# Hot-Start from Pixels: Low-Resolution Visual Tokens for Chinese Language Modeling

**Anonymous ACL submission**

## Abstract

Large language models typically represent Chinese characters as discrete index-based tokens, largely ignoring their visual form. For logographic scripts, visual structure carries semantic and phonetic information, which may aid prediction. We investigate whether low-resolution visual inputs can serve as an alternative for character-level modeling. Instead of token IDs, our decoder receives grayscale images of individual characters, with resolutions as low as $8 \times 8$ pixels. Remarkably, these inputs achieve 39.2% accuracy, comparable to the index-based baseline of 39.1%. Such low-resource settings also exhibit a pronounced *hot-start* effect: by 0.4% of total training, accuracy reaches above 12%, while index-based models lag at below 6%. Overall, our results demonstrate that minimal visual structure can provide a robust and efficient signal for Chinese language modeling, offering an alternative perspective on character representation that complements traditional index-based approaches.

## 1 Introduction

In Chinese, meaning arises not only through sequential context but also through the internal structure of characters. We human readers naturally pay attention to radicals, stroke layout, and overall shape when reading Chinese.

In contrast, the majority of mainstream models today process Chinese through sequences of symbolic character IDs, whereas English is typically tokenized into subword indices (Brown et al., 2020; Bommasani et al., 2021; Bender and Koller, 2020; Rust et al., 2021). While this index-based abstraction is effective for alphabetic writing systems, it may be suboptimal for Chinese.

This omission can have concrete consequences. Take the character 山 (  'mountain') as an example: its shape resembles small mountain peaks, thereby immediately conveying meaning to a human reader. The contrast between human perception and index-based modeling is clear here: while humans can directly leverage visual form, 山 is represented in an index-based model as an abstract token ID, stripped of its shape. Similarly, characters such as 灭 (extinguish) and 火 (fire) differ clearly in their visual structure, yet these differences are not readily accessible to the model at early stages of learning. It is akin to assembling a jigsaw puzzle with the image erased: the components remain, but the visual cues that facilitate interpretation are missing.

As one of the most representative logographic systems, Chinese treats visual form not as auxiliary, but as an essential part of meaning construction, carrying semantic and phonetic information through its visual structure. Such visual intuition extends beyond isolated characters to contextual prediction. In weak linguistic contexts, human readers naturally rely on glyph patterns to predict the following text. This suggests that visual form may serve as a more fundamental cue for logographic language processing, especially when semantic signals are sparse.

This motivates us to ask: *can language models effectively process Chinese characters—or more generally, visually structured information—instead of index-based tokens?* If visual structure indeed plays such a central role for humans, then abstracting Chinese characters into index-based tokens may not merely be an engineering choice, but a fundamental modeling assumption.

This leads to a fundamental choice in representation: index-based tokenization that relies on contextual embeddings alone, or vision-based processing that extracts character shape and structure through a visual encoder (Poznanski et al., 2025). These two paths frame our experimental comparison.

This architectural distinction has implications for representation topology. Index-based embeddings initialize as unstructured points in feature space, requiring the model to discover character

relationships purely from co-occurrence statistics. In contrast, visual encoders provide built-in geometrically meaningful spatial organization from the outset, potentially offering a structural prior that accelerates early learning. Importantly, such representations also make model behavior more interpretable, as predictions can be traced back to salient regions and structural patterns in the image.

This question is intriguing given recent advances in applying VLMs to optical character recognition (OCR) and document understanding tasks. Systems like DeepSeek-OCR (Wei et al., 2025) and Pix2Struct (Lee et al., 2023) demonstrate that textual content can be processed exclusively as images. However, our work differs fundamentally from these transcription-focused models: rather than transcribing visual glyphs back to symbols, we ask: can visual forms alone support linguistic prediction—in particular, predicting the next character based on visual context? This shifts the focus from recognition to language modeling itself, from symbol reconstruction to linguistic reasoning.

Our preliminary experiments would reveal that even heavily cropped or low-resolution character images retain sufficient structural information for prediction. For instance, low-resolution or partially cropped character images still allow models to identify the correct next character, reminiscent of how humans can still read small or degraded characters. Similar observations have been reported in recent studies (Li et al., 2025; Poznanski et al., 2025; Wei et al., 2024). These early observations motivate a systematic study of the sufficiency of visual forms in language modeling.

Exploring this question brings significant practical and scientific value. In resource-constrained environments, models that use visual structure can extract meaningful patterns from limited data, potentially achieving faster convergence than index-based alternatives. These considerations motivate a systematic investigation across multiple dimensions: visual sufficiency, early-stage learning dynamics, resolution sensitivity, and spatial robustness.

