

Beyond Isolated Words: Diffusion Brush for Handwritten Text-Line Generation

Gang Dai^{1*}, Yifan Zhang^{2,3*}, Yutao Qin^{1*}, Qiangya Guo¹, Shuangping Huang^{1,4†}, Shuicheng Yan³

¹South China University of Technology

²MiroMind AI ³National University of Singapore ⁴Pazhou Laboratory

eedaigang@mail.scut.edu.cn, yifan.zhang@miromind.ai, eehsp@scut.edu.cn

Abstract

Existing handwritten text generation methods primarily focus on isolated words. However, realistic handwritten text demands attention not only to individual words but also to the relationships between them, such as vertical alignment and horizontal spacing. Therefore, generating entire text line emerges as a more promising and comprehensive task. However, this task poses significant challenges, including the accurate modeling of complex style patterns—encompassing both intra- and inter-word relationships—and maintaining content accuracy across numerous characters. To address these challenges, we propose DiffBrush, a novel diffusion-based model for handwritten text-line generation. Unlike existing methods, DiffBrush excels in both style imitation and content accuracy through two key strategies: (1) content-decoupled style learning, which disentangles style from content to better capture intra-word and inter-word style patterns by using column- and row-wise masking; and (2) multi-scale content learning, which employs line and word discriminators to ensure global coherence and local accuracy of textual content. Extensive experiments show that DiffBrush excels in generating high-quality text-lines, particularly in style reproduction and content preservation. Code is available at <https://github.com/dailenson/DiffBrush>

1. Introduction

Handwritten text generation aims to automatically synthesize realistic handwritten text images that visually convey a user’s personal writing style (e.g., text slant, stroke width, ligatures) while ensuring the content readability. This task has broad applications, including assisting individuals with writing difficulties, accelerating handwritten font design, and enriching data for text recognizer. Most existing methods [1, 3, 7, 11, 12, 35, 44] focus on generating handwritten

* Authors contributed equally.

† Corresponding author

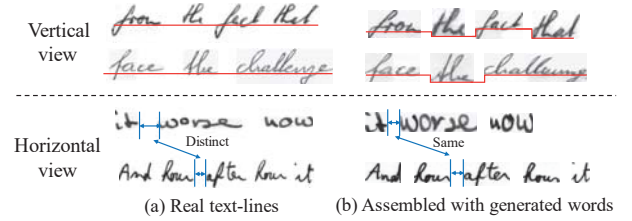


Figure 1. Comparison of handwritten text-lines (a) written by real writers, and (b) assembled with isolated words generated from One-DM [7]. The latter one applies fixed inter-word spacing due to the lack of spacing information in generated words. Red lines indicate the baseline (i.e., the reference line at the bottom of the characters), while blue lines highlight word spacing.

images at the word level, with few efforts [9, 23] exploring the generation of complete text lines. To bridge this gap, our work focuses on high-quality handwritten text-line generation with better control over both style and content.

Most previous state-of-the-art methods [7, 12, 40, 44] focus on handwritten word generation by using reference images from writers as style inputs and conditioning on character-wise labels or images for content. This allows for the synthesis of handwritten words with controllable styles and specific content. However, as shown in Figure 1, generating text at the word level cannot effectively capture the cohesive style of a complete text line: (1) Humans generally maintain consistent vertical alignment across words, while synthesized words often exhibit arbitrary vertical positioning. (2) Different writers have unique word spacing characteristics that are often lost in isolated word generation.

Direct approaches for text-line generation are relatively limited, with two notable GAN-based methods proposed. TS-GAN [9] optimizes a global content recognition loss based on the entire generated text-line image, primarily guiding content learning while implicitly influencing style learning. CSA-GAN [23], in contrast, leverages both a content recognition loss and a writer classification loss computed from the generated text-line image, thus better modeling style through writer identity supervision. However, both methods suffer from two key limitations: 1) **Ineffective style extraction**: Since both methods jointly optimize content and style from the same model output, the two aspects interfere with each other, preventing effective learn-

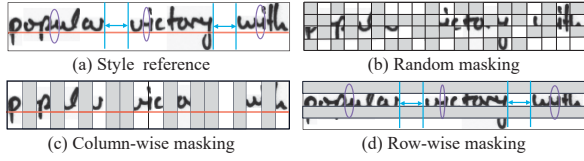


Figure 2. (a) This style reference exhibits rich style patterns, *e.g.*, ligatures, spacing, and vertical alignment alongside undesired content information: “popular victory with”. (b) Random masking disrupts both style features and content information. (c) Column-wise masking maintains style patterns (*e.g.*, character style and vertical alignment) while removing horizontal content information. (d) Row-wise masking preserves joins and spacing while disrupting vertical content.

ing of either. For instance, minimizing content recognition loss inevitably pushes these models to produce easily recognizable outputs with simplified styles (*e.g.*, regular fonts and standard strokes, as shown in Figure 9), ultimately hindering the faithful mimicking of diverse handwriting styles. 2) **Difficulty in maintaining character-level accuracy:** Ensuring the readability of text lines with numerous characters remains challenging. For instance, in datasets like IAM [36], where a single text line averages 42 characters—approximately six times the length of a typical word. Optimizing content loss at the text-line level encourages global correctness but may fail to preserve individual character accuracy, making it difficult to maintain content integrity across the entire generated text (cf. D_{CER} and D_{WER} columns in Table 1).

To generate handwritten text lines with improved content accuracy and style fidelity, we introduce **DiffBrush**, a novel diffusion-based approach. Our method incorporates two key strategies: (1) **content-decoupled style learning**, and (2) **multi-scale content learning**.

Content-decoupled style learning aims to disrupt content information of style references while preserving key style patterns. This eliminates content interference, thus achieving effective one-shot style learning (cf. Table 1 and Figure 6). A naive approach, such as random masking, fails as it disrupts both content and key style features like word spacing and vertical alignment (cf. Figure 2(b)). To address this, we propose column- and row-wise masking (cf. Figure 2(c), (d)), which selectively disrupts content while preserving critical style patterns. For style enhancement in vertical direction, as shown in Figure 3, we apply column-wise masking to the extracted style features, maintaining vertical alignment while removing horizontal content information. A Proxy-NCA loss [25, 38] is then used to enforce style consistency within writers while distinguishing different writers, enabling the vertical enhancing head to refine vertical alignment. For style enhancement in horizontal direction, row-wise masking preserves word and character spacing while disrupting vertical content, allowing the horizontal enhancing head to reinforce horizontal spacing patterns. This novel masking strategy effectively disentangles

style from content, enabling more accurate and independent style representation in handwritten text-line generation.

Multi-scale content learning seeks to enhance content accuracy at both global and local levels: at the global level, we preserve character order within a text line to maintain contextual relationships between characters, while at the local level, we ensure the structural correctness of each individual word. To achieve this, we develop a novel multi-scale content discriminator. The line content discriminator segments the text-line image and processes it with a 3D CNN [55] to capture global contextual relationships, encouraging the generator to maintain proper character sequencing. Meanwhile, the word discriminator employs an attention mechanism to isolate individual words and verify their content accuracy, guiding the generator to refine local text content. Our empirical results (cf. Figure 7) show that this multi-scale content discriminator significantly improves content accuracy without hindering style imitation quality.

Our main contributions include: (1) To the best of our knowledge, DiffBrush is among the first to leverage diffusion generative models for handwritten text-line generation. (2) DiffBrush introduces a novel content-decoupled style learning strategy that significantly enhances style imitation, along with a new multi-scale content learning strategy that boosts content accuracy. (3) Extensive experiments on two popular English handwritten datasets (cf. Table 1 and Figure 6) and one Chinese dataset (cf. Figure 11) demonstrate that DiffBrush significantly outperforms state-of-the-arts.

2. Related Work

Handwritten text generation methods are generally divided into online and offline: the former synthesizes dynamic stroke sequences, while the latter generates static text images. With the advancement of deep learning, Transformer decoders [6] and diffusion models [8, 34, 47] have been used for synthesizing online handwritten text. However, as highlighted in recent studies [3, 7, 44], online methods require temporal data (*e.g.*, coordinate points and writing orders) collected from a digital stylus pen and cannot synthesize stroke width, ink color like offline methods. In light of this, this paper focuses on offline handwriting generation.

The advent of Generative Adversarial Networks [20, 33] has accelerated the development of offline handwritten text generation. Early works [1, 11] use character labels as content inputs and random noise as style inputs to synthesize handwritten words with controllable content and random styles. To enhance style control, SLOGAN [35] conditions style inputs on fixed writer IDs but fails to mimic unseen styles. Unlike them, GANwriting [22] and HWT [3] employ CNN or transformer encoder to extract style features from style references and are thus capable of imitating any styles. Further, VATr [44] utilizes symbol images as content representations, enabling character generation beyond

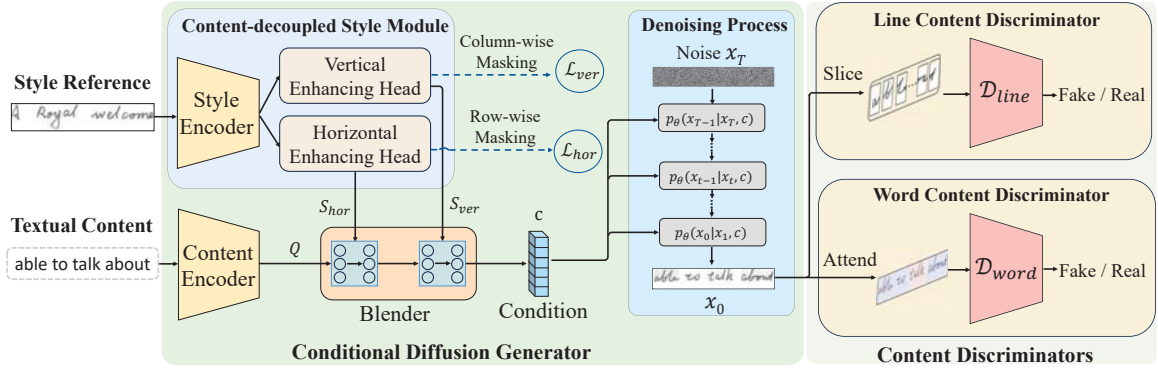


Figure 3. Method overview. Our DiffBrush consists of a conditional diffusion generator, a content-decoupled style module, and multi-scale content discriminators. The style module extracts two enhanced style features S_{hor} and S_{ver} , which are blended with the content representation Q from the content encoder to construct the condition vector c . This condition is then used to guide the denoising process to generate images. To enhance style learning, we explore column- and row-wise masking to effectively eliminate content interference in style modeling. The content discriminators provide feedback at both line and word levels, enhancing the readability of generated text. The slice operation divides the text-line image into horizontal segments, while the attend operation locates word positions in the text-line.

the training charset. In contrast to the above word-focused methods, TS-GAN [9] and CSA-GAN [23] are developed to synthesize handwritten text-lines. However, they struggle to produce satisfactory results due to design drawbacks in style learning and content supervision.

