DOI: 10.1111/cgf.70342

EUROGRAPHICS 2026 / B. Masia and J. Thies
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 2

TextFlux: An OCR-Free DiT Model for High-Fidelity Multilingual Scene Text Synthesis

Yu Xie [1,*], Jielei Zhang [1,*,†], Pengyu Chen [1], Weihang Wang [1], Longwen Gao [1], Peiyi Li [1], Qian Qiao [1], Zhouhui Lian [2]✉

[1] Bilibili Inc., China
[2] Wangxuan Institute of Computer Technology, Peking University, China
* Equal contribution † Project Leader ✉ Corresponding author

Figure 1: *Some examples of high-fidelity multilingual scene text images generated by our TextFlux.*

Abstract

Diffusion-based scene text synthesis has progressed rapidly, yet existing methods commonly rely on additional visual conditioning modules and require large-scale annotated data to support multilingual generation. In this work, we revisit the necessity of complex auxiliary modules and further explore an approach that simultaneously ensures glyph accuracy and achieves high-fidelity scene integration, by leveraging diffusion models' inherent capabilities for contextual reasoning. To this end, we introduce TextFlux, a DiT-based framework that enables multilingual scene text synthesis. The advantages of TextFlux can be summarized as follows: (1) OCR-free model architecture. TextFlux eliminates the need for OCR encoders that are specifically used to extract visual text-related features. (2) Strong multilingual scalability. TextFlux is effective in low-resource multilingual settings, and achieves strong performance in newly added languages with fewer than 1,000 samples. (3) Streamlined training setup. TextFlux is trained with only 1% of the training data required by competing methods. (4) Controllable multi-line text generation. TextFlux offers flexible multi-line synthesis with precise line-level control, outperforming methods restricted to single-line or rigid layouts. Extensive experiments and visualizations demonstrate that TextFlux outperforms previous methods in both qualitative and quantitative evaluations. Our code is available at https://github.com/yyyyyxie/textflux.

Keywords: Scene Text Synthesis, Diffusion Models, OCR-free, Image Editing, Multilingual Generation

Figure 2: *TextFlux addresses the common conflict between glyph accuracy and stylistic integration in scene text synthesis. Prior works often exhibit either glyph errors (first column) or poor visual fidelity and integration (second column). In contrast, TextFlux accurately renders complex and multi-line text with high fidelity to the scene context (third and fourth columns).*

1. Introduction

The synthesis of scene text in this work encompasses both *text reconstruction* and *text editing*, aiming to restore or modify textual content in natural images while preserving the visual fidelity of the scene. The challenges of this task can be categorized into two core aspects: first, ensuring the **"spelling" accuracy** of the generated text itself – that is, the correctness of its glyph structure; and second, **naturally and realistically** integrating the edited or generated text into the complex visual contexts of diverse target scenes.

To address the first core challenge (ensuring the accuracy of the glyph structure), existing methods [CHL*24a, TXH*23, ZL23, YGY*23, MDC*24] often introduce *specialized textual features* (such as explicit glyph information) as conditions. However, while leveraging such specialized textual features for strong, specific control does improve the accuracy of the generated glyphs, it tends to cause the generated text to appear merely "pasted on" and lack realistic integration with the scene, as shown in the second column of Fig. 2. To address this issue of overall visual fidelity (the second core challenge), some approaches [TGB24, WZZJ24, ZSL*24, DCC*25] attempt to establish independent controls for distinct visual attributes such as style, font, and color, injecting corresponding features as conditions. However, the inherent *diversity, complexity, and subjectivity* of text visual styles make it extremely difficult to construct a comprehensive universal representation for them. Moreover, some attributes, such as lighting and texture, are inherently hard to disentangle, greatly increasing the complexity of model design and training.

Considering the aforementioned challenges, this paper aims to explore a new approach to reconcile the conflict between glyph accuracy and realistic integration in scene text synthesis. We observe that current diffusion models [RBL*21, PEL*23, PX23, Lab24] already excel in maintaining overall contextual coherence and visual fidelity in inpainting tasks. The real challenge lies in enabling them to **"learn to spell" from scratch**, especially for complex character systems like Chinese with its intricate strokes. If the model inherently knew the specific details of glyph structures, it could theoretically generate text with high visual fidelity. Based on these insights, we depart from the traditional approach of feature-level conditioning and instead turn to the image's own spatial dimension: by directly providing a visual glyph reference, we transform the core task from "learning to spell" to **learning how to integrate this given glyph into the context with a scene-adaptive style**. This simplified learning objective allows the model to focus on the integration process by leveraging its inherent strengths, rather than on the complex task of "learning to spell" from scratch.

In this paper, we propose TextFlux, an OCR-free diffusion framework for multi-language scene text synthesis that eliminates the need for auxiliary OCR encoders to extract glyph or style features, as well as any OCR-related supervision losses. Built upon the state-of-the-art DiT-based Flux architecture [Lab24], TextFlux guides the model to adaptively infer and render harmonious text styles from the scene context. This approach circumvents the dilemmas faced by existing methods in the definition and control of various text visual attributes, offering a concise and efficient solution for generating high-fidelity, contextually consistent text. Furthermore, benefiting from the design of this new paradigm, TextFlux demonstrates strong capabilities in simultaneously editing multi-line text, handling multiple languages, rendering complex glyphs, and even achieving zero-shot generalization to characters not seen in the training set. To show more intuitive results, we also compared TextFlux with several recent commercial image generation models [Ope25, Goo25, WLZ*25].

Our main contributions can be summarized as follows:

- We propose TextFlux, an OCR-free diffusion framework for scene text synthesis. TextFlux introduces essential textual guidance by spatially integrating glyph-rendered visual cues, thereby eliminating the need for dedicated OCR encoders for various visual text attributes.
- We demonstrate that TextFlux achieves strong multilingual scalability, especially in low-resource languages, effectively synthesizing text across multiple languages and rapidly adapting to new, low-resource languages with minimal language-specific data.
- We enable flexible and controllable multi-line text synthesis through inherent spatial guidance, allowing precise line-level

control over content and position. Extensive experiments on multiple benchmarks demonstrate that TextFlux achieves state-of-the-art performance in multilingual scene text synthesis, outperforming existing methods in both visual fidelity and sequence accuracy.

