ICCV OCT 19-23, 2025 HONOLULU HAWAII

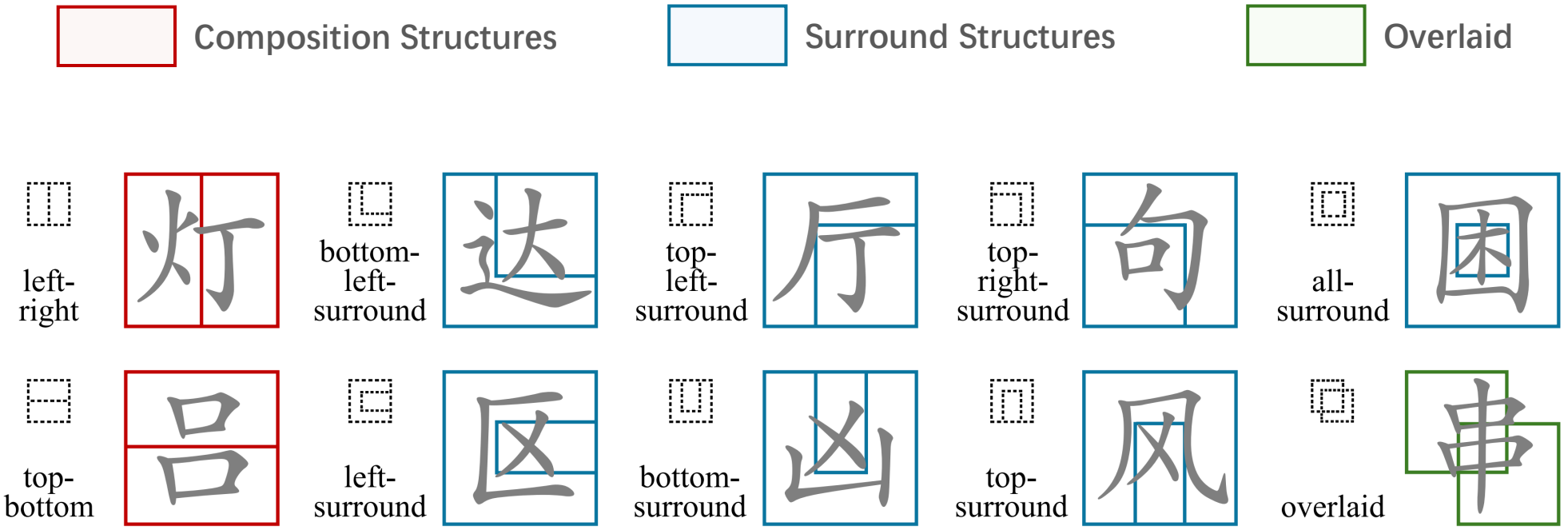
# MSA$^2$ : Multi-task Framework with Structure-Aware and Style-Adaptive Character Representation for Open-set Chinese Text Recognition