The rapid development of diffusion models [10, 19, 28, 58, 60] offers new potential for handwritten text generation. However, some early attempts [39, 67], which condition denoising process on fixed writer labels, cannot mimic unseen handwriting styles. To address this, DiffusionPen [40] and One-DM [7] extract style information from reference images, and then merge this information with the textual content to guide the denoising process. However, regarding text content readability, One-DM simply incorporates a text recognizer with a CTC decoder, while DiffusionPen neglects this challenge entirely. Different from them, we propose a novel multi-scale content learning strategy, significantly enhancing text content readability. Moreover, previous diffusion methods [7, 39, 40, 67] focus on generating isolated words whereas our DiffBrush aims for entire text-line generation. We discuss more related works about diffusion methods for general image generation in Appendix A.

3. Problem Statement and Preliminaries

Problem statement. Given a text string \mathcal{A} and a style reference s_i randomly sampled from an exemplar writer $w_i \in \mathcal{W}$, we aim to synthesize a handwritten text-line image x that captures the unique calligraphic style of w_i while accurately preserving the content of \mathcal{A} . Here, $\mathcal{A} = \{a_i\}_{i=1}^L$ represents a sequence of length L , where each a_i is a Unicode character, including lowercase and uppercase letters, digits, punctuation. The key challenges lie in accurately capturing handwriting styles, including both intra- and inter-word patterns from the style reference, while ensuring the readability of text-lines that typically contain numerous characters.

Conditional diffusion model. The diffusion model [17] generates realistic images by progressively denoising a random Gaussian noise input. To achieve controllable generation, the conditional diffusion model [4, 49, 62–64, 66] incorporates a condition signal c to guide the denoising process. Starting from pure Gaussian noise $x_T \sim \mathcal{N}(0, \mathcal{I})$, a denoising network p_θ iteratively refines the image over multiple timesteps to produce the target image x_0 . The network p_θ , typically based on a U-Net architecture [50], integrates c via cross-attention or adaptive modulation layers. The training objective minimizes the mean squared error (MSE) between the predicted and real images: $\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_t, c} [\|x_0 - x_{\text{real}}\|^2]$. By leveraging condition signals such as text prompts and reference images, conditional diffusion models enable fine-grained control over the generation.

4. Method

4.1. Overall Scheme

To generate handwritten text-lines with enhanced content accuracy and style fidelity, we propose DiffBrush, a novel conditional diffusion generation method. As shown in Figure 3, the architecture of DiffBrush consists of three main components: content-decoupled style module, conditional diffusion generator, and multi-scale content discriminators.

The content-decoupled style module ξ_{style} aims to better capture the text-line styles of exemplar writers. To achieve this, we introduce a content-decoupled style learning strategy (cf. Section 4.2), which leverages two novel content-masking techniques and a style learning loss $\mathcal{L}_{\text{style}}$ to enhance text-line style modeling (cf. Figure 3). The extracted style features are then fused with content features from a content encoder to form the condition vector c within a blender module. Both the content encoder and blender module are designed based on One-DM [7], with further extensions (cf. Appendix B for details). Guided by c , the condi-

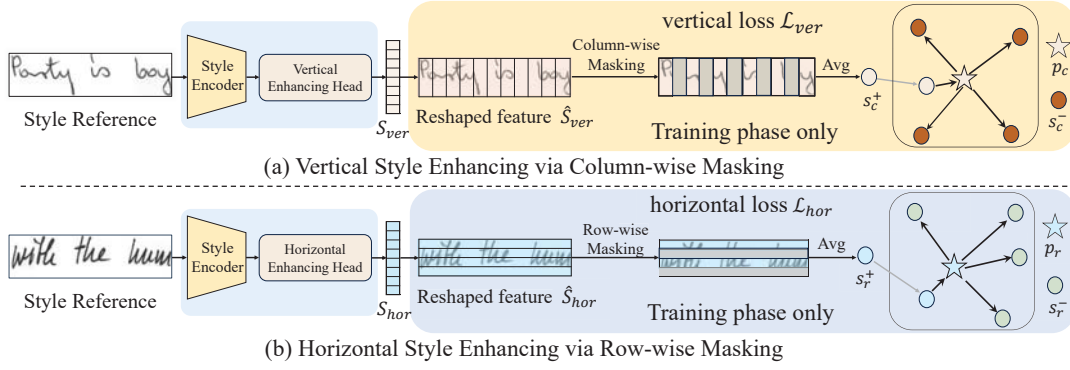


Figure 4. Style learning via column- and row-wise masking. (a) An improved vertical style feature S_{ver} is extracted by the vertical enhancing head, guided by a vertical loss \mathcal{L}_{ver} . Specifically, during training, we begin reshaping S_{ver} into spatial feature \hat{S}_{ver} and perform column-wise masking on \hat{S}_{ver} . After average pooling, s_c^+ is drawn closer to its corresponding writer proxy p_c , while the negatives s_c^- belonging to different writers are pushed away by p_c . (b) Similarly, a style feature S_{hor} is extracted by the horizontal enhancing head. We reshape S_{hor} and conduct row-wise masking. After pooling, s_r^+ is linked to its writer proxy p_r , and negatives s_r^- are pushed away.

tional diffusion generator \mathcal{G} performs denoising process to synthesize realistic handwritten text-line image x_0 .

However, training the generator \mathcal{G} with solely the diffusion loss \mathcal{L}_{diff} is insufficient to ensure the content readability of generated text lines. To address this, we introduce a multi-scale content learning strategy (cf. Section 4.3). Specifically, we develop a multi-scale discriminator \mathcal{D} to evaluate the content correctness at both the line and word levels, thus providing more fine-grained content supervision $\mathcal{L}_{content}$ for content adversarial learning between \mathcal{G} and \mathcal{D} .

To summarize, the overall training objectives of our Diff-Brush combines all three loss functions:

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{diff} + \mathcal{L}_{style} + \lambda \mathcal{L}_{content}, \quad (1)$$

where λ serves as a trade-off factor, and we empirically set it to 0.05 in training.

4.2. Content-decoupled Style Learning

As discussed in Section 1, text-line style learning is hindered by content interference, leading to ineffective style extraction. To address this, we propose to disrupt content information (by masking) to eliminate content interference during style learning, forcing the model to focus on essential style patterns. Specifically, as shown in Figure 3, style samples are first processed by a CNN-Transformer style encoder to obtain an initial style feature S with rich calligraphic attributes. To refine style representation, we introduce two dedicated style-enhancing heads, each incorporating a standard self-attention layer to extract fine-grained style features: S_{ver} (vertical) and S_{hor} (horizontal). These features are learned by using row- and column-wise masking for content disruption. The overall content-decoupled style loss \mathcal{L}_{style} is formulated as the sum of a vertical enhancing loss \mathcal{L}_{ver} and a horizontal enhancing loss \mathcal{L}_{hor} :

$$\mathcal{L}_{style} = \mathcal{L}_{ver} + \mathcal{L}_{hor}. \quad (2)$$

Vertical style enhancing via column-wise masking. The vertical enhancing head aims to enhance style learning in the vertical direction (e.g., vertical alignment patterns) via column-wise masking. Specifically, as shown in Figure 4(a), we perform masking on S_{ver} by first reshaping the sequential feature S_{ver} back into spatial feature $\hat{S}_{ver} \in \mathbb{R}^{h \times w \times c}$. We then divide \hat{S}_{ver} into several columns and randomly mask a subset of columns with equal probability to obtain an average feature $s_c \in \mathbb{R}^{h \times n \times c}$, where $n = w \cdot \rho$ and ρ is the masking ratio. After column-wise masking, we adopt a Proxy-NCA loss [25, 38] for style learning, which enforces style consistency within writers while distinguishing different writers. Specifically, our vertical enhancing loss \mathcal{L}_{ver} assigns a proxy to each writer, treating it as an anchor to cluster the masked style features of the same writer while pushing apart those of different writers:

$$\mathcal{L}_{ver} = \frac{1}{|P_c^+|} \sum_{p_c \in P_c^+} \log \left(1 + \sum_{s_c \in S_c^+} e^{-f_c^+} \right) + \frac{1}{|P_c|} \sum_{p_c \in P_c} \log \left(1 + \sum_{s_c \in S_c^-} e^{f_c^-} \right), \quad (3)$$

where $S_c = \{s_c^i\}_{i=1}^N$ is a batch of masked style features, P_c denotes the set of proxies of all writers, and P_c^+ refers to the set of writers present in the current batch. For each proxy p_c , S_c is divided into a positive set S_c^+ , consisting of s_c from the same writer as p_c , and a negative set $S_c^- = S_c - S_c^+$. The similarity between positive pairs is $f_c^+ = \alpha(g(s_c, p_c) - \delta)$ for $s_c \in S_c^+$, and that of negative pairs is $f_c^- = \alpha(g(s_c, p_c) + \delta)$ for $s_c \in S_c^-$, where $g(\cdot)$ is the cosine similarity, $\delta > 0$ is a margin and α is a scaling factor.

