InnoText: A Unified Model for Visual Text Generation and Editing

Haowei Liu[1,2*§], Runze He[3*], Jian Lu[4*], Ao Ma[2†‡], Run Ling[2], Ke Cao[2], Jiasong Feng[5], Wei Feng[2], Shuo Lu[6], Yexing Xu[2], Yun Wang[7], Jing Wang[1], and Zhanjie Zhang[8†]

[1] Shenzhen Campus of Sun Yat-Sen University
[2] JD.com, Inc.
[3] University of Chinese Academy of Sciences
[4] Chongqing University of Post and Telecommunications
[5] Beijing University of Technology
[6] Institute of Automation, Chinese Academy of Sciences
[7] City University of Hong Kong
[8] Zhejiang University
liu_haow@163.com

Fig. 1: Visual examples generated by our method for visual text generation and editing.

Abstract. Diffusion models have recently achieved remarkable success in high-fidelity image synthesis, yet their application to visual text gener-

*Equal contribution. †Corresponding author. ‡Project leader. §Conducted during internship.

ation and editing remains relatively underexplored. Unlike general image generation, visual text tasks demand precise structural regularity and legibility, which may pose additional challenges for small-scale text and non-Latin scripts such as Chinese. Existing UNet-based models often struggle to produce clear and coherent text, while DiT-based models, though more expressive, are typically limited to a single task, which may lead to redundant training pipelines, inconsistent visual styles, and reduced cross-task generalization. To address these challenges, we propose InnoText, a unified DiT-based framework capable of performing both text generation and editing within a single model. We introduce a Font Size-Aware Modulation (FSAM) module to enhance representations across font scales, a Small-Character Aware Augmentation strategy to improve fine-grained fidelity, and a Task-Specific Region Weighted Loss for adaptive optimization. To support training and evaluation, we also construct a high-quality bilingual (English-Chinese) visual text dataset covering diverse fonts, sizes, and backgrounds. Experimental results demonstrate that our method achieves superior generation accuracy and editing quality, producing visually appealing and realistic text images.

1 Introduction

Recent advances in diffusion models [5, 8, 22] have demonstrated profound capabilities in image generation [25, 33, 41]. However, extending these successes to visual text generation and editing remains a formidable challenge. Visual text generation aims to synthesize coherent text within images directly from prompts, a task we formulate as a simultaneous text-to-image (T2I) synthesis problem [27], as shown in Fig. 2. This paradigm contrasts with conventional unified models [30, 43] that treat generation as a text-guided image-to-image (TI2I) insertion task. While visual text editing modifies content while preserving visual context. However, prevailing UNet-based methods [27, 36] often suffer from structural distortions and localized artifacts (see Fig. 3). Although emerging Diffusion Transformer (DiT) architectures [19, 32] offer superior fine-grained modeling, most existing frameworks are restricted to single-task designs. Such isolation prevents the exploitation of cross-task synergies and necessitates redundant optimization. Furthermore, the inherent complexity of non-Latin scripts like Chinese, characterized by intricate glyph structures, poses significant hurdles for current models primarily optimized for English. This difficulty is exacerbated in micro-scale text scenarios where existing methods fail to maintain legibility. Moreover, the scarcity of high-resolution Chinese datasets [27] further constrains model generalization to diverse real-world conditions.

To address these systemic limitations, we propose InnoText, a unified DiT-based framework that reformulates visual text tasks through in-context learning. By designing specialized input paradigm, InnoText achieves architectural unification, enabling seamless switching between generation and editing within a single model. To mitigate the scale-sensitivity issue, we introduce the Font Size-Aware Modulation (FSAM) module. FSAM dynamically re-calibrates the latent feature space via a spatial size-map, ensuring scale-invariant fidelity across diverse

font dimensions. Furthermore, we propose a Small-Character Aware Augmentation (SCAA) strategy, acting as a foveal attention mechanism during training to distill sub-pixel structural priors of intricate glyphs. Finally, a Task-Specific Region Weighted Loss is implemented to balance global coherence in generation with localized precision in editing. Our approach not only establishes a versatile benchmark for bilingual visual text tasks but also provides a robust solution for preserving structural integrity in complex, multi-scale real-world scenarios.

In summary, the main contributions of InnoText are as follows:

- We propose InnoText, a novel unified framework based on the DiT architecture. By leveraging in-context learning to design specialized input patterns, InnoText seamlessly integrates visual text generation and editing into a single model, offering superior flexibility and broader applicability in real-world scenarios.
- We introduce a Font Size-Aware Modulation (FSAM) module that leverages a size map to adaptively enhance text regions of different scales. In addition, we propose a Small-Character Aware Augmentation and a Task-Specific Region Weighted Loss to improve generation quality.
- We construct a high-quality visual text dataset with diverse font styles, sizes, and background complexities. Extensive experiments demonstrate the superiority of our model over existing baselines in both visual text generation and editing tasks.

