This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SkyReels-Text: Fine-Grained Font-Controllable Text Editing for Poster Design

Yunjie Yu, Jingchen Wu, Junchen Zhu, Chunze Lin*, Guibin Chen
Skywork AI



Figure 1. **Font-controllable text editing.** SkyReels-Text modifies text within images using novel fonts guided by a single reference image.

Abstract

Artistic design, particularly poster design, often demands rapid yet precise modification of textual content while preserving visual harmony and typographic intent, especially across diverse font styles. Although modern image editing models have grown increasingly powerful, they still fall short in fine-grained, font-aware text manipulation, limiting their utility in professional workflows. To address this issue, we present SkyReels-Text, a novel font-controllable framework for precise poster text editing. Our method enables simultaneous editing of multiple text regions, each rendered in distinct typographic styles, while preserving the visual appearance of non-edited regions. Notably, our model requires neither font labels nor test-time fine-tuning: users can simply provide cropped glyph patches corresponding to their desired typography—even if the font is not included in any standard library. Extensive experiments on multiple benchmarks demonstrate that SkyReels-Text achieves state-of-the-art performance in both text fidelity and visual realism, offering unprecedented control over font families and stylistic nuances. This work bridges the gap between general-purpose image editing and professional-grade typographic design. Code and models are publicly available at https://github.com/SkyworkAI/SkyReels-Text.

1. Introduction

Poster design is a cornerstone of visual communication in advertising, film, events, and social media. A key requirement in professional workflows is the ability to rapidly revise textual content—changing slogans, brand names, dates, or multilingual captions—while preserving the original visual harmony, typographic identity, and aesthetic intent. This demands not only accurate character rendering but also fine-grained control over font families and stylistic nuances, especially when handling complex scripts such as

*Corresponding author

handwritten text.

Although modern diffusion-based image editing models (e.g., FLUX [17, 18], Qwen-image [39], Seedream 4.0 [34]) have achieved impressive results in general-purpose image manipulation, they remain fundamentally limited in their ability to perform fine-grained, font-aware text editing. These models can incorporate reference glyphs as visual context, but they often fail to correctly edit the textual content or accurately transplant the reference typography into the target region. As a result, either the semantic content is corrupted, or the generated text does not faithfully align with the desired position and layout—undermining the typographic precision essential for professional design. This limitation highlights a critical gap between current AI-based editing tools and the fine-grained control demanded by real-world design workflows.

Concurrently, another line of research has focused on developing specialized text editing approaches to address the challenge of accurate visual text generation. Models such as FLUX-Text [19] and TextFLUX [40] inject explicit visual priors (e.g., rendered text content, positional masks) into powerful DiT backbones. Despite their effectiveness in spelling correctness, a critical drawback remains: they lack a mechanism to accept an arbitrary font style extracted from a user-provided visual patch. Consequently, they offer no fine-grained control over the output typography and cannot guarantee that the edited text will precisely mimic a given reference font—a paramount requirement in professional workflows such as poster design.

In this paper, we present SkyReels-Text, a novel framework for fine-grained, font-controllable text editing that enables precise textual content modification while preserving typographic style. Our method introduces a dual-stream visual conditioning mechanism that leverages user-provided glyph patches as explicit visual references, eliminating the need for font labels or test-time fine-tuning. By formulating text editing as a regional modification task with targeted visual conditioning, our approach provides unambiguous priors that ensure both semantic accuracy and typographic fidelity. This design enables the simultaneous editing of multiple text regions with heterogeneous typographic styles, supporting the complex multilingual layouts common in professional poster design. Unlike previous approaches that depend on text prompts alone or internal glyph priors, SkyReels-Text utilizes explicit visual cues to achieve unprecedented control over font families, weights, and stylistic nuances, effectively meeting the needs of professional-grade typographic design.

The key contributions of this work are fourfold:

- A novel visual conditioning strategy that uses reference glyph patches to enable fine-grained, font-controllable text editing without requiring font labels.
- A VLM-based OCR system capable of recognizing and localizing ornamental or calligraphic fonts, which are highly challenging for conventional OCR methods.
- A dedicated pipeline integrating LLM text replacement, local editing, and dual-verification to ensure content-style decoupling, yielding a dataset of 100K aligned original—edited image pairs.
- A comprehensive evaluation on both public and in-house text editing benchmarks, achieving state-of-the-art performance in both semantic accuracy and typographic style fidelity.

2. Related Work

2.1. Image Editing Models

In recent years, with the advancement of image generation models, image editing has achieved increasingly sophisticated capabilities, including semantic manipulations such as object insertion, style transfer, and inpainting. FLUX.1 Kontext [18] extends the image generation model FLUX [17] into a robust image editing framework by treating the input image as contextual information. Meanwhile, Qwen-Image-Edit [39], built on the Qwen-Image foundation model, specializes in precise image manipulation with native bilingual (Chinese and English) text rendering capabilities. Recently, multimodal foundation models have unified vision and language to enable open-domain, instruction-guided image editing. Models such as Gemini 2.5 Flash Image [11], GPT-4o [25], and Seedream 4.0 [34] support image editing through natural language instructions and the fusion of multiple inputs, showcasing strong capabilities in multi-turn editing and large-scale conversational refinement. However, a key limitation of these general-purpose models is their primary focus on broad image editing tasks. Consequently, they lack the specific robustness required for reliable text editing, especially in font transfer scenarios. This fundamentally limits their capacity for the fine-grained, typographically precise editing demanded in professional artistic design workflows.