In this work, we systematically investigate a vision-token-based formulation of Chinese language modeling, processing low-resolution grayscale character images through a lightweight visual encoder fed into a standard language decoder architecture. Our findings reveal that visual form alone brings strong predictive power, providing a structural foundation that accelerates early stage learning, even in low-resource settings.

**Research Questions.** Our evaluation addresses the following research questions:
**RQ1 (Visual Sufficiency)**: Can visual inputs of Chinese characters alone suffice for character-level prediction?
**RQ2 (Early-Stage Dynamics)**: What are the learning trajectories of vision-token models?
**RQ3 (Resolution Sensitivity)**: How does predictive performance vary as image resolution decreases from high-fidelity to near-minimal levels?
**RQ4 (Spatial Robustness)**: Can vision-token models maintain accuracy when only partial character regions are visible?

**Main Contributions.** Our main contributions are sixfold, structured around our research questions:

- **For Methodology**, we propose a vision-token formulation for Chinese language modeling, replacing completely index-based tokens with character images processed through a visual encoder.
- **For RQ1**, we demonstrate that visual inputs alone achieve accuracy comparable to index-based baselines (39.2% vs. 39.1%), confirming that purely visual structure suffices for character-level prediction.
- **For RQ2**, we identify a *hot-start* effect: at only 0.4% of total training regimen, visual models reach 12.3% accuracy—more than double the baseline's 5.8%.
- **For RQ3 & RQ4**, we establish robustness under visual degradation: performance remains stable across resolutions from $8 \times 8$ to $96 \times 96$ pixels and under severe spatial cropping.
- **For Explainable NLP**, we analyze how visual tokens offer inherent interpretability: embedding spaces organize by morphological similarity, and gradient analysis traces predictions to salient pixel regions.

## 2 Methodology

### 2.1 Visual-Language Model Architecture

Figure 1 illustrates our processing pipelines for a concrete example—predicting the final character in "数据显示" (data shows) using a model checkpoint of the early-stage training.

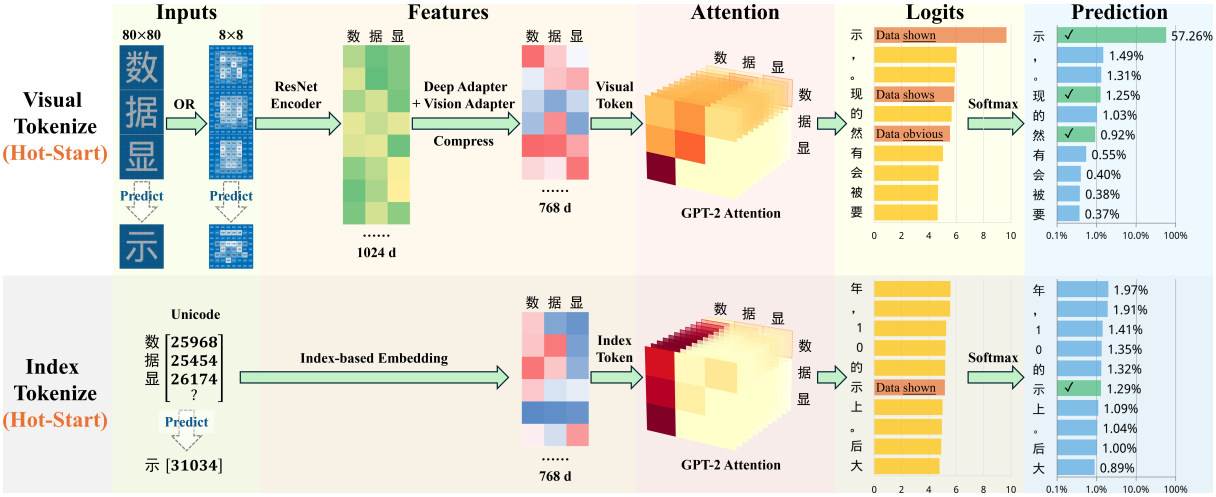



Figure 1: Model architecture and concrete processing example showing the prediction of the final character in the phrase "数据显示" (The data shown); numerical results demonstrate the *hot-start* phenomenon of visual tokens by 0.5% of the total training progress. Top: Visual-based training pipeline; numerical results are based on inputs of $8 \times 8$ character images. Bottom: Index-based training pipeline.

The diagram gives the two fundamentally different input paths while sharing the same language decoder–GPT-2-small-style (Radford et al., 2019) in our experiment, with $\sim 117M$ parameters. In the index-based path (bottom), each character is represented by a discrete token index. In the visual-based path (top), characters are first rendered as low-resolution grayscale images and passed through a lightweight visual encoder before being fed into the same decoder. Note that the figure visualizes the *hot-start* effect: after only 0.5% of training, the visual-based model already assigns much higher probability to the correct next character than its index-based counterpart and gives linguistically more plausible candidate rankings.