Yangfu Li, Hongjian Zhan, Qi Liu, Sun Li, Yu-jie Xiong, Yue Lu


華東師範大學
EAST CHINA NORMAL UNIVERSITY

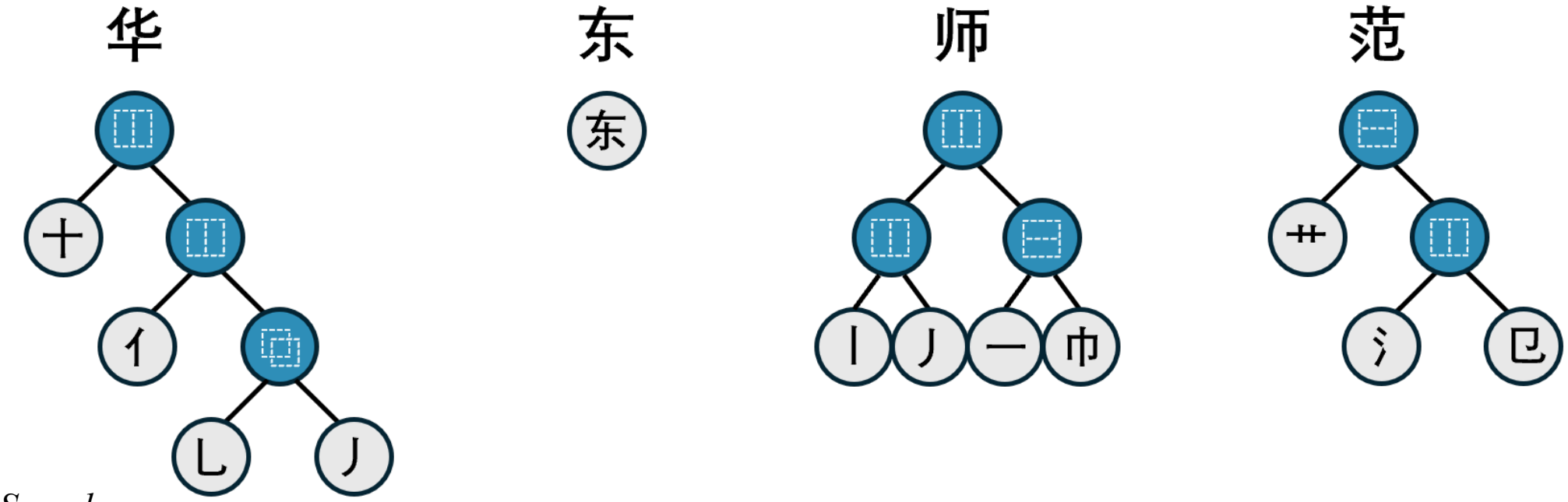
# Preliminaries

## Radical & Structure



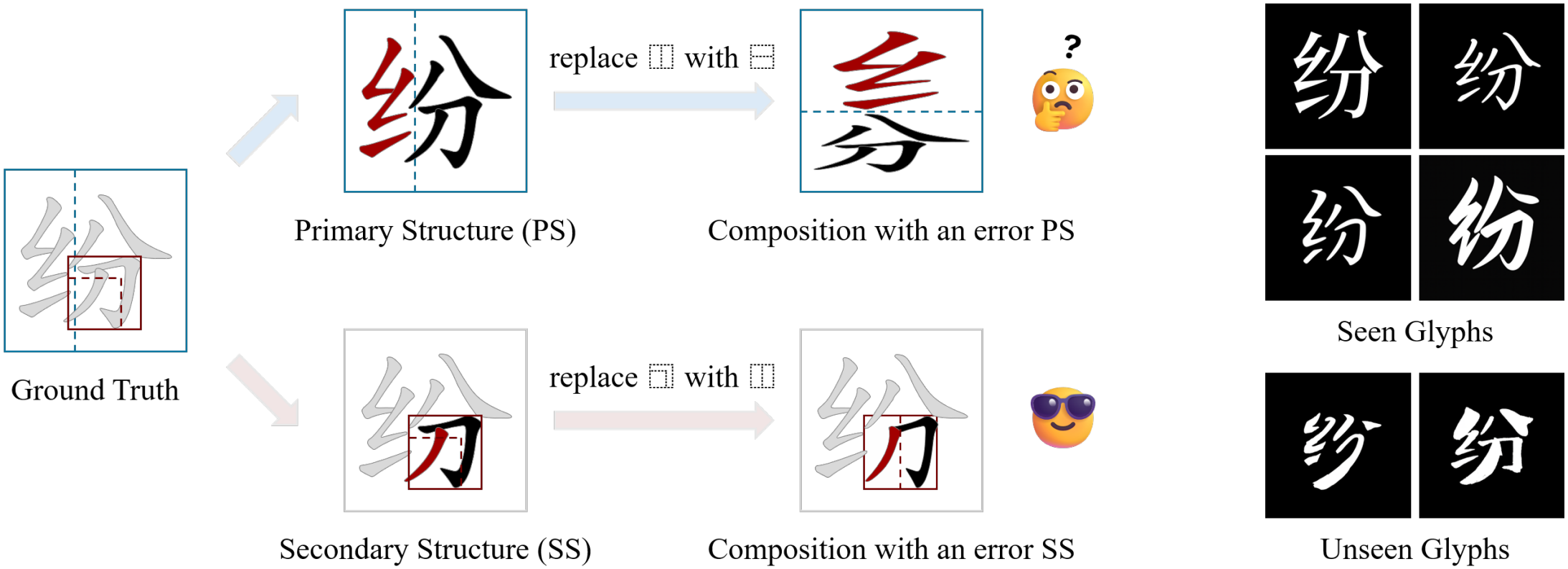
# Preliminaries

## Ideographic Description Sequence (i.e., IDS)



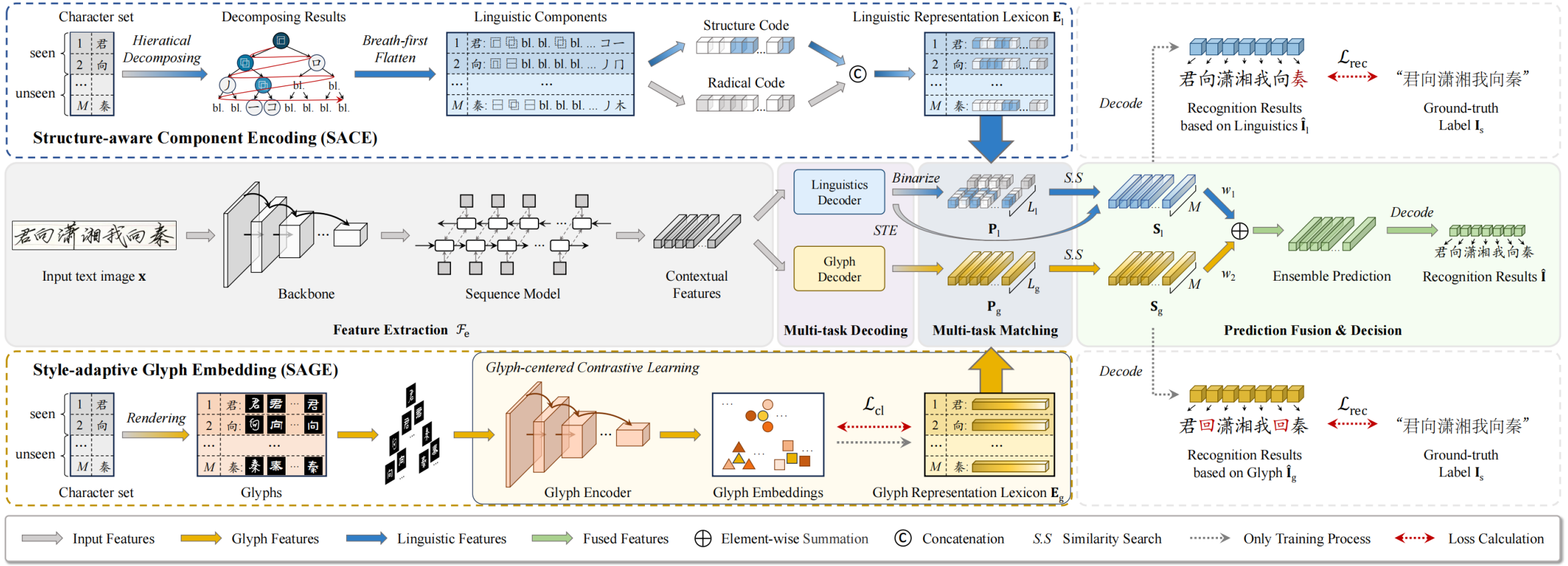
# Motivation

Recognize unseen characters like humans



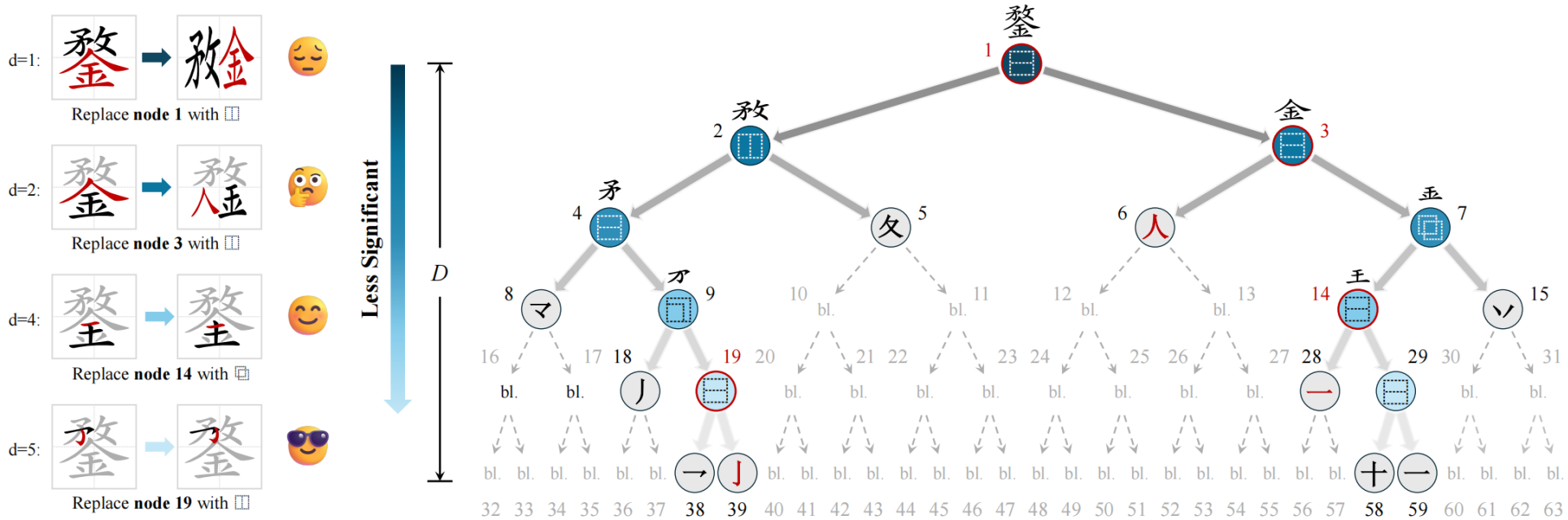