2 Related Work

2.1 Text to Image Generation

With the rise of diffusion models, researchers have begun exploring their potential in text-to-image generation [5, 8, 20]. The introduction of latent diffusion models [22] marked a key milestone by performing diffusion in latent space, significantly reducing computational cost while preserving visual detail. While early models like Stable Diffusion [20, 22] achieved strong performance in general image synthesis, recent transformer-based architectures, such as Diffusion Transformer [19], have further improved generation quality. Notable examples like SD3 [5] and Flux [8] leverage stronger architectures and techniques such as rectified flow [12] to enhance image generation quality and efficiency. These advances have broadened the applicability of DiT-based models to downstream

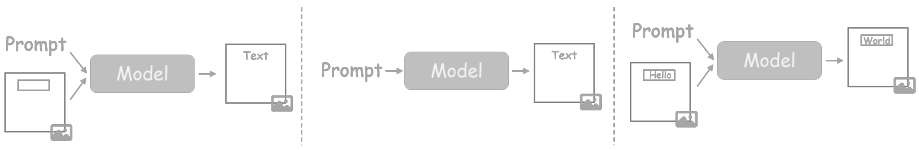

Fig. 2: (Left) Visual text generation in previous unified models [30,43]. (Middle) Visual text generation in AnyText [27] and ours. (Right) Visual text editing task.

Fig. 3: Visual text images generated by diffusion models based on UNet and DiT, respectively.

tasks including customized image generation [1, 21], visual editing [23, 40], and visual text synthesis [14, 32].

2.2 Controllable Diffusion Transformers

Traditional UNet-based generation methods, such as ControlNet [38], IP-Adapter [35], and T2I-Adapter [18], enable downstream applications by allowing the model to take reference images (e.g. depth maps, subject images) as conditional input [11, 21]. Recently, DiT-based conditional generation [10, 15, 23, 25, 29, 39] has also seen significant progress. These approaches typically encode visual conditions into tokens, which are then integrated with text tokens using the existing multi-modal attention mechanisms within the DiT architecture. This design allows for effective fusion of image-based conditions without requiring substantial architectural modifications.

2.3 Visual Text Generation and Editing

With the rapid development of diffusion models, generating high-quality and semantically accurate visual text images has become a key focus of research. Although recent work [2, 3, 34, 45] predominantly focused on synthesizing English-language text, recent studies, such as AnyText [27] and GlyphByT5 [13], have begun exploring the challenges of generating non-Latin character text, such as Chinese. Due to the limited language understanding capabilities of standard diffusion models for non-English languages, some approaches [6, 13, 16, 43] have proposed improvements to the text embedding process, for example, by leveraging large language models to enhance the semantic representation of multilingual

prompts. Other methods [7, 26, 27, 37] introduce glyph images as additional visual conditions, rendering the prompt text into font-based visual representations that are fed into the model to guide more accurate visual text generation.

To improve generation quality, recent state-of-the-art methods [4,9,14,28,32] increasingly adopt the DiT-based architecture as the backbone. However, most existing approaches are designed for either visual text editing [9, 32] or generation [4, 14, 28], and few are capable of jointly handling both tasks within a unified framework. To address this gap, we propose a DiT-based model specifically designed for visual text editing and generation, enabling a more flexible and accurate solution for real-world multilingual text synthesis.

3 Datasets

3.1 Overview

To overcome the scarcity of high-fidelity Chinese visual text data, we meticulously curated InnoText-30K, a bilingual dataset (20K Chinese, 10K English) designed for unified text tasks. Our primary contribution lies in bridging the quality gap between Chinese and English datasets through cross-lingual semantic mapping: we extended the English Lex-10K [42] into a high-quality Chinese counterpart of 10K samples using advanced generative pipelines, ensuring structural and aesthetic parity across languages. This is further augmented by a strategically selected 10K-sample subset from Anyword-3M [27], integrating diverse real-world and synthetic scenarios to enhance environmental robustness.

As illustrated in Fig. 4, InnoText-30K achieves a superior balance between visual aesthetics (ranking second only to Lex-10K) and perceptual naturalness (surpassing Lex-10K and matching Anyword-3M). These results validate that our curated data provides a high-quality, distributionally consistent foundational corpus for fine-tuning text-centric DiT models.

3.2 Pipeline

The InnoText-30K construction pipeline, illustrated in Fig. 4, prioritizes cross-lingual semantic fidelity and visual refinement. For the synthetic component, we initiate a knowledge transfer process by translating Lex-10K [42] captions into Chinese, followed by a semantic expansion phase using Intern-VL3 [44]. This stage leverages context-aware keyword substitution and prompt enrichment to synthesize expressive, high-entropy textual descriptors. These refined prompts are then fed into the Seedream [6] engine to generate Chinese image–text pairs, effectively bridging the representation gap in existing bilingual datasets.