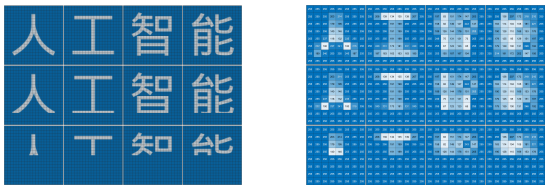
In particular, in the visual-based paths, visual inputs are passed through a ResNet encoder (He et al., 2016) and a Vision Adapter (Wu et al., 2019) before reaching the decoder embedding space, while index-based inputs are directly mapped into the decoder embeddings.

**Remark of Reverse OCR concern.** Our approach differs fundamentally from reversing an OCR (transcribing images back to symbols). Evidence against mere symbol reconstruction is: (1) the *hot-start* effect where visual models outperform index-based baselines early in training with minimal data; (2) the ability to discriminate subtle glyph differences; and (3) robustness under partial cropping. These patterns all suggest that the model learns structural understanding of character shapes rather than performing symbol mapping (Wu et al., 2019; Geirhos et al., 2020).

Chinese characters differ fundamentally from alphabetic systems: each character is a minimal unit where visual patterns carry information through local details, compositional elements, and global shape. Capturing these visual regularities provides a structural prior that facilitates more efficient learning in Chinese language modeling.

To operationalize this approach, we render each character as a grayscale image. In the *Vision-100%* mode, characters occupy approximately 80% of the image width and height. For example, at $8 \times 8$ resolution, the character occupies about $6.4 \times 6.4$ active pixels, leaving 10% border margins on each side. We intentionally include these border margins to simulate realistic reading conditions—mimicking how characters appear within documents with natural spacing.

Cropping removes portions while keeping resolution fixed; for instance, in *Vision-50%*, the top 50% of original pixels are retained, with the remainder filled with background. This design is motivated by an intuitive observation: humans can often recognize partially visible Chinese characters. Our cropping experiments test whether models similarly extract predictive signals from limited visual regions. Figure 2 illustrates "人工智能" at $80 \times 80$ and $8 \times 8$ pixels, showing full characters, top 80%, and top 50% crops.



(a) 80 × 80 pixels (b) 8 × 8 pixels

Figure 2: Heatmap visualization of character example cropping at two resolutions: low and high.

### 2.2 Training Objective

Models are trained to predict the next character conditioned on preceding inputs, minimizing the standard cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\log P(c_{t+1}^{(i)}|I_1^{(i)},\ldots,I_t^{(i)}), \tag{1}$$

where $N$ is the batch size, $T$ is the fixed sequence length, $c_{t+1}^{(i)}$ is the ground-truth character at position $t+1$ in sequence $i$, and $I_t^{(i)}$ denotes input representations (visual or index embeddings). We train models on the THUCNews dataset, a large-scale Chinese news corpus covering multiple domains, split into fixed-length sequences. Optimization employs AdamW with a fixed learning rate and batch size. For visual inputs in particular, gradients propagate through the projection and adapter into the shared decoder.

We evaluate three main input configurations: the *Index-based Model* baseline, which inputs token IDs; *Vision-100%* mode in which characters are represented as full images; and *Cropped-Vision* mode, where partial crops retain the top 80% (*Vision-80%*) or top 50% (*Vision-50%*) of the input images. Experiments across different image sizes, especially at low resolutions ($8\times8$ and $4\times4$), assess whether minimal visual information suffices for accurate prediction, and partial cropping examines reliance on distributed visual features rather than OCR-like reconstruction.

## 3 Experiments and Results

### 3.1 Experimental Setup

**Model Configurations.** We evaluate three main input configurations: the *Index-based Model* baseline (token IDs), *Vision-100%* mode (full images), and *Cropped-Vision* mode with partial crops (top 80% and 50%).

Experiments span resolutions from $4\times4$ to $80\times80$ pixels, with particular focus on low resolutions ($8\times8$ and $4\times4$) to test visual sufficiency. Partial cropping examines reliance on distributed visual features rather than exact glyph reconstruction.

**Dataset and Training.** We train models on the THUCNews dataset, based on historical data collected from the RSS subscription channels of Sina News between 2005 and 2011 (Guo et al., 2016). After filtering and cleaning, it consists of approximately 740,000 news articles in UTF-8 encoded plain text format. The corpus contains 100K sequences, which is 12.8M character instances, split into fixed-length sequences of 128 characters. Sequences consist primarily of Chinese characters with occasional English letters and punctuation (approximately 10%), reflecting real news text composition. We employ a quadratic curriculum learning strategy, where the number of training sequences grows as $5000 + 918.37\text{epoch} + 18.74\text{epoch}^2$. Under this curriculum, the first 100K sequences are seen a total of approximately 2.13M training instances across epochs. The dataset increases progressively while evaluating on a fixed validation set of 5K sequences. This approach balances early fast convergence with later exposure to the full dataset.

