



# BIRD: Bronze Inscription Restoration and Dating

**Wenjie Hua[1]** **Hoang Huu Nguyen[2]** **Gangyan Ge[3]**

[1] Wuhan University

[2] University of Illinois, Chicago

[3] Xinjiang University

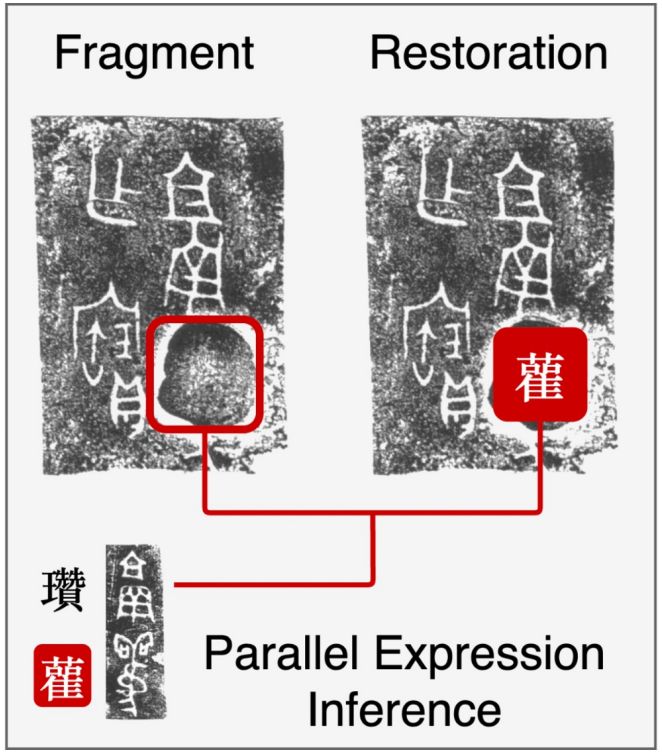



Figure a: Experts restore via *parallel expressions*

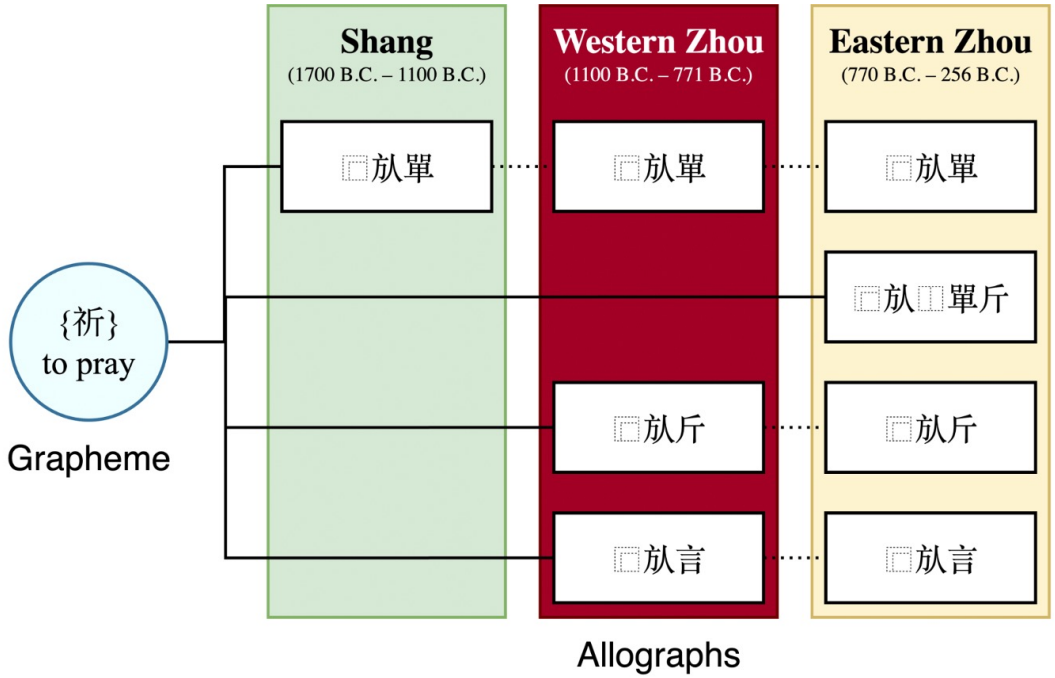
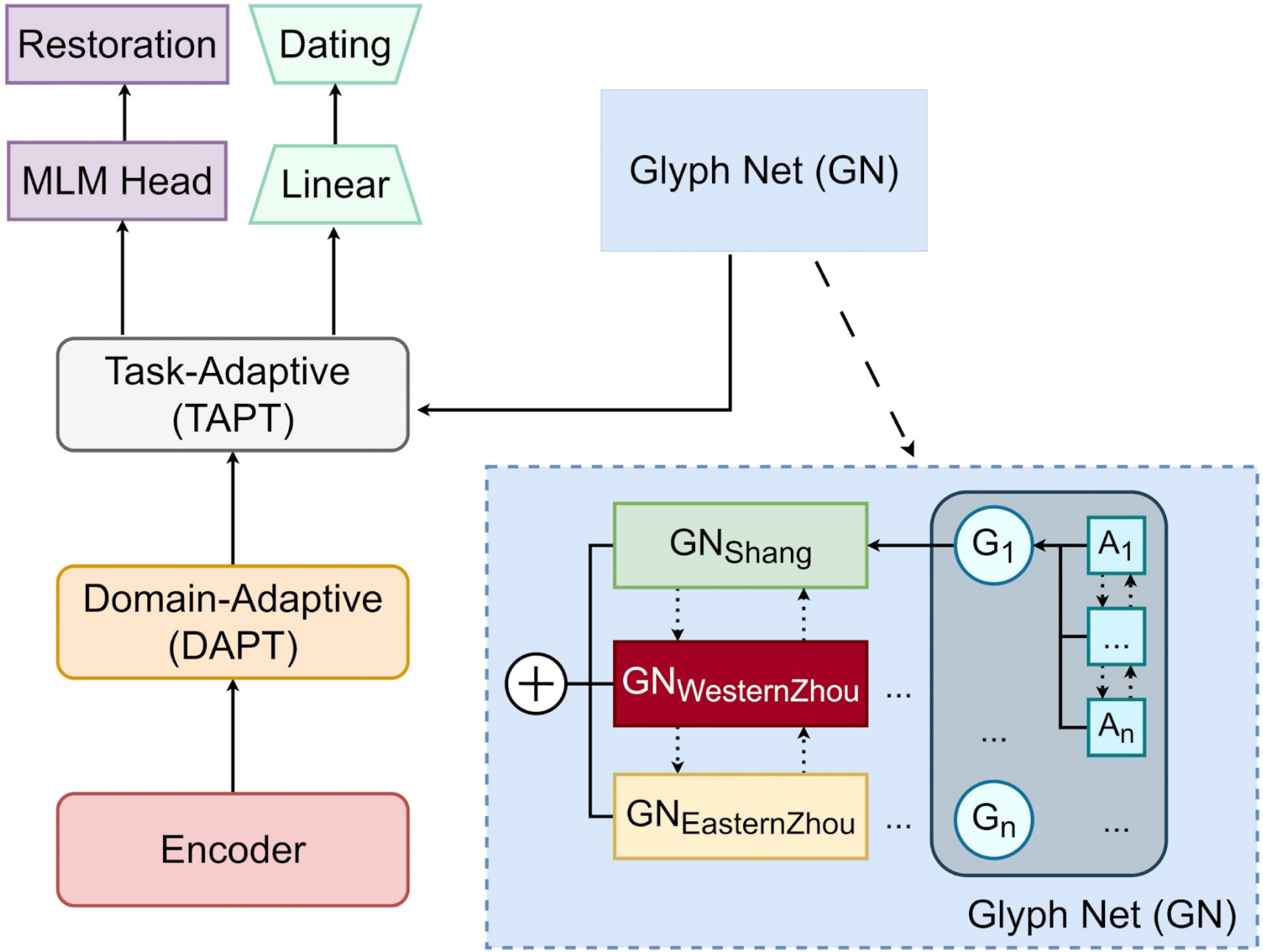