Horizontal style enhancing via row-wise masking. The horizontal enhancing head aims to enhance style learning in the horizontal direction (e.g., word and character spacing) via row-wise masking. As shown in Figure 4(b), the masking operation is conducted in a similar way as column-wise making. Specifically, we reshape the sequential feature S_{hor} back into a spatial feature $\hat{S}_{hor} \in \mathbb{R}^{h \times w \times c}$ and then conduct random row-wise masking to obtain $s_r \in$

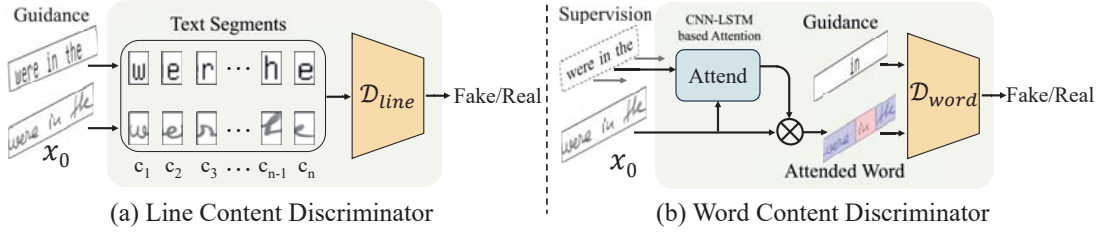


Figure 5. Illustration of the multi-scale content discriminators. (a) For the line content discriminator \mathcal{D}_{line} , we concatenate the generated text-line image x_0 and guidance image along the channel dimension, slice the result into n segments, and then process them with a 3D CNN [55] to integrate context information. Finally, the line discriminator \mathcal{D}_{line} evaluates each segment as real or fake. (b) For the word content discriminator \mathcal{D}_{word} , we utilize an attention module [52] to attend to individual words within the generated text line image x_0 , and then input these words, along with corresponding content guidance, into \mathcal{D}_{word} for realism discrimination.

$\mathbb{R}^{m \times w \times c}$, where $m = h \cdot \rho$. Our horizontal enhancing loss \mathcal{L}_{hor} pulls the masked style features of the same writer together while pushing those of different writers apart:

$$\mathcal{L}_{hor} = \frac{1}{|P_r^+|} \sum_{p_r \in P_r^+} \log \left(1 + \sum_{s_r \in S_r^+} e^{-f_r^+} \right) + \frac{1}{|P_r^-|} \sum_{p_r \in P_r^-} \log \left(1 + \sum_{s_r \in S_r^-} e^{f_r^-} \right), \quad (4)$$

where we assign a proxy p_r to each writer and link it with all masked results S_r . The similarity between positive pairs is $f_r^+ = \alpha(g(s_r, p_r) - \delta)$ for $s_r \in S_r^+$, and that of negative pairs is $f_r^- = \alpha(g(s_r, p_r) + \delta)$ for $s_r \in S_r^-$.

4.3. Multi-Scale Content Learning

Existing methods [7, 9, 12, 23, 44] rely on content recognition losses to enhance the content readability of generated handwriting images. However, these losses, applied at the text-line level, prioritize global correctness but often fail to ensure character-level accuracy. For instance, in datasets like IAM [36], where a single text line averages 42 characters—approximately six times the length of a typical word, maintaining content integrity across the entire generated text becomes challenging. To address this problem, we introduce a multi-scale content learning strategy that provides finer-grained content supervision at both global (line) and local (word) levels for content adversarial training. As shown in Figure 5, the line content discriminator \mathcal{D}_{line} evaluates the overall character order with a line discrimination loss \mathcal{L}_{line} , while the word content discriminator \mathcal{D}_{word} ensures character-level accuracy through a word discrimination loss \mathcal{L}_{word} . The overall multi-scale content loss is then formulated as:

$$\mathcal{L}_{content} = \mathcal{L}_{line} + \mathcal{L}_{word}. \quad (5)$$

Line content discriminator. As shown in Figure 5(a), given the generated image x_0 after diffusion and the content guidance I_{line} without style information, the line discriminator \mathcal{D}_{line} aims to determine whether the overall character order in x_0 matches that in I_{line} . Firstly, we set the wider of the I_{line} and x_0 as the benchmark, padding the

narrower one with white background pixels for width alignment. We then concatenate x_0 and I_{line} along the channel dimension, and then slice the concatenated result into n non-overlapping segments $\{c_i\}_{i=1}^n$ from left to right. Afterwards, a 3D CNN [55] based discriminator \mathcal{D}_{line} processes $\{c_i\}_{i=1}^n$ to incorporate global context information of characters, and determines whether this output is real or fake, providing feedback for the overall character order. The line discriminator loss \mathcal{L}_{line} is formulated as:

$$\mathcal{L}_{line} = \log(\mathcal{D}_{line}(I_{line}, x_{real})) + \log(1 - \mathcal{D}_{line}(I_{line}, x_0)). \quad (6)$$

Word content discriminator. Compared to the line discriminator \mathcal{D}_{line} , the word discriminator \mathcal{D}_{word} is designed to ensure that the text structure is correctly generated at the word level. However, accurately locating word positions within a whole text-line x_0 is non-trivial. Motivated by ASTER [52], we utilize an attention module with a CNN-LSTM architecture to obtain word positions.

As shown in Figure 5(b), given the generated image x_0 after diffusion, a CNN encoder first extracts spatial features $F_{map} \in \mathbb{R}^{h \times w \times c}$ from x_0 , which is flattened into sequential features $H \in \mathbb{R}^{l \times c}$, where $l = h \times w$. The LSTM decoder then takes x_0 and a start-of-sequence (SOS) token as input, sequentially outputting attention maps for character positions until the end-of-sequence (EOS) token is reached. The character-level attention maps are then concatenated into word-level attention maps $A = \{a_t\}_{t=1}^T$ (cf. Figure 12 in Appendix), where $a_t \in \mathbb{R}^{h \times w}$ and T denotes the number of words in the text-line. Based on the attention maps, we extract attended words $\{x_{word}^t\}_{t=1}^T$, with $x_{word}^t = a_t \cdot x_0$. Lastly, each x_{word} and its corresponding content guidance I_{word} are fed into \mathcal{D}_{word} for discrimination, which provides word-level content feedback for the generator to refine the local content readability. Specifically, the word discriminator loss \mathcal{L}_{word} is formulated as:

$$\mathcal{L}_{word} = \sum_{i=1}^T \log(\mathcal{D}_{word}(I_{word}^i, x_{real}^i)) + \sum_{i=1}^T \log(1 - \mathcal{D}_{word}(I_{word}^i, x_{word}^i)), \quad (7)$$

where i represents the i -th word in a text line.

Datasets	Method	Shot	HWD ↓	D _{CER} ↓	D _{WER} ↓	FID ↓	IS ↑	GS ↓
IAM	TS-GAN [9]	one	2.11	44.20	87.13	16.76	1.76	2.87×10^{-2}
	CSA-GAN [23]	few	2.25	42.27	84.14	13.52	1.74	1.62×10^{-2}
	VATr [44]	few	1.87	28.80	71.77	12.51	1.69	1.45×10^{-2}
	DiffusionPen [40]	few	1.72	54.75	84.70	10.24	1.83	6.42×10^{-3}
	One-DM [7]	one	1.80	20.91	54.27	10.60	1.82	8.42×10^{-3}
	Ours	one	1.41	8.59	28.60	8.69	1.85	2.35×10^{-3}
CVL	CSA-GAN [23]	few	1.72	41.64	72.02	8.71	1.48	6.71×10^{-2}
	VATr [44]	few	1.50	38.49	66.33	9.04	1.44	1.43×10^{-1}
	DiffusionPen [40]	few	1.32	55.94	88.36	11.90	1.59	5.08×10^{-2}
	One-DM [7]	one	1.47	32.42	63.35	11.95	1.46	1.29×10^{-1}
	Ours	one	1.06	20.92	36.38	7.57	1.70	2.96×10^{-2}

Table 1. Comparisons with baselines on handwritten text-line generation on IAM and CVL. All methods are trained on the same training set and evaluated using the same protocols. The ‘‘Shot’’ column indicates the number of text-line style references required for each method.

5. Experiments

5.1. Experimental Settings

Evaluation dataset. To evaluate our DiffBrush in generating handwritten text-line, we use the widely adopted handwriting datasets IAM [36] and CVL [26]. IAM contains 13,353 English text-line images belonging to 657 unique writers. Following the protocol of CSA-GAN [23], we use text-lines from 496 writers for training and the remaining 161 writers for testing. CVL dataset consists of handwritten text-lines from 310 writers in both English and German. For our experiments, we use the English portion, consisting of 11,007 text-lines, and follow the standard CVL split, with 283 writers for training and 27 for testing. In all experiments, we resize the images to a height of 64 pixels while preserving their aspect ratio, as done in previous works [7, 9, 23]. To manage varying widths, images with a width smaller than 1024 pixels are padded, whereas those exceeding 1024 are resized to a fixed size of 64×1024 . We also evaluate DiffBrush on popular Chinese dataset CASIA-HWDB (2.0–2.2) [32] (cf. Appendix E for more details.)

Evaluation metrics. 1) We use the newly proposed Handwriting Distance (HWD) [45], specifically designed for handwriting style evaluation. HWD computes the Euclidean distance between features extracted by a VGG16 network pre-trained on a large corpus of handwritten text images. 2) For content evaluation, we follow latest studies [7, 40, 41] by using the generated training sets from each method to train an OCR system [48] and report its recognition performance on the real test set in terms of CER and WER. We name these new evaluation metrics D_{CER} and D_{WER}, with further discussions provided in Appendix C. 3) We use Fréchet Inception Distance (FID) [15], Inception Score (IS) [51], and Geometry Score (GS) [24] to measure the visual quality of generated images. 4) We also conduct user studies to quantify the subjective quality of the generated handwritten text-line images in Appendix D.