2.2. Text Editing Models

Specialized text synthesis methods have sought to bridge the gap between general image generation and typographic precision. Early works like Imagen [31] and Ediff-i [2] employed large-scale language models (e.g., T5-XXL [5]) to improve textual spelling accuracy. Subsequent approaches introduced glyph-aware conditioning: GlyphDraw [20] and GlyphControl [42] used glyph images to guide character rendering, while TextDiffuser [4] leveraged OCR-derived segmentation masks. Multilingual extensions such as AnyText [37] and font-aware systems like DreamText [38] further enriched control over script and style. FLUX-Text [19] and TextFLUX [40], built on the FLUX foundation model, inject glyph priors (rendered glyphs and position masks)

into the DiT backbone to enhance its text generation capability. In contrast, our work departs from this paradigm: rather than depending on text prompts alone or internal glyph priors, SkyReels-Text leverages user-provided glyph patches as visual references to achieve font-controllable text editing.

3. Method

3.1. Preliminaries

Given a source poster image X_{src}, a set of text instances $T = \{(t_1, r_1, g_1), (t_2, r_2, g_2), \ldots, (t_n, r_n, g_n)\}$, and a text prompt y containing the target text provided by the user, our goal is to generate an edited image X_{out} where the original text regions are replaced with the target text in visually consistent fonts. Here, t_i represents the i-th target text to be edited within region r_i using the reference glyph g_i, and n denotes the total number of text instances to be edited.

3.2. Data Processing

Data Curation. Most text editing methods rely on synthetic data generated by rendering text with standard system fonts (e.g., Arial, SimSun) onto simple backgrounds. While scalable, this approach suffers from three critical limitations: (1) limited font diversity, excluding rich real-world styles such as calligraphic or custom designs; (2) unrealistic backgrounds, lacking the texture, lighting, and layout complexity found in professional designs; and (3) unnatural glyph-background integration, creating a significant domain gap with real-world editing scenarios.

To address these issues, we curate a high-quality dataset from real-world posters collected from design platforms and public repositories. We apply a multi-stage data curation pipeline: (i) filtering out low-resolution posters; (ii) discarding images without legible text; and (iii) employing a custom-trained vision-language model (VLM) as a robust OCR system to accurately detect and recognize text—particularly ornamental or calligraphic fonts that are highly challenging for conventional OCR methods. This real-world, data-centric approach ensures that our model learns authentic font-background interactions and generalizes well to professional design contexts.

VLM-Based OCR Model. Existing OCR systems, ranging from practical models such as PP-OCRv4 [27] and PP-OCRv5 [7] to the more powerful VLM-based approach PaddleOCR-VL [6], are typically optimized for clean, regular typefaces and scene text. Thus, they fail to generalize to ornate calligraphic scripts or custom-designed letterforms that exhibit irregular strokes and non-standard spatial layouts. Given the complexity of recognizing non-standard fonts, which may be classified as icons or patterns, we introduce a specialized OCR model to enable simultaneous analysis from both visual and semantic perspectives. Specifically, we fine-tune the Qwen2.5-VL 7B [1] model, which excels at parsing non-standard textual patterns within images. This adaptation allows the model to better handle diverse font geometries and artistic variations.

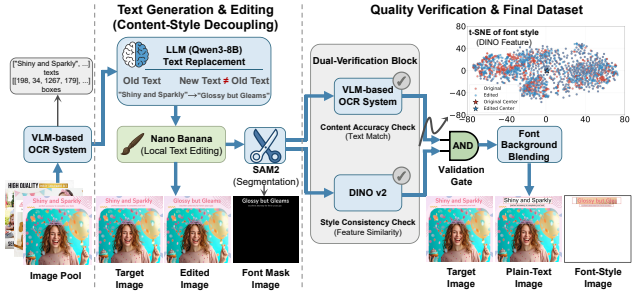


Figure 2. **Pipeline of content-style decoupling.** We utilize our VLM-based OCR and Qwen3-8B to obtain divergent text content, followed by local editing via Nano Banana and mask extraction via SAM2 to generate decoupled image pairs. A rigorous Dual-Verification Block ensures both content accuracy and style consistency. The t-SNE visualization of font styles verifies the style alignment as well as the high diversity of our dataset.

Data Generation with Content-Style Decoupling. A critical challenge in style transfer learning is content interference. Specifically, the model often conflates content and style features for certain characters within the text. To address this issue, we designed a dedicated data generation pipeline, as illustrated in Figure 2. First, we identify high-quality text instances in rich-text posters using our VLM-based OCR model. During this step, we filter out overlapping text regions. For the remaining text, we further employ Qwen3-8B [41] to generate semantically divergent replacement text. Crucially, we ensure the new word set is disjoint from the original text to maximize content decoupling. Next, we leverage the Nano Banana [11] editing model to replace the original content while preserving all typographical attributes, including font, color, and local alignment, thereby ensuring the retention of fine-grained style patterns. To rigorously validate the fidelity of this process, we implement a two-pronged verification strategy. First, we re-apply our VLM-based OCR model to confirm the content accuracy of the newly rendered text. Second, we extract DINO v2 features [26] from both the original and edited text regions to verify high feature-level similarity, confirming style preservation. This methodology produces perfectly aligned original–edited image pairs that share identical styles but contain maximally different content. Finally, this robust pipeline yields a high-quality dataset of 100K samples, comprising paired target images and reference fonts where the text content and style are completely decoupled.

