequency composites	∼175	圭, 衣, 青, 齿
	Total	676	

Level-2 reusable units. **Level-2 units** are composite structures formed from Level-1 primitives. In principle, any recursive combination of Level-1 primitives may serve as a candidate Level-2 unit. However, to avoid combinatorial explosion and maintain computational efficiency, SWPC-BNF adopts a controlled set of reusable units selected according to frequency and structural complexity.

These units function as **reusable macros**, enabling structural compression while preserving full recoverability. Typical examples include 隹 (in 谁, 推, 唯, 维, 雀, 鹤) and 相 (in 湘, 想, 箱, 厢, 霜, 缃). Once registered, such units can be recursively reused in higher-level structures, reducing representation depth and matching complexity.

This design balances *expressive completeness* through recursive compositionality and *computational efficiency* through frequency-guided reuse. Furthermore, Level-2 units can be dynamically extended as the character inventory grows, providing a scalable and adaptive structural vocabulary for SWPC-BNF.

4.2 Symbolic Representation of SWPC-BNF

Building upon SWPC 1.0 and SWPC 2.0, we introduce **SWPC-BNF**, a structure-constrained and recursively defined representation framework for Chinese characters. Unlike prior encoding schemes that rely on linear sequences or implicit structures, SWPC-BNF represents characters as *explicit hierarchical symbolic expressions* governed by a BNF-style grammar [16,17].

Structural categories. SWPC-BNF adopts **13 IDS-aligned structural categories**, covering the major spatial composition patterns found in modern Chinese character formation. These categories provide a compact and computation-oriented abstraction of Ideographic Description Sequences (IDS), while preserving the essential spatial relationships required for recursive structural modeling.

Table 2. Structural distribution of the manually verified 3866-record SWPC-BNF corpus across 13 IDS-aligned categories.

Cat.	Structure	Count	Ratio (%)	Avg. Code Len.	Avg. Units	IDS	Example
SG	Single	676	17.49	2.00	1.00	–	大
LR	Left-Right	2091	54.09	4.82	2.41	⿰	任
TB	Top-Bottom	671	17.36	4.91	2.46	⿱	芯
LMR	Left-Middle-Right	49	1.27	6.29	3.14	⿲	衍
TMB	Top-Middle-Bottom	82	2.12	6.39	3.20	⿳	裹
EF	Full Enclosure	13	0.34	4.15	2.08	⿴	回
ET	Top Enclosure	30	0.78	4.27	2.13	⿵	问
EB	Bottom Enclosure	1	0.03	4.00	2.00	⿶	凶
EL	Left Enclosure	10	0.26	4.60	2.30	⿷	区
DT	Top-left Diagonal	114	2.95	5.19	2.60	⿸	病, 厅
DB	Top-right Diagonal	31	0.80	4.39	2.19	⿹	旬, 戒
DL	Bottom-left Diagonal	90	2.33	4.60	2.30	⿺	赶, 这
HC	Hybrid/Complex	8	0.21	7.50	3.75	⿻	幽
Total	–	3866	100.00	–	–	–	–

These categories cover the principal IDS configurations while providing a more compact abstraction suitable for structural indexing, recursive representation, and similarity computation.

The distribution of the 3,866-record corpus is highly skewed toward a small number of dominant structural categories (Table 2). Left–right structures account for 54.09% of all records, followed by top–bottom structures (17.36%) and single-component forms (17.49%). Together, these three categories cover nearly 89% of the corpus. Such concentration enables efficient hierarchical pruning during structural retrieval and matching, substantially reducing the candidate search space before finer-grained component analysis.

Symbolic form. In SWPC-BNF, each Chinese character is represented as a structured object rather than an atomic symbol. The representation consists of a recursive structural expression together with a set of auxiliary attributes:

$$\mathcal{C} = \texttt{StructureExpr} \; ; \; [\texttt{Code}, \texttt{NComp}, \texttt{STK}, \texttt{IDS}] \tag{2}$$

where `StructureExpr` is a BNF-style structural expression, `Code` denotes the corresponding SWPC encoding, `NComp` is the number of structural components, `STK` is the stroke count, and `IDS` indicates the associated IDS-aligned structural category.