2. Related Work

2.1. Text-to-Image Synthesis

In recent years, diffusion models have achieved significant success across various tasks, especially in text-to-image synthesis [DN21, RBL*21, MWX*24, LLZ*24, HXL*24, GHL*25], image-to-image translation [SCC*22], and image editing [CMGS25, BHE23, HMT*22, FMW*24]. These successes demonstrate the superiority of diffusion models in the field of image generation. Emerged areas of exploration include Personalized Generation [RLJ*23, RLJ*24, CXL*25], Controllable text-to-image (T2I) Generation [ZRA23, LYK*24], LLM-assisted T2I [FZF*23], Style Transfer [WHSX21], and Safety Issues [LST*24, WGW*24]. To further enhance generation performance, recent studies integrate large-scale transformer architectures as the backbone of diffusion models, resulting in advanced models like DiT [PX23, Lab24, CYG*23]. Among these architectural innovations, Flux [Lab24], which is based on flow matching objectives [LCBH*22], has achieved state-of-the-art generation results and has been open-sourced. These advancements have subsequently fueled research into the control [TLY*24, YKJ*24] and acceleration [TXY*25, MTJ*25] of these new architectures.

2.2. Scene Text Synthesis

Despite the rapid development of diffusion models, these general methods often face limitations when generating scene text. Early researchers pointed out that text encoders play a crucial role in generating accurate text. To address this issue, Imagen [SCS*22], eDiff-I [BNH*22], and DeepFloyd [Dee23] utilized large-scale language models (e.g., T5-XXL [CHL*24b]) to optimize text spelling capabilities. UDiffText [ZL23] attempts to train a text encoder aligned with visual text features to replace the text encoder in CLIP [RKH*21], thereby enhancing the glyph-awareness. However, improvements on text encoders bring only limited gains in text rendering quality within diffusion models, especially for non-Latin scripts.

As a result, more scene text synthesis methods [ZCW*24, TXH*23, MDC*24, WLQ*25, GLL*25] are focused on designing specialized condition control modules specifically tailored to visual text. GlyphDraw [MZC*23] initially used glyph images as condition control and rendered characters at the center. GlyphControl [YGY*23] further extended this approach by spatially aligning the glyph rendering position with the actual text generation position. TextDiffuser [CHL*23] trained an additional OCR engine to generate segmentation masks, which are used for condition control. AnyText [TXH*23] inherited the design philosophy of condition control from GlyphControl and expanded it to multilingual versions. DreamText [WZZJ24] further introduced additional control conditions such as different fonts to enhance the rendering capability of visual text. Besides these methods, some approaches [FLW*25, ZSL*24] aim to reduce the difficulty of visual text editing by cropping the text to be edited and only processing text lines. Although these methods significantly improve character accuracy, they often sacrifice visual fidelity in text generation due to the lack of context integration with the entire image.

Although the various OCR encoders proposed by the aforementioned methods enhanced the effectiveness of scene text synthesis, they also led to architectural redundancy and optimization difficulties due to excessive condition control. Moreover, the overemphasis on OCR characteristics often results in a loss of fidelity. In the new wave of control and acceleration based on the latest DiT series of architectures [Lab24, PX23], this paper seeks to shift the paradigm away from using specialized condition control modules (OCR encoders) in scene text synthesis. Instead, it introduces a novel approach that leverages contextual information from the image itself to achieve scene-adaptive and visually coherent text generation.

2.3. Text Rendering in Recent Commercial Models

Recently, commercial image generation models have garnered increasing attention, exemplified by models such as Seedream2.0 [GHL*25], Seedream3.0 [GGG*25], GPT-Image-1 [Ope25], Qwen-Image-Edit [WLZ*25], and Nano Banana [Goo25]. These models have demonstrated remarkable capabilities in text rendering. However, they still confront several limitations. For instance, Nano Banana [Goo25] excels in English text rendering with high consistency but struggles with non-Latin languages like Chinese. Qwen-Image-Edit [WLZ*25] performs well in editing both English and Chinese text but exhibits clear limitations with Japanese and Korean, particularly in rendering complex glyphs and precisely locating small text. While GPT-Image-1 [Ope25] boasts relatively strong text rendering abilities, its consistency is often suboptimal. Consequently, specialized research remains essential for multilingual scene text synthesis. The experimental section of this paper further details a comparative analysis of our proposed method against these commercial models.

3. Methodology

3.1. Preliminary

While U-Net has been the dominant architecture in early diffusion models, recent works like FLUX-1 [Lab24], Stable Diffusion 3 [RBL*21], and PixArt [CYG*23] have explored the Transformer-based DiT architecture [PX23]. These DiT models scale well to larger sizes and demonstrate an improved ability to understand the overall context and relationships within the images. Notably, OmniControl [TLY*24] and In-Context LoRA [HWW*24] further suggest that DiT-based architectures inherently possess contextual reasoning capabilities. These insights motivate a new perspective on control mechanisms specifically for scene text synthesis, where contextual understanding plays a key role. Among the DiT-based architectures, FLUX-1-Fill-dev [Lab24] is an inpainting-oriented variant that supports flexible conditioning. In this design, the standard DiT input—noisy image tokens $\mathbf{X} \in \mathbb{R}^{N\times d}$ and text conditioning tokens $\mathbf{T}_c \in \mathbb{R}^{M\times d}$—is extended for the inpainting task by introducing masked image tokens $\mathbf{X}_i \in \mathbb{R}^{N\times d}$ and binary mask tokens $\mathbf{X}_m \in \mathbb{R}^{N\times 4d}$ resulting in an augmented visual sequence:

$$\mathbf{Z} = \mathrm{Concat}(\{\mathbf{X}, \mathbf{X}_i, \mathbf{X}_m\}, \mathrm{dim} = -1). \quad (1)$$