Implementation details. In all experiments, we use a randomly selected text-line sample as the style reference. In DiffBrush, both the style and content encoders are based

on a ResNet18, followed by 2 transformer encoder layers. Line discriminator uses three 3D convolution layers, and word discriminator has three 2D convolution layers. The model is trained for 800 epochs on eight RTX 4090 GPUs using the AdamW optimizer with a learning rate of 10^{-4} . We set masking ratio ρ to 0.5 after a grid search (cf. Table 5 in Appendix). We randomly drop the condition c with the probability 0.1 for classifier-free training [16]. During inference, we adopt a classifier-free guidance scale of 0.2 and use DDIM [53] with 50 steps to accelerate the process. More details are put in Appendix B.

Compared Methods. We compare DiffBrush with state-of-the-art handwritten text-line generation methods, including TS-GAN [9], CSA-GAN [23], and advanced word-level handwritten text generation approaches like VATr [44], DiffusionPen [40] and One-DM [7]. We provide implementation details of word-level methods in Appendix G.

5.2. Main Results

Styled handwritten text-line generation. We assess DiffBrush for generating handwritten text-line images with desired style and specific content. To quantify style similarity, following CSA-GAN [23], we generate text-line images for each method using style information from test set and content input from a subset of WikiText-103 [37]. We then calculate the HWD between the generated and real samples for each writer, and finally average the results.

The quantitative results in Table 1 show that DiffBrush outperforms all state-of-the-art methods on both IAM and CVL datasets. Specifically, it improves HWD by 18.02% ($1.72 \rightarrow 1.41$) on IAM and 19.69% ($1.32 \rightarrow 1.06$) on CVL compared to the second-best method, highlighting its superior style imitation ability. Moreover, DiffBrush achieves significantly lower D_{CER} and D_{WER} on both IAM and CVL datasets, demonstrating its advantage in content readability. In contrast, DiffusionPen [40] yields the highest D_{CER} and D_{WER} due to ineffective content supervision.

We provide qualitative results to intuitively explain the benefit of our DiffBrush in Figure 6. TS-GAN struggles to accurately capture the style patterns of reference sam-

Style samples	That boy! That damned fool boy! What does come home now, even if he did promise	There were few passengers on the plane and Gavin was together. The porter brought Gavin's bag out to the
TS-GAN	Success is not the destination, it's the journey, every step forward is a step toward growth.	Success is not the destination, it's the journey, every step forward is a step toward growth.
CSA-GAN	Success is not the destination, it's the journey, every step forward is a step toward growth.	Success is not the destination, it's the journey, every step forward is a step toward growth.
VATr	Success is not the destination, it's the journey, every step forward is a step toward growth.	Success is not the destination, it's the journey, every step forward is a step toward growth.
DiffusionPen	Success is not the destination, it's the journey, every step forward is a step toward growth.	Success is not the destination, it's the journey, every step forward is a step toward growth.
One-DM	Success is not the destination, it's the journey, every step forward is a step toward growth.	Success is not the destination, it's the journey, every step forward is a step toward growth.
Ours	Success is not the destination, it's the journey, every step forward is a step toward growth.	Success is not the destination, it's the journey, every step forward is a step toward growth.

Figure 6. Qualitative comparisons between our method and state-of-the-art approaches for handwritten text-line generation, conditioned on out-of-vocabulary (OOV) textual content and unseen styles from the IAM test dataset. We use the same guiding text, “Success is not the destination, it’s the journey, every step forward is a step toward growth.” for all methods, instructing them to generate the text in different handwriting styles. The red circles highlight missing characters or structural errors. Better zoom in 200%.

Style sample	<i>she had been sufficiently</i>	HWD↓	D _{CER} ↓	D _{WER} ↓
Base+ ε_{single} +without masking	<i>gave a longcous cough cough</i>	1.82	56.02	87.25
Base+ ε_{single} +random masking	<i>gave a coughcough a long rough</i>	1.75	55.41	86.42
Base+ ξ_{style}	<i>gave a bugh cough raucous</i>	1.47	54.64	84.33
Base+ ξ_{style} + \mathcal{D}_{word}	<i>gave a long raucous cough cough</i>	1.42	14.61	43.93
Base+ ξ_{style} + \mathcal{D}_{line}	<i>gave a long raucous cough</i>	1.44	15.28	43.31
Base+ ξ_{style} + \mathcal{D}_{line} + \mathcal{D}_{word}	<i>gave a long raucous cough</i>	1.41	8.59	28.60

Figure 7. Ablation studies of the content-decoupled style module (ξ_{style}) and the multi-scale content discriminators (i.e., \mathcal{D}_{line} and \mathcal{D}_{word}) based on IAM test set. ε_{single} denotes a single CNN-Transformer style encoder (ResNet18 followed by 3 transformer encoder layers). The red boxes highlight failure instances of structure preservation, whereas the blue box points out an incorrect repetitive word.

ples, like ink color and stroke width. CSA-GAN produces samples that lack style consistency, including inconsistent character slant, ink color, and stroke width. VATr has difficulty maintaining vertical alignment between words in the synthesized text lines. DiffusionPen struggles to ensure the content readability of generation results. One-DM occasionally generates text lines with missing or incorrect characters. Conversely, our DiffBrush excels at generating precise character details while maintaining overall consistency.

Style-agnostic text-line generation. We further evaluate DiffBrush’s ability to generate realistic handwritten text-line images, independent of style imitation. Following TS-GAN [9], each method generates 25k random text-line images to calculate FID against all training samples, and 5k random samples for GS calculation, compared with 5k samples from the test set. Besides, we generate the entire test set using each method and evaluate the results using the IS. As shown in Table 1, DiffBrush achieves the highest performance across FID, IS, and GS metrics on both IAM and CVL datasets, further demonstrating its ability to generate superior-quality handwritten text-line images.

5.3. Analysis

In this section, we conduct ablation studies to analyze our DiffBrush. We provide more analyses in Appendix, including generalization evaluation on various style backgrounds, **failure case analysis**, ablation results on masking ratio, enriching datasets to train recognizer, style interpolation results, style evaluation results in terms of WIER [12, 61], discussions about fine-grained style learning.

Quantitative evaluation of style module and content discriminators. We perform multiple ablation studies to analyze different components. Quantitative results in Figure 7 reveal that: (1) Compared to two variants—a basic style encoder without masking, and with random masking (best masking ratio 0.5)—our style module ξ_{style} significantly improves HWD by 19.23% (1.82 \rightarrow 1.47), and 16.00% (1.75 \rightarrow 1.47), respectively. This highlights the effectiveness of ξ_{style} in style learning. (2) The combination of the \mathcal{D}_{word} and \mathcal{D}_{line} leads to significant improvements in terms of D_{CER} and D_{WER} without reducing HWD. This is achieved by employing style-free conditional discriminators, which


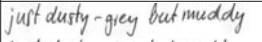
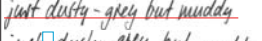
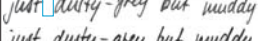
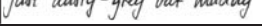
Style module in DiffBrush		HWD↓	D _{CER} ↓	D _{WER} ↓
Single style encoder		1.78	11.34	36.29
Ver. enhancing head only		1.63	10.92	35.04
Hor. enhancing head only		1.58	10.66	33.64
Our style module		1.41	8.59	28.60

Figure 8. Effect of the horizontal and vertical enhancing heads on IAM test set. Red lines highlight vertical alignment of words, while blue boxes denote the word spacing. “Single style encoder” is a basic CNN-Transformer encoder without masking.

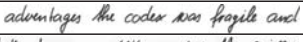
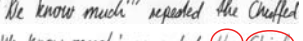
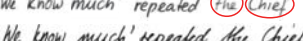
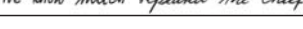
Style sample		HWD↓	D _{CER} ↓	D _{WER} ↓
Base+ ξ_{style}		1.47	54.64	84.33
Base+ ξ_{style} +CTC		1.67	16.53	50.48
Base+ ξ_{style} +D		1.41	8.59	28.60

Figure 9. Effect of discriminators \mathcal{D} and CTC recognizer [14] under their best trade-off factor. Red circles indicate handwritten texts with simplified style (i.e., regular texts with standard strokes).

focus solely on forcing the generator to enhance content readability, without impeding the learning of style.

Qualitative evaluation of style module and content discriminators. We conduct visual ablation experiments to further analyze each module in our DiffBrush. As shown in Figure 7, we observe that the first two basic baselines show clear drawbacks in both style imitation and content readability. Adding our style module significantly improves style reproduction, such as ink color and stroke width, but content readability remains poor. Adding \mathcal{D}_{word} alone improves the character details. However, it struggles to maintain overall content readability, leading to issues like word repetitions and shifts. Employing \mathcal{D}_{line} solely enhances overall content readability, but character detail issues still remain. In contrast, the best generation results are achieved when both line- and word-level discriminators are used.

Discussions about style learning. We conduct ablation study on the style module to analyze the differences between two style-enhancing heads. As shown in Figure 8, adding either the vertical or horizontal enhancing head improves text-line style quality in terms of HWD. The vertical head enhances the style imitation ability, particularly in maintaining consistent vertical word alignment. Meanwhile, the horizontal head also improves the style learning, like horizontal spacing patterns. These findings support our motivation that content-masking strategies in different directions help the effective style learning (cf. Figure 2). Finally, it is worth emphasizing that our style features contain complete style information as they are extracted from the entire style reference before any masking (cf. Figure 3).

Discussions on discriminators and CTC. The quantitative results in Figure 9 show that incorporating CTC recognizer [14] significantly reduces D_{CER} and D_{WER} , while also impairing the style evaluation (HWD). Visualization

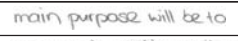
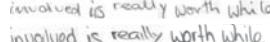

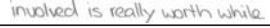

Style sample		HWD↓	D _{CER} ↓	D _{WER} ↓	FID↓
VATr+assembling		2.46	36.58	80.94	29.87
One-DM+assembling		2.37	26.50	62.79	26.15
DiffusionPen+assembling		2.19	28.64	71.25	25.24
DiffBrush (Ours)		1.41	8.59	28.60	8.69

Figure 10. Comparisons between directly generated text-lines and text-lines assembled by isolated words on IAM test set.