To ensure rigorous quality control, the synthetic data is integrated with an Anyword-3M [27] subset and subjected to a multi-stage curation bottleneck. First, an aesthetic scoring filter prunes samples with suboptimal composition. Subsequently, we execute precision OCR re-annotation using PPOCR-v4 to rectify spatial-textual misalignments. Finally, DeepSeek-VL2 [31] is employed for

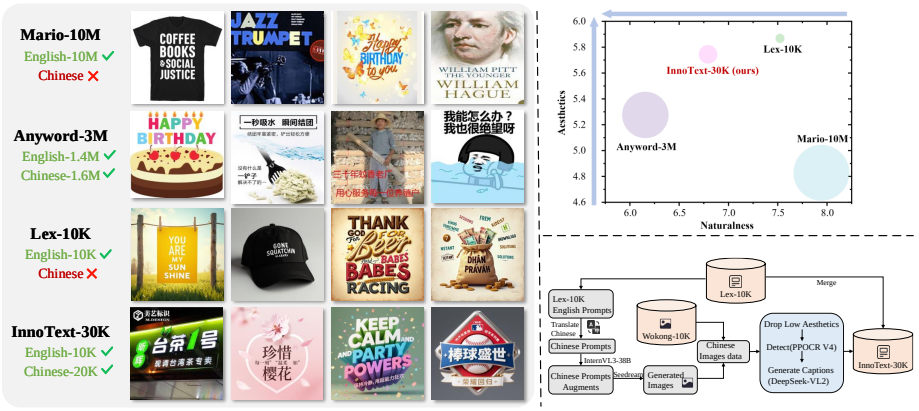

Fig. 4: Datasets overview. (Left) Representative cases from multiple datasets. (Top right) Comparison of aesthetics and naturalness across different datasets. The aesthetic scores are obtained using the Neural Image Assessment (NIMA) model [24], while the naturalness scores are evaluated using the Natural Image Quality Evaluator (NIQE) metric [17]. (Bottom right) Dataset processing pipeline.

vision-language recalibration, generating dense captions that synchronize visual content with OCR outputs for enhanced semantic grounding.

The resulting Chinese corpus is fused with Lex-10K to form our final dataset. For robust evaluation, we establish InnoText-Bench, a diagnostic suite of 1,500 images sampled across the three constituent subsets, ensuring a balanced assessment of model performance across diverse linguistic and aesthetic distributions.

4 Method

4.1 Overview

Our framework is based on the Flux.1 Fill dev [8], a powerful inpainting model derived from the Flux backbone. While retaining its strong generative capacity, we adapt this model to support both generation and editing tasks by carefully designing the input masking strategy.

For editing tasks, we define three input components: the masked image I_{masked}, the glyph image I_{glyph}, and the corresponding mask M. These are combined to form the model input in the spatial dimension as follows:

$$I_{\text{input}} = \text{Concat}([I_{\text{glyph}}, I_{\text{masked}}], \text{dim} = 0) \tag{1}$$

$$I_{\text{mask}} = \text{Concat}([\mathbf{0}, M], \text{dim} = 0) \tag{2}$$

For generation tasks, the model input is constructed by concatenating the glyph image with a fully black image (i.e., an empty canvas) along the spatial dimension:

$$I_{\text{input}} = \text{Concat}([I_{\text{glyph}}, \mathbf{0}], \text{dim} = 0) \tag{3}$$

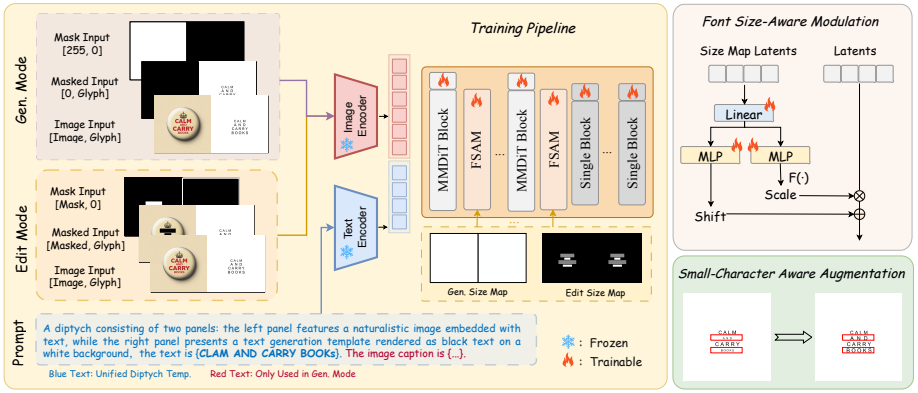

Fig. 5: Overview of InnoText. The model adopts different input paradigms and size maps for the editing and generation modes. In our framework, we incorporate a Font Size-Aware Modulation (FSAM) module that dynamically enhances the model's attention to text regions of varying font sizes. In addition, we adopt a Small-Character Aware Augmentation strategy, which selectively enlarges small-font text samples to improve the generation quality of fine-scale characters.