**Key Parameters.** The language decoder follows a GPT-2-small-style architecture with 12 layers, 768 hidden dimensions, pre-trained on UER (Zhao et al., 2019). In particular, optimization uses AdamW with learning rate $2\times10^{-4}$ (OneCycle scheduler, max $1.5\times10^{-3}$), batch size 128, weight decay 0.01, gradient clipping at 1.0, mixed precision (FP16), and early stopping (patience: 7 epochs).

**Resolution Spectrum.** Experiments span resolutions from $4\times4$ to $96\times96$ pixels, assessing whether minimal visual information suffices for accurate prediction. In particular, we include extreme low ($4\times4$, almost no human-recognizable cues), typical low ($8\times8$, essential structure retained), intermediate ($20\times20$ to $40\times40$, recognizable shapes), and high ($80\times80$ to $96\times96$, all details preserved) resolutions. This spectrum allows us to examine how much visual information is necessary for Chinese character modeling.

### 3.2 RQ1: Visual Sufficiency

RQ1 asks whether visual inputs alone suffice for character-level prediction. Note that Chinese contains over 5,500 distinct characters: randomly guessing yields only about 0.02% accuracy (1/5,500), while a unigram baseline—predicting the most frequent character according to dataset

statistics—achieves roughly 2%. In this context, our model's performance of 39% indicates that it captures substantial linguistic structure beyond simple frequency statistics.

Table 1 (first row, *Vision-100%* mode) provides the answer: even $8 \times 8$ inputs achieve 39.21% accuracy, matching the index-based baseline (39.10%). This confirms that minimal visual information suffices for accurate prediction.

### 3.3 RQ3: Resolution Sensitivity

RQ3 investigates how performance varies with resolution. Table 1 shows results across the resolution spectrum: from $4 \times 4$ (29.70%) to $80 \times 80$ (39.03%), with $8 \times 8$ achieving 39.21%—comparable to the index-based baseline (39.10%). This confirms that minimal resolution suffices once essential structure is preserved.

### 3.4 RQ4: Spatial Robustness

RQ4 examines robustness to spatial cropping. Table 1 shows that severe cropping causes minimal performance drops: at $8 \times 8$, *Vision-80%* (top 80%) achieves 39.18% and *Vision-50%* (top 50%) 38.63%.

Analysis of $8 \times 8$ input images reveals sparse pixel usage: *Vision-100%* uses $6 \times 6$ active pixels, *Vision-80%* $6 \times 5$, and *Vision-50%* $6 \times 3$. This explains the robustness—models extract essential structure even when peripheral regions are missing.

Figure 3 illustrates the "toast-center" effect: the central strokes (Toast-Center / Crumb) contain rich caracter information, while the outer layer (Crust) contributes less. Besides, blank space carries negligible information. This concentrated central structure might explain why models can maintain accuracy even under severe cropping.

While $80 \times 80$ images preserve all character details, their predictive advantage over $8 \times 8$ is statistically marginal. To quantify this, we computed adjusted standard errors accounting for sequence-level correlations (DEFF = 19.9, $\rho = 0.15$), yielding SE = 0.27–0.54% across accuracy levels. The small differences between high and low resolutions confirm that coarse structural cues—not fine-grained details—drive predictive performance, further supporting RQ1's sufficiency claim.

### 3.5 RQ2: Early-Stage Dynamics

RQ2 examines how vision-token models behave during initial training compared to index-based models.

| Mode | 4×4 Acc/PPL | 8×8 Acc/PPL | 20×20 Acc/PPL | 30×30 Acc/PPL | 80×80 Acc/PPL |
|---|---|---|---|---|---|
| *Vision-100%* | 29.70/85.33 | 39.21/46.59 | 39.16/45.83 | 39.14/48.73 | 39.03/49.41 |
| *Vision-80%* | 18.28/194.98 | 39.18/46.23 | 39.15/46.33 | 39.07/48.83 | 39.08/48.74 |
| *Vision-50%* | 2.10/2249.29 | 38.63/47.95 | 38.70/48.04 | 38.66/49.81 | 38.57/50.33 |
| *Index-based* | | 39.10/47.58 | | | |

Table 1: Accuracy (%) / PPL across resolutions. *Vision-100%*: full images; *Vision-80%*: top 80% crop; *Vision-50%*: top 50% crop. *Index-based Model*: baseline using discrete character indices.

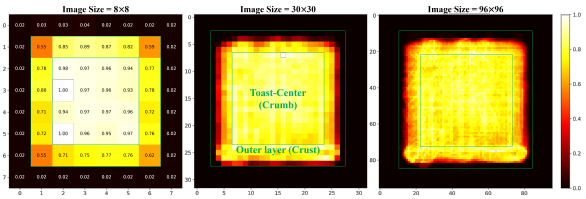



Figure 3: "toast-center effect": center strokes (blue box) receive more attention than outer pixels (red box).