SWPC-BNF adopts a dual-layer representation consisting of: (i) a **structural layer**, which explicitly describes the hierarchical composition of a character through recursive operators; (ii) a **coding layer**, which maps structural units to compact SWPC codes. The two layers are linked through a one-to-one correspondence between structural components and their symbolic encodings, ensuring both interpretability and machine-readable representation.

Definition 1 (SWPC-BNF Representation). A Chinese character $\mathcal{C}$ is represented as

$$\mathcal{C} = (\mathcal{S}, \mathcal{E}), \tag{3}$$

where $\mathcal{S}$ denotes a recursive structural expression generated by the SWPC-BNF grammar, and $\mathcal{E}$ denotes the corresponding symbolic encoding layer.

Definition 2 (Structural–Code Mapping). Let $\mathcal{S} = \langle s_1, s_2, \ldots, s_n \rangle$ be a structural expression. A structural–code mapping is a function

$$\phi : s_i \mapsto e_i \tag{4}$$

that assigns each structural unit s_i to a symbolic code e_i. The encoding layer $\mathcal{E}$ is obtained by applying ϕ to all components in $\mathcal{S}$ while preserving the hierarchical structure. This formulation separates structural composition from symbolic encoding, enabling explicit modeling of both glyph-level hierarchy and machine-readable representation. It transforms Hanzi from flat symbols into *interpretable structural objects*, enabling explicit modeling of composition and hierarchy.

4.3 Encoding Rules and Directional Operators

A central innovation is the use of **directional expansion operators** to resolve positional ambiguity of radicals. In linear encodings, the same radical may appear on different sides (e.g., 阝 in “阿” vs. “郑”), leading to representational ambiguity. SWPC-BNF disambiguates such cases explicitly.

BNF-style formalization. At the core, SWPC-BNF defines character structure using recursive production rules, following the same form as BNF [16,17]:

$$\langle S \rangle ::= \langle P \rangle \mid \texttt{LR}\langle S, S \rangle \mid \texttt{TB}\langle S, S \rangle \mid \texttt{LMR}\langle S, S, S \rangle \mid \texttt{TMB}\langle S, S, S \rangle \mid \texttt{EF}\langle S, S \rangle \mid \cdots \tag{5}$$

where $\langle P \rangle$ denotes a primitive unit (Level-1 or Level-2 component). Recursion terminates when all substructures are reduced to primitive units. This formulation explicitly captures the hierarchical and compositional nature of Chinese characters within a formal grammar framework.

Basic composition rules. For a standard left–right structure:

$$\texttt{LR<Left,Right>};[< Code_L, Code_R >; NComp; STK;\ ] \tag{6}$$

Similarly, top–bottom and enclosure structures are defined using TB, LMR, TMB, EF, ET, EB, EL, DT, DB, and DL forms.

Directional expansion operators. To encode asymmetrical structural refinement, we introduce two directional expansion operators. Both operators expand the *non-radical side*, while recording the position of the radical.

- `/` (forward expansion): the radical is located on the **left** in LR structures or on the **top** in TB structures; the expansion applies to the opposite side: (i) LR example: 阿 = `LR/< 阝, 可 >`; (ii) TB example: 苗 = `TB/< 艹, 田 >`.
- ‡ (reverse expansion): the radical is located on the **right** in LR structures or on the **bottom** in TB structures; the expansion applies to the opposite side: (i) LR example: 郑 = `LR‡< 关, 阝 >`; (ii) TB example: 想 = `TB‡<LR/< 木, 目 >, 心 >`.

Although the two operators perform the same functional role, complementary expansion of the non-radical side, they encode opposite directional asymmetries through the position of the radical. This makes SWPC-BNF more informative than a purely unordered component tree, since the operator itself preserves the structural relation between radical position and expansion direction.

Decomposition flexibility. SWPC-BNF does not assume a unique decomposition. Alternative representations may coexist. For example, 草 may be represented as `TMB/< 艹, 日, 十 >` or `TB/< 艹, 早 >`. The choice depends on structural standardization, character frequency, visual discriminability, and downstream computational requirements.