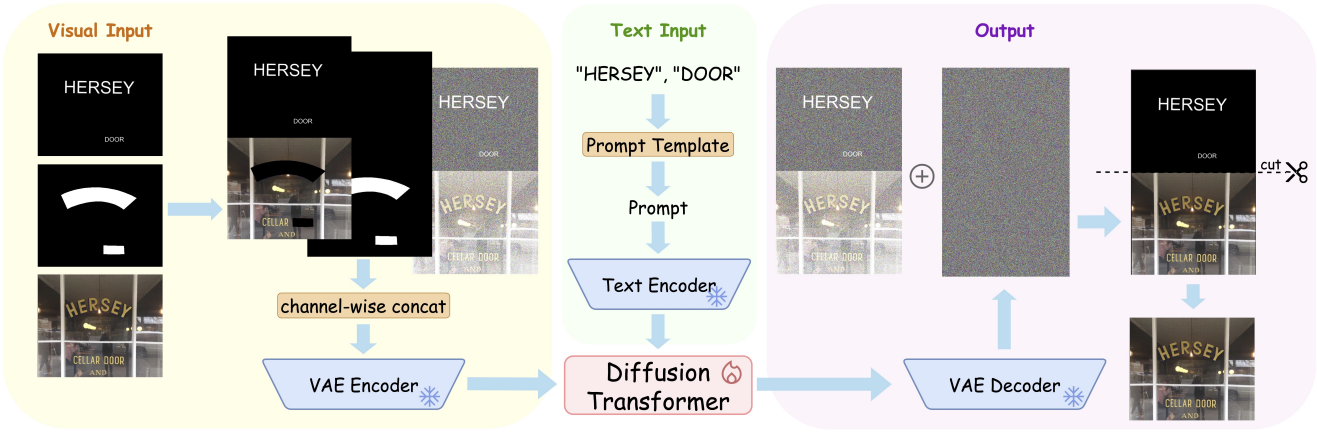

Figure 3: ***Overview of TextFlux.*** *We propose an OCR-free scene text synthesis method that spatially concatenates glyph-rendered text with the original image as model input, enabling the diffusion transformer to leverage its inherent context-awareness to render text in the masked regions.*

The sequence $\mathbf{Z}$, along with the text conditioning tokens $\mathbf{T}_c$, is then fed into the DiT blocks. This architecture serves as the foundation of TextFlux.

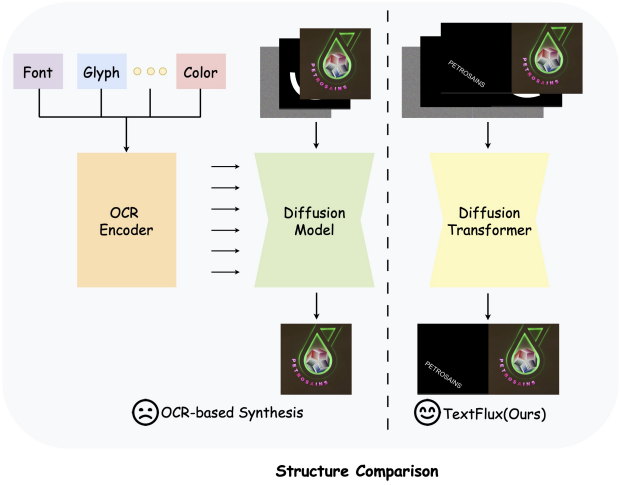

Figure 4: *Traditional methods employ OCR encoders to extract and inject various visual text features (e.g., font, glyph, color) as conditions. TextFlux streamlines the process by directly providing spatial glyph cues.*

3.2. Motivation

Recent diffusion-based scene text synthesis methods typically employ additional visual conditioning modules named OCR encoders, as shown in Fig. 4 (left). Although the design of these approaches seems reasonable, they possess several critical limitations. First, integrating diverse OCR encoders considerably increases model architecture complexity. The text-specific feature representations may be alien to the general pre-trained diffusion model, necessitating extensive learning from scratch and complicating optimization. Second, the aforementioned optimization difficulty often demands large-scale annotated datasets [TXH*23, CHL*23] and prolonged training, hindering scalability, especially for low-resource languages. Third, the use of OCR-based modules typically leads to the requirement of additional specialized loss functions [CHL*23, TXH*23], significantly increasing the implementation complexity and computational cost. Last but not least, the heavy reliance on these modules biases the model towards fitting the specific OCR representations. This could potentially lead the model to overlook the broader scene context, ultimately undermining the visual fidelity and natural integration of the synthesized text. It is similar to the "pasted-on" appearance discussed in the Introduction section.

In summary, the proposed TextFlux consists of the following advantages: 1) By eliminating the reliance on OCR encoders, both efficiency and architectural simplicity can be achieved. 2) Our training strategy focuses on enabling the diffusion model to adapt a provided glyph to the scene image context, which can be markedly simplified by our new paradigm. 3) We significantly lessen the dependency on large-scale annotated data, especially for multilingual settings. Thus, even in low-resource scenarios, excellent performance can be achieved with only minimal additional data (e.g., adapting to new languages). A key insight serves as the foundation for our proposed TextFlux's simplified paradigm: pretrained diffusion transformers inherently possess strong capabilities for contextual reasoning and visual understanding, which we leverage by concatenating glyphs spatially.

3.3. TextFlux

Based on the above-mentioned analyses, we utilize FLUX.1-Fill-dev [Lab24], an inpainting-oriented variant from the DiT family, to develop TextFlux, a scene text synthesis system that supports multilingual scenarios through an efficient concatenation scheme. The overall architecture is illustrated in Fig. 3. We describe the system from the perspective of input construction.

Model Input. Our method prepares the glyph-guided image input

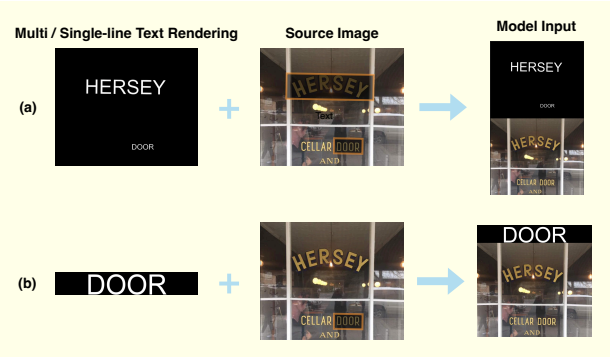

Figure 5: *Visualization of the multi-line layout used in multi-line text synthesis and the concatenation method used in single-line text synthesis.*

for the diffusion model. First, the target text is rendered as white foreground on a black background to create a binary glyph mask $\mathbf{I}_{\text{glyph}}$. Second, $\mathbf{I}_{\text{glyph}}$ is spatially concatenated with $\mathbf{I}_{\text{scene}}$ (e.g., either vertically or horizontally) to form the combined input $\mathbf{I}_{\text{concat}}$. This input structure enables the model to directly observe the precise glyph template alongside the full scene context.

Glyph Design. The design of the glyph mask ($\mathbf{I}_{\text{glyph}}$) is handled in two distinct ways depending on the text layout:

For multi-line text, its layout should be taken into account. For each text region in the image that is defined by polygon coordinate labels, we use Pillow, a widely used Python imaging library, to render the text within the bounding box of the polygon. Subsequent rotation and scaling operations are applied to precisely constrain the text within the specified polygonal area. The resulting rendered text is shown in Fig. 5(a).

For single-line text, complex layout handling is unnecessary. Therefore, the text is rendered directly onto a single-line image strip, as shown in Fig. 5(b). This approach not only reduces computational overhead during inference but also provides a clear glyph representation, which is particularly advantageous for small target regions.