$$I_{\text{mask}} = \text{Concat}([\mathbf{0}, \mathbf{255}], \text{dim} = 0) \tag{4}$$

We next provide a detailed description of the key components used in our model, including the Font Size-Aware Modulation module, the Small-Character Aware Augmentation strategy, and the Task-Specific Region Weighted Loss employed during training.

4.2 Font Size-Aware Modulation

Font Size-Aware Modulation (FSAM) enhances the model's sensitivity to text regions of varying font sizes through an adaptive nonlinear feature calibration mechanism. Unlike conventional normalization-based modulation that applies uniform affine transformations, FSAM dynamically adjusts scaling and shifting parameters conditioned on font-size cues, improving fine-grained representation across multiple text scales.

The core of FSAM lies in the size map latents, which encode spatially varying font-size priors into the feature space. Unlike a binary mask, the size map is a grayscale image that assigns distinct intensity values to each detected text region, where the gray level indicates the degree of attention corresponding to its font size. To construct the map, we first estimate the approximate font height F_h for each detected character region as:

$$F_h = \max\big(\min(h_{\text{bbox}}, w_{\text{bbox}}), 1\big), \tag{5}$$

where $h_{\text{bbox}} = \max(y) - \min(y)$ and $w_{\text{bbox}} = \max(x) - \min(x)$ denote the height and width of the region's bounding contour, respectively. To emphasize

small-text regions, we compute the inverse font height and assign it to the corresponding spatial positions in the size map. The resulting grayscale map captures local scale variations across the text layout and is normalized to the range $[0, 1]$ for inter-sample consistency.

As illustrated in Fig. 5, this size map is patchfied and reshaped into a token sequence S, which is linearly projected to obtain latent embeddings S' with the same dimensionality as the hidden states. These embeddings serve as size map latents, enabling cross-scale conditioning of hidden representations. Two parallel MLP branches then decode S' into modulation parameters for scale and shift operations, respectively.

The scale modulation $\gamma(s)$ is generated in two steps: the scale $\mathrm{MLP}(\mathrm{MLP}_{\mathrm{shift}})$ produces bounded outputs $s \in [0, 1]$ via a sigmoid activation, which are subsequently transformed by a nonlinear amplification function:

$$\gamma(s) = 1 + (\gamma_{\max} - 1) \cdot \sigma\big(k \cdot (s - 0.5)\big) \tag{6}$$

where σ is the Sigmoid function, k controls the sharpness of the response, and γ_{max} defines the maximum amplification ratio.

The shift modulation is generated by a separate MLP with a Tanh activation, scaled by a predefined parameter λ_{shift} that controls the modulation amplitude:

$$\beta(S') = \lambda_{\mathrm{shift}} \cdot \tanh(\mathrm{MLP}_{\mathrm{shift}}(S')) \tag{7}$$

This scaling parameter λ_{shift} provides flexible control over the intensity of the shift modulation, enabling more stable feature calibration across different text scales.

The final modulated hidden state is then computed as:

$$H' = H \cdot \gamma(s) + \beta(S') \tag{8}$$

where H represents the hidden states from the DiT block.

By explicitly coupling the hidden activations with size-aware latent modulation, FSAM allows the network to adaptively recalibrate its internal representations based on font-scale priors. This design effectively strengthens the attention to small-text regions while maintaining visual consistency across scales, thus enhancing the model's capability in both text editing and generation tasks.

4.3 Small-Character Aware Augmentation

To bolster the model's structural fidelity and recognition performance for diminutive text, we introduce the SCAA strategy. This approach adaptively recalibrates the model's focus toward fine-grained, intricate details via stochastic regional magnification while maintaining global spatial coherence.

Specifically, we categorize a character instance as small-scale if its estimated font height h falls below a predefined geometric threshold τ. For these identified regions, we execute a probabilistic spatial scaling defined by a base factor λ and a random perturbation δ:

$$R' = \text{Resize}(R, \lambda + \delta), \quad \delta \sim \mathcal{U}(-\epsilon, \epsilon), \tag{9}$$

where R denotes the original region and $\mathcal{U}(-\epsilon, \epsilon)$ introduces controlled variability. This stochasticity serves as a form of foveal augmentation, enriching the model's exposure to diverse resolution scales and mitigating representation degradation in micro-scale text.