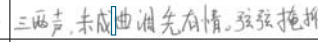


Method		HWD↓	D _{AR} ↑	D _{CR} ↑	FID↓
One-DM		1.09	81.99	82.80	17.38
Ours		0.73	96.24	96.65	7.87

Figure 11. Quantitative and qualitative comparisons with One-DM [7] on Chinese handwritten text-line generation. The blue boxes highlight the character spacing, while the red circles emphasize the incorrect character structures.

results in Figure 9 intuitively explain the reasons for this degradation. The CTC version tends to produce texts with simplified styles that differ significantly from style samples. Conversely, our discriminators enhance content readability while preserving style mimicry performance. We provide more experiment details and visual results in Appendix E. **Discussions on directly generated and assembled text-lines.** We conduct experiments to demonstrate the superiority of directly generated text lines over those obtained through concatenation. To this end, we apply the assembling strategy in DiffusionPen [40] to enable the official word-level generation baselines to generate text-line image (cf. Appendix G). The results in Figure 10 demonstrate the superiority of our method over those non-text line methods. More experimental results are provided in Appendix H.

Applications to Chinese text-line generation. We assess DiffBrush’s ability to generate Chinese scripts, a challenging task due to thousands of character categories and their complex structures. Both quantitative and qualitative results are provided on Figure 11. We observe our DiffBrush effectively handles Chinese handwritten text-lines in terms of style imitation and content fidelity. More experimental details and visualizations are put in Appendix E.

6. Conclusion

In this paper, we introduce DiffBrush, a novel diffusion model tailored for handwritten text-line generation. To the best of our knowledge, this is among the first exploration of diffusion models for this task. For better style learning and content guidance, we propose a content-decoupled style learning strategy that significantly enhances style imitation and multi-scale content discriminators that supervise textual content at both the line and word levels while preserving style imitation performance. Promising results on three widely-used handwritten datasets verify the effectiveness of our DiffBrush. In the future, we plan to explore its potential to support other generative tasks, such as font generation.

Acknowledgments The research is partially supported by National Natural Science Foundation of China (No.62176093, 61673182), National Key Research and Development Program of China (No.2023YFC3502900), Guangdong Emergency Management Science and Technology Program (No.2025YJKY001).

References

- [1] Eloi Alonso, Bastien Moysset, and Ronaldo Messina. Adversarial generation of handwritten text images conditioned on sequences. In *International Conference on Document Analysis and Recognition*, pages 481–486, 2019. 1, 2
- [2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*, 2(3):8, 2023. 12
- [3] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Mubarak Shah. Handwriting transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1086–1094, 2021. 1, 2
- [4] Changjian Chen, Fei Lv, Yalong Guan, Pengcheng Wang, Shengjie Yu, Yifan Zhang, and Zhuo Tang. Human-guided image generation for expanding small-scale training image datasets. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 3
- [5] Yi Chen, Heng Zhang, and Cheng-Lin Liu. Improved learning for online handwritten chinese text recognition with convolutional prototype network. In *International Conference on Document Analysis and Recognition*, pages 38–53, 2023. 14
- [6] Gang Dai, Yifan Zhang, Qingfeng Wang, Qing Du, Zhu-liang Yu, Zhuoman Liu, and Shuangping Huang. Disentangling writer and character styles for handwriting generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5977–5986, 2023. 2
- [7] Gang Dai, Yifan Zhang, Quhui Ke, Qiangya Guo, and Shuangping Huang. One-shot diffusion mimicker for handwritten text generation. In *European Conference on Computer Vision*, 2024. 1, 2, 3, 5, 6, 8, 12, 13, 14, 15, 25, 26
- [8] Ayan Das, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. Chirodiff: Modelling chirographic data with diffusion models. In *International Conference on Learning Representations*, 2023. 2
- [9] Brian L. Davis, Bryan S. Morse, Brian L. Price, Chris Tensmeyer, Curtis Wigington, and Rajiv Jain. Text and style conditioned gan for the generation of offline-handwriting lines. In *British Machine Vision Conference*, 2020. 1, 3, 5, 6, 7
- [10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3, 12
- [11] Sharon Fogel, Hadar Averbuch-Elor, Sarel Cohen, Shai Mazor, and Roei Litman. Scrabblegan: Semi-supervised varying length handwritten text generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4324–4333, 2020. 1, 2
- [12] Ji Gan, Weiqiang Wang, Jiaxu Leng, and Xinbo Gao. Higan+: handwriting imitation gan with disentangled representations. *ACM Transactions on Graphics*, 42(1):1–17, 2022. 1, 5, 7, 16
- [13] Shengjie Gong, Haojie Li, Jiapeng Tang, Dongming Hu, Shuangping Huang, Hao Chen, Tianshui Chen, and Zhuoman Liu. Monocular and generalizable gaussian talking head animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5523–5534, 2025. 14
- [14] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning*, pages 369–376, 2006. 8, 13, 14, 15
- [15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 6, 14
- [16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv*, 2022. 6, 12
- [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3, 12
- [18] Lei Hu and Shuangping Huang. Enhancing table structure recognition via bounding box guidance. In *International Conference on Pattern Recognition*, pages 209–225, 2024. 12
- [19] Lei Hu, Zhiyong Gan, Ling Deng, Jinglin Liang, Lingyu Liang, Shuangping Huang, and Tianshui Chen. Replaycad: Generative diffusion replay for continual anomaly detection. *arXiv*, 2025. 3
- [20] Hongxiang Huang, Daihui Yang, Gang Dai, Zhen Han, Yuyi Wang, Kin-Man Lam, Fan Yang, Shuangping Huang, Yongge Liu, and Mengchao He. Aagtgan: Unpaired image translation for photographic ancient character generation. In *ACM International Conference on Multimedia*, pages 5456–5467, 2022. 2
- [21] Shuangping Huang, Yu Luo, Zhenzhou Zhuang, Jin-Gang Yu, Mengchao He, and Yongpan Wang. Context-aware selective label smoothing for calibrating sequence recognition model. In *ACM International Conference on Multimedia*, pages 4591–4599, 2021. 13
- [22] Lei Kang, Pau Riba, Yaxing Wang, Marçal Rusinol, Alicia Fornés, and Mauricio Villegas. Ganwriting: content-conditioned generation of styled handwritten word images. In *European Conference on Computer Vision*, pages 273–289, 2020. 2
- [23] Lei Kang, Pau Riba, Marçal Rusinol, Alicia Fornés, and Mauricio Villegas. Content and style aware generation of text-line images for handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 8846–8860, 2021. 1, 3, 5, 6
- [24] Valentin Khulkov and Ivan Oseledets. Geometry score: A method for comparing generative adversarial networks. In *International conference on machine learning*, pages 2621–2629, 2018. 6

- [25] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3238–3247, 2020. 2, 4
- [26] Florian Kleber, Stefan Fiel, Markus Diem, and Robert Sablatnig. Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. In *International Conference on Document Analysis and Recognition*, pages 560–564, 2013. 6
- [27] Daiyuan Li, Guo Chen, Xixian Wu, Zitong Yu, and Mingkui Tan. Face anti-spoofing with cross-stage relation enhancement and spoof material perception. *Neural Networks*, 175: 106275, 2024. 12
- [28] Jinglin Liang, Jin Zhong, Hanlin Gu, Zhongqi Lu, Xingxing Tang, Gang Dai, Shuangping Huang, Lixin Fan, and Qiang Yang. Diffusion-driven data replay: A novel approach to combat forgetting in federated class continual learning. In *European Conference on Computer Vision*, pages 303–319, 2024. 3
- [29] Wenhui Liao, Jiapeng Wang, Zening Lin, Longfei Xiong, and Lianwen Jin. Pptser: A plug-and-play tag-guided method for few-shot semantic entity recognition on visually-rich documents. In *Findings of the Association for Computational Linguistics*, pages 10522–10539, 2024. 12
- [30] Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. Doclayllm: An efficient multi-modal extension of large language models for text-rich document understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4038–4049, 2025.
- [31] Zihao Lin, Jinrong Li, Gang Dai, Tianshui Chen, Shuangping Huang, and Jianmin Lin. Contrastive representation enhancement and learning for handwritten mathematical expression recognition. *Pattern Recognition Letters*, 186:14–20, 2024. 12
- [32] Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. Casia online and offline chinese handwriting databases. In *International Conference on Document Analysis and Recognition*, pages 37–41, 2011. 6, 14
- [33] Zhuoman Liu, Wei Jia, Ming Yang, Peiyao Luo, Yong Guo, and Mingkui Tan. Deep view synthesis via self-consistent generative network. *IEEE Transactions on Multimedia*, 24: 451–465, 2021. 2
- [34] Troy Luhman and Eric Luhman. Diffusion models for handwriting generation. *arXiv*, 2020. 2
- [35] Canjie Luo, Yuanzhi Zhu, Lianwen Jin, Zhe Li, and Dezhi Peng. Slogan: handwriting style synthesis for arbitrary-length and out-of-vocabulary text. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8503–8515, 2022. 1, 2
- [36] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5: 39–46, 2002. 2, 5, 6
- [37] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv*, 2016. 6
- [38] Yair Movshovitz-Attias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *IEEE International Conference on Computer Vision*, pages 360–368, 2017. 2, 4
- [39] Konstantina Nikolaidou, George Retsinas, Vincent Christlein, Mathias Seuret, Giorgos Sfikas, Elisa Barney Smith, Hamam Mokayed, and Marcus Liwicki. Wordstylist: Styled verbatim handwritten text generation with latent diffusion models. In *International Conference on Document Analysis and Recognition*, pages 384–401, 2023. 3, 13
- [40] Konstantina Nikolaidou, George Retsinas, Giorgos Sfikas, and Marcus Liwicki. Diffusionpen: Towards controlling the style of handwritten text generation. In *European Conference on Computer Vision*, 2024. 1, 3, 6, 8, 13, 15
- [41] Konstantina Nikolaidou, George Retsinas, Giorgos Sfikas, and Marcus Liwicki. Rethinking htg evaluation: Bridging generation and recognition. In *European Conference on Computer Vision workshop*, 2024. 6, 13
- [42] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975. 15
- [43] Wenjie Peng, Hongxiang Huang, Tianshui Chen, Quhui Ke, Gang Dai, and Shuangping Huang. Globally correlation-aware hard negative generation. *International Journal of Computer Vision*, pages 1–22, 2024. 12
- [44] Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara. Handwritten text generation from visual archetypes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22458–22467, 2023. 1, 2, 5, 6, 12, 14, 15
- [45] Vittorio Pippi, Fabio Quattrini, Silvia Cascianelli, and Rita Cucchiara. HWD: A novel evaluation score for styled handwritten text generation. In *British Machine Vision Conference*, pages 7–9, 2023. 6, 16
- [46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 12
- [47] Min-Si Ren, Yan-Ming Zhang, Qiu-Feng Wang, Fei Yin, and Cheng-Lin Liu. Diff-writer: A diffusion model-based stylized online handwritten chinese character generator. In *International Conference on Neural Information Processing*, pages 86–100, 2023. 2
- [48] George Retsinas, Giorgos Sfikas, Basilis Gatos, and Christophoros Nikou. Best practices for a handwritten text recognition system. In *International Conference on Document Analysis and Recognition workshop*, pages 247–259, 2022. 6, 13, 14, 16
- [49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3, 12
- [50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing*