Figure b: Glyph family of a Chinese character

- **Low-resource** setting
- **Allographs** split the data into many forms

- **BIRD** 41k
- NLP-ready fully encoded bronze inscription corpus
- **Glyph Net (GN)**: grapheme-allograph & families

| Dataset | Ava. | Dedup. | Filt. | Enc. | Chron. |
|---|---|---|---|---|---|
| Jihewang | ✗ | ✗ | ✗ | Partial | ✓ |
| AS DABII | ✗ | ✗ | ✗ | Partial | ✓ |
| AS Lexicon | ✗ | ✗ | ✗ | Partial | ✓ |
| **BIRD** | ✓ | ✓ | ✓ | Full | ✓ |


Restoration
Dating
MLM Head
Linear
Glyph Net (GN)
Task-Adaptive (TAPT)
Domain-Adaptive (DAPT)
Encoder
GN_Shang
GN_WesternZhou
GN_EasternZhou
G1
An
Gn
Glyph Net (GN)

| Model | Params | E@1 | E@5 | E@10 | F@1 | F@5 | F@10 |
|---|---|---|---|---|---|---|---|
| BiLSTM | 20M | 39.02 | 42.98 | 53.10 | **57.41** | 57.63 | 62.50 |
| SikuRoBERTa | 109M | **49.47** | **65.20** | **70.15** | 54.32 | **68.05** | **73.07** |
| mBERT | 110M | 43.55 | 58.57 | 63.71 | 46.93 | 61.28 | 65.92 |
| XLM-Base | 278M | 43.51 | 58.35 | 62.94 | 44.28 | 59.49 | 64.03 |
| XLM-Large | 550M | 45.64 | 60.92 | 64.91 | 47.16 | 61.17 | 65.36 |

Table a: Restoration performance comparison of backbone models under the unified GN setting

| Model | Params | Dynasty | | Hier-Dyn | | Hier-Per | |
|---|---|---|---|---|---|---|---|
| | | Acc | F1 | Acc | F1 | Acc | F1 |
| SVM | 0.08M | 75.31 | 49.44 | 76.32 | 42.67 | 58.55 | 49.43 |
| SikuRoBERTa | 109M | **86.42** | **77.83** | **84.21** | **54.32** | **67.11** | 62.91 |
| mBERT | 110M | 84.57 | 74.77 | 82.24 | 53.12 | 63.82 | 58.63 |
| XLM-Base | 278M | 79.01 | 50.34 | 80.92 | 51.32 | 62.50 | 57.34 |
| XLM-Large | 550M | 84.01 | 74.60 | 81.58 | 53.12 | 65.13 | **62.96** |

Table b: Dating performance comparison of backbones under the unified glyph-biased sampling

| Model | Scenario | E@1 ↑ | E@5 ↑ | E@10 ↑ | F@1 ↑ | F@5 ↑ | F@10 ↑ |
|---|---|---|---|---|---|---|---|
| SikuRoBERTa | Baseline | 0.236 | 0.377 | 0.440 | 0.244 | 0.395 | 0.458 |
| | DAPT_only | 0.260 | 0.423 | 0.494 | 0.253 | 0.432 | 0.512 |
| | TAPT_Bias | 0.483 | 0.626 | 0.676 | 0.544 | 0.678 | 0.731 |
| | TAPT_GN | **0.495** | **0.652** | **0.702** | 0.543 | 0.681 | **0.731** |
| | TAPT_GN_Bias | 0.492 | 0.638 | 0.686 | **0.554** | **0.688** | 0.729 |
| | TAPT_from_DAPT | 0.485 | 0.636 | 0.685 | 0.535 | 0.681 | 0.729 |
| | TAPT_only | 0.488 | 0.639 | 0.684 | 0.539 | 0.681 | 0.723 |
| mBERT | Baseline | 0.112 | 0.224 | 0.282 | 0.093 | 0.205 | 0.267 |
| | DAPT_only | 0.148 | 0.283 | 0.353 | 0.139 | 0.278 | 0.355 |
| | TAPT_Bias | 0.427 | 0.572 | 0.622 | 0.464 | 0.617 | 0.665 |
| | TAPT_GN | **0.436** | **0.586** | **0.637** | **0.469** | 0.613 | 0.659 |
| | TAPT_GN_Bias | 0.424 | 0.583 | 0.635 | 0.466 | **0.618** | **0.665** |
| | TAPT_from_DAPT | 0.431 | 0.574 | 0.623 | 0.464 | 0.607 | 0.657 |
| | TAPT_only | 0.427 | 0.570 | 0.613 | 0.465 | 0.606 | 0.648 |
| XLM-Base | Baseline | 0.122 | 0.195 | 0.234 | 0.112 | 0.187 | 0.228 |
| | DAPT_only | 0.161 | 0.279 | 0.337 | 0.151 | 0.270 | 0.332 |
| | TAPT_Bias | 0.432 | 0.572 | 0.622 | **0.454** | 0.598 | 0.644 |
| | TAPT_GN | **0.435** | **0.584** | **0.629** | 0.443 | 0.595 | 0.640 |
| | TAPT_GN_Bias | 0.429 | 0.583 | 0.626 | **0.454** | **0.608** | **0.651** |
| | TAPT_from_DAPT | 0.434 | 0.583 | 0.629 | 0.447 | 0.595 | 0.639 |
| | TAPT_only | 0.424 | 0.557 | 0.602 | 0.434 | 0.568 | 0.614 |
| XLM-Large | Baseline | 0.140 | 0.225 | 0.265 | 0.132 | 0.208 | 0.257 |
| | DAPT_only | 0.178 | 0.321 | 0.382 | 0.166 | 0.312 | 0.384 |
| | TAPT_Bias | 0.453 | 0.595 | 0.640 | **0.479** | **0.615** | 0.656 |
| | TAPT_GN | **0.456** | **0.609** | **0.649** | 0.472 | 0.612 | 0.654 |
| | TAPT_GN_Bias | 0.454 | 0.598 | 0.645 | 0.476 | 0.609 | **0.657** |
| | TAPT_from_DAPT | 0.456 | 0.600 | 0.648 | 0.471 | 0.604 | 0.649 |
| | TAPT_only | 0.435 | 0.577 | 0.621 | 0.442 | 0.584 | 0.622 |