During inference, this scaling is symmetrically applied to editing tasks to ensure train-test consistency, thereby enhancing the fidelity of localized edits. Conversely, for text generation, we bypass this scaling to preserve the original font-size distribution of the input reference, ensuring a seamless and natural visual output.

4.4 Task-Specific Region Weighted Loss

During training, we employ the flow matching loss as the primary optimization objective to guide the model in generating predictions that are consistent with the ground-truth images under conditional inputs. Since our model simultaneously addresses both editing and generation tasks, we introduce task-specific loss formulations tailored to different input configurations, namely the combinations of I_{input} and I_{mask}. This design enables effective unification and cooperative optimization across tasks.

$$\mathcal{L}_{\text{total}} = \begin{cases} \|S \cdot \mathcal{L}_{\text{FM}}\|_2^2, & \text{if editing task} \\ \mathcal{L}_{\text{FM}}, & \text{if generation task} \end{cases} \tag{10}$$

For the editing task, where only partial text regions require modification, we apply a region-weighted loss guided by a font size map. Higher weights are assigned to masked or modified areas to emphasize precise local refinement. In contrast, for the generation task, where the model synthesizes full-text images from scratch, we adopt a global loss to supervise the entire output, ensuring holistic structural consistency and overall visual fidelity.

This unified yet differentiated loss strategy aligns the optimization objectives with the specific requirements of each task, enhancing the model's robustness and performance across diverse visual text scenarios.

5 Experiments

5.1 Experiments Setting

Implementation Details. Our base model is Flux-Fill [8]. During training, we alternate between the editing and generation tasks with an equal probability of 0.5. We use a batch size of 2 and optimize the model using the Prodigy optimizer with a weight decay of 0.01. The LoRA rank is set to 128. Training is conducted on 8 NVIDIA A100 GPUs for a total of 30,000 steps. For small-character aware

Table 1: Quantitative comparison of editing and generating tasks, the best results are **bolded**. More results can be found in the supplementary materials.

Tasks	Methods	English			Chinese		
		Sen. Acc ↑	NED ↑	LPIPS ↓	Sen. Acc ↑	NED ↑	LPIPS ↓
Edit	Flux-Fill [8]	0.1342	0.2481	0.1650	0.0191	0.0443	0.1276
	Anytext [27]	0.6839	0.8632	0.1234	0.6410	0.8209	0.1093
	Anytext2 [26]	0.7893	**0.9082**	0.1739	0.7079	0.8419	0.1509
	TextFlux [32]	0.7732	0.8908	0.0838	0.7164	0.8498	**0.0567**
	Ours	**0.7988**	0.9016	**0.0786**	**0.7257**	**0.8579**	0.0591
Gen.	Flux-Fill [8]	0.0126	0.0635	0.6232	0.0057	0.0200	0.6402
	Anytext [27]	0.4707	0.7509	0.6463	0.4584	0.6440	0.6575
	Anytext2 [26]	0.5638	0.7712	0.6330	0.5262	0.6636	0.5979
	TextFlux [32]	0.0071	0.0623	0.5968	0.0187	0.0470	0.6527
	Ours	**0.6586**	**0.8066**	**0.4787**	**0.5907**	**0.6837**	**0.5311**

augmentation, the size threshold λ and scaling factor ϵ are set to 1.5 and 0.3, respectively. All images used in the experiments were resized to 512×512 pixels.

Evaluation Metrics. Following previous works, we evaluate both textual fidelity and perceptual quality for visual text editing and generation tasks. The evaluation metrics include Sentence Accuracy (Sen. Acc), which measures exact text correctness, Normalized Edit Distance (NED), which reflects character-level similarity between predictions and ground truth, and Learned Perceptual Image Patch Similarity (LPIPS), which assesses perceptual visual quality. The AnyText-Benchmark is used to evaluate editing performance, while InnoText-Bench is used for generation. Together, these metrics comprehensively measure the model's ability to preserve textual accuracy and visual realism.

Comparative Methods. We compare our method with several representative baselines covering both UNet-based and DiT-based diffusion architectures. The UNet-based methods include AnyText [27] and AnyText2 [26], which support both visual text editing and generation within a unified framework. The DiT-based methods include Flux-Fill [8] and TextFlux [32], both designed for text editing tasks. All comparative methods are evaluated using their officially released pretrained weights.

5.2 Quantitative Results

As shown in Table 1, InnoText demonstrates consistently strong performance in visual text editing, ranking among the top methods across all evaluation metrics in both English and Chinese. It achieves the highest sentence accuracy and the lowest LPIPS, indicating its superior capability to preserve textual semantics and maintain high visual fidelity. Compared with representative baselines such as TextFlux and AnyText2, InnoText delivers more precise text reconstruction and better structural consistency, particularly in challenging Chinese text scenarios.