Interpretability. Unlike implicit representations in neural models, each SWPC-BNF expression corresponds to a *human-interpretable structural description*, enabling transparent reasoning and verification.

4.4 Recursive Structural Expansion

A key innovation of SWPC-BNF is **recursive structural expansion**, which leverages BNF-style compositionality to represent complex characters compactly.

Recursive compression via Level-2 units. With Level-2 units (Sec. 4.1), frequently occurring substructures can be reused as atomic components. For example, the character 瀛 (氵 + 嬴) can be represented as: (i) Full expansion: `LR/< 氵, TMB/< 亡, 口,LMR< 月, 女, 凡 >>>`; (ii) Recursive form: `LR/< 氵, 嬴 >`.

This transformation illustrates how recursive reuse reduces structural complexity while preserving full recoverability.

Efficiency gains. We evaluate efficiency using three metrics: number of leaf components (N_{leaf}), BNF depth (D), and matching cost proxy ($C \approx N_{\text{leaf}} \cdot D$). For 瀛: N_{leaf}: $6 \rightarrow 2$ ($\downarrow$67%); Depth D: $3 \rightarrow 2$ ($\downarrow$33%); Cost C: $18 \rightarrow 4$ ($\downarrow$78%).

These reductions demonstrate that recursive encoding transforms character representation from a combinatorial expansion problem into a **hierarchically constrained structure**.

Template instantiation and vocabulary expansion. A representative example is the triplication template

$$\langle S_{\text{trip}}\rangle ::= \texttt{TB/<}\langle X\rangle\texttt{,LR/<}\langle X\rangle\texttt{,}\langle X\rangle\texttt{>>}, \quad \langle X\rangle \in \mathcal{P} \cup \mathcal{M}, \tag{7}$$

where $\mathcal{P}$ denotes the set of Level-1 primitive units and $\mathcal{M}$ denotes the set of reusable macro units; X denotes a primitive unit or reusable macro. Different instantiations generate structurally related characters, e.g.,

$$\text{叒} : X = \texttt{Cc}(\text{又}), \quad \text{淼} : X = \texttt{Il}(\text{水}), \quad \text{鑫} : X = \texttt{Qa}(\text{金}).$$

To evaluate the generative capability of this template, we collected 53 known triplication characters from modern and historical Hanzi resources. Among them, only five (众, 森, 品, 晶, and 矗) appear in the current 3866-record SWPC-BNF corpus.

Applying the template to the primitive units and reusable components contained in the corpus yields 53 candidate triplication characters. Among these, 40 are not explicitly included in the current database (75.47%), demonstrating substantial out-of-vocabulary (OOV) generation capability. Examples include 焱 (three fires), 淼 (three waters), 磊 (three stones), and the traditional forms 鑫 and 聶. In addition, eight generated triplication characters require traditional-character primitives (e.g., 雲 and 龍), whose inclusion would further enable the generation of rare forms such as 靐 and 龘.

More generally, a structural template in SWPC-BNF is a recursive production rule:

$$X^{(k+1)} = \mathcal{T}(X_1^{(k)}, X_2^{(k)}, \ldots, X_m^{(k)}), \tag{8}$$

where $X_i^{(k)}$ may be either a primitive unit or a reusable structure generated at a lower level. Thus, templates function as generative rules rather than static features:

$$X \rightarrow \mathcal{T}(X) \rightarrow \mathcal{T}'(\mathcal{T}(X)).$$

For example, `Y= 叒 =TB/<Cc, LR/<Cc, Cc>>` is generated from `X= 又 (Cc)` by the triplication template, then reused in `桑 = 叒 + 木 (Ss)=TB‡<Y,Ss>`, and further reused in `嗓 = 口 (Kf)+ 桑 = LR/<Kf,TB‡<Y,Ss>>`.

This mechanism has two important implications. First, vocabulary growth is modeled as *structured generation* rather than flat enumeration. Newly encountered characters can often be represented as compositions of existing primitives and templates. Second, highly complex or rare characters can be decomposed into reusable structural units, preserving interpretability while reducing representation complexity.

