GLYPH-TO-IMAGE: SEMANTIC CHINESE GLYPH GENERATION VIA TEXT-TO-IMAGE MODELS

Anonymous authors
Paper under double-blind review

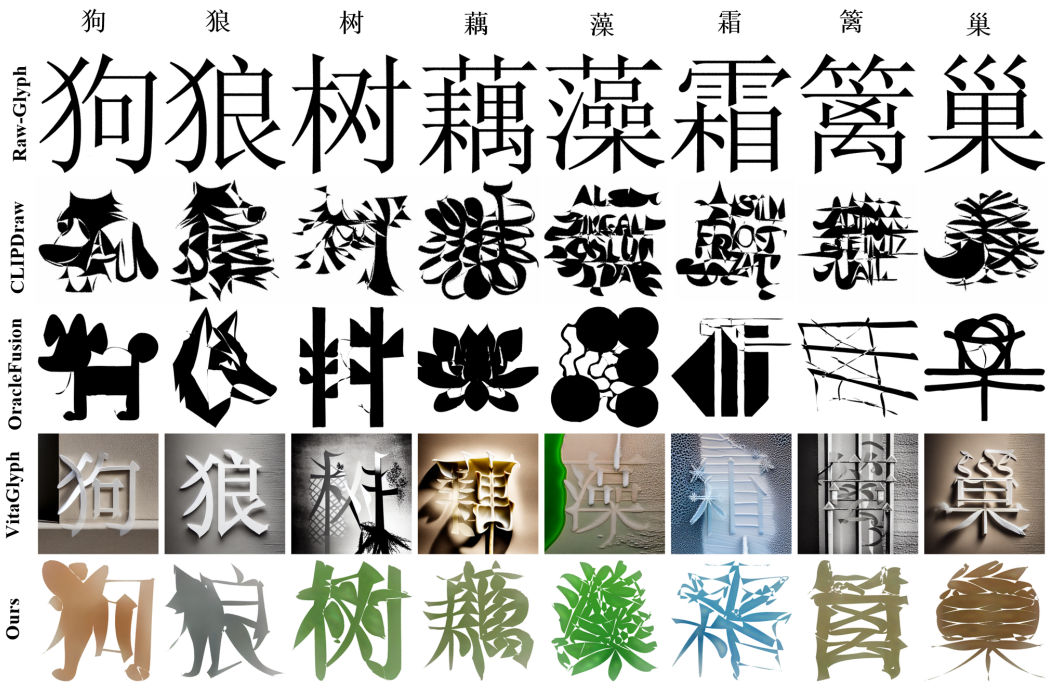

Figure 1: **Glyph-to-Image.** Different from existing works on semantic typography, artistic typography, and text-to-drawing, such as ClIPDraw (Frans et al., 2022), OracleFusion (Li et al., 2025), and VitaGlyph (Feng et al., 2026), which either generate semantically rich images at the expense of glyph fidelity or produce stylized fonts without rich semantic content, our approach simultaneously preserves both glyph structure and semantic content, while producing coherent color and texture.

ABSTRACT

Chinese characters combine symbolic identity with highly structured visual forms, making them a natural medium for semantic visual expression. However, automatically transforming prescribed Chinese glyphs into concept-bearing images while preserving their character identity remains challenging, since unconstrained generative models provide strong semantics but limited structural control. We introduce Glyph-to-Image, a task that transforms editable Chinese glyphs according to a target concept. Starting from native cubic Bézier outlines, Glyph-to-Image jointly optimizes glyph geometry and a learnable appearance field through differentiable rendering. To stabilize semantic appearance formation, we introduce an Appearance Low-Pass Filter (ALPF) that constrains the bandwidth of both the rendered appearance and its updates. To preserve local glyph structure during semantic deformation, we further introduce Glyph Consistency Regularization (GCR), which regularizes neighboring Bézier segments with respect to local similarity transformations. Beyond single-character generation, Glyph-to-Image naturally extends to multi-character generation. We establish benchmarks, develop evaluation metrics, and reproduce representative methods from semantic and artistic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

typography as baselines. Extensive experiments show that the proposed Glyph-to-Image achieves a better balance between semantic expressiveness and character readability than existing approaches.

1 INTRODUCTION

Writing systems can be broadly categorized into phonographic and logographic systems, with alphabetic scripts such as English primarily encoding phonological information and Chinese characters primarily encoding morphemes (Daniels & Bright, 1996). Chinese characters are highly organized visual structures, whose geometric organization makes them a natural medium for semantic visual design, with applications in artistic typography, branding, and visual communication.

To bridge the gap between linguistic meaning and glyph appearance, semantic typography generates glyphs whose visual forms reflect their semantic meanings. Recent approaches such as Word-As-Image (Iluz et al., 2023) and DS-Fusion (Tanveer et al., 2023) show that generative priors can transform textual forms into semantically expressive visual objects while retaining a degree of readability. However, most existing works are predominantly designed for Latin letters or rasterized text and are not well suited to the structural properties of Chinese glyphs.

On the other hand, artistic glyph generation focuses on synthesizing character-consistent glyphs under a desired typographic style (Feng et al., 2026). These methods primarily target font-style generation and transfer, rather than injecting visual semantics into existing glyphs. For Chinese characters, most existing approaches lack explicit constraints on character glyphs, often causing the generated glyphs to deviate substantially from the original glyphs.

Chinese characters contain dense internal structures and strong dependencies among neighboring strokes, making their identity particularly sensitive to local geometric deformation. At the same time, geometry alone provides limited capacity for expressing semantic cues such as color, texture, and surface appearance. Unconstrained appearance optimization introduces an additional challenge, as powerful generative priors may encode semantics through fragmented or high-frequency patterns rather than coherent visual structure. These characteristics create a fundamental trade-off between semantic expressiveness and glyph preservation.

In this work, we introduce Glyph-to-Image, a task that transforms prescribed Chinese glyphs into concept-bearing visual forms while preserving their character identity. A comparison of our task with other related tasks is shown in Figure 2. Given a glyph and a semantic concept, the goal is to produce an editable glyph that visually conveys the target concept while remaining recognizable as the initial glyph. Each source glyph is represented by its native cubic Bézier geometry together with a learnable appearance field. Through differentiable rendering, semantic gradients from pretrained text-to-image models are directly propagated to both geometric and appearance parameters, allowing semantic information to be expressed through shape and color while retaining an explicit and editable glyph representation. We also introduce the Appearance Low-Pass Filter (ALPF) and Glyph Consistency Regularization (GCR) to constrain the frequency bandwidth of appearance and preserve local geometric structure, respectively, enabling appearance and geometric adaptation without excessive degradation. We establish benchmarks covering Chinese expressions ranging from one to four characters and compare our approach with generative, vector-optimization, and semantic-typography baselines. Both quantitative and qualitative results demonstrate that our approach effectively balances semantic alignment and glyph readability.

Contributions: 1) Glyph-to-Image Paradigm: We introduce the paradigm of semantic Chinese glyph generation, leveraging the generalization capability of pretrained text-to-image models to transform Chinese glyphs into concept-bearing visual forms while preserving character identity. **2) Flexible Glyph-to-Image Pipeline:** We propose a flexible glyph-to-image generation pipeline based on joint glyph-semantic optimization, applicable to both single and multiple Chinese characters, with the Appearance Low-Pass Filter (ALPF) and Glyph Consistency Regularization (GCR) modules enabling appearance adaptation and geometric deformation, respectively. **3) Comprehensive Benchmarks:** We establish comprehensive Chinese semantic character generation benchmarks for both single-character and multi-character glyph-to-image generation, together with appropriate evaluation metrics and representative baselines.