As shown in Table 2, even with 4,096 training sequences, visual models show substantial gains: with $40 \times 40$ pixel inputs, accuracy reaches 13.06%, tripling the index-based baseline's 4.30%.

| Sample Trained | Baseline | 8×8 Vision | 40×40 Vision |
|---|---|---|---|
| 4,096 | 4.30% | 4.19%∼ (-0.11%) | 13.06%∼(+8.76%) |
| 6,152 | 4.61% | 5.57%∼(+0.96%) | 14.7%∼(+10.09%) |
| 8,200 | 5.84% | 12.34%∼(+6.5%) | 15.46%∼(+9.62%) |
| 10,248 | 8.45% | 13.94%∼(+5.49%) | 15.92%∼(+7.47%) |

Table 2: Hot-start progression across training stages. Higher resolutions show earlier and stronger advantages that gradually narrow over time.

Remarkably, we observe a pronounced *hot-start* effect. At 0.4% of total training (8,200 sequences), $8 \times 8$ visual inputs achieve 12.34% accuracy, more than double the index-based baseline's 5.84%. In addition, as illustrated in Figure 1, during this hot-start stage, visual models not only achieve higher next-character accuracy, but also assign higher probabilities to plausible top candidates, suggesting that the learned representations support linguistically reasonable predictions beyond the single argmax.

This early advantage persists with visual models maintaining a consistent lead. For instance, by 16,441 sequences, $8 \times 8$ visual models achieve 15.65% accuracy, while the baseline reaches only 13.33%. This sustained advantage underscores that visual structure provides not just an initial shortcut, but a persistent edge that the index-based baseline struggles to match.

It is also important to note that the onset timing of this advantage correlates directly with input resolution. Higher-resolution models reach their hot-start phase earlier: at only 0.2% total training time, $40 \times 40$ inputs already achieve 13.06% ac-

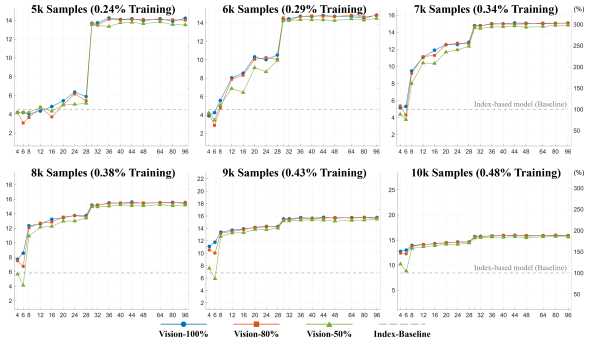



Figure 4: Validation accuracy across image resolutions at selected early training checkpoints (5k–10k samples), plotted on the index-based baseline scale (dashed line).

curacy, while the baseline remains at 4.30%. This pattern suggests that richer visual detail accelerates the extraction of structural regularities, giving higher-resolution models a decisive head start in learning.

Recall that our training employs a quadratic curriculum: the dataset size grows quadratically across epochs, meaning models see progressively more data as training advances. As a result, these early sequence counts ($\sim 10^4$) represent an extremely data-scarce regime where models have access to only a tiny fraction of the eventual corpus—making the observed hot-start advantage particularly meaningful.

Figure 4 further vizualizes that, across early training checkpoints (5k–10k samples), visual inputs maintain a consistent performance advantage, supporting the observed hot-start effect.

As noted in Section 3.4, the central strokes (toast-center) contain concentrated structural information, which helps explain the hot-start effect.

### 3.6 Ablation Studies

To understand which components drive performance, we conducted three ablation studies using $8 \times 8$ resolution as our testbed. In each case, training data and hyperparameters remain identical.

**Dual-Encoder Ablation.** Following DeepSeek-OCR's architecture (Wei et al., 2025), we test a dual-encoder that concatenates features from both ResNet and Vision Transformer (ViT). Comparisons with show that ResNet alone provides most of the structural benefit. We hypothesize that ViT is better suited for inter-character spatial relationships, adding marginal gains primarily in high-resolution settings.