5 Structural Similarity and Efficiency Analysis

Before presenting quantitative results, we briefly compare SWPC-BNF with existing approaches (Table 3). Unlike IDS, which is descriptive but not computationally efficient, and symbolic encodings such as Wubi and SWPC 2.0, which sacrifice structural transparency for compactness, SWPC-BNF integrates explicit hierarchy, recursive composition, and machine-processable encoding within a unified framework.

Table 3. Comparison of representational paradigms for structural modeling of Hanzi

Property	**IDS**	**SWPC 2.0/Wubi**	**SWPC-BNF (proposed)**
Explicit structure modeling	High (tree-like)	Low (flattened sequence)	High (BNF hierarchy)
Computational usability	Low (descriptive only)	High (symbolic codes)	High (symbolic + structure)
Recursive reuse	No	No	Yes (Level-2 units)
Complex character handling	Poor (deep trees)	Moderate (code ambiguity)	Strong (compact & interpretable)
Example (羸)	⿳吂⿲月 贝凡	`Ydkfeemjmu (SWPC)` `YEMY (Wubi-98)`	`TMB/< 亡, 口,LMR< 月, 女, 凡 >>` → `TMB/<Yd,Kf,LMR<Ee,Mj,Mu>>`

This combination is essential for similarity analysis. Structural similarity in Hanzi depends on both components and their spatial organization. Sequence-based encodings obscure these relations, while IDS is not directly computation-ready. SWPC-BNF resolves this by providing a *structure-aware and computation-ready representation.*

Consequently, similarity in the SWPC-BNF space is defined over aligned structural expressions rather than flat sequences. Characters sharing radicals, phonetic components, or recursive substructures form coherent clusters, while unrelated characters remain well separated, forming the basis for the quantitative analysis in the following subsection.

5.1 Similarity Analysis in the SWPC-BNF Space

To evaluate whether SWPC-BNF induces a meaningful structural representation space, we performed a similarity analysis on 115 representative characters spanning the major structural categories and several recursive character families (e.g., 青", 可", 藿", and 羸"). The purpose of this experiment is not large-scale corpus evaluation, but visualization of the structural organization induced by SWPC-BNF representations. The full 3866-record database is used for corpus statistics, structural distribution analysis, and template-instantiation evaluation, whereas the 115-character subset is used only for visualization. Each Hanzi is

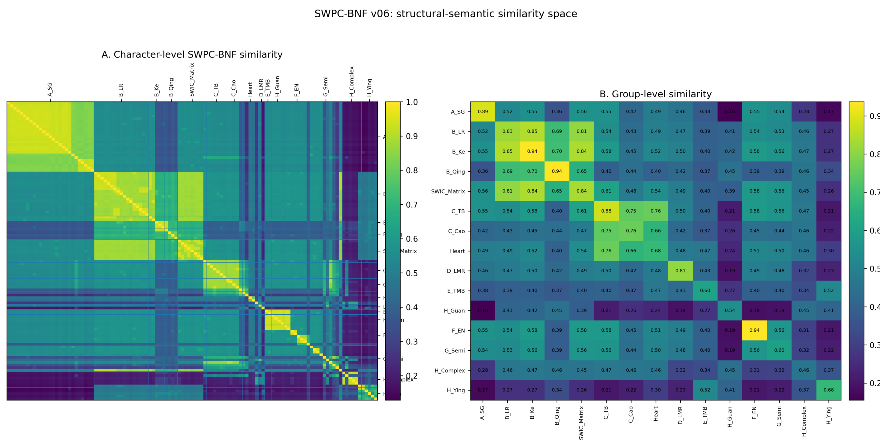

Fig. 1. Structural similarity patterns induced by SWPC-BNF. (A) Character-level heatmap illustrating structural clustering among representative characters. (B) Cluster organization of recursive families (e.g., 青”, 可”, 雚”, 嬴”), demonstrating hierarchical consistency and component reuse.

represented according to Equation 2. Pairwise similarity is defined as:

$$\mathrm{Sim}(c_i, c_j) = w_s S_{\text{struct}} + w_c S_{\text{code}} + w_n S_{\text{comp}} + w_v S_{\text{visual}} \tag{9}$$

where S_{struct} captures BNF-level alignment, $w_s = 0.40$; S_{code} measures SWPC similarity, w_c=0.30; S_{comp} measures overlap between structural components, w_n=0.15; and S_{visual} is approximated via stroke-based features, w_v=0.15. The weights were fixed *a priori* to emphasize structural information and were not optimized on any downstream task.