# Methods

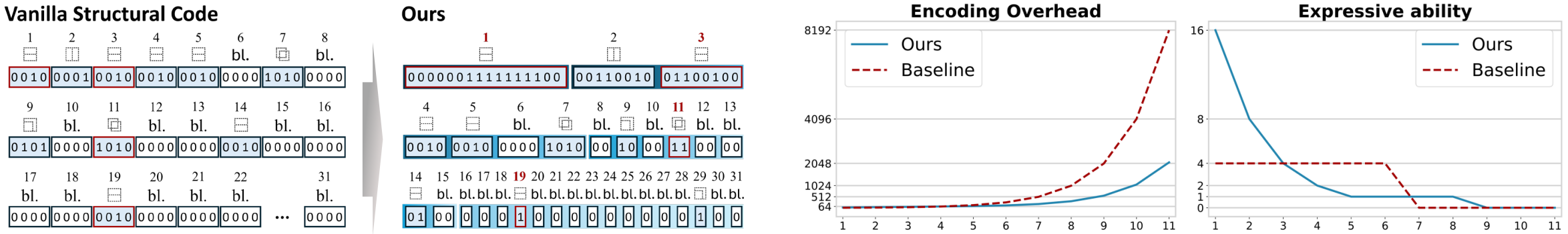
## Overview



# Methods

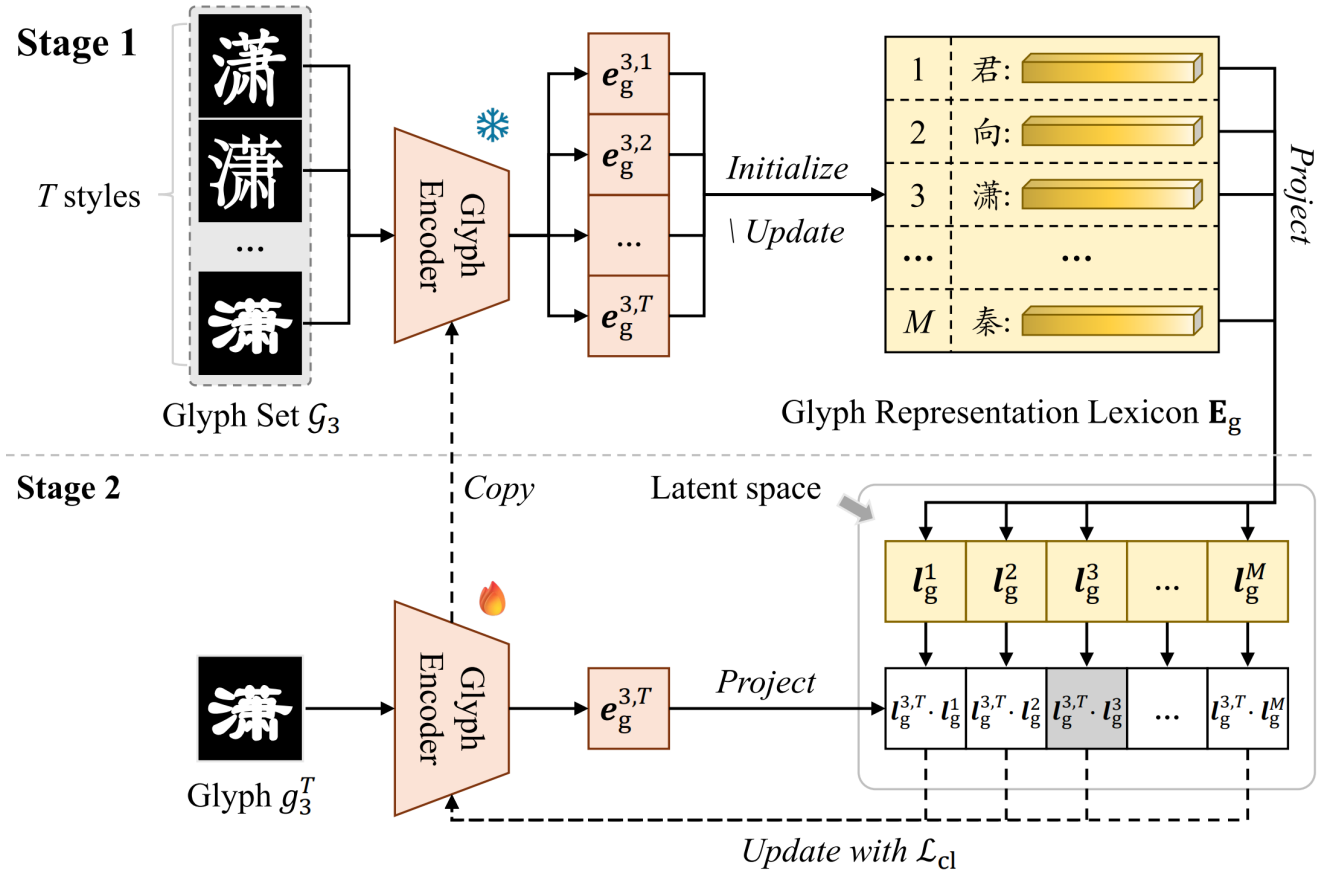
## Structure-Aware Component Encoding (SACE)



# Methods

## Style-Adaptive Glyph Embedding (SAGE)




### Style Embedding

$$\boldsymbol{e}_{\mathrm{g}}^{m,t} = \mathcal{F}_{\mathrm{g}}(g_m^t), t \in [1, 2, \ldots, T]$$

### Glyph Embedding

$$\hat{\boldsymbol{e}}_{\mathrm{g}}^{m} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{e}_{\mathrm{g}}^{m,t}$$

### Contrastive Loss

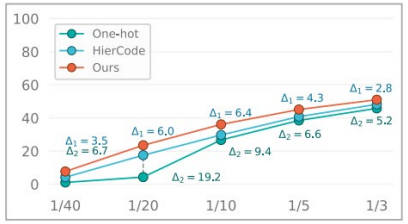
$$\mathcal{L}_{\mathrm{cl}} = -\frac{1}{MT}\sum_{i=1}^{M}\sum_{t=1}^{T} \log \frac{\exp(\boldsymbol{l}_{\mathrm{g}}^{i,t} \cdot \boldsymbol{l}_{\mathrm{g}}^{i})}{\sum_{j=1,j\neq i}^{M} \exp(\boldsymbol{l}_{\mathrm{g}}^{i,t} \cdot \boldsymbol{l}_{\mathrm{g}}^{j})}$$