Prompt design. Following the paradigm of in-context learning in diffusion models [HWW*24], we additionally provide a descriptive text prompt to accompany each input image. The prompt is designed to clarify the roles of the two concatenated images and the target text content. It follows the template: *"The pair of images highlights some white words on a black background, as well as their style on a real-world scene image. [IMAGE1] is a template image rendering the text, with the words {words}; [IMAGE2] shows the text content {words} naturally and correspondingly integrated into the image."* Here, "*{words}*" is replaced by the actual text to be rendered. During training, this prompt guides the model to understand the semantic relationship between the glyph template and the scene image.

Consequently, by spatially concatenating $\mathbf{I}_{\text{glyph}}$ and $\mathbf{I}_{\text{scene}}$ into a unified input $\mathbf{I}_{\text{concat}}$, TextFlux offers a direct and information-rich visual guidance mechanism. Then, this conditional image is concatenated with a noise image and a mask to form the complete input for the diffusion model, as described in Equation 1. This design enables the model to concentrate on its well-developed pre-trained capabilities for contextual understanding and visual fusion, facilitating the efficient synthesis of high-quality scene text.

3.4. Model Training and Inference

To train the model, we adopt a flow-matching objective as introduced in the Flux framework [Lab24]. Given a clean latent representation $\mathbf{x}_0$, a noise vector $\mathbf{z}_1 \sim \mathcal{N}(0, \mathbf{I})$, and a noise scale σ_t associated with the random time step t, the noisy latent input is generated by convex interpolation:

$$\mathbf{x}_t = (1 - \sigma_t)\,\mathbf{x}_0 + \sigma_t\,\mathbf{z}_1. \tag{2}$$

The model is trained to predict the velocity between $\mathbf{x}_0$ and $\mathbf{z}_1$, with the training loss defined as:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{z}_1}\left[\omega_t \cdot \|\hat{\mathbf{v}}_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\mathbf{z}_1 - \mathbf{x}_0)\|_2^2\right], \tag{3}$$

where $\hat{\mathbf{v}}_\theta$ is the model prediction, ω_t is a time-dependent weighting factor, and $\mathbf{c}$ includes the conditioning features such as the concatenated image, text prompt embeddings, and inpainting mask features. No additional perceptual loss is used, keeping the training objective simple and stable.

During the inference stage, as illustrated in Fig. 3, the user provides three inputs: a scene image to be edited, a binary mask indicating the target text region, and the desired text content. The pipeline automatically generates a glyph-based template image, concatenates it with the input scene image, and feeds the result into the model. The output image is cropped to remove the template region, resulting in the final edited scene image.

3.5. Implementation Details

Our method is built on the pre-trained FLUX.1-Fill-de v, a latent rectified flow transformer model for image synthesis. For training, we set the batch size to 1 and use the gradient accumulation of 8. We employ the AdamW optimizer with a constant learning rate of 2e-5, running for 30,000 iterations in total. Since resolution is critical for scene text tasks, we develop a specialized data augmentation approach by resizing the image's longer side to 512, 640, 768, 896, or 1024, thus obtaining input images of various resolutions. During training, we directly mix data from different languages.

We train two versions of TextFlux. The first one employs full-parameter training on two A100 (80 GB) GPUs; to support this computational scale, we utilize DeepSpeed ZeRO-Stage 2 for memory optimization. The second version is trained via LoRA on a single A100 (80 GB) GPU with a LoRA rank of 128.

4. Experiment

4.1. Datasets and Evaluation Metrics

Datasets. In previous studies, large-scale datasets are commonly employed for multilingual visual text generation tasks. For instance, the AnyWord-3M [TXH*23] dataset contains approximately three million publicly sourced multilingual images, while

Table 1: ***Quantitative comparison of multi-line text synthesis metrics against baselines.*** *We use Sequence Accuracy (SeqAcc) as the main evaluation metric to measure recognition correctness. The best scores are highlighted in bold. The second-best results are underlined. FID and LPIPS are computed on the ReCTS dataset. The User Study (US) is conducted to capture human evaluations regarding the overall quality of generated images (score range: 0–10). Detailed FID and LPIPS results on other datasets, as well as additional Normalized Edit Distance (NED) results, are provided in the appendix.*

Method	SeqAcc-Recon (%)↑				SeqAcc-Editing (%)↑				FID↓	LPIPS↓	US↑
	AnyWord(EN)	AnyWord(CH)	TotalText	ReCTS	AnyWord(EN)	AnyWord(CH)	TotalText	ReCTS			
Flux [Lab24]	43.0	9.3	29.5	4.8	11.6	0.0	11.5	0.0	18.25	0.1431	4.5
AnyText [TXH*23]	14.8	24.1	6.5	20.6	13.7	19.2	4.6	18.5	22.57	0.4095	3.8
AnyText2 [TGB24]	23.7	28.1	15.5	25.2	17.0	24.2	15.0	23.6	21.75	0.3054	4.3
TextFlux(LoRA)	76.7	50.8	62.3	56.6	61.1	32.8	35.4	32.1	12.09	0.1038	7.4
TextFlux	**77.3**	**61.4**	**62.9**	**64.1**	**63.8**	**40.7**	**36.2**	**37.2**	**11.02**	**0.0975**	**8.0**

Table 2: *Quantitative comparison of single-line text generation metrics against baselines, where '*' indicates results obtained by generating single-line text using a multi-line layout.*

Method	SeqAcc-Recon (%)↑				SeqAcc-Editing (%)↑			
	AnyWord(EN)	AnyWord(CH)	TotalText	ReCTS	AnyWord(EN)	AnyWord(CH)	TotalText	ReCTS
TextDiffuser-2 [CHL*24a]	57.7	2.9	27.1	3.8	39.4	0.0	21.8	0.0
UdiffText [ZL23]	68.6	6.2	36.7	7.3	58.5	0.0	31.7	0.0
DreamText [WZZJ24]	69.6	6.5	36.7	7.6	58.5	0.0	31.8	0.0
AnyText [TXH*23]	34.8	31.7	11.4	36.2	30.9	28.4	10.5	30.4
AnyText2 [TGB24]	45.1	35.9	20.5	41.5	42.0	37.5	21.3	34.6
Flux [Lab24]	53.6	6.6	63.2	10.0	42.1	0.0	41.6	0.0
Flux-Text [LBD*25]	**81.0**	**71.9**	**66.4**	68.5	54.9	50.8	44.4	44.5
TextFlux(LoRA)*	80.1	52.7	66.1	63.4	55.5	36.8	**45.0**	37.2
TextFlux*	80.3	62.3	65.3	68.5	56.2	48.2	41.9	40.6
TextFlux(LoRA)	78.9	58.8	65.0	69.3	56.8	41.0	42.0	48.4
TextFlux	79.3	64.3	66.0	**71.9**	**59.0**	**52.2**	44.1	**51.5**

the MARIO-10M [CHL*23] dataset comprises around ten million images that are primarily in English, though a small portion may include other languages. In contrast to these large-scale datasets, we use a relatively small training set of 30,405 images: approximately 10,000 in English, 15,000 in Chinese, and 1,000 each for Japanese, Korean, French, German, and Italian. Specifically, the English data primarily come from MLT2017 [MLT19], TotalText [CCL20], and CTW1500 [CCL20] training sets commonly used in OCR-related tasks [XQG*24, YZZ*23]; the Chinese data are mainly derived from the ReCTS [ReC19] and RCTW [RCT17] training sets; the remaining languages are obtained from the MLT2019 [MLT19] competition data.