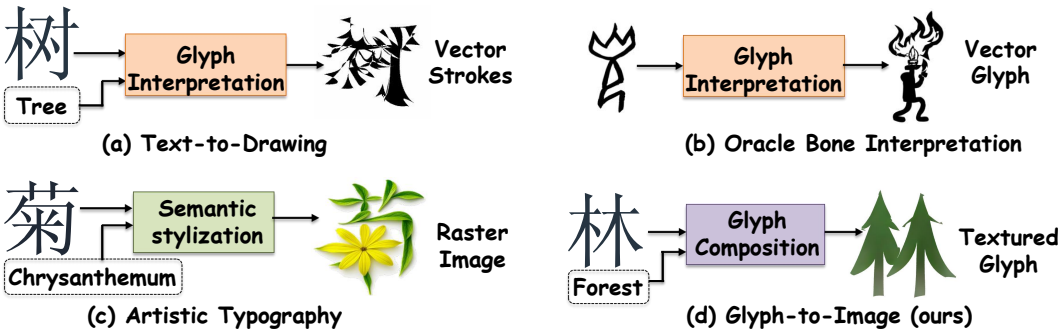

Figure 2: **Comparison of different Chinese character image generation tasks.**Text-to-drawing (Frans et al., 2022) focuses on semantic interpretation without preserving the initial glyph structure. Oracle bone interpretation (Li et al., 2025) primarily targets oracle bone script, while artistic typography (Feng et al., 2026) aim to generate artistic fonts rather than deforming glyphs according to semantic concepts. In contrast, our glyph-to-image task generates semantically deformed glyphs with rich textures while preserving their character identity.

2 Related Work

Semantic Typography: Semantic typography aims to embed visual meaning into textual form while preserving linguistic readability, with representative approaches discussed below. Word-As-Image (Iluz et al., 2023) uses pretrained diffusion priors to guide semantic deformation of vector letter outlines while maintaining legibility and font style. DS-Fusion (Tanveer et al., 2023) further combines diffusion-based stylization with glyph-preserving supervision to introduce richer visual appearance. Rather than directly deforming contours, Anything to Glyph (Wang et al., 2023a) constructs artistic glyphs from semantically related visual elements under spatial constraints. TypeDance (Xiao et al., 2024) explores interactive type-imagery fusion at stroke, character, and multi-character levels, while Dynamic Typography (Liu et al., 2025) extends semantic deformation to temporally coherent animated text using a text-to-video prior. For logographic scripts, OracleFusion (Li et al., 2025) combines semantic reasoning with structural constraints for oracle-bone-script interpretation, and VitaGlyph (Feng et al., 2026) performs controllable artistic generation by decomposing characters into semantic subject and surrounding regions. Despite these advances, balancing flexible semantic transformation with fine-grained character-structure preservation remains challenging.

Artistic Glyph Generation: Glyph and font generation focuses on synthesizing character-consistent glyphs under a desired typographic style. DM-Font (Cha et al., 2020) introduces a dual-memory architecture for few-shot font generation, while MX-Font (Park et al., 2021) uses localized experts to capture fine-grained style variations across character regions. Tang et al. (2022) model stroke- and component-level local styles for complex glyph transfer, and XMP-Font (Liu et al., 2022) learns multi-scale structural and stylistic representations through cross-modality pretraining. Component-aware supervision (Kong et al., 2022) further improves local structural preservation during font synthesis. More recent methods emphasize editable representations. DeepVecFont-v2 (Wang et al., 2023b) models complex vector-outline sequences with Transformers, VecFontSDF (Xia et al., 2023) reconstructs vector glyphs through signed-distance representations, and FontDiffuser (Yang et al., 2024) applies diffusion models to one-shot font generation. VecFusion (Thamizharasan et al., 2024) combines raster and vector diffusion for high-quality vector font synthesis, FontAnimate (Fu et al., 2025) formulates few-shot font transfer as a dynamic generation process, and VecGlypher (Huang et al., 2026) directly generates editable vector glyphs from textual or visual conditions. These methods mainly target font-style generation and transfer rather than injecting visual semantics into an existing glyph.

Icon Generation: Icon generation studies how semantic concepts can be represented as compact and editable visual symbols. DeepSVG (Carlier et al., 2020) learns hierarchical generative representations for structured SVG graphics, while Im2Vec (Reddy et al., 2021) learns vector graphics from raster supervision without requiring ground-truth vector annotations. With the emergence of large generative models, VectorFusion (Jain et al., 2023) transfers semantic knowledge from a frozen text-to-image diffusion model to SVG parameters through differentiable rendering and score distillation,

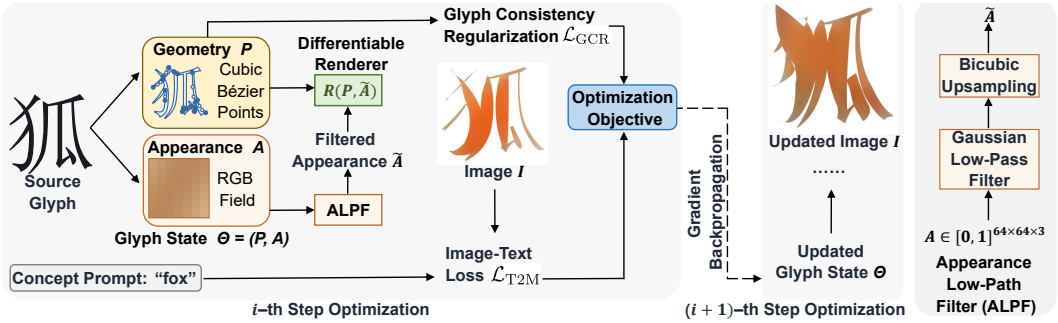

Figure 3: **Overview of the Glyph-to-Image framework.** The source glyph is represented by native cubic Bézier geometry and a learnable appearance field. A frozen text-to-image model provides semantic guidance through an image-text loss with differentiable rendering, while the Appearance Low-Pass Filter (ALPF) controls the appearance bandwidth and Glyph Consistency Regularization (GCR) preserves local glyph structure. The glyph geometry and appearance are iteratively optimized to produce the final semantic glyph representation.

and DiffSketcher (Xing et al., 2023) similarly optimizes Bézier primitives under diffusion guidance for text-conditioned vector sketches. IconShop (Wu et al., 2023a) directly models text-conditioned vector-icon generation with an autoregressive Transformer, while SVGDreamer (Xing et al., 2024) improves text-guided SVG synthesis through diffusion-based vector optimization. Recent methods increasingly combine language models with explicit vector representations. Chat2SVG (Wu et al., 2025) constructs structured SVGs with language models and refines them using generative visual priors, StarVector (Rodriguez et al., 2025) directly generates SVG code from text or images, and LLM4SVG (Xing et al., 2025) improves complex vector-graphics understanding and generation through SVG-specific semantic representations. These methods provide strong semantic generation capabilities, but generally do not need to preserve the identity of a predefined symbolic character.

3 Method

Given the source glyph and textual concept prompt, we leverage a pretrained text-to-image model to gradually transform the glyph toward the target concept while preserving its character identity. The pipeline of our approach is shown in Figure 3, and its key components are described below.

3.1 Semantic Glyph Deformation

Glyph-Semantic Optimization : Given a character c and a concept prompt y, we extract its native cubic Bézier outline and optimize control points $P = \{p_i\}_{i=1}^{N}$ together with a learnable RGB appearance field $A \in [0,1]^{r\times r\times 3}$. We denote the glyph-semantic state by $\Theta = (P, A)$ and its rendered image by $I = \mathcal{R}(P, A)$, where $\mathcal{R}$ is differentiable with respect to both geometry and appearance.

The rendered image is encoded into the latent space using a pretrained probability flow model, such as Flow Matching, together with its corresponding image encoder, which provides a text-conditioned semantic objective $\mathcal{L}_{\mathrm{T2M}}(I, y)$. Gradients are propagated only to the explicit glyph representation through the differentiable renderer, while the text-to-image model provides semantic guidance rather than serving as a trainable glyph generator. Let $g_I = \nabla_I \mathcal{L}_{\mathrm{T2M}}$ denote the semantic gradient in image space. Through the renderer Jacobians, the semantic signal is propagated to the geometry and appearance parameter spaces as follows:

$$\nabla_P \mathcal{L}_{\mathrm{T2M}} = J_P^\top g_I, \qquad \nabla_A \mathcal{L}_{\mathrm{T2M}} = J_A^\top g_I, \tag{1}$$

where J_P and J_A are the Jacobians of $\mathcal{R}$ with respect to P and A, respectively. This semantic guidance enables geometry to encode shape-level cues, while appearance captures complementary color and surface information.

Appearance Low-Pass Filter: Directly optimizing a dense appearance field gives the semantic prior enough freedom to encode concept-related signals through unstable high-frequency patterns. We

therefore introduce Appearance Low-Pass Filter (ALPF), which constrains appearance bandwidth in both the forward and backward paths.