Figure 1 shows the resulting similarity matrix. Clear block structures emerge, indicating strong intra-group coherence and relatively low inter-group similarity. Recursive families such as 青”, 可”, 雚”, and 嬴” form coherent subclusters despite differences in surface appearance, suggesting that SWPC-BNF successfully captures underlying structural relationships.

Characters sharing structural templates or major components form compact clusters. Representative examples include the 青” family (清, 情, 晴, 请), the 艹” group (草, 花, 苗, 苦), and the “可” family (何, 河, 柯, 珂, 坷), whose pairwise similarities typically exceed 0.7. Although the experiment uses a representative subset of 115 characters, the observed clustering behavior is consistent with the large-scale structural statistics obtained from the full 3,866-record SWPC-BNF database.

5.2 Efficiency Analysis via Recursive Representation

Building on the illustrative example in Sec. 4.4, we now evaluate efficiency at the dataset level. As shown in Fig. 2, recursive representation consistently reduces structural complexity across 115 Chinese characters.

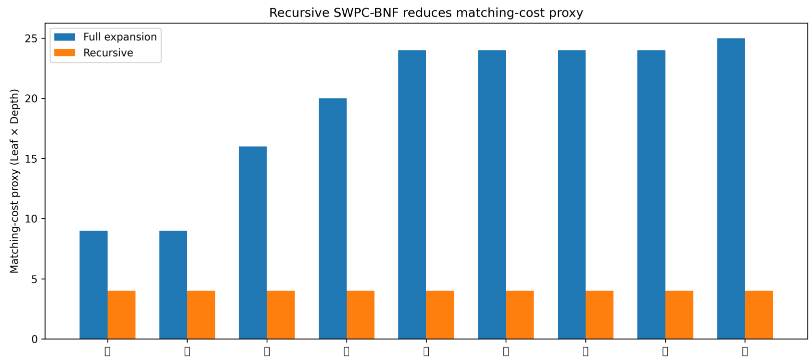

Fig. 2. Efficiency comparison between full expansion and recursive representation in SWPC-BNF. The cost proxy $C \approx N_{\text{leaf}} \cdot D$ demonstrates consistent and scalable reduction across structural complexity levels when reusable substructures (e.g., “嬴” and “藿”) are applied.

Comparison setup. We compare two strategies: (i) **Full expansion**, where all components are explicitly decomposed; and (ii) **Recursive representation**, where Level-2 units are reused as atomic structures.

Overall results. Across the dataset, recursive encoding reduces N_{leaf} by $\approx 56\%$, depth D by $\approx 47\%$, and cost proxy C by $\approx 76\%$. These gains are more pronounced for structurally complex characters with reusable substructures.

Scalability. Recursive encoding ensures bounded representation depth and sublinear growth in matching complexity, making large-scale character sets computationally tractable.

Structural advantage. Unlike tree- or graph-based approaches that fully expand all nodes, SWPC-BNF employs **controlled recursion**, enabling compact representation while preserving structural integrity.

Summary. Together with the clustering results in Fig. 1, these findings show that SWPC-BNF achieves both structural fidelity and computational efficiency. By reducing representation depth and matching cost, it transforms character processing into a **structure-constrained optimization problem**.

6 Discussion

This work introduces SWPC-BNF as a structure-constrained and recursive framework for Chinese character representation. Beyond the empirical results presented in Section 5, several broader implications emerge.

From symbolic encoding to structural representation. Traditional Chinese character processing typically treats Hanzi as atomic symbols or flat code sequences. In contrast, SWPC-BNF models characters as recursive structural objects generated by explicit composition rules. By integrating SWPC with Backus–Naur Form (BNF), character representations become both machine-processable and structurally interpretable, enabling direct access to compositional information that is typically hidden in conventional encodings or embeddings.