For validation, we use the test set provided in [TXH*23] from the AnyWord-3M dataset, which includes 1,000 English and 1,000 Chinese images. To further evaluate our method under more challenging conditions, we additionally include two harder test sets: TotalText [CCL20] test set for English, featuring 300 images with curved and arbitrarily shaped text, and the ReCTS [ReC19] test set for Chinese, consisting of 2,000 real-world images with diverse and complex layouts. These datasets provide a more rigorous benchmark for assessing the robustness and generalization capability of our method, particularly under complex and diverse text conditions. Additionally, to showcase more intuitive results from TextFlux, we compared it with some existing closed-source and open-source commercial diffusion models in Sec 4.3, such as GPT-Image-1 [Ope25], Nano Banana [Goo25], and Qwen-Image-Edit [WLZ*25]. The results are presented in Table 3, and all test results were evaluated manually.

Evaluation. We evaluate our method on two tasks: scene text reconstruction and scene text editing. In scene text reconstruction, the text image is reconstructed by rendering text in the masked region using the words directly from the ground truth text labels. In scene text editing, the original words in the labels are replaced with a random word. For evaluation, we use off-the-shelf scene text recognition (STR) models to calculate recognition accuracy, primarily measured by Sentence Accuracy (Sen. Acc), with additional analysis using Normalized Edit Distance (NED) provided in the appendix.

To further evaluate the difference between synthetic and real images, we use Frechet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS) to assess the visual fidelity of the generated images. In addition, we conduct a user study, where participants are asked to rate the generated results on a scale from 0 to 10, based on overall visual quality and realism. The averaged user scores serve as a subjective evaluation to complement the quantitative metrics.

Figure 6: ***Comparison of scene text synthesis methods:*** *AnyText, AnyText2, and our TextFlux. More results are available in the appendix.*

4.2. Quantitative and Qualitative Results

Quantitative results. In our experiments, we adopt the evaluation metrics outlined in Section 4.1. Multi-line text synthesis presents unique challenges, including stronger contextual interference, potential mask region overlap, and difficulties in precise positional alignment. Therefore, we provide metrics separately for multi-line (Table 1) and single-line (Table 2) scenarios. The single-line results were obtained by randomly sampling three text instances from the multi-line dataset examples and generating them individually.

As shown by the multi-line metrics in Table 1, our method consistently outperforms the baseline approaches across all metrics and four benchmark datasets. Even the lightweight LoRA-tuned version surpasses all baselines, demonstrating the effectiveness and adaptability of our approach. When fully trained, our model particularly excels in Chinese text synthesis. On the SeqAcc-Recon metric, it achieves scores of 61.4 for AnyWord(CH) and 64.1 for ReCTS. Furthermore, on the more difficult SeqAcc-Editing metric, its performance on Chinese text, scoring 40.7 on AnyWord(CH) and 37.2 on ReCTS, also substantially exceeds that of baseline methods. Note that methods such as TextDiffuser-2 [CHL*24a], UDiffText [ZL23], and DreamText [WZZJ24] are excluded from Table 1 as they are inherently restricted to single-line generation. Turning to the single-line metrics in Table 2, we are able to include these single-line specialized methods and the concurrent work Flux-Text [LBD*25] for a more comprehensive comparison. As observed, while the concurrent Flux-Text shows strong performance in English reconstruction likely due to similar backbone architectures, TextFlux achieves state-of-the-art results in multilingual settings and editing tasks (e.g., surpassing FluxText by 7.0 points on ReCTS Editing). Furthermore, regarding the rendering strategy, employing the single-line text rendering approach achieves a 10.9% improvement in accuracy on the ReCTS editing task compared to the multi-line text rendering approach. This gain primarily stems from the clearer and more stable supervisory signals provided by the single-line text rendering approach, especially when the masked regions are particularly small. Notably, while the base Flux model is incapable of generating Chinese text, exhibiting zero accuracy on this task, the application of our method unlocks its multilingual text generation capabilities, achieving performance significantly superior to existing approaches.

Qualitative results. Fig. 6 shows multilingual text synthesis results generated by TextFlux under various challenging conditions, such as complex backgrounds, curved text, and handwritten styles. The visualizations demonstrate that TextFlux significantly outperforms existing methods in terms of character accuracy and image fidelity. In most cases, the generated results are nearly indistinguishable from real images. Additionally, we demonstrate the zero-shot ca-

Figure 7: ***Comparison of our method TextFlux with commercial image generation models: GPT-Image-1, Nano banana, and Qwen-Image-Edit.*** *GPT-Image-1 shows poor consistency and may modify text outside the desired editing area. Nano banana exhibits significant limitations with non-Latin languages. Qwen-Image-Edit performs well in English and Chinese but demonstrates clear limitations in localization and other non-Latin languages. TextFlux demonstrates leading performance in scene text synthesis.*

pability in the appendix, which can render languages not included in the training set, such as minority languages.

We showcase zero-shot visualization results in Fig. 8, where the model, tasked with generating text unseen during training, consistently demonstrates strong text rendering capabilities. These results suggest that our model does not merely memorize and reproduce trained glyphs but has instead learned a more generalizable and profound capability: to stylistically fuse any given visual glyph reference with the scene context. This generalizable capability is also key to TextFlux's efficiency in handling multilingual text and its strong adaptability to low-resource languages.