Fig. 6: Qualitative comparison on editing tasks. Our method generates text that blends seamlessly with complex backgrounds or non-flat background(first and second rows), maintains high visual quality across varying font sizes (third row), and produces coherent results under diverse text layouts, such as vertically arranged text (fourth row). More results can be found in the supplementary materials.

Table 2: Ablation studies on visual text generation and editing, the best results are **bolded**.

Methods	Edit			Generation		
	Sen. Acc ↑	NED ↑	LPIPS ↓	Sen. Acc ↑	NED ↑	LPIPS ↓
Anytext-FT	0.5794	0.7138	0.1009	0.4622	0.6513	0.6186
Anytext2-FT	0.5882	0.7341	0.0532	0.5358	0.6801	0.5539
w/o FSAM	0.5672	0.7048	0.0604	0.5508	0.6479	0.5928
w/o SCAA	0.5963	0.7196	0.0497	0.5615	0.6418	0.5539
w/o $\mathcal{L}_{TSRW}$	0.6180	**0.7533**	0.0509	0.5892	0.6592	0.5671
w/ all	**0.6374**	0.7528	**0.0484**	**0.5907**	**0.6837**	**0.5311**

For visual text generation, InnoText likewise surpasses all competing methods, attaining the best sentence accuracy and the lowest LPIPS scores. These results highlight its strong ability to synthesize text that is both semantically accurate and visually coherent, confirming the effectiveness of our unified framework in handling diverse generation and editing tasks with consistent quality.

Fig. 7: Qualitative comparison on generation task. The visual text images generated by our method exhibit high aesthetic quality (first row), demonstrate appropriate integration with the background and higher background quality (second row), and achieve relative satisfactory results even for small-sized text (third and fourth rows). The corresponding prompt is provided in the supplementary materials.

5.3 Qualitative Results

We conducted qualitative analyses to validate the effectiveness of our method in both tasks. As shown in Fig. 6, Flux-Fill fails to generate Chinese text due to its lack of Chinese encoding capability. Although AnyText and AnyText2 can produce Chinese characters, their generation quality noticeably deteriorates in complex scenes or when handling text of varying sizes. TextFlux demonstrates limited fidelity in character details, with certain strokes appearing distorted or incorrectly formed.

As shown in Fig. 7, for visual text generation, our method produces images with superior background quality and text rendering fidelity compared with competing approaches. The generated text not only maintains high semantic accuracy, but also exhibits enhanced visual aesthetics, achieving a more natural and visually pleasing integration between text and background.

5.4 Ablation Studies

The results of our ablation studies are shown in Table 2. We evaluate the performance of AnyText and AnyText2 after fine-tuning on the InnoText-30K dataset,

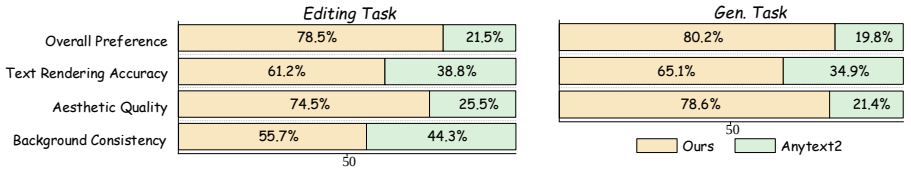

Fig. 8: Results of Human Perceptual Evaluation.

and further analyze the impact of each module through ablation studies. Previous methods such as AnyText and AnyText2 show performance improvements after fine-tuning on the InnoText-30K dataset. However, their results still remain inferior to our method.

Among all modules, the Font Size-Aware Modulation (FSAM) contributes the most, as its removal increases the LPIPS score from 0.5311 to 0.5928, particularly degrading generation quality. Excluding the Small Character Aware Augmentation consistently lowers sentence accuracy and visual similarity, especially in editing, underscoring its importance in enhancing small-text clarity. Removing the Task-Specific Region Weighted Loss leads to moderate performance degradation, including a 1.93% drop in editing accuracy, indicating its role in task-aware optimization. The best overall performance is achieved when all components are combined, demonstrating their complementary contributions to both visual quality and text generation accuracy.

5.5 Human Evaluation

To rigorously assess the perceptual superiority of InnoText, we conducted a comprehensive double-blind Side-by-Side (SbS) user study involving 15 participants with diverse backgrounds. We curated a balanced evaluation set comprising 100 cases (50 for generation and 50 for editing) and compared our model against the state-of-the-art AnyText2. Respondents were tasked with evaluating paired outputs based on four critical dimensions: text rendering accuracy (structural integrity and OCR accuracy), aesthetic quality (visual realism and texture quality), background consistency (seamless integration between text and background, particularly for editing), and overall preference. To ensure statistical objectivity, the order of images was randomized, and participants were blinded to the underlying model identities.