- and Computer Assisted Intervention, pages 234–241, 2015. 3
- [51] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 29, 2016. 6
- [52] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048, 2018. 5
- [53] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 6
- [54] Tong-Hua Su, Tian-Wen Zhang, De-Jun Guan, and Hu-Jie Huang. Off-line recognition of realistic chinese handwriting using segmentation-free strategy. *Pattern Recognition*, 42(1):167–182, 2009. 14
- [55] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision*, pages 4489–4497, 2015. 2, 5, 12
- [56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 12
- [57] Qiu-Feng Wang, Fei Yin, and Cheng-Lin Liu. Handwritten chinese text recognition by integrating multiple contexts. *IEEE transactions on pattern analysis and machine intelligence*, 34(8):1469–1481, 2011. 14
- [58] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. In *International Conference on Learning Representations*. OpenReview.net, 2023. 3, 12
- [59] Tianming Xie and Wenxiong Kang. A random-binding based bio-hashing template protection method for palm vein recognition. *IEEE Transactions on Information Forensics and Security*, 2025. 13
- [60] Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8196–8206, 2024. 3, 12
- [61] Jan Zdenek and Hideki Nakayama. Handwritten text generation with character-specific encoding for style imitation. In *International Conference on Document Analysis and Recognition*, pages 313–329, 2023. 7, 16
- [62] Yifan Zhang and Bryan Hooi. Hipa: enabling one-step text-to-image diffusion models via high-frequency-promoting adaptation. *arXiv*, 2023. 3
- [63] Yifan Zhang, Daquan Zhou, Bryan Hooi, Kai Wang, and Jia-ashi Feng. Expanding small-scale datasets with guided imagination. In *Advances in Neural Information Processing Systems*, pages 76558–76618, 2023.
- [64] Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv*, 2025. 3
- [65] Zhiyuan Zhang, Xiaofan Li, Zhihao Xu, Wenjie Peng, Zijian Zhou, Miaoqing Shi, and Shuangping Huang. Mpdrive: Improving spatial understanding with marker-based prompt learning for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12089–12099, 2025. 12
- [66] Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. Memo: Memory-guided diffusion for expressive talking video generation. *arXiv*, 2024. 3
- [67] Yuanzhi Zhu, Zhaohai Li, Tianwei Wang, Mengchao He, and Cong Yao. Conditional text image generation with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14235–14245, 2023. 3
- [68] Zhenzhou Zhuang, Zonghao Liu, Kin-Man Lam, Shuangping Huang, and Gang Dai. A new semi-automatic annotation model via semantic boundary estimation for scene text detection. In *International Conference on Document Analysis and Recognition*, pages 257–273, 2021. 13

Beyond Isolated Words: Diffusion Brush for Handwritten Text-Line Generation

Supplementary Material

We organize our supplementary material as follows.

- In Appendix A, we review the related work of diffusion methods in general image generation.
- In Appendix B, we describe more implementation details.
- In Appendix C, we provide more discussions about content evaluation metrics, *i.e.* D_{CER} and D_{WER} .
- In Appendix D, we conduct user study experiments.
- In Appendix E, we put the experimental details and visual results for Chinese handwritten text-line generation.
- In Appendix F, we provide more ablation experiments: 1) Comparing the content discriminators and CTC recognizer. 2) More visual ablation results of the style module and content discriminators, 3) More visual ablation results of vertical enhancing and horizontal enhancing heads, 4) More ablation results on discriminators architecture, and 5) The effect of masking ratio ρ .
- In Appendix G, we provide the implementation details of word-level generation methods.
- In Appendix H, we provide more discussions about directly generated text-lines and assembled text-lines.
- In Appendix I, we provide more style evaluation results.
- In Appendix J, we explore the generalization of our DiffBrush to style images with various backgrounds.
- In Appendix K, we explore the downstream application of enriching datasets for training more robust recognizers.
- In Appendix L, we discuss the fine-grained style learning.
- In Appendix M, we provide failure case analysis.
- In Appendix N, we present results from visual style interpolation experiments.
- In Appendix O, we present extensive visual results for English and Chinese handwritten text-line generation.

A. More related work

Image diffusion Diffusion models such as Denoising Diffusion Probabilistic Model (DDPM) [17] and Latent Diffusion Model (LDM) [49] have shown great success in image generation. For example, guided diffusion [10] and classifier-free diffusion [16] condition the image synthesis on class labels. Some text-to-image diffusion methods like Stable-diffusion [49] and DALL-E3 [2] further employ CLIP [46] to convert text descriptions into comprehensive representations, thereby producing impressive results. Very recently, some methods [58, 60] combine adversarial learning with diffusion using a discriminator to enhance generation quality. Unlike these GAN-diffusion approaches that simply distinguish between real and generated images, our two-level content discriminators are specifically designed to provide content supervision at both the line and word levels.

B. More implementation details

Content encoder. Following VATr [44] and One-DM [7], we render the text string \mathcal{A} into Unifont images. The strength of Unifont is its ability to represent all Unicode characters, allowing our method to accept any user-provided string input. We then input the rendered images into a CNN-Transformer content encoder to obtain an informative content feature $Q = \{q_i\}_{i=1}^L \in \mathbb{R}^{L \times c}$.

Blender. Motivated by One-DM [7], content Q and style features S_{ver} , S_{hor} are fused in the blender with 6 transformer decoder layers [18, 27, 29–31, 43, 56, 65] across two stages. Initially, S_{ver} serves as the key/value vectors, while Q serves as the query vector that attends to S_{ver} in the first three layers to produce a fused vector. Then, this fused vector becomes a new query vector, attending to S_{hor} in the last three layers as the guiding condition $c \in \mathbb{R}^{L \times c}$.

Multi-scale content discriminators. Before being fed into the line content discriminator, the generated image x_0 and the content guidance image I_{line} are concatenated along the channel dimension, and the resulting tensor is divided into n segments, as described in Section 4.3. We set $n=32$ to ensure that each segment approximately covers a single character. Specifically, as mentioned in Section 5, the total width of a text-line image is adjusted to 1024 pixels after data pre-processing. Dividing it into non-overlapping 32 parts yields segments with a width of 32 pixels, closely matching the average character width in the dataset.

Assume the divided segments $C \in \mathbb{R}^{n \times c \times h \times w}$. A 3D CNN [55] employs sliding window operations across both the spatial dimensions (*i.e.*, h and w) and the temporal dimension (*i.e.*, n) to capture the global contextual information of characters. This representation is then passed to the line content discriminator, which evaluates whether the overall character order in the generated image x_0 matches that of the content guidance image I_{line} .

We pre-train the attention module of the word content discriminator on the training set, enabling it to accurately attend to word positions (cf. Figure 12). Its parameters are then frozen during the training of the entire DiffBrush.

Masking strategy. Our masking strategy involves randomly masking rows or columns in feature maps (cf. Figure 4). The number of masked elements is determined by the sampling rate ρ , which consequently controls both the quantity and size of the masked features. Given $\hat{S} \in \mathbb{R}^{h \times w \times c}$, column-wise masking selects $(w \times \rho)$ tensors of size $h \times 1 \times c$, while row-wise masking selects $(h \times \rho)$ tensors of size $1 \times w \times c$.



Figure 12. Visualization of attention maps for each word in a handwritten text-line image.

Conditional diffusion generator. To conserve GPU memory and accelerate the training time, following Wordstyle [39] and One-DM [7], we streamline the U-Net by reducing the number of ResNet [59] blocks and attention heads and take the diffusion process into the latent space. Specifically, we adopt a powerful, pre-trained Variational Autoencoder (VAE) of Stable Diffusion (1.5) to convert the image into latent space. During the training phase, we freeze the parameters of VAE and set $T = 1000$ steps, and forward process variances are set to constants increasing linearly from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$.

More training details. The proposed conditional diffusion generator \mathcal{G} and the multi-scale discriminators \mathcal{D} engage in an adversarial learning process: \mathcal{G} seeks to synthesize realistic images that \mathcal{D} cannot distinguish from real ones based on content, while \mathcal{D} assess the content at both the line and word scales. The readability of the generated images improves through two adversarial losses, \mathcal{L}_{line} and \mathcal{L}_{word} , which further enhances generation quality in terms of content accuracy. In summary, the overall training objectives for the conditional diffusion generator, and the multi-scale discriminators are defined as:

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{diff} + \mathcal{L}_{style} + \lambda \mathcal{L}_{content}, \quad (8)$$

$$\mathcal{L}_{\mathcal{D}} = -\mathcal{L}_{content}, \quad (9)$$

Our DiffBrush is trained with 800 epochs, as described in Section 5. During the first 750 epochs, we optimize the model using only \mathcal{L}_{ver} , \mathcal{L}_{hor} and \mathcal{L}_{diff} . In the final 50 epochs, we retain these loss functions and further introduce \mathcal{D}_{line} and \mathcal{D}_{word} to enhance the readability of the generated images x_0 . Specifically, we use the conditional diffusion generator to perform 5 denoising steps, generating a handwritten image x_0 with coarse content structure. We then input x_0 to the multi-scale content discriminators to obtain content supervision at both line and word levels. The total training time for our DiffuBrush is approximately 4 days.