We optimize A at a resolution of 64×64. During the forward pass, the appearance field A is reconstructed using antialiased bicubic upsampling followed by Gaussian low-pass filtering. During the backward pass, the appearance gradient is further filtered with a Gaussian kernel. The forward and backward processes are formulated as follows:

$$\tilde{A} = G_f * U(A), \qquad \tilde{g}_A = G_b * \nabla_A \mathcal{L}_{\mathrm{T2M}}, \tag{2}$$

where U denotes bicubic antialiased upsampling, and G_f and G_b denote Gaussian low-pass kernels.

Both operations exhibit low-pass responses in the frequency domain. The forward filter limits the high-frequency appearance content exposed to the frozen prior, while the gradient filter suppresses high-frequency components that semantic guidance can write back into the learnable appearance field. The ALPF stabilizes semantic appearance formation without introducing an additional loss.

3.2 Glyph Consistency Regularization

Although semantic guidance induces useful shape deformations, unconstrained geometry updates may over-stretch thin strokes, introduce anisotropic distortions, or destabilize dense local structures. We therefore introduce Glyph Consistency Regularization (GCR), which preserves the local organization of neighboring Bézier segments while allowing moderate, similarity-preserving deformations.

For two adjacent Bézier segments sharing an endpoint, we construct the ordered seven-point patch

$$X_i = [P_i, C_{i1}, C_{i2}, P_{i+1}, C_{i+1,1}, C_{i+1,2}, P_{i+2}] \in \mathbb{R}^{m\times 2}, \tag{3}$$

where $m = 7$, P_i denotes the endpoint of the i-th segment, P_{i+1} is the shared endpoint, and C_{i1} and C_{i2} denote its two control points. Let X_i^0 and X_i^t denote the initial and current configurations of the i-th patch, respectively. Patches are constructed only between adjacent segments belonging to the same closed path, including the closing seam, and never across different paths. To make the comparison invariant to translation and isotropic scaling, we normalize each patch as follows:

$$\mu(X) = \frac{1}{m}\sum_{j=1}^{m} X_j, \qquad \bar{X} = \frac{X - \mu(X)}{\|X - \mu(X)\|_F + \epsilon}. \tag{4}$$

We then align the current patch to its initialization using the optimal two-dimensional rotation

$$R_i^* = \arg\min_{R\in SO(2)} \left\| \bar{X}_i^t R - \bar{X}_i^0 \right\|_F^2, \tag{5}$$

and define the normalized local distortion as

$$e_i = \frac{1}{\sqrt{m}} \left\| \bar{X}_i^t R_i^* - \bar{X}_i^0 \right\|_F. \tag{6}$$

Rather than penalizing every geometric change, the GCR introduces a deformation margin τ and only suppresses distortions that exceed this tolerance:

$$\mathcal{L}_{\mathrm{GCR}} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{ReLU}(e_i - \tau)^2, \tag{7}$$

where K denotes the number of valid local patches. Since translation and isotropic scaling are removed before alignment and the optimal rotation is explicitly fitted, e_i is invariant to local similarity transformations. The GCR permits translation, rotation, and uniform scaling while penalizing excessive shear, anisotropic distortion, and local collapse. The margin τ leaves moderate semantic deformations unpenalized rather than forcing the glyph to remain close to its initial geometry.

3.3 Glyph-to-Image Generation

Single-Character Generation: The pipeline of single-character glyph-to-image generation is shown in Algorithm 1. The final objective combines the text-conditioned semantic objective and the glyph consistency regularization:

$$\mathcal{L} = \mathcal{L}_{\mathrm{T2M}}(\mathcal{R}(P, \tilde{A}), y) + \lambda \mathcal{L}_{\mathrm{GCR}}, \tag{8}$$

Algorithm 1 Single-Character Glyph-to-Image Generation

Input: Character c, Prompt y, Iterations T, GCR parameters λ_g, τ, filters G_f, G_b
Output: Optimized State Θ^*, Rendered Image I

Extract initial Bézier geometry P^0, initialize $A^0 \in [0,1]^{r\times r\times 3}$, $r = 64$
Store reference patches $\{\bar{X}_i^0\}_{i=1}^K$ using Eqs. (3-4)
for $t = 0, \ldots, T-1$ **do**
    Forward Process: $\tilde{A}^t \leftarrow G_f * U(A^t)$; $I \leftarrow \mathcal{R}(P^t, \tilde{A}^t)$
    Evaluate $\mathcal{L}_{\mathrm{T2M}}(I, y)$ via the pretrained text-to-image model.
    Compute $\mathcal{L}_{\mathrm{GCR}}$ from P^t and the reference patches using Eqs. (4-7)
    Gradient Backpropagation with $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{T2M}} + \lambda\mathcal{L}_{\mathrm{GCR}}$; $\tilde{g}_A \leftarrow G_b * \nabla_A \mathcal{L}$
    Adam update: $(P^t, A^t) \rightarrow (P^{t+1}, A^{t+1})$ using $(\nabla_P \mathcal{L}, \tilde{g}_A)$, respectively
end for
Optimized State $\Theta^* \leftarrow (P^T, A^T)$, Rendered Image $I \leftarrow \mathcal{R}(P^T, G_f * U(A^T))$

where λ is the weight of the regularization.

Multi-Character Generation: For a multi-character expression $w = (c_1, \ldots, c_n)$, all constituent characters receive the same word-level concept y but are optimized independently:

$$\Theta_i^* = \arg\min_{\Theta_i} \mathcal{L}(\Theta_i; c_i, y). \tag{9}$$

The rendered images of the individually optimized glyphs are arranged in their original order.

4 Experiments

We first construct the Chinese glyph-to-image dataset, design evaluation metrics from different perspectives, and reproduce representative baselines for character image generation. We then evaluate our approach on these benchmarks and compare it with state-of-the-art methods. Finally, we conduct ablation studies and parameter analyses to further investigate the effectiveness of our approach.

4.1 Evaluation Metrics and Benchmark

Glyph-to-Image Dataset: We construct separate glyph-to-image benchmarks for single-character and multi-character Chinese expressions. The single-character subset contains 90 glyph–concept pairs and is divided into three structural-complexity groups: *simple*, *medium*, and *complex*, with 30 instances in each group. The multi-character subset contains two to four character expressions, with 90 expression–concept pairs for each length, resulting in 270 instances.

Evaluation Metrics: We evaluate glyph-to-image generation from three aspects: semantic alignment, image quality, and glyph readability. CLIP-T (Hessel et al., 2021) measures the semantic alignment between the generated image and the target concept, while HPSv2 (Wu et al., 2023b) provides an independent perceptual score for image-text preference. Glyph readability is measured by the confidence score assigned to the target character by a fixed character recognition model using PaddleOCR (Du et al., 2020).

To combine semantic alignment and glyph readability, we normalize the CLIP-T score using the lower and upper bounds obtained from the initial glyph and text-to-image generation as references:

$$S = \frac{\tilde{S} - l}{u - l}. \tag{10}$$

where $\tilde{S}$ is the raw CLIP-T score of a particular sample, l and u are the average CLIP-T scores of the initial glyph and text-to-image generation on the validation set, respectively. The normalized score S is clipped to the range $[0, 1]$.

We then define the Semantic–Glyph Balance Score (SGBS) for each sample as

$$\mathrm{SGBS} = \sqrt{S \cdot P_{\mathrm{ocr}}}, \tag{11}$$

where P_{ocr} is the confidence score predicted by PaddleOCR. SGBS is high only when both semantic alignment and glyph readability are preserved.

Method	CLIP-T ↑	HPSv2 ↑	P_{ocr} ↑	S ↑	SGBS ↑
Raw Glyph	0.1618	0.2347	**0.9952**	0.0708	0.1484
Text-to-Image	0.2256	**0.2509**	0.0002	0.7759	0.0040
CLIPDraw (Frans et al., 2022)	**0.2318**	0.2452	0.0192	**0.8187**	0.0393
DS-Fusion (Tanveer et al., 2023)	0.2122	0.2469	0.0019	0.6065	0.0045
OracleFusion (Li et al., 2025)	0.2168	0.2361	0.0790	0.6645	0.0610
VitaGlyph (Feng et al., 2026)	0.1765	0.2408	0.3952	0.3011	0.1585
Ours	0.1993	0.2410	0.4054	0.4463	**0.2189**

Table 1: **Quantitative comparison of single-character glyph-to-image.** Higher is better for all metrics. SGBS measures the joint balance between semantic alignment and glyph readability.