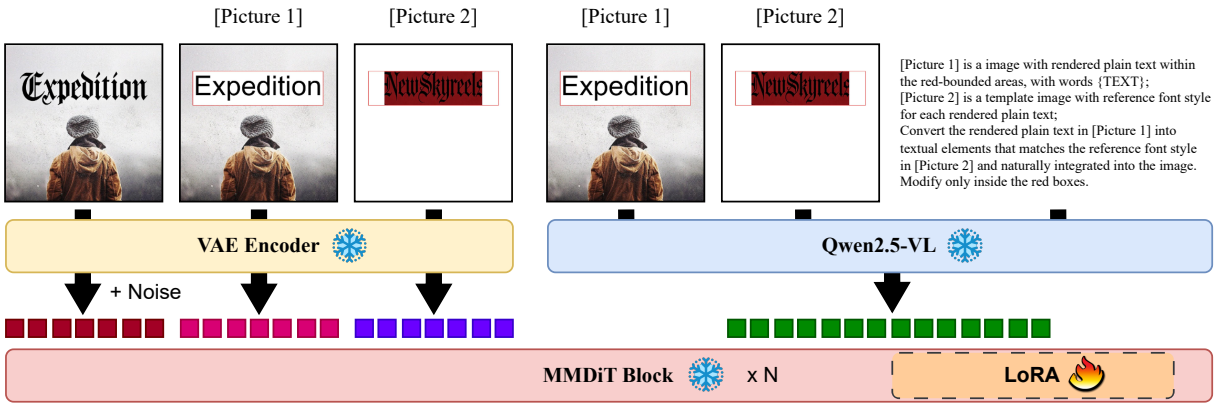


Figure 3. **Overview of the proposed method.** SkyReels-Text employs a dual-stream visual conditioning mechanism to decouple textual content and typographic style. Specifically, the VAE encoder extracts visual features from the plain-text reference and the glyph map, which are then concatenated with the noisy image latent. Simultaneously, Qwen2.5-VL processes the multimodal editing instructions. The combined sequence is fed into MMDiT blocks equipped with trainable LoRA modules, enabling zero-shot style transfer.

3.3. Font-Controllable Text Editing

General image editing models typically rely on textual prompts to modify text within an image. However, such prompt-driven approaches suffer from inherent limitations in fine-grained precision, often resulting in failed modifications, particularly the inaccurate reproduction of original typographic styles. To address this, we formulate font-controllable text editing as a regional modification task, where the model is conditioned to modify a user-specified region r_i based on a target text t_i and a reference glyph g_i.

To overcome the limitations of purely text-based guidance, we propose a diffusion model with a dual-stream visual conditioning mechanism that provides explicit, unambiguous priors for precise typographic control. As demonstrated in Figure 3, we decouple content and style into two distinct conditions. First, we generate a plain-text reference X_{text} (Picture 1) by inpainting the masked text regions in the original input with the target text t_i, thereby establishing the desired textual content and layout within the visual context. Second, we create a dedicated glyph map X_{glyph} (Picture 2) by rendering the corresponding glyph g_i onto a canvas matching the input size. By conditioning the model on both X_{text} and X_{glyph}, we ensure that the generated output achieves precise alignment with the intended semantic content and the specified typographic style.

Instead of injecting glyph priors via parameter-heavy embedding modules or ControlNet, we directly use a frozen VAE encoder to extract features from the plain-text reference X_{text} and the glyph map X_{glyph}. This simplifies the training process while ensuring effective extraction and compression of visual information. These VAE features are subsequently injected into the model by concatenating them with the noisy image latent z_t along the sequence dimension:

$$Z_{\text{in}} = \text{Concat}(z_t, \text{VAE}(X_{\text{text}}), \text{VAE}(X_{\text{glyph}})).$$

This design allows the model to directly attend to text content and font exemplars, enabling zero-shot style transfer.

3.4. Text-Region Weighted Loss

Generally, text regions occupy only a small fraction of the overall image, typically less than 10%, which means that applying uniform optimization across all pixels can hinder the convergence of text generation during training. To ensure high fidelity in the edited text regions while preserving background integrity, we introduce a text-aware weighted reconstruction loss. Let X_{gt} be the ground-truth edited image and $\hat{X}$ be the model prediction. We compute the standard diffusion loss (e.g., $\mathcal{L}_2$ or flow matching loss) but reweight it spatially using the text mask M:

$$\mathcal{L} = \mathbb{E}\left[\|X_{\text{gt}} - \hat{X}\|_2^2 \odot (1 + \lambda \cdot M)\right],$$

where $\lambda > 0$ is a hyperparameter that amplifies the loss in text regions, and M is a tight text-region mask generated by SAM2 [30]. This simple yet effective strategy focuses optimization on glyph accuracy without requiring a separate perceptual encoder or multi-stage training. In practice, we set $\lambda = 5$, which yields an optimal balance between text fidelity and background consistency.