Table 3: *SeqAcc-Editing results on 55 challenging, manually annotated images, where the test set features characteristics such as complex glyphs, small text, and arbitrarily shaped text.*

Method	SeqAcc-Editing(%)
GPT-Image-1 [Ope25]	60.0
Nano Banana [Goo25]	36.7
Qwen-Image-Edit [WLZ*25]	38.2
TextFlux(LoRA)	67.2
TextFlux	71.0

4.3. Comparative Analysis with Commercial Models

We manually evaluated 55 images, encompassing five languages: English, Chinese, Japanese, Korean, and German, with 15 images for Chinese and 10 images for each of the remaining languages. All evaluations were conducted manually, primarily serving as a stress test to assess model performance under challenging conditions. For English, which is generally considered simpler, the test set included 5 arbitrarily-shaped texts and 5 small texts. The Chinese set comprised 5 text synthesis tasks of normal difficulty, 5 with complex glyphs, and 5 with small texts. For languages other than English and Chinese, images of general difficulty were used. The

Table 4: *SeqAcc-Recon results on the ReCTS and TotalText datasets using different training strategies.*

Strategy	ReCTS	TotalText
No Concat + LoRA	5.2	29.8
Concat + No Train	9.2	26.2
Concat + LoRA	54.6	62.3
Concat + Full-Param	**64.1**	**62.9**

Table 5: *Evaluating different text encoders based on SeqAcc-Recon results when provided with empty input prompts.*

CLIP	T5	ReCTS	TotalText
✓	✓	**64.1**	**62.9**
✗	✓	64.0	62.7
✓	✗	63.8	55.5
✗	✗	63.6	55.1

Figure 8: ***Zero-shot synthesis of unseen scripts and characters.*** *The results include rare Chinese characters not present in training and also demonstrate successful generation in Mongolian (Cyrillic script) and Russian, which are languages the model has never seen. These results highlight the generalization ability of TextFlux to novel glyphs.*

Figure 9: *Effectiveness of style adaptation. By injecting the identical glyph mask into different contexts, TextFlux renders significantly different styles, proving its capability for high-fidelity contextual reasoning beyond simple glyph pasting.*

final visual results and quantitative metrics are presented in Fig. 7 and Table 3, respectively.

As shown in Fig. 7, although commercial image generation models demonstrate strong performance, they still face several challenges. For instance, while GPT-Image-1 [Ope25] demonstrates good character accuracy, the visual results in Fig. 7 clearly indicate a severe degradation in image consistency, with some originally correct characters even being erroneously altered. Qwen-Image-Edit [WLZ*25] exhibits strong performance in Chinese and English but shows lower accuracy for other languages. Nano Banana [Goo25], while demonstrating exceptional English editing capabilities, performs less robustly across other Latin-script languages. Overall, TextFlux demonstrates strong performance under these stress-test conditions. On one hand, this highlights the importance of masks for precise text localization: mask-free methods struggle with small text, and prompt-only generation in blank regions is often unreliable. On the other hand, it reveals that current commercial image generation models remain largely confined to simple cases. Specialized research is therefore still essential for tackling more challenging tasks, such as multilingual scene text synthesis.

It is worth noting that during evaluation, a sample is considered positive if the target text to be edited is correctly modified—even if other parts of the image contain errors. For instance, even though GPT-Image-1 may erroneously alter characters elsewhere, it is still counted as a positive sample as long as the specific characters requiring prediction are correctly generated. To ensure robustness in evaluation, we generate two outputs for each input image. A sample is considered positive if at least one of the two generations correctly edits the target text.

4.4. Ablation Study

Effectiveness of Concatenation Strategies We first analyze the impact of the proposed concatenation strategy and different fine-tuning approaches, with results presented in Table 4. (1) Training directly on the original images without concatenation (No concat + LoRA) achieves a very low sequence accuracy of 5.2% on ReCTS, failing to generate readable Chinese text, indicating the base model's limitation. (2) Using the concatenation strategy without further training (Concat + No train) shows a basic ability to render Chinese. (3) Applying LoRA fine-tuning after concatenation (Concat + LoRA) achieves remarkable performance), highlighting

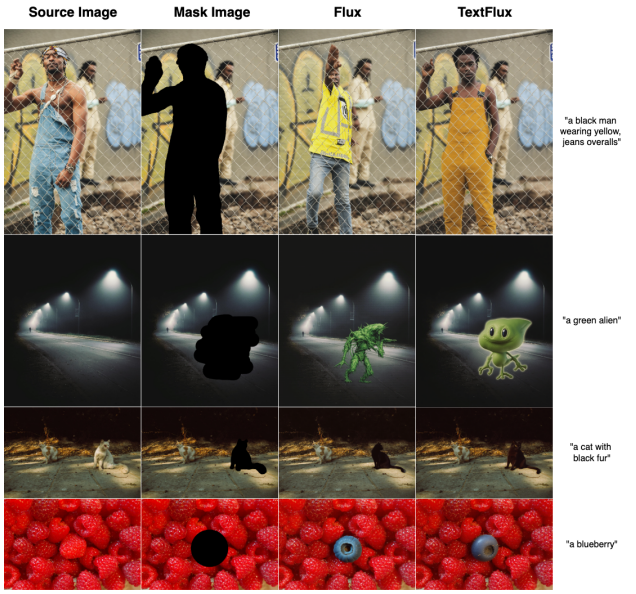

Figure 10: *Visualization of general inpainting tasks using Flux and TextFlux under the same prompt and mask conditions. Prompt texts are shown on the right.*

the effectiveness of glyphs as contextual cues even with limited parameter updates. (4) Full-parameter fine-tuning (Concat + Full-Param) yields the best results, confirming the strategy's scalability and its ability to fully unlock multilingual capabilities.

Impact of Text Encoders on Text Rendering Quality We investigate the necessity of text encoders for text rendering in TextFlux, given its primary reliance on visual contextual reasoning. While prior works often emphasize the importance of powerful language modeling in text-to-image generation tasks, we aim to revisit this assumption in the specific context of multilingual visual text generation. Therefore, we train the model by setting the prompts of CLIP or T5 to empty during training to examine the role of textual guidance. Interestingly, as shown in Table 5, our results reveal that removing either the CLIP or T5 encoder individually leads to only marginal changes in rendering performance for non-Latin scripts. For Latin-based languages, removing the T5 encoder results in a 7.4% performance drop, but the overall rendering quality remains at a high level. These findings suggest that for diffusion models equipped with strong contextual reasoning capabilities, high-quality text rendering can be achieved solely guided by visual context. For future work aiming to further enhance non-Latin text generation capabilities, developing a more efficient text encoder for non-Latin scripts remains a potential research avenue.