# Main Results

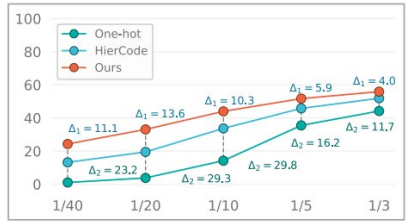
## Text Recognition

| Methods | Venue | Scene | Web | Doc | Handw | Avg |
|---|---|---|---|---|---|---|
| CRNN [27] | PAMI'16 | 53.4 | 54.5 | 97.5 | 46.4 | 67.0 |
| ASTER [28] | PAMI'18 | 54.6 | 52.3 | 93.1 | 38.9 | 64.7 |
| MORAN [20] | PR'19 | 51.7 | 49.5 | 95.4 | 39.6 | 64.3 |
| SAR [10] | AAAI'19 | 62.5 | 54.1 | 94.2 | 33.7 | 67.3 |
| SRN [40] | CVPR'20 | 60.1 | 52.3 | 96.7 | 18.0 | 65.0 |
| SEED [25] | CVPR'20 | 49.6 | 46.3 | 93.7 | 32.1 | 61.2 |
| MASTER [19] | PR'21 | 62.8 | 52.1 | 84.4 | 26.9 | 56.6 |
| TransOCR [2] | CVPR'21 | 63.3 | 62.3 | 96.9 | 53.4 | 72.8 |
| ABINet [7] | CVPR'21 | 64.4 | 67.4 | 97.2 | 54.8 | 74.1 |
| SVTR-B [6] | IJCAI'22 | 71.7 | 73.8 | 98.2 | 52.2 | 75.2 |
| SVTR-L [6] | | 72.2 | 74.1 | 98.1 | 53.6 | 75.5 |
| CCR-CLIP [42] | ICCV'23 | 71.3 | 69.2 | 98.3 | 60.3 | 75.8 |
| MSA$^{2,\dagger}$ | - | **73.7** | **76.5** | **99.2** | **61.5** | **77.1** |
| $\Delta$ | - | +1.5 | +2.4 | +0.9 | +1.2 | +1.3 |
| One-hot | - | 60.3 | 60.2 | 92.8 | 54.1 | 70.0 |
| HierCode [48] | PR'25 | 63.7 | 66.2 | 98.2 | 56.3 | 74.2 |
| MSA$^2$ | - | **65.9** | **69.4** | **98.7** | **59.2** | **75.4** |
| $\Delta_1$ | - | +2.2 | +3.2 | +0.5 | +2.9 | +1.2 |
| $\Delta_2$ | | +5.6 | +9.2 | +5.9 | +5.1 | +5.4 |

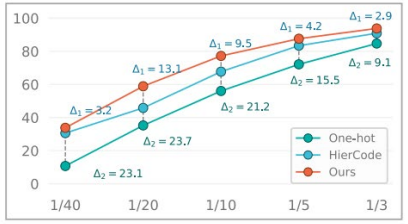
$^\dagger$ applies the same configuration of backbone as [42] for fair comparisons.



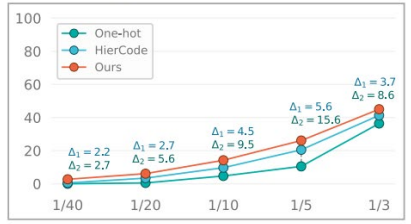

(a) Scene Dataset




(b) Web Dataset




(c) Document Dataset




(d) Handwriting Dataset

| Char | Methods | ICDAR | BCTR | | | |
|---|---|---|---|---|---|---|
| | | Line | Scene | Web | Doc | Handw |
| Ch. | One-hot | 93.35 | 82.09 | 79.57 | 98.64 | 91.65 |
| | HierCode [48] | 94.53 | 83.41 | 83.39 | 99.71 | 92.35 |
| | MSA$^2$ | **95.54** | **85.85** | **85.64** | **99.73** | **94.38** |
| | $\Delta_1$ | +1.01 | +2.44 | +2.25 | +0.02 | +2.03 |
| | $\Delta_2$ | +2.19 | +3.76 | +6.07 | +1.09 | +2.73 |
| NCh. | One-hot | 85.56 | 90.24 | 84.67 | 99.37 | 86.59 |
| | HierCode [48] | 85.65 | 90.27 | 85.19 | 99.54 | 86.61 |
| | MSA$^2$ | **87.73** | **92.78** | **88.02** | **99.79** | **89.22** |
| | $\Delta_1$ | +2.08 | +2.51 | +2.83 | +0.25 | +2.61 |
| | $\Delta_2$ | +2.17 | +2.54 | +3.35 | +0.42 | +2.63 |

# Main Results

## Character Recognition

| Methods | Venue | Representation | | Handwritten/% ($m$ for classes) | | | | | | Scene/% ($m$ for classes) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Linguistics | Glyph | 500 | 1000 | 1500 | 2000 | 2755 | Full | 500 | 1000 | 1500 | 2000 | 3150 |
| DenseRAN [32] | ICFHR'18 | ✓ | | 1.70 | 8.44 | 14.71 | 19.51 | 30.68 | 96.66 | 0.15 | 0.54 | 1.60 | 1.95 | 5.39 |
| HDE [1] | PR'20 | ✓ | | 4.90 | 12.77 | 19.25 | 25.13 | 33.49 | 97.14 | 0.82 | 2.11 | 3.11 | 6.96 | 7.75 |
| Chen *et al.* [3] | IJCAI'21 | ✓ | ✓ | 5.60 | 13.85 | 22.88 | 25.73 | 37.91 | 96.73 | 1.54 | 2.54 | 4.32 | 6.82 | 8.61 |
| CUE [21] | PR'23 | ✓ | | 7.43 | 15.75 | 24.01 | 27.04 | 40.55 | 96.96 | - | - | - | - | - |
| SideNet [13] | PR'24 | ✓ | | 5.10 | 16.20 | 33.80 | 44.10 | 50.30 | - | - | - | - | - | - |
| HierCode [48] | PR'25 | ✓ | | 6.22 | 20.71 | 35.39 | 45.67 | 56.21 | 97.68 | 1.67 | 2.59 | 4.54 | 7.02 | 9.13 |
| MSA$^2$ | - | ✓ | ✓ | **8.24** | **26.13** | **40.67** | **51.44** | **60.17** | **98.85** | **2.05** | **3.11** | **4.98** | **7.65** | **9.68** |
| $\Delta$ | - | - | | +2.02 | +5.96 | +5.28 | +5.73 | +3.96 | +1.17 | +0.38 | +0.52 | +0.44 | +0.63 | +0.55 |