3.5. Distillation for Fast Inference

To accelerate inference without sacrificing font-controllable editing quality, we distill SkyReels-Text using the Improved Distribution Matching Distillation (DMD2) paradigm [43]. We adapt DMD2 to our architecture, producing an 8-step

Table 1. Performance comparison on SkyReels-Text benchmark.

Method	Sen. Acc↑	NED↑	Spatial↑	DINO↑	B-PSNR↑
Nano Banana [11]	0.7290	0.9195	0.7011	0.8125	28.78
Seedream 4.0 [34]	0.7772	0.9348	0.6844	0.7629	25.50
FLUX-Kontext-Pro [18]	0.6390	0.8458	0.5063	0.8130	27.21
Qwen-Image-Edit [39]	0.7760	0.9337	0.5845	0.8209	25.88
FLUX-Text [19]	0.8266	0.9352	0.7503	0.6679	**34.53**
Calligrapher [21]	0.6404	0.8811	0.7281	0.7162	24.26
SkyReels-Text	**0.8334**	**0.9502**	**0.7506**	**0.8503**	34.17

student model that closely matches the teacher's output distribution while preserving fine-grained font awareness. This reduces the number of sampling steps and eliminates the need for classifier-free guidance, achieving a speedup of over $10\times$ during inference.

4. Experiments

4.1. Implementation Details

We adopt the Qwen-Image-Edit [39] model as our baseline to ensure high-fidelity and reliable editing performance. To preserve the strong capabilities of the underlying foundation model, we fine-tune it using LoRA with a rank of 64. We set the batch size to 64, employ the AdamW optimizer with a learning rate of 10^{-4}, and train for 2 epochs, consuming about 512 A100-hours. For the VLM-based OCR model, we fine-tune Qwen2.5-VL 7B [1] with a learning rate of 10^{-4} using approximately 72 A100-hours, enabling precise and pixel-localized detection of textual content.

4.2. Evaluation Dataset

We evaluate the performance of SkyReels-Text across artistic, general, and handwritten text synthesis using one in-house benchmark and two categories of public benchmarks. Specifically, we meticulously curate a benchmark of 200 diverse text samples encompassing a wide range of typographic styles, referred to as the SkyReels-Text Benchmark. For assessing the accuracy and visual fidelity of synthesized text, we utilize the AnyText benchmark [37], which comprises 1,000 images from the Wukong [12] (Chinese) and LAION [33] (English) datasets. To validate the model's generalization capabilities for handwritten text generation, we employ two widely used datasets: IAM [22] and CVL [16]. For the IAM dataset, we select text lines from 161 writers following the protocol established by CSA-GAN [14]. For the CVL dataset, we construct the test set from 27 writers following the standard CVL split.

4.3. Evaluation Metrics

For the SkyReels-Text benchmark, we evaluate performance across four key dimensions: (1) Text fidelity: Sentence Accuracy (Sen. Acc.) and Normalized Edit Distance (NED = $1 -$ Edit Distance), computed using our fine-tuned Qwen2.5-VL 7B model; (2) Layout alignment: spatial IoU between OCR-detected and ground-truth text boxes; (3) Font style adherence: DINO similarity between generated and reference text regions; and (4) Background preservation: B-PSNR, computed on non-text regions to quantify structural fidelity. For the AnyText benchmark [37], we adopt Sen. Acc. and NED (evaluated via DuGuangOCR [23] for fair comparison) for text accuracy, along with Fréchet Inception Distance (FID) [35] and Learned Perceptual Image Patch Similarity (LPIPS) [45] for distributional and perceptual image quality. For handwritten text generation, we use Handwriting Distance (HWD) [29] to assess style fidelity, alongside Inception Score (IS) [32] and Geometry Score (GS) [15] for overall visual quality.

4.4. Comparison with State-of-the-Art Methods

Quantitative Results. We evaluate SkyReels-Text on our comprehensive benchmark against the current state-of-the-art (SOTA) commercial and open-source methods, including Nano Banana [11], Seedream 4.0 [34], FLUX-Kontext-Pro [18], Qwen-Image-Edit [39], FLUX-Text [19], and Calligrapher [21]. As shown in Table 1, SkyReels-Text achieves high text fidelity with a Sen. Acc. of 83.34% and an NED of 95.02%. Crucially, it demonstrates high layout alignment and font style adherence, as evidenced by the spatial IoU and DINO scores, respectively. Furthermore, the model's ability to maintain background integrity, quantified by B-PSNR, is comparable to that of FLUX-Text [19]. Overall, compared to these competing methods, SkyReels-Text achieves SOTA performance, highlighting its superiority in text quality and font consistency.