**Training Strategy Ablation.** We compare joint training (visual components + decoder) versus freezing the decoder during adapter training. Joint training gives significantly better results, indicating that end-to-end optimization is crucial.

| Category | Setting | Acc (%) | Δ (pp) | Sig. |
|---|---|---|---|---|
| *Main Results* ($8 \times 8$) | | | | |
| Baseline | *Index-based* | 39.10 | — | Ref. |
| Vision | *Vision-100%* | 39.21 | +0.11 | $p$=0.77 |
| Vision | *Vision-80%* (80%) | 39.18 | +0.08 | $p$=0.81 |
| Vision | *Vision-50%* (50%) | 38.63 | -0.47 | $p$=0.09 |
| *Ablations (vs Vision-100%)* | | | | |
| Encoder | ViT | 38.45 | -0.76 | $p < 0.001$ |
| Training | Frozen decoder | 36.78 | -2.43 | $p < 0.001$ |
| Architecture | No adapter | 37.12 | -2.09 | $p < 0.001$ |
| *Hot-Start (10K samples)* | | | | |
| Early | *Index-based* | 6.45 | — | Ref. |
| Early | *Vision-100%* | 14.65 | +8.20 | $p < 0.001$ |
| Early | *Vision-80%* | 14.70 | +8.25 | $p < 0.001$ |
| Early | *Vision-50%* | 14.38 | +7.93 | $p < 0.001$ |
| *Statistical Validation* | | | | |
| Design | DEFF=19.9 | — | — | $\rho$=0.15 |
| Samples | $n_{\text{eff}}$=31.9K | — | — | 4.46× SE |
| CI | Width ±0.537pp | — | — | Δ=0.11pp |

Table 3: Consolidated results. All tests use adjusted SE (DEFF=19.9). Main: *Visual-based Model*s match text baseline. Ablations: CNN>ViT (Δ=-0.76pp), joint training vital (Δ=-2.43pp). Hot-start: 2.27× faster early learning. Stats: sequence correlations accounted for.

**Architecture Ablation.** We remove the visual adapter entirely (direct projection to decoder) causes performance degradation, particularly in the hot-start phase.

Results summarized in Table 3 show that ResNet with joint training achieves the best performance. Notably, hot-start advantages persist across all ablations, supporting that structural learning—not just specific architectural choices—is the key driver.

### 3.7 Summary of Visual Advantages

Across resolutions and ablations, three insights emerge. Low-resolution inputs alone capture essential character structure. Visual cues accelerate early stage training, producing a *hot-start* effect. Finally, models remain robust under severe degradation, demonstrating that coarse structural cues suffice. These results establish that visual representations provide a robust and sample-efficient alternative to index-based inputs for Chinese language modeling.

## 4 Interpretability Analysis

**All analyses in this section use the $8 \times 8$ resolution setting at the hot-start phase (with 10k training sequences).** This setting isolates essential structural information while removing fine-grained visual details, allowing us to examine how minimal visual cues enable early linguistic discrimination. We emphasize that these analyses are post-hoc and

| | *Index-based Model* | *Visual-based Model* | Ratio |
|---|---|---|---|
| **L2-Normalized Euclidean Distance (mean / 95% CI)** | | | |
| similar pairs | 1.41 / [1.40,1.44] | 1.20 / [1.10,1.33] | 1.20:1 |
| ⺾ | 1.41 / [1.39,1.44] | 1.20 / [1.09,1.32] | 1.18:1 |
| 扌 | 1.41 / [1.39,1.44] | 1.20 / [1.09,1.31] | 1.18:1 |
| **Cosine Similarity** | | | |
| similar pairs | 0.01 / [-0.05,0.04] | 0.28 / [0.14,0.55] | 30.4:1 |
| ⺾ | 0.001 / [-0.03,0.04] | 0.27 / [0.14,0.40] | 225.5:1 |
| 扌 | 0.002 / [-0.03,0.04] | 0.27 / [0.15,0.40] | 152.2:1 |

Table 4: Embedding distance analysis at $8 \times 8$ resolution (L2-normalized). Vision embeddings show consistently higher similarity: $\sim$1.2$\times$ closer in Euclidean space and $\sim$30$\times$ higher cosine similarity for visually similar pairs. Within radicals, vision embeddings maintain strong structural organization while index-based embeddings show near-random alignment.

exploratory in nature, intended to provide qualitative insights rather than formal interpretability guarantees.

### 4.1 Analysis of Embedding Space Geometry

We compare index-based and vision-based embeddings in their high-dimensional spaces (768D and 1024D respectively) to examine whether visual similarity and radical-based structure are reflected in the learned embedding spaces.To compare embeddings across different dimensionalities, we normalize all vectors to unit length ($\hat{v} = v/||v||_2$) before computing distances.

To ensure coverage of different character structures, we include representative character samples from our dataset in three categories: indecomposable characters with simple glyphs (e.g., 戊/戌), left-right composition with radicals on the left (e.g., 扌), and top-bottom composition with radicals on the top (e.g., ⺾). Within each category, we consider characters that are of high-frequency, visually confusable, or otherwise structurally informative.

We use the following metrics: Euclidean distance for absolute separation, cosine similarity for angular alignment, with 95% confidence intervals (CI). Results are summarized in Table 4.