Method	Simple			Medium			Complex		
	CLIP-T ↑	P_{ocr} ↑	SGBS ↑	CLIP-T ↑	P_{ocr} ↑	SGBS ↑	CLIP-T ↑	P_{ocr} ↑	SGBS ↑
Raw Glyph	0.1655	**0.9975**	0.1814	0.1577	**0.9976**	0.0835	0.1623	**0.9904**	**0.1805**
Text-to-Image	0.2272	0.0006	0.0090	0.2181	0.0000	0.0017	**0.2317**	0.0000	0.0013
CLIPDraw	**0.2349**	0.0238	0.0457	**0.2306**	0.0329	0.0605	0.2299	0.0008	0.0117
DS-Fusion	0.2135	0.0055	0.0121	0.2256	0.0000	0.0011	0.1973	0.0000	0.0004
OracleFusion	0.2085	0.2269	0.1531	0.2242	0.0017	0.0130	0.2175	0.0085	0.0170
VitaGlyph	0.1826	0.4846	0.2207	0.1674	0.4954	0.1316	0.1796	0.2055	0.1234
Ours	0.1892	0.7019	**0.3866**	0.2001	0.4075	**0.1737**	0.2087	0.1068	0.0963

Table 2: **Complexity-wise quantitative comparison for single-character glyph-to-image.** Results on simple, medium, and complex Chinese characters.

For multi-character expressions, all metrics are first computed for each constituent glyph and then averaged across characters.

Baselines: We construct two reference baselines, Raw Glyph and Text-to-Image. Raw Glyph directly renders the original SimSun character as a structural reference, while Text-to-Image generates an unconstrained image from the semantic prompt as a semantic reference. Besides the reference baselines, we reproduce CLIPDraw (Frans et al., 2022), DS-Fusion (Tanveer et al., 2023), OracleFusion (Li et al., 2025), and VitaGlyph Feng et al. (2026) for comparison.

Implementation Details: All vector-based methods start from SimSun glyph outlines represented as native cubic Bézier paths and rendered using a differentiable rasterizer. The text-to-image model is based on the SD3.5 rectified-flow model (Esser et al., 2024). Each glyph is optimized for 500 iterations, with the appearance field optimized at a resolution of 64×64. For ALPF, Gaussian filtering uses a kernel size of 9 and a standard deviation of 2.0 in the forward pass, and a kernel size of 5 and a standard deviation of 1.0 for the appearance gradient. For GCR, we set the deformation tolerance to $\tau = 0.025$ and the regularization weight to $\lambda = 2$. Geometry and appearance parameters are optimized using separate Adam parameter groups.

4.2 Evaluations

Evaluation of Single-Character Generation: Table 1 reports the quantitative comparison. The results reveal a clear trade-off between semantic expressiveness and glyph readability. Raw Glyph nearly perfectly preserves character identity, but provides little semantic alignment. At the other extreme, Text-to-Image and CLIPDraw achieve substantially stronger semantic scores, but largely lose the prescribed glyph identity. DS-Fusion exhibits a similar tendency, producing semantically relevant appearance with very low character recognition. In contrast, our approach maintains substantially higher readability while retaining meaningful semantic alignment. These results show that combining structured geometric deformation with controlled semantic appearance provides an effective balance between concept expression and glyph preservation.

Table 2 further evaluates performance across different levels of glyph complexity. Our method achieves the highest SGBS on simple and medium-complexity characters. As glyph complexity



Figure 4: **Multi-character semantic glyph composition.** Glyph-to-Image generates semantically coherent multi-character expressions while preserving recognizable structures of individual glyphs.

Method	CLIP-T ↑	HPSv2 ↑	P_{ocr} ↑	S ↑	SGBS ↑
Raw Glyph	0.1392	0.1985	**0.9979**	0.0107	0.0378
Text-to-Image	**0.2425**	**0.2520**	0.0001	0.8632	0.0059
CLIPDraw (Frans et al., 2022)	0.2316	0.2339	0.0100	**0.7818**	0.0508
DS-Fusion (Tanveer et al., 2023)	0.2201	0.2366	0.0001	0.6677	0.0041
OracleFusion (Li et al., 2025)	0.2080	0.2235	0.0569	0.5734	0.0884
VitaGlyph (Feng et al., 2026)	0.1692	0.2211	0.4458	0.2524	0.2137
Ours	0.1867	0.2202	0.4454	0.3404	**0.2584**

Table 3: **Quantitative comparison on the multi-character benchmark.** Results are averaged over two-, three-, and four-character expressions. Higher is better for all metrics.

increases, semantic alignment remains relatively stable, while glyph readability becomes more sensitive to structural density. This trend reflects the increasing difficulty of preserving dense local structures during semantic deformation.

Evaluation of Multi-Character Generation: Table 3 quantitatively evaluates multi-character glyph-to-image generation. Figure 4 shows that the generated glyphs of multi-character exhibit consistent concept-level visual characteristics, yielding compositions with a coherent semantic theme rather than visually isolated characters.

4.3 Ablations and Analysis

Ablation Studies: We evaluate the contributions of ALPF and GCR through quantitative and qualitative comparisons. As shown in Table 4, our approach achieves the best balance between semantic alignment and glyph readability. Removing ALPF primarily affects appearance optimization, making semantic appearance formation less stable and reducing the SGBS. Meanwhile, removing GCR allows the geometry to deform more freely, which can slightly improve semantic alignment but weakens character readability. Removing both components leads to further degradation. These results support the intended roles of ALPF in regulating appearance optimization and GCR in preventing excessive geometric distortion.

Figure 5 further illustrates these effects. Without ALPF, appearance optimization produces visible high-frequency and repetitive patterns. Without GCR, stronger object-like deformation emerges, but local stroke organization is more easily disrupted. Removing both components combines these failure modes. In contrast, our approach maintains coherent semantic appearance while preserving recognizable glyph structure. These results indicate that ALPF stabilizes semantic appearance formation, whereas GCR controls excessive geometric deformation.

Sensitivity of Deformation Tolerance τ: The tolerance τ determines how much local deformation is allowed before GCR becomes active, while λ controls the strength of the resulting structural penalty. As shown in Figure 6, a zero tolerance imposes overly strict structural constraints and limits its semantic adaptation, whereas larger values allow greater geometric deformation but may weaken glyph identity. We set $\tau = 0.025$, which provides the best qualitative balance between structural preservation and semantic deformation. The results also remain visually stable under nearby toler-



Figure 5: **Qualitative component ablation.** ALPF stabilizes appearance, while GCR limits geometric distortion.

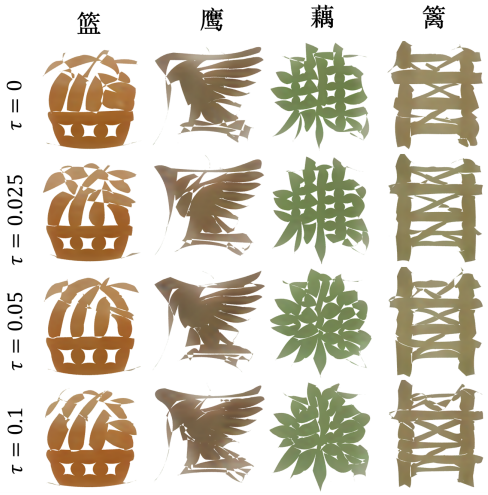

Figure 6: **Effect of deformation tolerance τ.** Qualitative comparison across different GCR tolerance settings.

Variant	CLIP-T ↑	P_{ocr} ↑	SGBS ↑
w/o ALPF & GCR	0.1987	0.3183	0.1784
w/o ALPF	0.1978	0.3986	0.2067
w/o GCR	**0.2049**	0.3512	0.1938
Ours	0.1993	**0.4054**	**0.2189**

Table 4: **Component ablation.** Higher is better for all metrics.

Setting	CLIP-T ↑	P_{ocr} ↑	SGBS ↑
$\lambda = 0.5$	**0.2037**	0.3663	0.2025
$\lambda = 1$	0.2013	0.3894	0.2134
$\lambda = 2$	0.1993	0.4054	**0.2189**
$\lambda = 4$	0.1968	0.4207	0.2155
$\lambda = 8$	0.1931	**0.4372**	0.2091

Table 5: **Effect of GCR weight λ.**

ance values, indicating that GCR operates within a reasonably robust range rather than depending on a narrowly tuned threshold.