Next, we evaluate our model on the AnyText benchmark and compare it against the current SOTA methods, including DiffSTE [13], TextDiffuser [4], DiffUTE [3], AnyText [37], TextCtrl [44], AnyText2 [36], FLUX-Text [19], and Calligrapher [21]. As shown in Table 2, SkyReels-Text achieves the best performance across all metrics (Sen. Acc., NED, FID, and LPIPS) on both the Chinese and English datasets. Specifically, SkyReels-Text achieves an English Sen. Acc. of 85.36% and a Chinese Sen. Acc. of 77.10%, surpassing FLUX-Text by 1.37% and 7.50%, respectively. Regarding the NED metric, it also attains the highest scores of 94.06% for English and 87.64% for Chinese. Furthermore, SkyReels-Text yields the lowest FID and LPIPS scores, indicating superior distribution-level and perceptual alignment with real images.

Finally, we evaluate our model on handwritten text datasets. Although we do not explicitly train on dedicated handwritten data, our model can still generate text images that faithfully mimic the target handwritten style when provided with reference images. As shown in Table 3, we compare our model against the latest handwritten text generation models. SkyReels-Text achieves highly com-

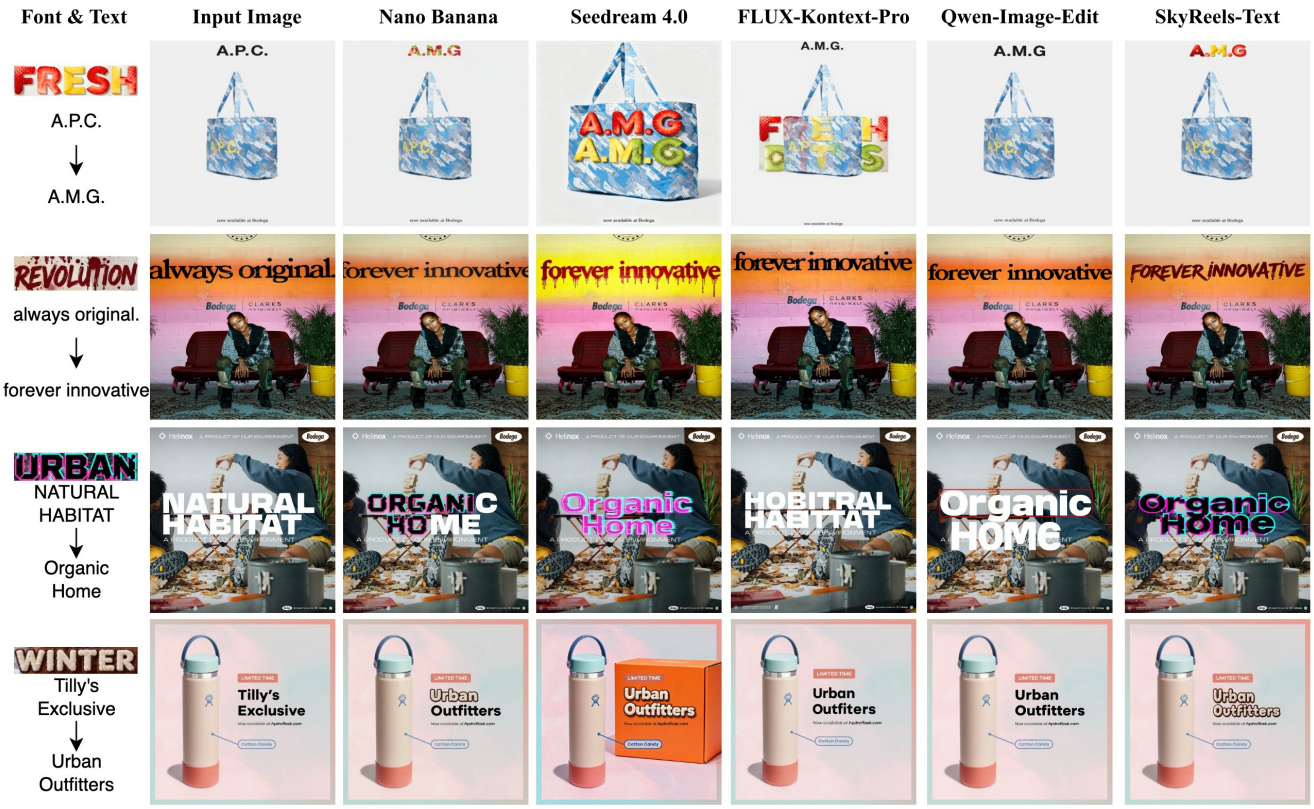


Figure 4. **Comparison with SOTA commercial image editing models in single-font editing.** The leftmost column presents the reference font image and the source/target plain text. SkyReels-Text produces edits that faithfully follow the target typography while preserving the background structure and content intact.

Table 2. Performance comparison of different methods on AnyText benchmark for English and Chinese datasets.