We see that vision embeddings are consistently closer than index-based embeddings: $\sim$1.2$\times$ closer in L2-normalized Euclidean distance for visually similar pairs (mean ratio 1.20:1, 95% CI [1.02, 1.58]), with significantly higher cosine similarities (mean ratio 30.4:1, 95% CI [-14.64, 204.54]). Within radicals, vision embeddings show strong positive cosine similarities (扌: 0.27$\pm$0.13, ⺾: 0.27$\pm$0.13) while index-based embeddings remain near-zero (0.001-0.002).

Vision embeddings actively separate confusable pairs, mapping visually similar but semantically different characters further apart. Cosine similarity ratios range from -15.86 to 331.36, indicating both intra-radical attraction and intentional separation of confusable characters. The 95% CIs confirm structured organization in vision embeddings across all character structure types (indecomposable, left-right, and top-bottom), while index-based embeddings show no significant structure. This organization is consistent with the *hot-start* effect, where *Visual-based Model*s gain early predictive advantage.

PCA visualization (Figure 5) further illustrates this structured organization. We observe regional clustering of visually similar characters in the embedding space, with confusable pairs intentionally separated—consistent with the adaptive discrimination shown in Table 4.

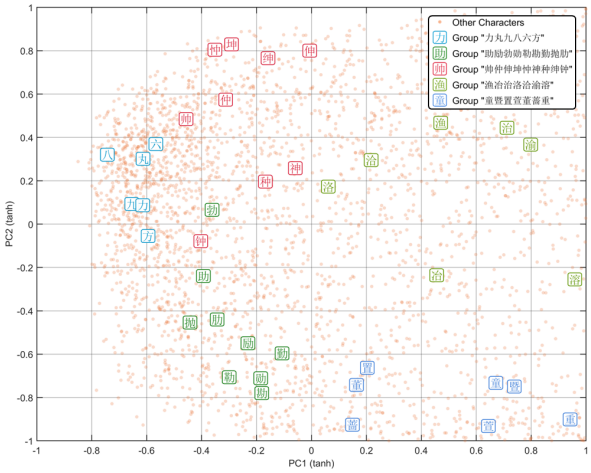



Figure 5: PCA of vision-based embeddings reveals localized spatial groupings of visually similar Chinese characters.

### 4.2 Behavior on Visually Similar Characters

We examine model sensitivity to minimal glyph differences using carefully constructed weak-context sentences. Representative predictions are shown in Table 5.

We observe that vision embeddings show enhanced sensitivity relative to ID-based embeddings, distinguishing confusable pairs like 土/士 (Example 1) and 人/入 (Example 8). Visual priors enable non-trivial preferences even with minimal semantic cues (Examples 5 日/目 and 7 入/人), reflecting the *hot-start* effect through larger probability differences. Remarkably, $8 \times 8$ visual inputs preserve sufficient structure for fine-grained distinctions (Examples 2 士/土 and 6 目/日), while ID-based models often require stronger contextual support. Across all examples, visual-based models

| ID | Sentence → Candidates (✓/×) | Model | P(%) | Choice |
|---|---|---|---|---|
| 1 | 下雨天鞋子上很容易沾上泥 | Vision | 0.05/0.00 | 土✓ |
| | → 土/土 | Text | 0.00/0.01 | 土× |
| 2 | 他是一个边境战 | Vision | 0.01/0.00 | 士✓ |
| | → 士/土 | Text | 0.00/0.01 | 土× |
| 3 | 这地板的材料是实 | Vision | 0.00/0.23 | 本× |
| | → 木/本 | Text | 0.00/0.31 | 本× |
| 4 | 别忘了拿作业 | Vision | 0.04/0.00 | 本✓ |
| | → 本/木 | Text | 0.04/0.00 | 本✓ |
| 5 | 昨天周六，今天是星期 | Vision | 0.19/0.12 | 日✓ |
| | → 日/目 | Text | 0.00/0.01 | 目× |
| 6 | 这个广告牌很醒 | Vision | 0.09/0.08 | 目✓ |
| | → 目/日 | Text | 0.03/0.00 | 目✓ |
| 7 | 这个房间非请莫 | Vision | 0.04/0.02 | 入✓ |
| | → 入/人 | Text | 0.06/0.10 | 人× |
| 8 | 介绍一下，这位是我的爱 | Vision | 8.63/0.00 | 人✓ |
| | → 人/入 | Text | 0.06/0.25 | 入× |

Table 5: Predicted probabilities for visually similar character pairs at sentence end. Sentences are selected to provide minimal but coherent semantic context. Four pairs of cases show model disagreement, highlighting differences between vision-based and index-based predictions.

exhibit enhanced discriminative power at sentence-end prediction, strategically organizing confusable characters in embedding space to support early-stage prediction advantages.