Sensitivity of Regularization Weight λ: We study the effect of the regularization strength by varying λ. As shown in Table 5, increasing λ improves glyph readability but gradually reduces semantic alignment, revealing a clear trade-off between structural preservation and semantic flexibility. The SGBS score peaks at $\lambda = 2$ with a score of 0.2189 and decreases under stronger regularization, as excessive structural constraints limit semantic deformation.

5 Conclusion

We introduce Glyph-to-image, an interesting task that transforms editable Chinese glyphs into semantically expressive visual forms while preserving character identity. We propose an inference-time approach which jointly optimizes native cubic Bézier geometry and a learnable appearance field under a frozen text-to-image model. The Appearance Low-Pass Filter (ALPF) stabilizes semantic appearance formation, while Glyph Consistency Regularization (GCR) preserves local glyph structure during semantic deformation. We establish benchmarks for both single- and multi-character text-to-image generation. Our approach achieves a strong balance between semantic fidelity and character readability across different benchmarks. Ablation studies further verify the complementary roles of ALPF and GCR. We hope this work helps bridge the gap between linguistic meaning and glyph appearance in logographic writing systems such as Chinese characters.

Limitations: For simple characters with only a few strokes, our method may provide weaker semantic guidance as the limited number of glyph parameters restricts the freedom for semantic adaptation.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and assistance with code development. All AI-assisted content, mathematical derivations, experimental implementations, and generated text were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not introduce new training data or collect human-subject data. The proposed method inherits potential risks and biases from the underlying pretrained generative models.

REPRODUCIBILITY STATEMENT

We provide the complete mathematical formulation, optimization procedure, and implementation details required to reproduce the proposed method. Hyperparameter settings, evaluation protocols, and additional experimental results are provided in the main paper and appendix. Code will be released upon publication.

REFERENCES

Alexandre Carlier, Martin Danelljan, Alexandre Alahi, and Radu Timofte. Deepsvg: A hierarchical generative network for vector graphics animation. *Advances in Neural Information Processing Systems*, 33:16351–16361, 2020.

Junbum Cha, Sanghyuk Chun, Gayoung Lee, Bado Lee, Seonghyeon Kim, and Hwalsuk Lee. Few-shot compositional font generation with dual memory. In *European Conference on Computer Vision*, pp. 735–751. Springer, 2020.

Peter T Daniels and William Bright. *The world's writing systems*. Oxford University Press, 1996.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Kailai Feng, Yabo Zhang, Haodong Yu, Zhilong Ji, Jinfeng Bai, Hongzhi Zhang, and Wangmeng Zuo. VitaGlyph: Vitalizing artistic typography with flexible dual-branch diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8220–8230. IEEE, 2026.

Kevin Frans, Lisa Soros, and Olaf Witkowski. ClipDraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems*, 35: 5207–5218, 2022.

Bin Fu, Zixuan Wang, Kainan Yan, Shitian Zhao, Qi Qin, Jie Wen, Junjun He, and Peng Gao. Fontanimate: High quality few-shot font generation via animating font transfer process. In *IEEE/CVF International Conference on Computer Vision*, pp. 16015–16025. IEEE, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

Xiaoke Huang, Bhavul Gauri, Kam Woh Ng, Tony Ng, Mengmeng Xu, Zhiheng Liu, Weiming Ren, Zhaochong An, Zijian Zhou, Haonan Qiu, et al. Vecglypher: Unified vector glyph generation with language models. *arXiv preprint arXiv:2602.21461*, 2026.

Shir Iluz, Yael Vinker, Amir Hertz, Daniel Berio, Daniel Cohen-Or, and Ariel Shamir. Word-as-image for semantic typography. *ACM Transactions on Graphics*, 42(4):1–11, 2023.

Ajay Jain, Amber Xie, and Pieter Abbeel. Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1911–1920. IEEE, 2023.

Yuxin Kong, Canjie Luo, Weihong Ma, Qiyuan Zhu, Shenggao Zhu, Nicholas Yuan, and Lianwen Jin. Look closer to supervise better: One-shot font generation via component-based discriminator. In *IEEE/CVF conference on computer vision and pattern recognition*, pp. 13472–13481. IEEE, 2022.

Caoshuo Li, Zengmao Ding, Xiaobin Hu, Bang Li, Donghao Luo, AndyPian Wu, Chaoyang Wang, Chengjie Wang, Taisong Jin, Seven Shu, et al. OracleFusion: Assisting the decipherment of oracle bone script with structurally constrained semantic typography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19893–19902, 2025.

Wei Liu, Fangyue Liu, Fei Ding, Qian He, and Zili Yi. Xmp-font: Self-supervised cross-modality pre-training for few-shot font generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7895–7904. IEEE, 2022.

Zichen Liu, Yihao Meng, Hao Ouyang, Yue Yu, Bolin Zhao, Daniel Cohen-Or, and Huamin Qu. Dynamic typography: Bringing text to life via video diffusion prior. In *IEEE/CVF International Conference on Computer Vision*, pp. 14787–14797. IEEE, 2025.

Song Park, Sanghyuk Chun, Junbum Cha, Bado Lee, and Hyunjung Shim. Multiple heads are better than one: Few-shot font generation with multiple localized experts. In *IEEE/CVF International Conference on Computer Vision*, pp. 13880–13889. IEEE, 2021.

Pradyumna Reddy, Michael Gharbi, Michal Lukac, and Niloy J Mitra. Im2vec: Synthesizing vector graphics without vector supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7342–7351, 2021.

Juan A Rodriguez, Abhay Puri, Shubham Agarwal, Issam H Laradji, Pau Rodriguez, Sai Rajeswar, David Vazquez, Christopher Pal, and Marco Pedersoli. Starvector: Generating scalable vector graphics code from images and text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16175–16186. IEEE, 2025.

Licheng Tang, Yiyang Cai, Jiaming Liu, Zhibin Hong, Mingming Gong, Minhu Fan, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Few-shot font generation by learning fine-grained local styles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7885–7894. IEEE, 2022.

Maham Tanveer, Yizhi Wang, Ali Mahdavi-Amiri, and Hao Zhang. Ds-fusion: Artistic typography via discriminated and stylized diffusion. In *IEEE/CVF International Conference on Computer Vision*, pp. 374–384. IEEE, 2023.

Vikas Thamizharasan, Difan Liu, Shantanu Agarwal, Matthew Fisher, Michaël Gharbi, Oliver Wang, Alec Jacobson, and Evangelos Kalogerakis. Vecfusion: Vector font generation with diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7943–7952. IEEE, 2024.

Changshuo Wang, Lei Wu, Xiaole Liu, Xiang Li, Lei Meng, and Xiangxu Meng. Anything to glyph: Artistic font synthesis via text-to-image diffusion model. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023a.

Yuqing Wang, Yizhi Wang, Longhui Yu, Yuesheng Zhu, and Zhouhui Lian. Deepvecfont-v2: Exploiting transformers to synthesize vector fonts with higher quality. In *IEEE/CVF conference on computer vision and pattern recognition*, pp. 18320–18328. IEEE, 2023b.

Ronghuan Wu, Wanchao Su, Kede Ma, and Jing Liao. Iconshop: Text-guided vector icon synthesis with autoregressive transformers. *ACM Transactions on Graphics*, 42(6):1–14, 2023a.

Ronghuan Wu, Wanchao Su, and Jing Liao. Chat2svg: Vector graphics generation with large language models and image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23690–23700. IEEE, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023b.

Zeqing Xia, Bojun Xiong, and Zhouhui Lian. Vecfontsdf: Learning to reconstruct and synthesize high-quality vector fonts via signed distance functions. In *IEEE/CVF conference on computer vision and pattern recognition*, pp. 1848–1857. IEEE, 2023.

Shishi Xiao, Liangwei Wang, Xiaojuan Ma, and Wei Zeng. Typedance: Creating semantic typographic logos from image through personalized generation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2024.

Ximing Xing, Chuang Wang, Haitao Zhou, Jing Zhang, Qian Yu, and Dong Xu. Diffsketcher: Text guided vector sketch synthesis through latent diffusion models. *Advances in Neural Information Processing Systems*, 36:15869–15889, 2023.