Visual Analysis of the "Copy-Paste" Effect. We introduce Fig. 9 as a qualitative ablation study to validate the generation mechanism. The figure demonstrates that, even when provided with identical style-less glyph inputs, TextFlux generates a diverse range of results, ranging from simple to highly stylized, by adaptively responding to the complexity of different scene contexts. This provides intuitive evidence that our method effectively leverages the model's robust contextual reasoning capabilities for high-fidelity integration, rather than merely performing a simple "copy-paste" operation.

4.5. Qualitative Comparison with Flux on General Inpainting Tasks

To further assess the scalability of TextFlux, we evaluate its general inpainting capability against the original Flux model. Specifically, we select the first few sample images from the evaluation benchmark provided in the official Flux codebase [Lab24] and compare TextFlux with Flux under identical mask and prompt conditions. As shown in Fig. 10, TextFlux achieves inpainting performance on par with Flux in handling various types of inpainting tasks.

Specifically, in the first line of the Fig. 10, TextFlux can accurately understand the prompt "a black man wearing yellow, jeans overalls" and perform a natural and reasonable clothing replacement. The generated result even surpasses the original Flux in terms of visual style and background consistency. In the reconstruction of imaginary objects (such as "a green alien"), detail restoration (such as replacing with "a blueberry"), and animal editing tasks (such as "a cat with black fur"), the generation quality of TextFlux is also comparable to Flux.

These results show that although TextFlux is designed for text image synthesis tasks, its adaptation ability in general inpainting scenarios is still preserved. This lays a foundation for extending the method in this paper to broader multi-modal image editing tasks in the future.

5. Conclusion and Limitations

In this paper, we propose TextFlux, an OCR-free method that leverages the inherent capabilities of diffusion models to address the intrinsic conflict between generating precise glyph structures and achieving contextually consistent styles. The method not only offers architectural simplicity and significant data efficiency, but also demonstrates strong performance across various aspects, including multilingual support, multi-line editing, complex glyph rendering, and zero-shot generalization. This enables straightforward extension to a wider range of low-resource languages, thereby laying the groundwork for enhanced language accessibility in scene text synthesis.

However, our method still has some limitations. First, although only approximately 1% of typical training data is required, training a Flux-based model remains computationally expensive (about four days of training on two 80GB A100 GPUs). Future work could explore more efficient training strategies, such as parameter-efficient fine-tuning, sparse training, or architecture search, to reduce computational costs. Second, the performance of our TextFlux is still unsatisfactory in the task of scene text synthesis for cursive languages, where character representations may differ based on their positions or connections (such as Arabic and Hindi). Addressing this may require more fine-grained guidance, such as stroke-based guides, to better capture the dynamic nature of these scripts. We are planning to address them in our future work.

References

[BHE23] Brooks T., Holynski A., Efros A. A.: Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2023), pp. 18392–18402. 3

[BNH*22] Balaji Y., Nah S., Huang X., Vahdat A., Song J., Zhang Q., Kreis K., Aittala M., Aila T., Laine S., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324* (2022). 3

[CCL20] Ch'ng C.-K., Chan C. S., Liu C.-L.: Total-text: toward orientation robustness in scene text detection. *International Journal on Document Analysis and Recognition (IJDAR) 23*, 1 (2020), 31–52. 6

[CHL*23] 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* (2023), 9353–9387. 3, 4, 6

[CHL*24a] 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* (2024), Springer, pp. 386–402. 2, 6, 7

[CHL*24b] Chung H. W., Hou L., Longpre S., Zoph B., Tay Y., Fedus W., Li Y., Wang X., Dehghani M., Brahma S., et al.: Scaling instruction-finetuned language models. *Journal of Machine Learning Research 25*, 70 (2024), 1–53. 3

[CMGS25] Chen L., Mao Q., Gu Y., Shou M. Z.: Edit transfer: Learning image editing via vision in-context relations. *arXiv preprint arXiv:2503.13327* (2025). 3

[CXL*25] Chen H., Xu X., Li W., Ren J., Ye T., Liu S., Chen Y.-C., Zhu L., Wang X.: Posta: A go-to framework for customized artistic poster generation. *arXiv preprint arXiv:2503.14908* (2025). 3

[CYG*23] Chen J., Yu J., Ge C., Yao L., Xie E., Wu Y., Wang Z., Kwok J., Luo P., Lu H., Li Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426* (2023). 3

[DCC*25] 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). 2

[Dee23] DeepFloyd: Github link: https://githubcom/deep-floyd/if, 2023. URL: https://github.com/deep-floyd/IF. 3

[DN21] Dhariwal P., Nichol A.: Diffusion models beat gans on image synthesis. *Advances in neural information processing systems 34* (2021), 8780–8794. 3

[FLW*25] Fang Z., Lyu P., Wu J., Zhang C., Yu J., Lu G., Pei W.: Recognition-synergistic scene text editing. *arXiv preprint arXiv:2503.08387* (2025). 3

[FMW*24] Feng K., Ma Y., Wang B., Qi C., Chen H., Chen Q., Wang Z.: Dit4edit: Diffusion transformer for image editing. *arXiv preprint arXiv:2411.03286* (2024). 3

[FZF*23] Feng W., Zhu W., Fu T.-j., Jampani V., Akula A., He X., Basu S., Wang X. E., Wang W. Y.: Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems 36* (2023), 18225–18250. 3

[GGG*25] Gao Y., Gong L., Guo Q., Hou X., Lai Z., Li F., Li L., Lian X., Liao C., Liu L., et al.: Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346* (2025). 3

[GHL*25] 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). 3

[GLL*25] Gao Y., Lin Z., Liu C., Zhou M., Ge T., Zheng B., Xie H.: Postermaker: Towards high-quality product poster generation with accurate text rendering. *arXiv preprint arXiv:2504.06632* (2025). 3

[Goo25] Google: Gemini 2.5 flash image. https://ai.google.dev/gemini-api/docs/image-generation#gemini, 2025. 2, 3, 6, 8, 9

[HMT*22] Hertz A., Mokady R., Tenenbaum J., Aberman K., Pritch Y., Cohen-Or D.: Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022). 3

[HWW*24] Huang L., Wang W., Wu Z.-F., Shi Y., Dou H., Liang C., Feng Y., Liu Y., Zhou J.: In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775* (2024). 3, 5

[HXL*24] Hu X., Xu K., Liu B., Liu Q., Fei H.: Amo sampler: Enhancing text rendering with overshooting. *arXiv preprint arXiv:2411.19415* (2024). 3

[Lab24] Labs B. F.: Flux: Official inference repository for flux.1 models, 2024. Accessed: 2024-11-12. URL: https://github.com/black-forest-labs/flux. 2, 3, 4, 5, 6, 10