# Ablations

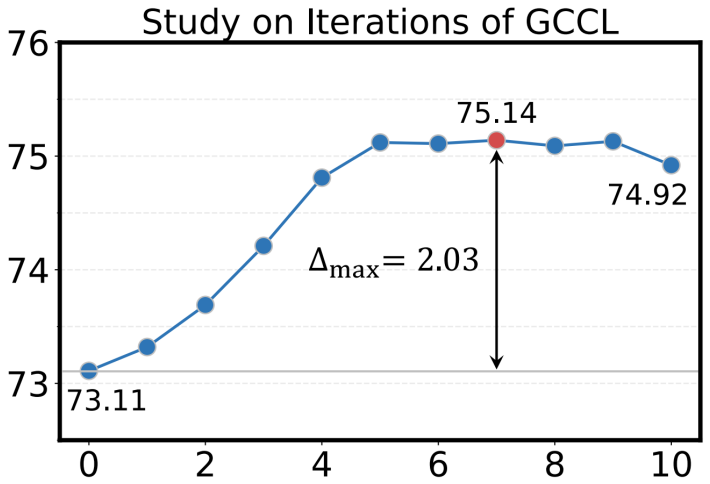
## Recognition Task

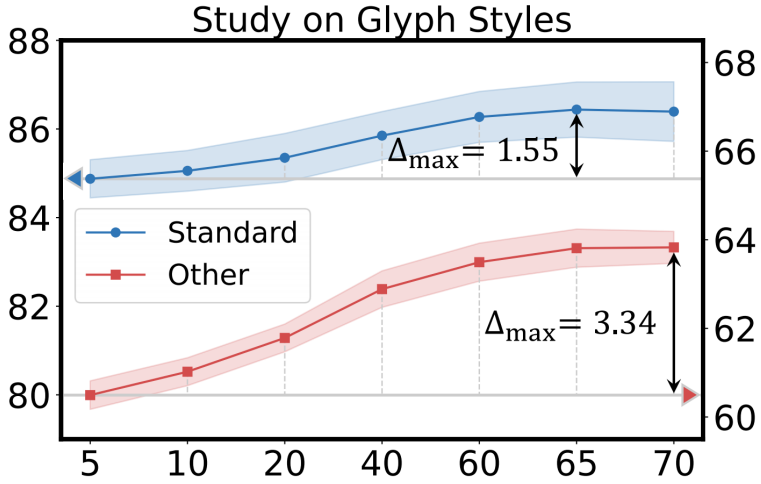
| Char | Linguistics | Glyph | Standard | Other |
|---|---|---|---|---|
| Compound Characters | ✓ | | 91.24 | 88.38 |
| | | ✓ | 90.95 | 87.93 |
| | ✓ | ✓ | **91.95** | **89.82** |
| Basic Characters | ✓ | | 92.11 | 88.57 |
| | | ✓ | 93.25 | 89.26 |
| | ✓ | ✓ | **93.43** | **90.41** |

## Structure Code

| $L_{S_1}$ | $L_{S_2}$ | $L_{S_3}$ | $L_{S_4}$ | $L_{S_5}$ | $L_S$ | AR | RR |
|---|---|---|---|---|---|---|---|
| Random Structure Code | | | | | 124 | 94.03 | 92.42 |
| 4 | - | - | - | - | 124 | 96.47 | 94.23 |
| 8 | - | - | - | - | 128 | 96.60 | 94.41 |
| 16 | 4 | - | - | - | 136 | 96.78 | 94.64 |
| 16 | 8 | 4 | 4 | - | 144 | 97.03 | 94.86 |
| 16 | 16 | - | - | - | 160 | 96.94 | 94.76 |
| 16 | 8 | 8 | - | - | 160 | 96.96 | 94.79 |
| 16 | 8 | 2 | - | - | 136 | 96.82 | 94.45 |
| 16 | 8 | - | 2 | 4 | 128 | 97.12 | 94.89 |
| 16 | 8 | - | 1 | - | 120 | 97.08 | 94.84 |
| 16 | 8 | - | 2 | 2 | 96 | 97.24 | 95.05 |
| 16 | 8 | - | 2 | 1 | 80 | **97.35** | **95.12** |