As illustrated in Fig. 8, InnoText consistently achieves a dominant lead over the baseline across both task categories, with its overall preference rate exceeding 78%. Quantitative analysis of the feedback indicates that our DiT-based architecture effectively resolves the low-level precision vs. high-level harmony trade-off, which often plagues UNet-based frameworks. Notably, in Chinese text scenarios, our model was frequently cited for its superior handling of intricate stroke connectivity and font-style consistency, while the competitors often exhibited blurred contours or structural artifacts. These results underscore the

Fig. 9: Representative failure cases.

robustness of InnoText in generating high-fidelity visual text that aligns with human perceptual expectations in complex real-world environments.

5.6 Failure Cases Studies

We analyze representative failure cases, as illustrated in Fig. 9. For the editing task, overly large mask regions may occasionally result in local text duplication, likely due to ambiguous spatial constraints when the editable area covers extensive contextual content. For small and densely structured characters, minor stroke-level inaccuracies can still appear, especially under extreme font scales. In the generation task, redundant characters and complex Chinese glyphs with high stroke density remain relatively challenging. Despite the font size-aware modulation and small-character aware augmentation improving multi-scale and fine-grained representations, precise structural modeling of repetitive patterns and intricate stroke layouts is not yet fully resolved.

These observations suggest that future work could further enhance spatial controllability and character-level structural consistency within unified DiT-based text rendering frameworks.

6 Conclusion

In this paper, we present InnoText, a unified Diffusion Transformer framework that harmonizes visual text generation and editing within a cohesive paradigm. By integrating Font Size-Aware Modulation (FSAM) and Small-Character Aware Augmentation (SCAA), our approach alleviates long-standing bottlenecks in micro-scale text representation and complex non-Latin script modeling. We further introduce a Task-Specific Region Weighted Loss to balance localized precision with global semantic coherence. Coupled with the InnoText-30K datasets, a

high-fidelity bilingual dataset, our work establishes a rigorous benchmark for visual text research. Extensive quantitative and qualitative evaluations, including a blind user study, demonstrate that InnoText consistently achieves state-of-the-art performance across diverse scenarios. Ultimately, this work not only offers a robust solution for high-precision typographic synthesis but also provides significant insights into the development of scale-aware and multi-modal foundational models for real-world creative applications.

References

1. Cao, K., Wang, J., Ma, A., Feng, J., He, X., Ling, R., Liu, H., Lu, J., Feng, W., Wang, H., et al.: Relactrl: Relevance-guided efficient control for diffusion transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 2598–2606 (2026)
2. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser: Diffusion models as text painters. Advances in Neural Information Processing Systems **36**, 9353–9387 (2023)
3. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser-2: Unleashing the power of language models for text rendering. In: European Conference on Computer Vision. pp. 386–402. Springer (2024)
4. Du, N., Chen, Z., Chen, Z., Gao, S., Chen, X., Jiang, Z., Yang, J., Tai, Y.: Textcrafter: Accurately rendering multiple texts in complex visual scenes. arXiv preprint arXiv:2503.23461 (2025)
5. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
6. Gong, L., Hou, X., Li, F., Li, L., Lian, X., Liu, F., Liu, L., Liu, W., Lu, W., Shi, Y., et al.: Seedream 2.0: A native chinese-english bilingual image generation foundation model. arXiv preprint arXiv:2503.07703 (2025)
7. Jiang, B., Yuan, Y., Bai, X., Hao, Z., Yin, A., Hu, Y., Liao, W., Ungar, L., Taylor, C.J.: Controltext: Unlocking controllable fonts in multilingual text rendering without font annotations. arXiv preprint arXiv:2502.10999 (2025)
8. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
9. Lan, R., Bai, Y., Duan, X., Li, M., Sun, L., Chu, X.: Flux-text: A simple and advanced diffusion transformer baseline for scene text editing. arXiv preprint arXiv:2505.03329 (2025)
10. Li, Y., Li, X., Zhang, Z., Bian, Y., Liu, G., Li, X., Xu, J., Hu, W., Liu, Y., Li, L., et al.: Ic-custom: Diverse image customization via in-context learning. arXiv preprint arXiv:2507.01926 (2025)
11. Ling, R., Cao, K., Lu, J., Ma, A., Liu, H., He, R., Wang, C., Xu, R., Shao, Y., Zhang, Z., et al.: Mofu: Scale-aware modulation and fourier fusion for multi-subject video generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 7033–7041 (2026)
12. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
13. Liu, Z., Liang, W., Liang, Z., Luo, C., Li, J., Huang, G., Yuan, Y.: Glyph-byt5: A customized text encoder for accurate visual text rendering. In: European Conference on Computer Vision. pp. 361–377. Springer (2024)