Ximing Xing, Haitao Zhou, Chuang Wang, Jing Zhang, Dong Xu, and Qian Yu. Svgdreamer: Text guided svg generation with diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4546–4555. IEEE, 2024.

Ximing Xing, Juncheng Hu, Guotao Liang, Jing Zhang, Dong Xu, and Qian Yu. Empowering llms to understand and generate complex vector graphics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19487–19497. IEEE, 2025.

Zhenhua Yang, Dezhi Peng, Yuxin Kong, Yuyi Zhang, Cong Yao, and Lianwen Jin. Fontdiffuser: One-shot font generation via denoising diffusion with multi-scale content aggregation and style contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 6603–6611, 2024.

Supplementary Contents

A Additional Experiments **13**

A.1 Effect of Appearance Low-Pass Filtering 14

A.2 Generalization across Fonts 14

A.3 Semantic Controllability 14

A.4 Hard Cases 14

A.5 Glyph-to-Image Visualizations 14

A.6 Visualizations of Optimization Trajectories 14

B Glyph-to-Image Dataset **14**

B.1 Single-Character 15

B.2 Multi-Character 16

A Additional Experiments

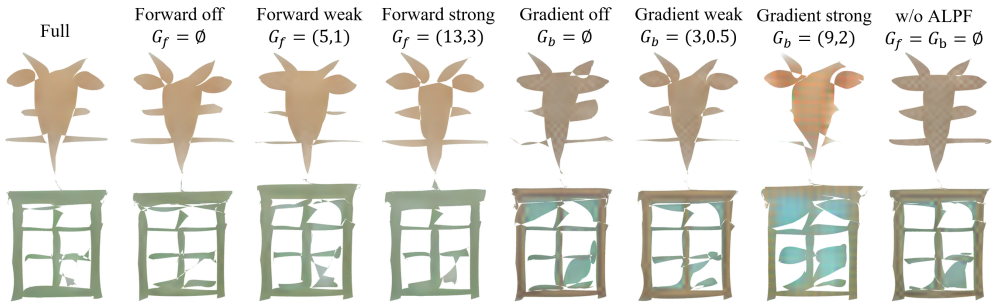

Figure S1: **Effect of appearance low-pass filtering.** We vary the forward appearance filter G_f and backward gradient filter G_b, where each pair denotes (k, σ).



Figure S2: **Generalization across fonts.** Qualitative results initialized from different Chinese fonts.

A.1 Effect of Appearance Low-Pass Filtering

We analyze ALPF by independently varying the forward appearance filter G_f and the backward gradient filter G_b. The selected setting uses $G_f = (9, 2)$ and $G_b = (5, 1)$, where each pair denotes (k, σ). As shown in Figure S1, insufficient filtering introduces irregular local patterns and structured color artifacts, whereas overly strong filtering suppresses semantic appearance formation. The selected setting provides a better balance, yielding smoother and more stable appearance optimization.

A.2 Generalization across Fonts

We further evaluate Glyph-to-Image under different Chinese font initializations. Figure S2 shows results on representative font styles, while Figure S3 extends the evaluation to eight additional fonts. For compact presentation, we denote them as F1–F8: F1 (LiuJianMaoCao), F2 (LongCang), F3 (MaShanZheng), F4 (ZhiMangXing), F5 (LXGWWenKai), F6 (HuaGuangBuDouTi), F7 (AnJingChenXingShu), and F8 (LuoGuoChengMaoBiXiaoXing).

Across different font initializations, the generated glyphs preserve font-specific variations in stroke shape and overall structure while incorporating consistent semantic cues. These results suggest that the proposed optimization is not tied to a specific font style and generalizes across a range of Chinese glyph initializations.

A.3 Semantic Controllability

We further evaluate the response of Glyph-to-Image to different semantic concepts while keeping the source glyph fixed. As shown in Figure S4, the same character can be adapted toward distinct concepts through corresponding changes in geometry and appearance, while retaining its recognizable glyph structure. This demonstrates that the learned representation is responsive to semantic conditioning rather than tied to a fixed visual interpretation.

A.4 Hard Cases

Figure S5 presents representative challenging cases. The extent of semantic adaptation can vary with the structural flexibility of the source glyph. Under the same semantic concept, simple characters with fewer strokes generally exhibit more limited geometric deformation, since fewer glyph parameters are available to accommodate semantic changes. As a result, their semantic expression can be more constrained than that of structurally richer glyphs.

A.5 Glyph-to-Image Visualizations

We provide additional qualitative comparisons on diverse Chinese glyph-concept pairs in Figure S6.

A.6 Visualizations of Optimization Trajectories

Figures S7 and S8 visualize the optimization trajectories of single- and multi-character generation. As optimization proceeds, the source glyph structures are progressively adapted toward the target concepts, with semantic geometry and appearance gradually emerging over iterations. For multi-character expressions, each constituent glyph follows the same independent optimization process under the shared semantic concept, showing similar progressive adaptation across different expression lengths.

B Glyph-to-Image Dataset

We provide the complete composition of the Glyph-to-Image Dataset in Tables S1–S4. The dataset consists of separate single-character and multi-character subsets, where each sample pairs a Chinese glyph or expression with an English semantic concept.

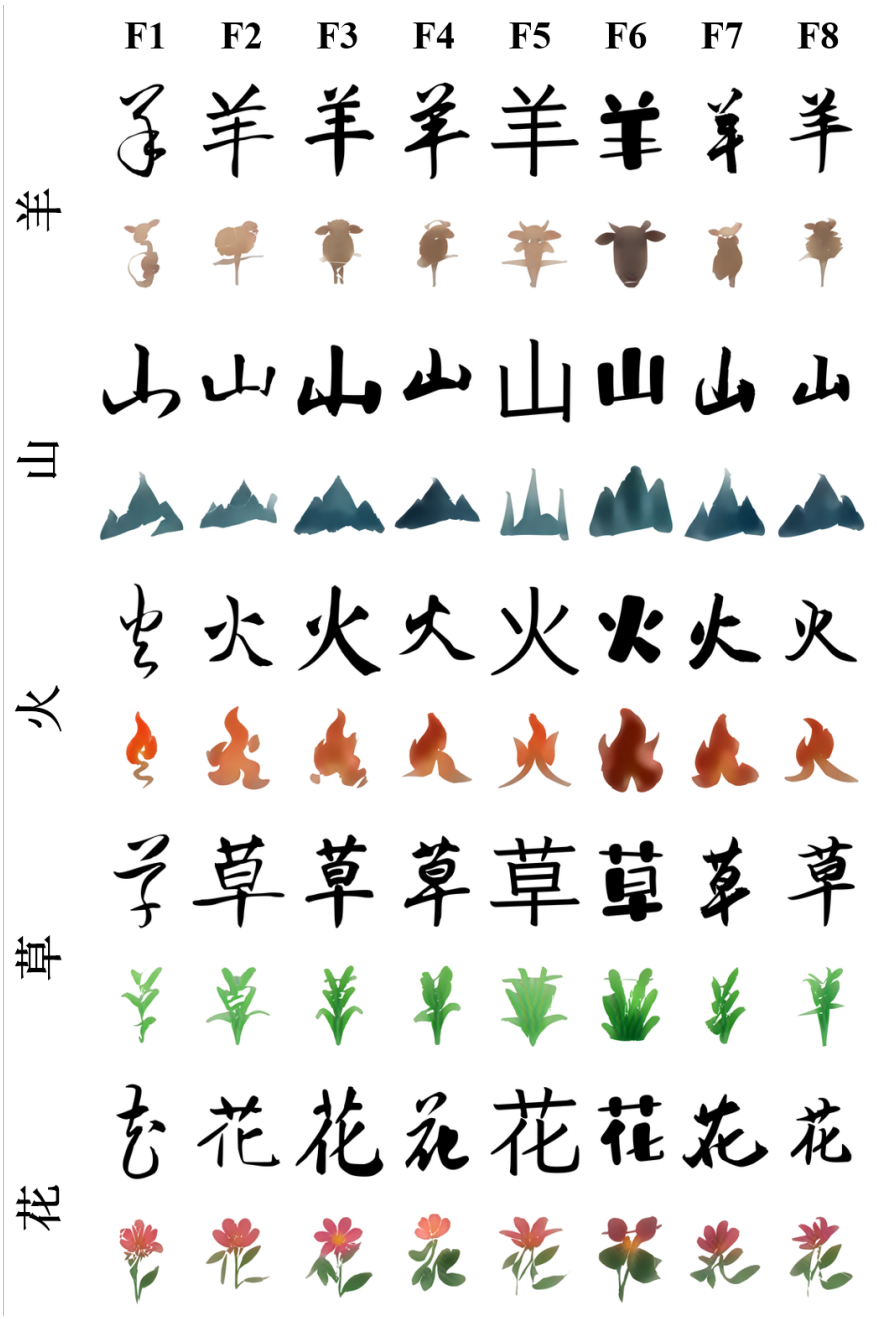

Figure S3: **Generalization across different fonts.** Results under eight font initializations (F1–F8).