Methods	English				Chinese			
	Sen. Acc↑	NED↑	FID↓	LPIPS↓	Sen. Acc↑	NED↑	FID↓	LPIPS↓
DiffSTE [13]	0.4523	0.7814	52.74	0.1816	0.0363	0.1226	57.49	0.1276
TextDiffuser [4]	0.5176	0.7618	29.76	0.1564	0.0559	0.1218	34.19	0.1252
DiffUTE [3]	0.4054	0.7005	25.35	0.1640	0.2978	0.5744	29.08	0.1745
Anytext [37]	0.6843	0.8588	21.59	0.1106	0.6476	0.8210	20.01	0.0943
TextCtrl [44]	0.5853	0.8146	35.73	0.1978	0.3580	0.6084	49.79	0.2298
Anytext2 [36]	0.7915	0.9100	29.76	0.1734	0.7022	0.8420	26.52	0.1444
FLUX-Text [19]	0.8419	0.9400	13.85	0.0729	0.7132	0.8510	13.68	0.0541
Calligrapher [21]	0.7761	0.9011	22.41	0.0851	0.4369	0.6884	24.93	0.0797
Baseline	0.8292	0.8708	35.62	0.2499	0.6735	0.8323	19.26	0.1239
SkyReels-Text	**0.8536**	**0.9406**	**6.12**	**0.0246**	**0.7710**	**0.8764**	**5.44**	**0.0192**

petitive results compared to models specifically designed and trained on handwritten data. In particular, compared to DiffBrush [9], our method attains lower HWD scores of 1.32 and 0.89 on the IAM and CVL datasets, respectively—indicating superior style imitation capabilities. This demonstrates that SkyReels-Text can generate highly realistic handwritten text line images without any additional training, relying solely on the provided reference samples.

Qualitative Results. To further evaluate our model, we present qualitative comparisons against both commercial and open-source models. First, we select four SOTA commercial image editing models: Nano Banana, Seedream 4.0,



Figure 5. **Comparison with SOTA open-source models.** Visual results of scene text editing for both English and Chinese.

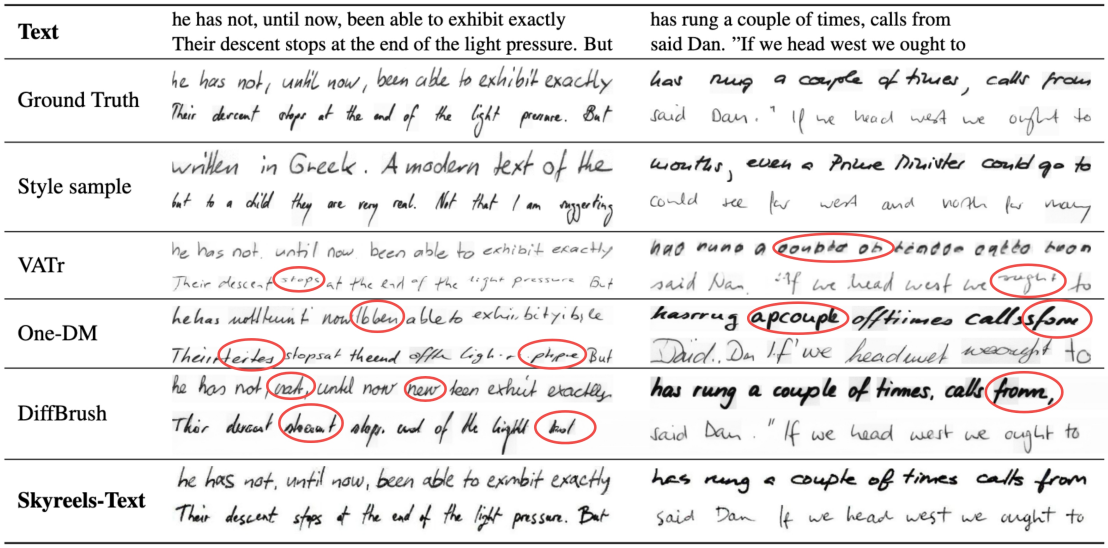


Figure 6. **Handwritten text generation.** Qualitative comparison with SOTA approaches using challenging exemplars.

FLUX-Kontext-Pro, and Qwen-Image-Edit. To demonstrate the superiority of SkyReels-Text, we evaluate both single-font and multi-font editing, as shown in Figure 4 and the supplement. SkyReels-Text successfully transfers reference font styles and performs accurate text edits in both settings, whereas the competing models either fail to apply the intended edits or struggle to preserve the background content. Second, we compare our model against open-source models. As shown in Figure 5, SkyReels-Text preserves the original font style while accurately modifying the textual content, achieving precise typographic consistency for both English and Chinese. Evaluations on samples beyond the poster domain demonstrate our model's robustness to diverse real-world images.

In addition to evaluating the text editing, we present a detailed comparison focusing on pure font editing. Specifically, we select challenging handwritten fonts as exemplars. As shown in Figure 6, although our model is not explicitly trained on handwritten datasets, it successfully generalizes to handwritten text using only a single reference. Even

Table 3. Performance comparison on IAM and CVL datasets. For simplicity, we omit $\times 10^{-3}$ in all GS metrics.