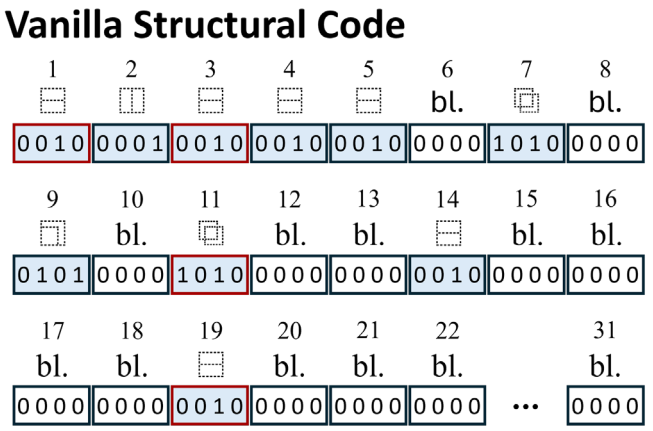
## GCCL

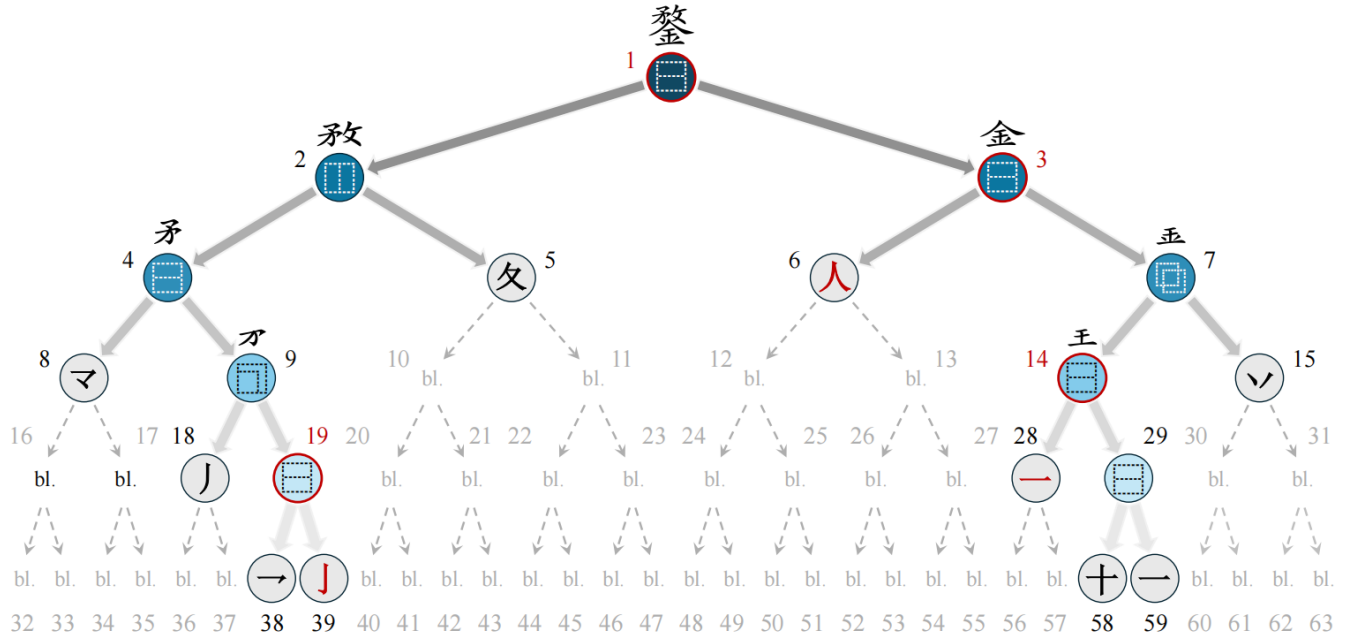


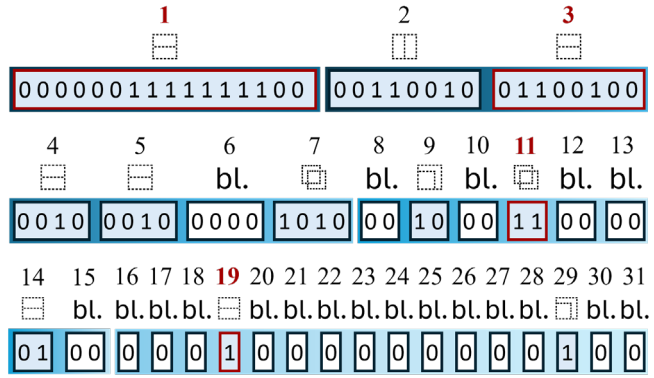



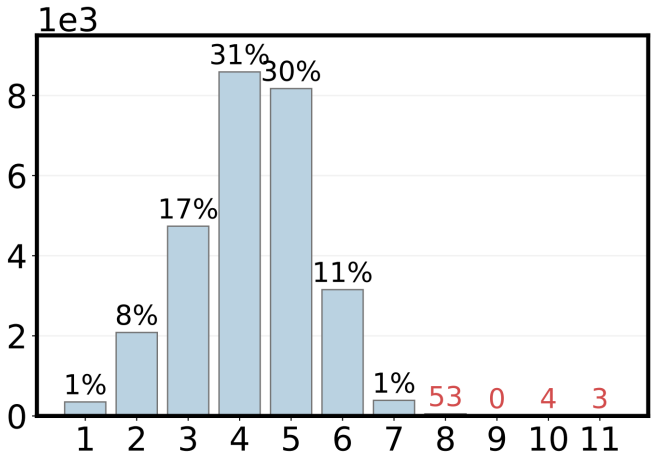
# In-depth Analysis

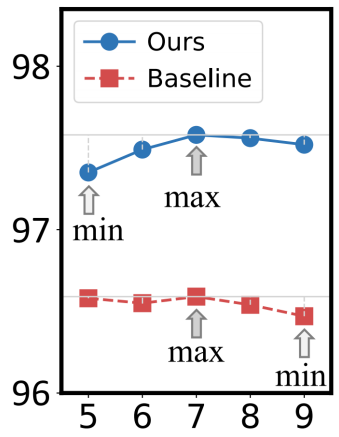
## How does SACE benefit CTR?