B.1 SINGLE-CHARACTER

The single-character subset contains 90 glyph–concept pairs. According to the structural complexity of the source glyphs, we divide them into three groups: *simple*, *medium*, and *complex*, with 30 characters in each group. Table S1 lists the complete single-character subset, with the three columns corresponding to the three complexity levels.

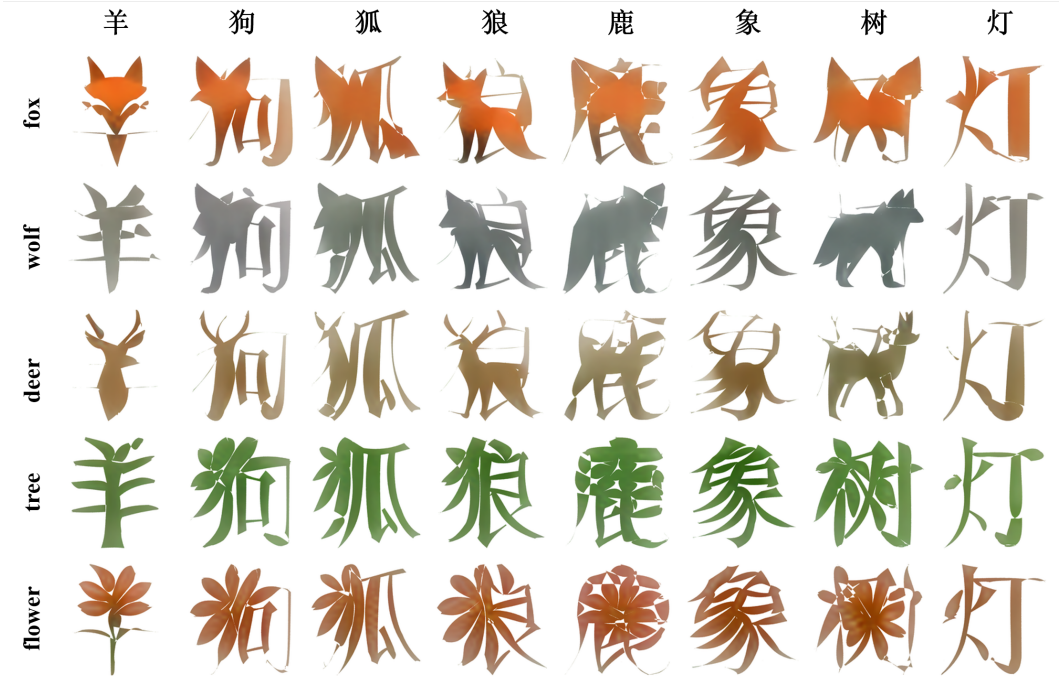

Figure S4: **Extended semantic controllability.** Additional results for different source glyphs under multiple semantic concepts.

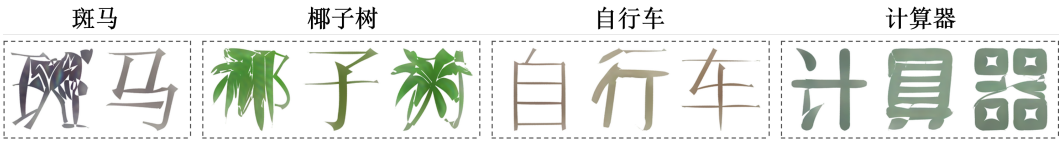

Figure S5: **Representative hard cases.** Examples where source glyphs with limited structural flexibility exhibit more constrained semantic deformation.

B.2 Multi-Character

The multi-character subset contains two-, three-, and four-character expressions. Each length contains 90 expression–concept pairs, resulting in 270 multi-character samples in total. The complete two-, three-, and four-character subsets are provided in Tables S2, S3, and S4, respectively.



Figure S6: **Visualizations and Comparisons of Glyph-to-Image.** The same source glyph under different semantic concepts.



Figure S7: **Single-character optimization trajectories.** Intermediate results at different optimization steps illustrate the progressive semantic adaptation of individual glyphs.



Figure S8: **Multi-character optimization trajectories.** Rows show two-, three-, and four-character expressions, and columns show their intermediate results at different optimization steps.

ID	Glyph	Concept	ID	Glyph	Concept	ID	Glyph	Concept
1	人	person	31	狗	dog	61	鹰	eagle
2	刀	knife	32	猫	cat	62	燕	swallow
3	弓	bow	33	猪	pig	63	蝶	butterfly
4	马	horse	34	鸡	chicken	64	蜂	bee
5	牛	cow	35	鸭	duck	65	龟	turtle
6	羊	sheep	36	鹅	goose	66	鲸	whale
7	鸟	bird	37	兔	rabbit	67	蟹	crab
8	鱼	fish	38	狼	wolf	68	虾	shrimp
9	木	wood	39	狐	fox	69	鳄	crocodile
10	石	stone	40	鹿	deer	70	鹤	crane
11	山	mountain	41	熊	bear	71	藕	lotus root
12	水	water	42	象	elephant	72	蕉	banana plant
13	火	fire	43	树	tree	73	蘑	mushroom
14	云	cloud	44	林	forest	74	薯	potato
15	雨	rain	45	桥	bridge	75	藻	algae
16	风	wind	46	船	boat	76	礁	reef
17	花	flower	47	房	house	77	露	dew
18	草	grass	48	屋	house	78	霞	rosy clouds
19	竹	bamboo	49	窗	window	79	霜	frost
20	瓜	melon	50	桌	table	80	鞭	whip
21	果	fruit	51	椅	chair	81	缆	cable
22	舟	boat	52	杯	cup	82	壶	pot
23	门	door	53	书	book	83	橱	cabinet
24	车	car	54	笔	pen	84	篱	fence
25	衣	clothes	55	钟	clock	85	墙	wall
26	伞	umbrella	56	镜	mirror	86	瓣	petal
27	豆	bean	57	鞋	shoes	87	巢	nest
28	米	rice	58	帽	hat	88	瓢	ladle
29	茶	tea	59	裙	skirt	89	篮	basket
30	月	moon	60	灯	lamp	90	檐	eaves

Table S1: **Complete single-character subset of the Glyph-to-Image Dataset.** The subset contains 90 glyph–concept pairs, divided into simple, medium, and complex groups with 30 samples each.