[LBD*25] Lan R., Bai Y., Duan X., Li M., Jin D., Xu R., Nie D., Sun L., Chu X.: Flux-text: A simple and advanced diffusion transformer baseline for scene text editing. *arXiv preprint arXiv:2505.03329* (2025). 6, 7

[LCBH*22] Lipman Y., Chen R. T., Ben-Hamu H., Nickel M., Le M.: Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022). 3

[LLZ*24] Liu Z., Liang W., Zhao Y., Chen B., Liang L., Wang L., Li J., Yuan Y.: Glyph-byt5-v2: A strong aesthetic baseline for accurate multilingual visual text rendering. *arXiv preprint arXiv:2406.10208* (2024). 3

[LST*24] Li H., Shen C., Torr P., Tresp V., Gu J.: Self-disco vering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 12006–12016. 3

[LYK*24] Li M., Yang T., Kuang H., Wu J., Wang Z., Xiao X., Chen C.: Controlnet++: Improving conditional controls with efficient consistency feedback: Project page: liming-ai. github. io/controlnet_plus_plus. In *European Conference on Computer Vision* (2024), Springer, pp. 129–147. 3

[MDC*24] Ma J., Deng Y., Chen C., Du N., Lu H., Yang Z.: Glyphdraw2: Automatic generation of complex glyph posters with diffusion models and large language models. *arXiv preprint arXiv:2407.02252* (2024). 2, 3

[MLT19] Icdar 2019 robust reading challenge on multi-lingual scene text detection and recognition. https://rrc.cvc.uabes/?ch=15, 2019. 6

[MTJ*25] Ma N., Tong S., Jia H., Hu H., Su Y.-C., Zhang M., Yang X., Li Y., Jaakkola T., Jia X., et al.: Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732* (2025). 3

[MWX*24] 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* (2024), vol. 38, pp. 4296–4304. 3

[MZC*23] Ma J., Zhao M., Chen C., Wang R., Niu D., Lu H., Lin X.: Glyphdraw: Seamlessly rendering text with intricate spatial structures in text-to-image generation. *arXiv preprint arXiv:2303.17870* (2023). 3

[Ope25] OpenAI: Gpt image-1. https://openai.com/index/introducing-4o-image-generation/, 2025. 2, 3, 6, 8, 9

[PEL*23] 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). 2

[PX23] Peebles W., Xie S.: Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (2023), pp. 4195–4205. 2, 3

[RBL*21] Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B.: High-resolution image synthesis with latent diffusion models, 2021. arXiv:2112.10752. 2, 3

[RCT17] Icdar2017 competition on reading chinese text in the wild. https://rctwvlrlab.net/dataset, 2017. 6

[ReC19] Icdar 2019 robust reading challenge on reading chinese text on signboard. https://rrc.cvc.uab.es/?ch=12, 2019. 6

[RKH*21] Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., et al.: Learning transferable visual models from natural language supervision. In *International conference on machine learning* (2021), PmLR, pp. 8748–8763. 3

[RLJ*23] Ruiz N., Li Y., Jampani V., Pritch Y., Rubinstein M., Aberman K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2023), pp. 22500–22510. 3

[RLJ*24] Ruiz N., Li Y., Jampani V., Wei W., Hou T., Pritch Y., Wadhwa N., Rubinstein M., Aberman K.: Hyperdreambooth:Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2024), pp. 6527–6536. 3

[SCC*22] Saharia C., Chan W., Chang H., Lee C., Ho J., Salimans T., Fleet D., Norouzi M.: Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings* (2022), pp. 1–10. 3

[SCS*22] Saharia C., Chan W., Saxena S., Li L., Whang J., Denton E. L., Ghasemipour K., Gontijo Lopes R., Karagol Ayan B., Salimans T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems 35* (2022), 36479–36494. 3

[TGB24] Tuo Y., Geng Y., Bo L.: Anytext2: Visual text generation and editing with customizable attributes. *arXiv preprint arXiv:2411.15245* (2024). 2, 6

[TLY*24] Tan Z., Liu S., Yang X., Xue Q., Wang X.: Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098* (2024). 3

[TXH*23] Tuo Y., Xiang W., He J.-Y., Geng Y., Xie X.: Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054* (2023). 2, 3, 4, 5, 6

[TXY*25] Tan Z., Xue Q., Yang X., Liu S., Wang X.: Ominicontrol2: Efficient conditioning for diffusion transformers. *arXiv preprint arXiv:2503.08280* (2025). 3

[WGW*24] Wu Z., Gao H., Wang Y., Zhang X., Wang S.: Universal prompt optimizer for safe text-to-image generation. *arXiv preprint arXiv:2402.10882* (2024). 3

[WHSX21] Wu X., Hu Z., Sheng L., Xu D.: Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 14618–14627. 3

[WLQ*25] Wang T., Liu T., Qu X., Wu C., Liu L., Hu X.: Glyphmastero: A glyph encoder for high-fidelity scene text editing. *arXiv preprint arXiv:2505.04915* (2025). 3

[WLZ*25] Wu C., Li J., Zhou J., Lin J., Gao K., Yan K., Yin S.-m., Bai S., Xu X., Chen Y., et al.: Qwen-image technical report. *arXiv preprint arXiv:2508.02324* (2025). 2, 3, 6, 8, 9

[WZZJ24] Wang Y., Zhang W., Zhou C., Jin C.: High fidelity scene text synthesis. *arXiv preprint arXiv:2405.14701* (2024). 2, 3, 6, 7

[XQG*24] Xie Y., Qiao Q., Gao J., Wu T., Fan J., Zhang Y., Zhang J., Sun H.: Dntextspotter: Arbitrary-shaped scene text spotting via improved denoising training. *arXiv preprint arXiv:2408.00355* (2024). 6

[YGY*23] 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* (2023), 44050–44066. 2, 3

[YKJ*24] Yu S., Kwak S., Jang H., Jeong J., Huang J., Shin J., Xie S.: Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940* (2024). 3

[YZZ*23] Ye M., Zhang J., Zhao S., Liu J., Liu T., Du B., Tao D.: Deepsolo: Let transformer decoder with explicit points solo for text spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 19348–19357. 6

[ZCW*24] Zhang L., Chen X., Wang Y., Lu Y., Qiao Y.: Brush your text: Synthesize any scene text on images via diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2024), vol. 38, pp. 7215–7223. 3

[ZL23] Zhao Y., Lian Z.: Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. *arXiv preprint arXiv:2312.04884* (2023). 2, 3, 6, 7

[ZRA23] 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* (2023), pp. 3836–3847. 3

[ZSL*24] 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* (2024), 138569–138594. 2, 3