Recursion, scalability, and vocabulary growth. A key contribution of SWPC-BNF is its use of recursive composition and reusable structural units. Rather than treating vocabulary expansion as the addition of new atomic symbols, SWPC-BNF represents new characters through combinations of existing primitives, macros, and templates. This transforms dictionary growth from a flat enumeration problem into a structured generative process. The manually verified 3,866-record database demonstrates that large character inventories can be represented within a unified recursive framework while preserving recoverability and interpretability.

Rare characters and compositional generalization. The template-based design of SWPC-BNF provides a natural mechanism for handling rare and out-of-vocabulary characters. Patterns such as [`TB/<X,LR/<X,X>>`] can be instantiated by different components to generate structurally related character families. Consequently, unseen characters may be interpreted as compositions of known primitives and templates rather than entirely new symbols. This suggests a pathway toward structure-driven generalization beyond fixed vocabularies.

Toward structure-aware language modeling. SWPC-BNF can be viewed as a structured representation layer that bridges human character-formation knowledge and computational modeling. By explicitly encoding hierarchy, spatial relations, and recursive composition, it provides a complementary perspective to purely sequence-based approaches. More broadly, it illustrates how symbolic structure can be incorporated into machine-readable representations, supporting interpretable and structure-aware processing of logographic languages.

7 Conclusion and Future Work

This paper proposes SWPC-BNF, a recursive grammar-based framework for structure-constrained representation of Chinese characters.

To support large-scale evaluation, we constructed a manually verified SWPC-BNF database containing 3,866 records covering the 3,500 most commonly used Hanzi and all 13 IDS-aligned structural categories. Experimental results demonstrate clear structural clustering, hierarchical consistency, and substantial reductions in representation depth and matching cost through recursive reuse.

Methodologically, SWPC-BNF introduces a grammar-based representation paradigm in which Chinese characters are modeled as recursive structural expressions rather than isolated symbols. The framework supports explicit structural composition, reusable macros, and dynamic templates, enabling interpretable and recoverable representations while maintaining computational efficiency.

More broadly, this work suggests that Chinese character processing can be reformulated as a structure-constrained problem. By representing characters through recursive composition rules, vocabulary growth becomes a structured generative process rather than a flat enumeration problem, providing a natural mechanism for handling rare and out-of-vocabulary characters.

Several limitations remain. The current study focuses on representation and structural modeling rather than downstream recognition tasks. In addition, the similarity metric employs fixed weights and has not been optimized for task-specific applications. Further validation on OCR, vision-language, and language-modeling tasks will require dedicated datasets and evaluation protocols.

Future work will focus on: (i) extending the SWPC-BNF database to larger character inventories and historical character sets; (ii) developing structure-constrained retrieval, matching, and indexing algorithms; (iii) investigating template-based OOV generalization; and (iv) integrating SWPC-BNF with neural architectures for structure-aware language and multimodal modeling.

Acknowledgements

This work was partially supported by the Brazilian National Council for Scientific and Technological Development (CNPq). We thank the developers of publicly available LLM services (ChatGPT, DeepSeek, Gemini) for assistance with language refinement.

References

1. OpenAI: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023). `https://doi.org/10.48550/arXiv.2303.08774`
2. Google: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023). `https://doi.org/10.48550/arXiv.2312.11805`
3. Guo, D., Yang, D., Zhang, H. et al.: DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. In: Nature, vol. 645, pp. 633-638, (2025). `https://doi.org/10.1038/s41586-025-09422-z`
4. Chen, Z., Wu, J., et al.: InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In: Proceedings of the IEEE/CVF, pp. 24185-24198, (2024). `https://doi.org/10.1109/CVPR52733.2024.01111`
5. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S. et al.: On the Opportunities and Risks of Foundation Models. In: arXiv preprint arXiv:2108.07258 (2021). `https://doi.org/10.48550/arXiv.2108.07258`
6. Dong, Y., Wu, B., Ma, J. and Li, X.: Graph-based radical structure tree representation for zero-shot Chinese character recognition. In: Pattern Recognition, vol. 177, p. 113314, (2026). `https://doi.org/10.1016/j.patcog.2026.113314`
7. Lowe, David G. "Distinctive image features from scale-invariant keypoints." In: International journal of computer vision, vol. 60, no. 2, p. 91-110, (2004). `https://doi.org/10.1023/B:VISI.0000029664.99615.94`