ID	Glyph	Concept	ID	Glyph	Concept	ID	Glyph	Concept
1	狮子	lion	31	玫瑰	rose	61	水杯	drinking cup
2	老虎	tiger	32	荷花	lotus flower	62	茶壶	teapot
3	熊猫	panda	33	菊花	chrysanthemum	63	雨伞	umbrella
4	狐狸	fox	34	兰花	orchid	64	书包	backpack
5	斑马	zebra	35	梅花	plum blossom	65	铅笔	pencil
6	河马	hippopotamus	36	桃花	peach blossom	66	毛笔	calligraphy brush
7	犀牛	rhinoceros	37	牡丹	peony	67	尺子	ruler
8	骆驼	camel	38	竹子	bamboo	68	剪刀	scissors
9	松鼠	squirrel	39	松树	pine tree	69	锤子	hammer
10	海豹	seal	40	柳树	willow tree	70	扳手	wrench
11	海狮	sea lion	41	枫树	maple tree	71	钥匙	key
12	鲨鱼	shark	42	椰树	palm tree	72	眼镜	eyeglasses
13	海豚	dolphin	43	苹果	apple	73	帽子	hat
14	章鱼	octopus	44	香蕉	banana	74	裙子	skirt
15	乌贼	squid	45	葡萄	grapes	75	衬衫	shirt
16	水母	jellyfish	46	西瓜	watermelon	76	沙发	sofa
17	龙虾	lobster	47	草莓	strawberry	77	桌子	table
18	螃蟹	crab	48	樱桃	cherry	78	椅子	chair
19	企鹅	penguin	49	柠檬	lemon	79	衣柜	wardrobe
20	孔雀	peacock	50	芒果	mango	80	枕头	pillow
21	鸽子	pigeon	51	台灯	desk lamp	81	汽车	car
22	天鹅	swan	52	手表	wristwatch	82	火车	train
23	麻雀	sparrow	53	闹钟	alarm clock	83	飞机	airplane
24	燕子	swallow	54	相机	camera	84	轮船	ship
25	鹦鹉	parrot	55	手机	smartphone	85	山峰	mountain peak
26	蜗牛	snail	56	电脑	computer	86	河流	river
27	蝴蝶	butterfly	57	键盘	keyboard	87	瀑布	waterfall
28	蜜蜂	bee	58	鼠标	computer mouse	88	海浪	ocean wave
29	蚂蚁	ant	59	耳机	headphones	89	云朵	cloud
30	蚯蚓	earthworm	60	音箱	speaker	90	雪花	snowflake

Table S2: **Complete two-character subset of the Glyph-to-Image Dataset.** The subset contains 90 two-character expression–concept pairs.

ID	Glyph	Concept	ID	Glyph	Concept	ID	Glyph	Concept
1	长颈鹿	giraffe	31	向日葵	sunflower	61	降落伞	parachute
2	北极熊	polar bear	32	蒲公英	dandelion	62	潜水艇	submarine
3	大熊猫	giant panda	33	仙人掌	cactus	63	机器人	robot
4	变色龙	chameleon	34	郁金香	tulip	64	望远镜	telescope
5	啄木鸟	woodpecker	35	牵牛花	morning glory	65	显微镜	microscope
6	丹顶鹤	red-crowned crane	36	含羞草	mimosa plant	66	照相机	camera
7	火烈鸟	flamingo	37	三叶草	clover	67	摄像机	video camera
8	猫头鹰	owl	38	圣诞树	Christmas tree	68	电视机	television
9	眼镜蛇	cobra	39	百合花	lily	69	洗衣机	washing machine
10	响尾蛇	rattlesnake	40	康乃馨	carnation	70	电冰箱	refrigerator
11	甲壳虫	beetle	41	薰衣草	lavender	71	微波炉	microwave oven
12	萤火虫	firefly	42	火龙果	dragon fruit	72	电饭煲	rice cooker
13	独角仙	rhinoceros beetle	43	猕猴桃	kiwifruit	73	吸尘器	vacuum cleaner
14	金丝猴	golden snub-nosed monkey	44	哈密瓜	Hami melon	74	电风扇	electric fan
15	扬子鳄	Chinese alligator	45	菠萝蜜	jackfruit	75	加湿器	humidifier
16	大白鲨	great white shark	46	椰子树	coconut palm	76	饮水机	water dispenser
17	锤头鲨	hammerhead shark	47	苹果树	apple tree	77	咖啡机	coffee machine
18	座头鲸	humpback whale	48	无花果	fig	78	打印机	printer
19	抹香鲸	sperm whale	49	西红柿	tomato	79	扫描仪	scanner
20	绿海龟	green sea turtle	50	冰淇淋	ice cream	80	计算器	calculator
21	寄居蟹	hermit crab	51	自行车	bicycle	81	摩天轮	Ferris wheel
22	帝王蟹	king crab	52	摩托车	motorcycle	82	过山车	roller coaster
23	招潮蟹	fiddler crab	53	救护车	ambulance	83	红绿灯	traffic light
24	枯叶蝶	dead leaf butterfly	54	消防车	fire truck	84	信号灯	signal light
25	凤尾蝶	swallowtail butterfly	55	越野车	off-road vehicle	85	垃圾桶	trash bin
26	金龟子	scarab beetle	56	小轿车	sedan	86	消防栓	fire hydrant
27	竹节虫	stick insect	57	面包车	van	87	电话亭	telephone booth
28	螳螂虾	mantis shrimp	58	直升机	helicopter	88	珊瑚礁	coral reef
29	食蚁兽	anteater	59	滑翔机	glider	89	钟乳石	stalactite
30	穿山甲	pangolin	60	热气球	hot-air balloon	90	火山口	volcanic crater

Table S3: **Complete three-character subset of the Glyph-to-Image Dataset.** The subset contains 90 three-character expression–concept pairs.

ID	Glyph	Concept	ID	Glyph	Concept	ID	Glyph	Concept
1	现炸油条	fried dough sticks	31	软毛扫把	broom	61	家用门锁	door lock
2	芝麻汤圆	sesame-filled glutinous rice balls	32	盒装抽纸	tissue box	62	床头台灯	bedside lamp
3	香脆薯条	French fries	33	洗衣肥皂	laundry soap bar	63	电吹风机	hair dryer
4	芝士汉堡	cheeseburger	34	马桶刷子	toilet brush	64	手持熨斗	clothes iron
5	意式披萨	pizza	35	家用牙刷	toothbrush	65	电动剃刀	electric shaver
6	奶油蛋糕	cream cake	36	浴室毛巾	bath towel	66	游戏手柄	game controller
7	黄油面包	butter bread	37	运动球鞋	sneaker	67	充电插头	power adapter
8	酥皮蛋挞	egg tart	38	棉布手套	cotton gloves	68	五号电池	AA battery
9	香甜玉米	corn on the cob	39	针织围巾	knitted scarf	69	烟雾探头	smoke detector
10	清蒸红薯	steamed sweet potato	40	运动发带	sports headband	70	家用门铃	doorbell
11	糖炒栗子	roasted chestnuts	41	皮革腰带	leather belt	71	小型音箱	small speaker
12	冰糖葫芦	candied hawthorn skewer	42	长筒棉袜	long cotton socks	72	小型风扇	small electric fan
13	珍珠奶茶	bubble tea	43	透明雨衣	transparent raincoat	73	共享单车	shared bicycle
14	热巧克力	hot chocolate	44	牛仔长裤	jeans	74	城市公交	city bus
15	罐装可乐	canned cola	45	随身钱包	wallet	75	快递货车	delivery truck
16	矿泉水瓶	mineral water bottle	46	金属拉链	metal zipper	76	儿童滑板	skateboard
17	爆米花桶	popcorn in a tub	47	线圈本子	spiral notebook	77	折叠童车	folding baby stroller
18	夹心饼干	sandwich cookie	48	塑料直尺	plastic ruler	78	拉杆皮箱	rolling suitcase
19	水果软糖	fruit gummies	49	绘图橡皮	drawing eraser	79	宠物牵绳	pet leash
20	香脆麻花	fried dough twist	50	彩色蜡笔	crayons	80	公园秋千	playground swing
21	平底煎锅	frying pan	51	金属书签	metal bookmark	81	儿童皮球	playground ball
22	陶瓷饭碗	ceramic bowl	52	透明胶带	transparent adhesive tape roll	82	彩色风筝	colorful kite
23	竹制筷子	bamboo chopsticks	53	塑料夹板	clipboard	83	红色龙虾	red lobster
24	厨房菜刀	Chinese kitchen cleaver	54	纸质信封	paper envelope	84	彩色水母	colorful jellyfish
25	木制砧板	wooden cutting board	55	落地衣架	freestanding coat rack	85	黑色蚂蚁	black ant
26	保鲜饭盒	food storage box	56	方格地毯	checkered rug	86	金色麦穗	golden wheat ear
27	厨房围裙	kitchen apron	57	遮光窗帘	curtains	87	木制小屋	wooden cabin
28	洗碗海绵	dishwashing sponge	58	布艺靠垫	fabric cushion	88	红色灯塔	red lighthouse
29	沥水菜篮	vegetable colander	59	家用插座	electrical wall socket	89	玻璃温室	glass greenhouse
30	平板拖把	flat mop	60	墙面挂钩	wall hook	90	古老城堡	old castle

Table S4: **Complete four-character subset of the Glyph-to-Image Dataset.** The subset contains 90 four-character expression–concept pairs.