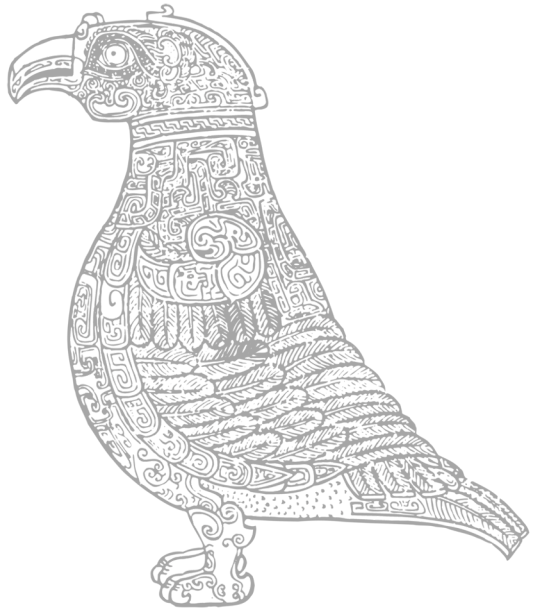
| Mask Position | Gold | Pred@1 | Top5 |
|---|---|---|---|
| 01 | 室 | 廟 | 廟 室 宮 寢 廷 |
| 02 | 王 | 王 | 王 公 君 伯 尹 |
| 03 | 芾 | 芾 | 芾 純 衡 衣 韍 |
| 05 | 命 | 於 | 於 于 揚 無 多 |
| 06 | 于 | 于 | 于 揚 穆 於 侑 |
| 07 | 年 | 年 | 年 人 世 壽 歲 |

Table a: Top-1 and -5 predictions versus gold characters

| Mask Position | Top10 Predictions |
|---|---|
| 04 | 鑾 旂 舄 筆 U+3AC3 金 矢 黃 弓 璋 |
| 08 | 介 伯 市 限 客 期 制 政 宰 人 |
| 15 | 之 外 一 若 内 賜 邑 大 下 又 |
| 16 | 賜 折 喬 杜 乘 造 擇 柞 之 于 |
| 17 | 則 許 弗 不 人 亦 也 而 帛 乃 |
| 18 | 則 曰 不 弗 許 告 厥 多 有 用 |
| 28 | 其 厥 若 越 乃 我 以 汝 如 余 |

Table b: Completions for undeciphered positions

Figure Hu Ding inscription



**A curated dataset**

**+ allograph-aware modeling**

**+ glyph-biased sampling**

**= BIRD**

Future: phonology & multimodal signals (vessel shape, motifs); broaden GN; larger generative backbones