Datasets	Method	Shot	HWD↓	IS↑	GS↓
IAM	TS-GAN [10]	one	2.11	1.76	28.7
	CSA-GAN [14]	few	2.25	1.74	16.2
	VATr [28]	few	1.87	1.69	14.5
	DiffusionPen [24]	few	1.72	1.83	6.42
	One-DM [8]	one	1.80	1.82	8.42
	DiffBrush [9]	one	1.41	1.85	2.35
	SkyReels-Text	one	**1.32**	**1.90**	**1.26**
CVL	CSA-GAN [14]	few	1.72	1.48	671
	VAtR [28]	few	1.50	1.44	143
	DiffusionPen [24]	few	1.32	1.59	50.8
	One-DM [8]	one	1.47	1.46	129
	DiffBrush [9]	one	1.06	1.70	**29.6**
	SkyReels-Text	one	**0.89**	**1.71**	31.0

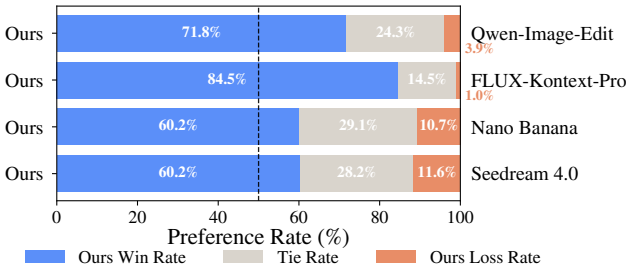


Figure 7. **User preference study.** SkyReels-Text outperforms all competitors in human preference.

compared to methods specifically designed and trained for handwritten text generation, SkyReels-Text achieves superior preservation of the font style and generates text that more closely aligns with the ground truth. We provide additional comparisons on Chinese handwritten text in the supplement. This strong generalization and zero-shot transfer capability obviates the need for extensive domain-specific fine-tuning, offering a flexible, efficient, and scalable solution for generating stylized text images.

User Study. To further validate the effectiveness of SkyReels-Text, we conducted a user study comparing our model with four commercial image editing systems: Qwen-Image-Edit, FLUX-Kontext-Pro, Nano Banana, and Seedream 4.0. Figure 7 indicates that SkyReels-Text consistently outperforms all competitors. In particular, it achieves clear advantages over Qwen-Image-Edit and FLUX-Kontext-Pro, with win rates of 71.8% and 84.5%, respectively. When compared to Nano Banana and Seedream 4.0, SkyReels-Text still maintains a substantial advantage, achieving a 60.2% win rate against a mere ∼10% loss rate. These results underscore that SkyReels-Text provides a more reliable and user-favored editing experience than leading commercial alternatives.

Table 4. Ablation experiments for Font Style Reference (FSR) and weight λ of Text-Region Weighted Loss. We use **boldface** and underlining to denote the best and second-best values.

FSR /λ	Sen. Acc↑	NED↑	Spatial↑	DINO↑	B-PSNR↑
×	**0.9327**	**0.9734**	0.6986	0.6995	33.61
✓	0.8334	0.9502	**0.7506**	**0.8503**	**34.17**
$\lambda = 0$	0.7998	0.9378	0.7501	0.8473	**34.28**
$\lambda = 1$	0.8126	0.9392	0.7492	0.8484	34.25
$\lambda = 2$	0.8267	0.9426	0.7505	0.8496	34.21
$\lambda = 5$	0.8334	0.9502	**0.7506**	**0.8503**	34.17
$\lambda = 10$	**0.8363**	**0.9558**	0.7503	0.8502	33.13

4.5. Ablation Studies

We conduct ablations on the SkyReels-Text benchmark to evaluate our key components: the font style reference and the text-region weighted loss.

Font Style Reference. As shown in Table 4, incorporating the Font Style Reference (FSR) substantially enhances the model's ability to align with the target style, as evidenced by a higher DINO score. In contrast, the setting without FSR yields better Sen. Acc and NED for two reasons: first, the task becomes simpler under single-condition generation; second, current OCR systems struggle to parse highly stylized text when FSR is enabled. Furthermore, the improved spatial score with FSR is driven by the fact that the reference image—cropped from the original—provides implicit spatial cues that facilitate accurate placement of generated text relative to the ground truth layout.

Text-Region Weighted Loss. Table 4 also evaluates the impact of the weight λ in the text-region weighted loss. Since text occupies only a small portion of the image, setting $\lambda > 0$ is essential to direct optimization toward glyph fidelity. Increasing λ leads to consistent improvements in Sen. Acc and NED, confirming the benefit of emphasizing text regions. While $\lambda = 10$ achieves the highest text accuracy, $\lambda = 5$ provides the optimal balance, maximizing spatial IoU and DINO scores—critical indicators of layout and style adherence. Thus, we set $\lambda = 5$ for the best trade-off between text convergence and background fidelity.

5. Conclusion

In this paper, we introduce SkyReels-Text, a novel framework for fine-grained, font-controllable text editing. It enables users to simultaneously modify multiple text regions, each guided by a user-provided reference glyph. We demonstrated that SkyReels-Text achieves state-of-the-art zero-shot performance while preserving background integrity, entirely bypassing the need for font labels or additional training. This work not only provides a highly practical solution for real-world design workflows but also paves the way for fine-grained, reference-driven visual generation.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 5

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2023. 2

[3] Haoxing Chen, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Changhua Meng, Huijia Zhu, Weiqiang Wang, et al. Diffute: Universal text editing diffusion model. *Advances in Neural Information Processing Systems*, 36:63062–63074, 2023. 5, 6