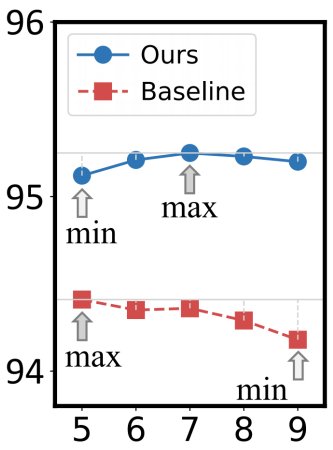




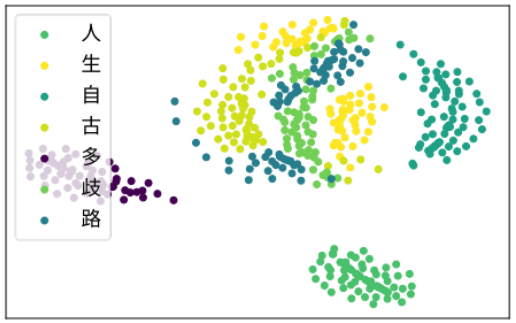



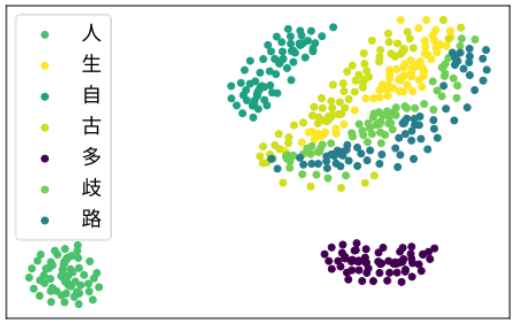



# In-depth Analysis

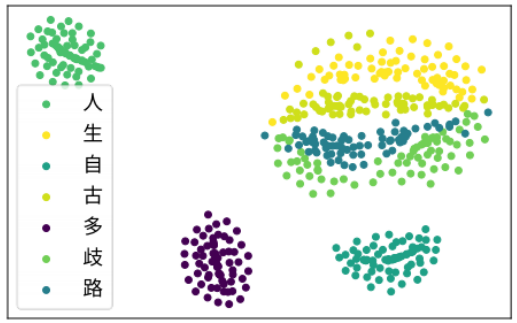
## How does SAGE benefit CTR?



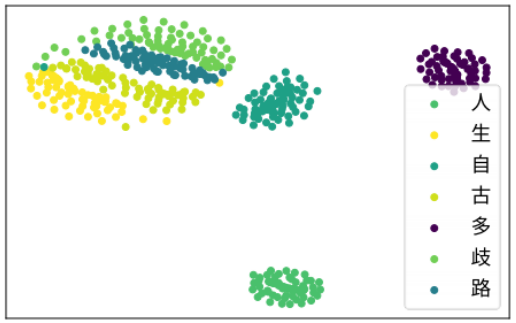

w/o Iteration



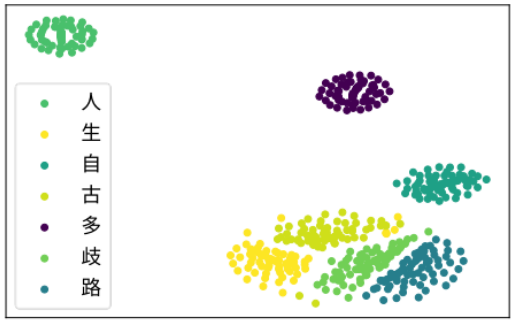

Iteration = 1



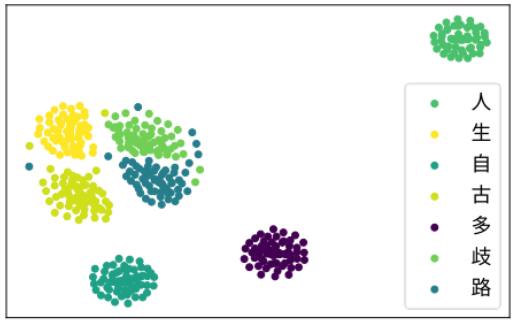

Iteration = 2



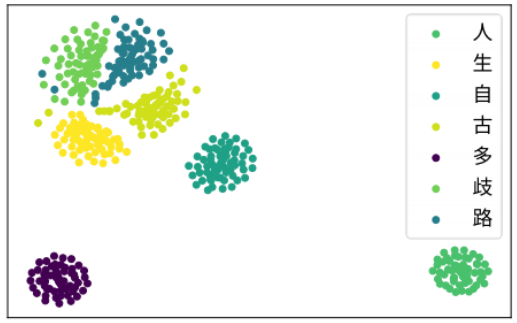

Iteration = 3



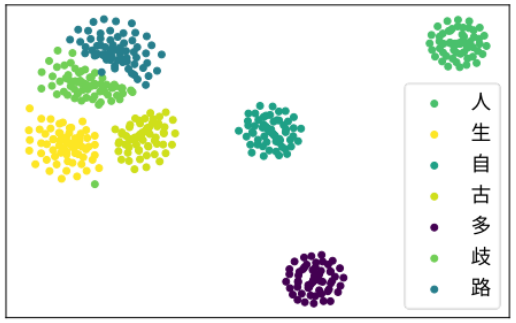

Iteration = 4




Iteration = 5




Iteration = 6




Iteration = 7

# In-depth Analysis

## Key takeaways

- Structure-aware coding reduces linguistic redundancy.
- Style-adaptive glyph learning generalizes across fonts.
- Linguistic × glyph fusion improves open-set character recognition.

## Limitations

- Relies on linguistic components (e.g., IDS).
- Does not model connections among same-type structures across varying depths.

ICCV HONOLULU HAWAII
OCT 19-23, 2025

**MSA$^2$ : Multi-task Framework with Structure-Aware and Style-Adaptive Character Representation for Open-set Chinese Text Recognition**

# Thank you !

*Contact with us: yfli_cee@stu.ecnu.edu.cn*

華東師範大學
EAST CHINA NORMAL UNIVERSITY

More Details