C. More discussions about D_{CER} and D_{WER}

To better evaluate the content quality of generated results, we use D_{CER} and D_{WER} as content evaluation metrics, following the recent works [7, 40, 41]. It is worth emphasizing that D_{CER} is also referred to as HTG_{HTR} in [41]. The difference lies in the fact that HTG_{HTR} measures character error rate (CER) at the word level, whereas D_{CER} measures it at the text-line level.

More specifically, the implementation details of D_{CER} and D_{WER} are as follows: 1) Each method is employed to generate a complete training set, which is then used to train an OCR system [48]. The system is built on a CNN-LSTM architecture with a CTC loss [14, 21, 68]. 2) The character error rate (CER) and word error rate (WER) are evaluated on the real test set, aiming to achieve recognition performance as close as possible to that obtained with a real training set.

As noted in [40, 41], the intuition behind this experiment is that a handwriting generation method achieving or exceeding the performance of the real IAM dataset demonstrates two crucial abilities: 1) The generated handwritten text images have accurate content. 2) The generated samples exhibit diverse styles. Although the first criterion is crucial, focusing only on it while overlooking the second criterion can lead to biased evaluations. For instance, if a generation method favors producing easily recognizable handwritten texts with simplified styles (cf. red circles in Figure 14), it might achieve high content accuracy. However, such low-diversity results are not satisfactory to us. To address this, we use D_{CER} and D_{WER} to provide a more comprehensive evaluation of the content quality in the generated samples, since D_{CER} and D_{WER} simultaneously take the aforementioned two factors into account.

D. User studies

User preference study. We invite human participants with postgraduate education backgrounds to evaluate the visual quality of synthesized handwritten text images, focusing on style imitation. The generated samples are from our method and other state-of-the-art approaches. In each round, we randomly select a writer from the IAM dataset and use their handwritten text-line sample as style guidance, along with identical text as content guidance, to direct all methods in generating candidate samples. Participants are presented with one text-line from the exemplar writer as a style reference and multiple candidates generated by different methods. They are asked to select the candidate that best matches the reference in style. This process is repeated 30 times, yielding 900 valid responses from 30 volunteers. As shown in Figure 13, our method receives the most user preferences, demonstrating its superior quality in style imitation.

User plausibility study. We conduct a user plausibility

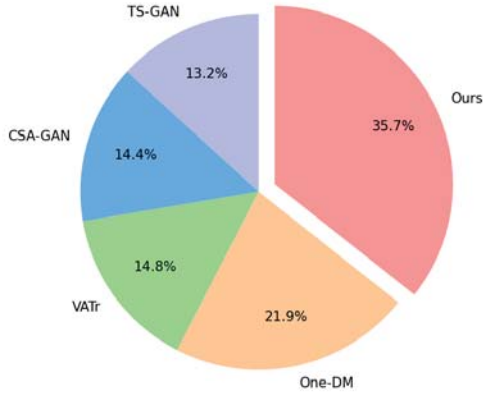


Figure 13. User preference study with a comparison to state-of-the-art methods on handwritten text-line generation.

Actual	Predicted		Classification Accuracy
	Real	Fake	
Real	27.22	22.78	49.11
Fake	28.11	21.89	

Table 2. Confusion matrix(%) from the user plausibility study. The classification accuracy of 49.11% suggests that users struggle to differentiate between handwritten text-line images generated by our DiffBrush and real ones.

study to assess whether the text-line images generated by DiffBrush are indistinguishable from real handwriting samples. In this study, participants are first shown 30 examples of authentic handwritten text-line samples. They are then asked to classify each image they see as either real or synthetic, with the images being randomly selected from both genuine samples and those generated by our method. In total, 30 participants provide 900 valid responses. The results, shown as a confusion matrix in Table 2, report a classification accuracy close to 50%, suggesting the task becomes equivalent to random guessing. This indicates that text-line images generated by our method are nearly indistinguishable from real samples.

E. Chinese handwritten text-line generation

In this section, we evaluate DiffBrush’s capability to generate scripts with thousands of character categories and complex character structures, such as Chinese. For this purpose, we perform experiments on the widely used Chinese handwritten text-line dataset CASIA-HWDB (2.0–2.2) [32].

Dataset. CASIA-HWDB (2.0–2.2) consists of 52,230 Chinese text-lines belonging to 1,019 different writers. Following the standard split of CASIA-HWDB (2.0–2.2), we use text-lines from 816 writers for training and remaining

DiffBrush	λ	HWD ↓	D _{CER} ↓	D _{WER} ↓
w/ CTC [14]	1	2.01	19.62	52.64
	0.1	1.67	16.53	50.48
	0.01	1.69	17.12	51.34
	0.001	1.68	17.29	50.72
w/ Discriminators	0.1	1.43	9.87	31.96
	0.05	1.41	8.59	28.60
	0.01	1.45	10.40	32.54

Table 3. Quantitative ablation results for the CTC recognizer and content discriminator variants of our DiffBrush on IAM test set. λ denotes the trade-off factor (cf. Eq. (1) in Section 4.1).

203 writers for testing. As mentioned in Appendix B, since Unifont [7, 44] can encode all Unicode characters, we still employ it to convert Chinese strings into textual content images. Similarly, in our experiments, all images are adjusted to 64×1024 pixels.

Evaluation Metrics. Similarly, we use HWD and FID [13, 15] to evaluate style imitation and visual quality of generated handwritten images, respectively. We still use the OCR system [48] to measure the content quality of generated results. Unlike English handwriting generation, we evaluate the OCR system’s recognition performance on a real test set using two widely adopted metrics in Chinese text-line recognition [5, 54, 57]: Accuracy Rate (AR) and Correctness Rate (CR). Similarly, we name them D_{AR} and D_{CR} .

Qualitative comparison. We provide qualitative results in Figure 26 and Figure 27. To ensure fair comparisons, both DiffBrush and One-DM [7] are conditioned on the same text contents and style samples. We observe that the handwritten text lines generated by our DiffBrush (rows of “Ours”) exhibit styles most similar to the reference samples, particularly in terms of character spacing and ink color, while preserving accurate character structures.

F. More ablation results

F.1. More ablation on discriminators and CTC

To further evaluate the impact of the proposed content discriminators, we replace the multi-scale content discriminator in DiffBrush with the standard CTC recognizer [14] adopted in One-DM [7] and conduct experiments to compare them. The quantitative results in Table 3 show that our discriminator version performs better. In Figure 14, we further visualize their best λ results on IAM test set. We observe that CTC tends to simplify handwriting styles, while the discriminator version better preserves reference styles.

F.2. More ablation results of ξ_{style} and \mathcal{D}

In Figure 19, we provide more qualitative results of the style module ξ_{style} and content discriminators (*i.e.*, \mathcal{D}_{line} and \mathcal{D}_{word}) on IAM dataset. From these results, we can observe that ε_{single} versions exhibit notable flaws in both

Text Content	obstacle overcome is a testament to your resilience,
Style sample	<i>will mean a rise in the cost of living</i>
Base+ ξ_{style}	<i>obstacle overcome is a testament to your resilience</i>
Base+ ξ_{style} +CTC	<i>obstacle overcome is a testament to your resilience,</i>
Base+ ξ_{style} +D	<i>obstacle overcome is a testament to your resilience,</i>
Text Content	muttered together.' Do you say that you poor
Style sample	<i>there was one of the most dangerous jobs con-</i>
Base+ ξ_{style}	<i>muttered together.' Do you say that you poor</i>
Base+ ξ_{style} +CTC	<i>muttered together.' Do you say that you poor</i>
Base+ ξ_{style} +D	<i>muttered together.' Do you say that you poor</i>

Figure 14. Qualitative comparisons between the discriminators \mathcal{D} and the CTC recognizer [14]. Red circles indicate handwritten texts with simplified style (*i.e.*, regular texts with standard strokes).

Discriminators	HWD ↓	D _{CER} ↓	D _{WER} ↓
2D \mathcal{D}_{line} + 2D \mathcal{D}_{word}	1.44	13.18	39.64
3D \mathcal{D}_{line} + 3D \mathcal{D}_{word}	1.43	9.58	30.27
Ours (3D \mathcal{D}_{line} + 2D \mathcal{D}_{word})	1.41	8.59	28.60

Table 4. Architecture ablation results on content discriminators.

content accuracy and style imitation, including inconsistencies in stroke color, thickness, and word spacing compared to the style samples. Incorporating ξ_{style} significantly improves style-related flaws; however, the readability of the generated text remains unsatisfactory. The introduction of \mathcal{D}_{line} ensures that the overall character order in the generated text closely matches the content reference, greatly improving content accuracy. Nevertheless, certain character details remain imperfect. After adding \mathcal{D}_{word} , incorrect character structures are effectively corrected, thus further enhancing the readability of the generated text-line.

F.3. Architecture ablation on discriminators

Our line content discriminator assesses global character order, while the word discriminator verifies local text accuracy. To this end: 1) A 3D CNN processes segmented character fragments to learn global character context. 2) A 2D CNN focuses on the content information of the attended words. In Table 4, an ablation study on IAM dataset show that: 1) Replacing \mathcal{D}_{line} with a 2D CNN significantly increases D_{CER} and D_{WER}, highlighting the limitations of using 2D CNNs for full text lines. 2) Switching \mathcal{D}_{word} to a 3D CNN yields no significant change, as a 2D CNN is sufficient for attended words with fewer characters.