[4] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023. 2, 5, 6

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 2

[6] Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*, 2025. 3

[7] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025. 3

[8] Gang Dai, Yifan Zhang, Quhui Ke, Qiangya Guo, and Shuangping Huang. One-dm: One-shot diffusion mimicker for handwritten text generation. In *European Conference on Computer Vision*, pages 410–427. Springer, 2024. 8

[9] Gang Dai, Yifan Zhang, Yutao Qin, Qiangya Guo, Shuangping Huang, and Shuicheng Yan. Beyond isolated words: Diffusion brush for handwritten text-line generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19054–19064, 2025. 6, 8

[10] Brian Davis, Chris Tensmeyer, Brian Price, Curtis Wigington, Bryan Morse, and Rajiv Jain. Text and style conditioned gan for generation of offline handwriting lines. *arXiv preprint arXiv:2009.00678*, 2020. 8

[11] Google. Gemini 2.5 flash image. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image, 2025. 2, 3, 5

[12] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022. 5

[13] Jiabao Ji, Guanhua Zhang, Zhaowen Wang, Bairu Hou, Zhifei Zhang, Brian Price, and Shiyu Chang. Improving diffusion models for scene text editing with dual encoders. *arXiv preprint arXiv:2304.05568*, 2023. 5, 6

[14] Lei Kang, Pau Riba, Marcal Rusinol, Alicia Fornes, and Mauricio Villegas. Content and style aware generation of text-line images for handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 8846–8860, 2021. 5, 8

[15] Valentin Khrulkov and Ivan Oseledets. Geometry score: A method for comparing generative adversarial networks. In *International conference on machine learning*, pages 2621–2629. PMLR, 2018. 5

[16] Florian Kleber, Stefan Fiel, Markus Diem, and Robert Sablatnig. Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. In *2013 12th international conference on document analysis and recognition*, pages 560–564. IEEE, 2013. 5

[17] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2

[18] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2, 5

[19] Rui Lan, Yancheng Bai, Xu Duan, Mingxing Li, Dongyang Jin, Ryan Xu, Lei Sun, and Xiangxiang Chu. Flux-text: A simple and advanced diffusion transformer baseline for scene text editing. *arXiv preprint arXiv:2505.03329*, 2025. 2, 5, 6

[20] Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. Glyphdraw: Seamlessly rendering text with intricate spatial structures in text-to-image generation. *arXiv preprint arXiv:2303.17870*, 2023. 2

[21] Yue Ma, Qingyan Bai, Hao Ouyang, Ka Leong Cheng, Qiuyu Wang, Hongyu Liu, Zichen Liu, Haofan Wang, Jingye Chen, Yujun Shen, and Qifeng Chen. Calligrapher: Freestyle text image customization. *arXiv preprint arXiv:2506.24123*, 2025. 5, 6

[22] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5(1): 39–46, 2002. 5

[23] ModelScope. Duguangocr. https://modelscope.cn/models/iic/cv_convnextTiny_ocr-recognition-general_damo/summary, 2023. 5

[24] Konstantina Nikolaidou, George Retsinas, Giorgos Sfikas, and Marcus Liwicki. Diffusionpen: Towards controlling the style of handwritten text generation. In *European Conference on Computer Vision*, pages 417–434. Springer, 2024. 8

[25] OpenAI. Introducing 4o image generation. https://openai.com/index/introducing-4o-image-generation, 2025. 2

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[27] PaddlePaddle. Pp-ocrv4. https://github.com/PaddlePaddle/PaddleOCR/blob/release/2.7/doc/doc_ch/PP-OCRv4_introduction.md, 2023. 3

[28] Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara. Handwritten text generation from visual archetypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22458–22467, 2023. 8

[29] Vittorio Pippi, Fabio Quattrini, Silvia Cascianelli, and Rita Cucchiara. Hwd: A novel evaluation score for styled handwritten text generation. In *Proceedings of the British Machine Vision Conference*, 2023. 5

[30] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4

[31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 5

[33] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[34] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. 2, 5

[35] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, 2020. Version 0.3.0. 5

[36] Yuxiang Tuo, Yifeng Geng, and Liefeng Bo. Anytext2: Visual text generation and editing with customizable attributes. *arXiv preprint arXiv:2411.15245*, 2024. 5, 6

[37] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 5, 6

[38] Yibin Wang, Weizhong Zhang, Honghui Xu, and Cheng Jin. Dreamtext: High fidelity scene text synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28555–28563, 2025. 2

[39] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 2, 5

[40] Yu Xie, Jielei Zhang, Pengyu Chen, Ziyue Wang, Weihang Wang, Longwen Gao, Peiyi Li, Huyang Sun, Qiang Zhang, Qian Qiao, et al. Textflux: An ocr-free dit model for high-fidelity multilingual scene text synthesis. *arXiv preprint arXiv:2505.17778*, 2025. 2

[41] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[42] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36:44050–44066, 2023. 2

[43] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024. 4

[44] Weichao Zeng, Yan Shu, Zhenhang Li, Dongbao Yang, and Yu Zhou. Textctrl: Diffusion-based scene text editing with prior guidance control. *Advances in Neural Information Processing Systems*, 37:138569–138594, 2024. 5, 6

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5