8. Weigang, L., Ramos, R.M., Brom, P.C. and Li, D.L.: Threshold study for Hanzi image recognition: defining character and component limits in Chinese, Japanese, and Korean script processing. In: International Journal of Asian Language Processing, vol. 35, no. 01, p. 2450011, (2025). https://doi.org/10.1142/S2717554524500116
9. Wu, S.-J., Yang, C.-Y. and Hsu, J.Y.-j.: Calligan: Style and structure-aware chinese calligraphy character generator. In: arXiv preprint arXiv:2005.12500 (2020). https://doi.org/10.48550/arXiv.2005.12500
10. Jian, A., Qiu, W., Wang, X., Wang, P., Hao, Y., Pei, J., Wei, Y., Peng, Y. and Song, X.: CSVQA: A Chinese Multimodal Benchmark for Evaluating STEM Reasoning Capabilities of VLMs. In: arXiv preprint arXiv:2505.24120 (2025). https://doi.org/10.48550/arXiv.2505.24120
11. Wang, Y., Liu, Y., Yu, F., Huang, C., Li, K., Wan, Z., Che, W. and Chen, H.: Cvlue: A new benchmark dataset for chinese vision-language understanding evaluation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 8, pp. 8196-8204, (2025). https://doi.org/10.1609/aaai.v39i8.12345
12. Tam, Z.R., Pai, Y.-T., Lee, Y.-W. and Chen, Y.-N.: Vistw: Benchmarking vision-language models for traditional chinese in taiwan. In: arXiv preprint arXiv:2503.10427 (2025). https://doi.org/10.48550/arXiv.2503.10427
13. Weigang, L. and Brom, P.C.: Paradox of poetic intent in back-translation: evaluating the quality of large language models in Chinese translation. In: Frontiers of Information Technology & Electronic Engineering, vol. 26, no. 11, pp. 2176-2203, (2025). https://doi.org/10.1631/FITEE.2400123
14. Weigang, L., Silva, N.: A study of parallel neural networks. In: Proceedings of the Int. Joint Conference on Neural Networks (IJCNN'99), pp. 1113-1116, (1999).
15. Weigang, L., Marinho, M.C., Li, D.L. and De Oliveira, V.V.: Six-writings multimodal processing with pictophonetic coding to enhance Chinese language models. In: Frontiers of Information Technology & Electronic Engineering, vol. 25, no. 1, pp. 84-105, (2024). https://doi.org/10.1631/FITEE.2300384
16. Backus, J.W.: The syntax and semantics of the proposed international algebraic language of the Zurich ACM-GAMM conference. In: Proceedings of the International Conference on Information Processing, (1959).
17. Naur, P. et al.: Revised report on the algorithmic language ALGOL 60. In: Communications of the ACM, vol. 3, no. 5, pp. 299-314, (1960). https://doi.org/10.1145/367236.367262
18. Zhang, J., Du, J. and Dai, L.: Radical analysis network for learning hierarchies of Chinese characters. In: Pattern Recognition, vol. 103, p. 107305, (2020). https://doi.org/10.1016/j.patcog.2020.107305
19. Li, Z., Huang, Y., Peng, D., He, M. and Jin, L.: Sidenet: Learning representations from interactive side information for zero-shot chinese character recognition. In: Pattern Recognition, vol. 148, p. 110208, (2024). https://doi.org/10.1016/j.patcog.2023.110208
20. Weigang, L. and Costa, A.J.P.V.: LLM-OCR-SWPC: Hanzi's Cognitive Advantage and Token Efficiency in Multimodal AI. In: Proceedings of 2026 18th Int. Conf. on Computer Research and Development, pp. 255-260, Singapore, (2026).
21. Ministry of Education of the People's Republic of China and National Language Commission of the People's Republic of China: Table of Radicals (2022 Revision). Standards Press of China, Beijing (2022).