### 4.3 Pixel-Level Importance Analysis via Gradient Back-propagation

To understand which visual features matter most for prediction, we perform gradient-based analysis (Simonyan et al., 2014; Aflalo et al., 2022). For character images $\{I_1, \ldots, I_n\}$ and target $c_{n+1}$, we compute character-level importance $S_k = \sum_{i,j} \left| \frac{\partial y_{c_{n+1}}}{\partial I_{k,i,j}} \right|$, where higher $S_k$ indicates that input character $I_k$ is more important for predicting the target. Pixel-level importance is normalized as $H_{k,i,j} = (\left| \frac{\partial y_{c_{n+1}}}{\partial I_{k,i,j}} \right| - \min)/(\max - \min)$.

Testing on the confusable pairs from Section 4.2 reveals that the model assigns high importance to semantically relevant input characters (e.g., "泥" when predicting 土) and near-zero to irrelevant ones ("的"). Table 6 shows similar attention intensities across character regions, suggesting that the model distributes attention broadly rather than focusing exclusively on any single area. This balanced pattern explains why our cropping experiments (*Vision-80%* and *Vision-50%*) remain effective: the model can extract predictive signals from various character subregions.

The observed gradient patterns offer preliminary support for our "toast-center" conjecture though verifying its role in the *hot-start* phase requires further temporal analysis in future work.

Together, the analyses above illustrate how our vision-token representations make model behavior more inspectable. They link language predictions to explicit visual structure in ways that are difficult to access in index-based models.

| Region | Avg. Intensity | Std. Dev. |
|---|---|---|
| Upper Half | 0.087 | 0.014 |
| Lower Half | 0.081 | 0.014 |
| Left Half | 0.085 | 0.016 |
| Right Half | 0.083 | 0.012 |

Table 6: Average attention intensity across character regions.

## 5 Conclusion

In this work, we challenge the dominant index-based paradigm for Chinese language modeling by asking: *Can language modeling rely solely on visual form?* Our results answer affirmatively, especially for early-stage or low-resource scenarios.

Our investigation yields affirmative answers to our four research questions. **RQ1 (Visual Sufficiency)** is confirmed: visual inputs achieve accuracy comparable to index-based baselines (39.2% vs. 39.1%). **RQ2 (Early-Stage Dynamics)** reveals a pronounced *hot-start* effect: visual models reach 12.3% accuracy within 0.4% of training time—more than double the baseline's 5.8%. **RQ3 (Resolution Sensitivity)** demonstrates that even $8 \times 8$ pixels retain sufficient structure, while **RQ4 (Spatial Robustness)** confirms effectiveness under severe cropping (top 50% retained).

This advantage stems from fundamental differences in representation topology. Unlike index-based embeddings—which begin as unstructured points in feature space—visual embeddings inherit spatial organization from the encoder, providing a structural regularity that accelerates early learning.

Beyond performance gains, visual representations naturally provide some interpretability: characters sharing radicals tend to cluster while visually confusable pairs are often separated. Further, gradient-based analysis gives how the pixel regions contribute to predictions. This structure emerges from the visual input itself, without requiring auxiliary objectives or post-hoc tools.

Overall, visual structure provides not merely an alternative input format, but a sample-efficient inductive bias for Chinese language modeling. This points toward architectures natively designed for visual glyph processing and learning strategies that leverage visual structure for efficient training.

## Limitations

Our study has several limitations that point to future work. First, our experiments use standard font rendering. Testing handwritten or stylized characters would assess robustness. The GPT-2–style decoder is relatively small; scaling to larger architectures may give different visual utilization patterns. Additionally, our single-character image processing differs from full-paragraph OCR approaches. Extending to multi-character or paragraph-level images remains a valuable direction. Finally, applying this approach to other logographic systems could further prove its generality.

In summary, while current limitations exist, our findings demonstrate decisively that low-resolution visual glyphs are not just a viable substitute, but a cognitively richer starting point for Chinese language models. In contrast to recent surveys highlighting modality collapse in vision–language models (Sim et al., 2025), our results show that when visual input is isolated and structurally constrained, models can reliably exploit visual form rather than bypass it.

## References


Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. 2022. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21374–21383.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, and 95 others. 2021. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2:665–673.

Zhipeng Guo, Yu Zhao, Yabin Zheng, Xiance Si, Zhiyuan Liu, and Maosong Sun. 2016. THUCTC: An efficient chinese text classifier. GitHub repository.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. 2025. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. *Preprint*, arXiv:2506.05218.

Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. 2025. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *Preprint*, arXiv:2502.18443.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025. Can VLMs actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470, Vienna, Austria. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *Preprint*, arXiv:2409.01704.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *Preprint*, arXiv:2510.18234.

Wei Wu, Yuxian Meng, Fei Wang, Qinghong Han, Muyu Li, Xiaoya Li, Jie Mei, Ping Nie, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for chinese character representations. *ArXiv*, abs/1901.10125.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. UER: An open-source toolkit for pre-training models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 241–246, Hong Kong, China. Association for Computational Linguistics.