14. Lu, R., Zhang, Y., Liu, J., Wang, H., Song, Y.: Easytext: Controllable diffusion transformer for multilingual text rendering. arXiv preprint arXiv:2505.24417 (2025)
15. Ma, A., Feng, J., Cao, K., Wang, J., Wang, Y., Zhang, Q., Zhang, Z.: Lay2story: extending diffusion transformers for layout-togglable story generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16102–16111 (2025)
16. Ma, L., Yue, T., Fu, P., Zhong, Y., Zhou, K., Wei, X., Hu, J.: Chargen: High accurate character-level visual text generation model with multimodal encoder. arXiv preprint arXiv:2412.17225 (2024)
17. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
18. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
19. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
20. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
21. Qin, Y., Cao, K., Liu, H., Ma, A., Li, F., Zhu, H., Zhang, Z., Ling, R., Feng, W., He, X., et al.: Innoads-composer: Efficient condition composition for e-commerce poster generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 32988–32999 (2026)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
23. Song, W., Jiang, H., Yang, Z., Quan, R., Yang, Y.: Insert anything: Image insertion via in-context editing in dit. arXiv preprint arXiv:2504.15009 (2025)
24. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE transactions on image processing **27**(8), 3998–4011 (2018)
25. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. arXiv preprint arXiv:2411.15098 (2024)
26. Tuo, Y., Geng, Y., Bo, L.: Anytext2: Visual text generation and editing with customizable attributes. arXiv preprint arXiv:2411.15245 (2024)
27. Tuo, Y., Xiang, W., He, J.Y., Geng, Y., Xie, X.: Anytext: Multilingual visual text generation and editing. arXiv preprint arXiv:2311.03054 (2023)
28. Wang, H., Xu, Y., Li, Y., Li, J., Zhang, C., Wang, J., Yang, K., Chen, Z.: Reptext: Rendering visual text via replicating. arXiv preprint arXiv:2504.19724 (2025)
29. Wang, J., Ma, A., Cao, K., Zheng, J., Feng, J., Zhang, Z., Pang, W., Liang, X.: Wisa: World simulator assistant for physics-aware text-to-video generation. Advances in Neural Information Processing Systems **38**, 5388–5416 (2026)
30. Wang, Y., Zhang, W., Xu, H., Jin, C.: Dreamtext: High fidelity scene text synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28555–28563 (2025)
31. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)

32. Xie, Y., Zhang, J., Chen, P., Wang, Z., Wang, W., Gao, L., Li, P., Sun, H., Zhang, Q., Qiao, Q., et al.: Textflux: An ocr-free dit model for high-fidelity multilingual scene text synthesis. arXiv preprint arXiv:2505.17778 (2025)
33. Xu, Y., Feng, W., Zhang, S., Wang, H., Qin, Y., Li, Y., Ma, A., Luo, Y., Wang, L., Ren, X., et al.: Design your ad: Personalized advertising image and text generation with unified autoregressive models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 472–483 (2026)
34. Yang, Y., Gui, D., Yuan, Y., Liang, W., Ding, H., Hu, H., Chen, K.: Glyphcontrol: Glyph conditional control for visual text generation. Advances in Neural Information Processing Systems **36**, 44050–44066 (2023)
35. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
36. Zeng, W., Shu, Y., Li, Z., Yang, D., Zhou, Y.: Textctrl: Diffusion-based scene text editing with prior guidance control. Advances in Neural Information Processing Systems **37**, 138569–138594 (2024)
37. Zhang, B., Gao, Z., Qu, Y., Xie, H.: How control information influences multilingual text image generation and editing? Advances in Neural Information Processing Systems **37**, 6884–6904 (2024)
38. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
39. Zhang, Y., Yuan, Y., Song, Y., Wang, H., Liu, J.: Easycontrol: Adding efficient and flexible control for diffusion transformer. arXiv preprint arXiv:2503.07027 (2025)
40. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. arXiv preprint arXiv:2504.20690 (2025)
41. Zhang, Z., Ma, A., Cao, K., Wang, J., Liu, S., Ma, Y., Cheng, B., Leng, D., Yin, Y.: U-stydit: Ultra-high quality artistic style transfer using diffusion transformers. arXiv preprint arXiv:2503.08157 (2025)
42. Zhao, S., Wu, Q., Li, X., Zhang, B., Li, M., Qin, Q., Liu, D., Zhang, K., Li, H., Qiao, Y., et al.: Lex-art: Rethinking text generation via scalable high-quality data synthesis. arXiv preprint arXiv:2503.21749 (2025)
43. Zhao, Y., Lian, Z.: Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. In: European conference on computer vision. pp. 217–233. Springer (2024)
44. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
45. Zhu, Y., Liu, J., Gao, F., Liu, W., Wang, X., Wang, P., Huang, F., Yao, C., Yang, Z.: Visual text generation in the wild. In: European Conference on Computer Vision. pp. 89–106. Springer (2024)