F.4. More ablation results on style learning

More visual ablation results are provided in Figure 20 to analyze the effect of vertical enhancing and horizontal enhancing heads. The findings indicate that incorporating the

Masking ratio	HWD ↓	D _{CER} ↓	D _{WER} ↓
0	1.72	56.52	87.15
0.25	1.64	56.48	86.99
0.50	1.47	54.64	84.33
0.75	1.58	55.29	88.34

Table 5. Effect of the masking ratio ρ . The results are derived from our DiffBrush without using multi-scale content discriminators.

vertical enhancing head enhances the model’s style imitation capabilities, notably in maintaining vertical alignment between words. Similarly, adding the horizontal enhancing head also improves the learning of style patterns, especially in preserving horizontal word spacing. The integration of both heads is crucial for our DiffBrush to produce high-quality results that faithfully replicate the writing styles of the reference samples.

F.5. Analysing the effect of masking ratio

We analyze ρ by removing content discriminators from DiffBrush. As shown in Table 5, $\rho = 0.5$ yields the best performance on IAM test set.

G. Comparisons with non-text line methods

As shown in Figure 1, directly **assembling isolated words** from official word-level generation models leads to unnatural and low-quality text-line results. We thus retrain these non-text line methods (*i.e.*, VATr [44], DiffusionPen [40] and One-DM [7]) on text-line dataset to enable them to directly generate text lines for fair comparisons. It is worth emphasizing that the official text-line generation scheme in DiffusionPen [40] also employs an **assembly-based strategy**. This involves resizing each generated word image to ensure consistent character width, followed by a concatenation operation with a fixed space.

H. Discussion on generated and assembled line

We further conduct more experiments on IAM dataset to demonstrate the superiority of directly generated text lines. We utilize our DiffBrush to directly generate handwritten text-line images. For the assembled lines, we employ One-DM [7] to produce isolated words, which are then concatenated into text lines using statistical methods. More specifically, given a text-line style reference, we first employ the Otsu algorithm [42] to compute a binary mask of the text-line image, effectively separating words from the background. We then calculate the average spacing m between words and determine whether the centroids of the words are aligned along a horizontal or skewed line. With this statistical information, we concatenate the synthesized words from One-DM, ensuring that the spacing between words is

Method	HWD ↓	D _{CER} ↓	D _{WER} ↓	FID ↓
One-DM + post-processing	2.17	24.81	62.08	23.92
DiffBrush (Ours)	1.41	8.59	28.60	8.69

Table 6. Quantitative comparisons between directly generated and assembled text-lines on the IAM test set.

Text Content	else in sight to supplant
Style sample	
One-DM+post-processing	
DiffBrush(Ours)	
Text Content	a pose of police arrived
Style sample	
One-DM+post-processing	
DiffBrush(Ours)	
Text Content	as president of a union
Style sample	
One-DM+post-processing	
DiffBrush(Ours)	

Figure 15. Qualitative comparisons between directly generated and assembled text-lines on the IAM test set.

m , and that the centroids of the words maintain the consistent vertical alignment patterns as the text-line reference.

Quantitative results in Table 6 indicate that the text-lines generated by our DiffBrush significantly outperform assembled text-lines in terms of style evaluation (HWD), content evaluation (D_{CER} and D_{WER}), visual quality evaluation (FID). These results demonstrate the advantages of direct generation. From qualitative results in Figure 15, we can observe that our directly generated text-lines exhibit more consistent stroke colors, uniform character sizes, and vertical word alignment patterns that more closely match the style samples. These findings further underscore the superiority of directly generating text lines.

I. Discussions on text-line style evaluation

We did not use the writer classifier from previous works [12, 61] because the models were designed for evaluating **word-level text style** and the IAM split details for training the classifier were missing, making fair and reproducible evaluation difficult. Instead, we adopt the open-source HWD metric [45], which offers two advantages: (1) it ensures reproducibility, and (2) it is pre-trained on large-scale handwritten data and proven effective in **text-line style** evaluation. Following the comment, we further include WIER [12, 61] for style evaluation. To this end, we randomly split the standard IAM test set into 80% for classifier [12] training

Method	TS-GAN	CSA-GAN	VATr	DiffusionPen	One-DM	Ours
WIER (%) ↓	96.03	82.14	76.96	73.85	70.92	59.77

Table 7. Style evaluation on the IAM test set.

Training Data	CER ↓	WER ↓	Improve. (%) ↑
Real	5.78	21.76	-
CSA-GAN + Real	5.39	19.89	6.74
VATr + Real	5.08	19.31	12.11
One-DM + Real	4.99	18.51	13.67
DiffBrush (Ours) + Real	4.62	16.86	20.07

Table 8. Handwritten text-line recognition on different training data. Improvement rate refers to CER performance gain achieved by incorporating synthetic data into the training process compared to using only the real training set.

and 20% for validation. The best classifier is then used to evaluate the generation results. As shown in Table 7, our DiffBrush continues to achieve the best style imitation performance.

J. Generalization to more style backgrounds

To assess whether DiffBrush can effectively generalize to different style backgrounds, we condition it on eight complex and realistic backgrounds. The generated results are shown in Figure 21 and Figure 22. We find that our DiffBrush still generates high-quality handwritten text-line images, further demonstrating the robustness of our DiffBrush.

K. Application for training robust recognizer

A key application of handwritten text-line generation models is to enrich the training dataset, facilitating the training of more robust recognizers. To this end, we combine the IAM training set generated by various methods with the real training set to create a new mixed dataset. We then train an OCR system [48] using this mixed dataset and report its performance on the real IAM test set. We present the quantitative results in Table 8. These results clearly show that the additional synthetic data contributes to improving the recognizer’s performance. Among all methods, our approach achieves the greatest performance improvement, with an improvement rate of 20.07%.

L. Discussions on fine-grained style learning

Our method effectively models them for the following reasons: 1) Our content-masking strategy preserves key fine-grained features, including character-level details (cf. green circles in Figure 2 (c)) and stroke-level patterns (cf. purple circles in Figure 2 (d)). 2) Following prior study [6], our model uses character-level content as queries in a cross-attention mechanism, enabling the style-content blender to adaptively attend to fine-grained style cues within the reference images, as shown in Figure 17.

Text Content	"other" fish (+8 to -11) & "other" vegetables
Style sample	Federal Government that the financial burden
VATr	"other" fish (+8 to -11) "other" vegetables
One-DM	"other" fish (+8 to -11) "other" vegetables
Ours	"other" fish (+8 to -11) ② "other" vegetables
Text Content	θ, μ stand for the angle, momentum parameter.
Style sample	bases in this country. In open letter
VATr	θ, μ stand for the angle, momentum parameter.
One-DM	θ, μ stand for the angle, momentum parameter.
Ours	③④ stand for the angle, momentum parameter.

Figure 16. Failure cases. The red circles highlight character structure errors. Better zoom in 200%.

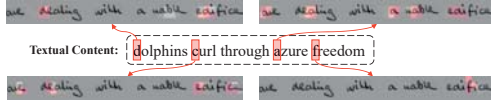


Figure 17. Visualization results of attention maps.

Style A	memorial, replied with a	Style C	most of them shopgirls in overalls.
The only way to do great work		The only way to do great work	
The only way to do great work		The only way to do great work	
The only way to do great work		The only way to do great work	
The only way to do great work		The only way to do great work	
The only way to do great work		The only way to do great work	
Style B	The NePren Rhodora conference in	Style D	had been arrested or convicted since

Figure 18. Style interpolation results between different individual handwriting styles on IAM dataset. The results are computed over both style features S_{ver} and S_{hor} . Better zoom in 200%.

M. Analysis of failure cases

We find that DiffBrush occasionally generates structurally incorrect characters when low-frequency characters from the training set are used as content conditions. This includes punctuation marks and Greek letters, as highlighted by the red circles in Figure 16. A simple yet effective solution is to employ a data oversampling strategy, increasing the frequency of these characters during training.

N. Style interpolation

To further explore the latent space learned by our style module, we conduct linear style interpolation experiments between different writers and display the generated handwritten text-line images in Figure 18. From these visual results, we find that the generated text-line images smoothly transition from one style to another, in terms of character slant,

and stroke thickness, while strictly preserving their original textual content. These results further confirm that our DiffBrush successfully generalizes to a meaningful style latent space, rather than simply memorizing style patterns from individual handwriting samples.

O. More generation results

Figure 23- Figure 27 present qualitative comparisons between our DiffBrush and previous state-of-the-art methods for multilingual handwritten text-line generation, covering both English and Chinese. The extensive visual results demonstrate that our DiffuBrush excels in both style imitation and structural preservation of generated multilingual text-lines, highlighting its superior performance.

Text content	If you want something you've never had,
Style sample	were infinitely soothing after city noises,
Base+ ϵ_{single} +without masking	If you want yewat, something, yewat, never had,,
Base+ ϵ_{single} +random masking	If you want nva vomaninyox live nee had had
Base+ ξ_{style}	If you want le something, you've never had, nva,
Base+ $\xi_{style} + D_{line}$	If you want something you've never had,
Base+ $\xi_{style} + D_{line} + D_{word}$	If you want something you've never had,
Text content	act as if what you do makes a difference, it truly does.
Style sample	long for the right man for you to come
Base+ ϵ_{single} +without masking	act as, if what takes, a, takes ab makes, truly does,
Base+ ϵ_{single} +random masking	act as if you do makes a difference, it truly does
Base+ ξ_{style}	act as if what you do you do makes as if truly does.
Base+ $\xi_{style} + D_{line}$	act as if what you do makes a difference, it truly does.
Base+ $\xi_{style} + D_{line} + D_{word}$	act as if what you do makes a difference, it truly does.
Text content	If you can dream it, you can achieve it, start believing.
Style sample	inequity of Miss Cheesecake well-high bathing
Base+ ϵ_{single} +without masking	If you can dream, than it you dream start believing,
Base+ ϵ_{single} +random masking	If you can dream than it, you dream believing
Base+ ξ_{style}	If you can see it, it, you dream start believing
Base+ $\xi_{style} + D_{line}$	If you can dream it, you can, achieve it start believing,
Base+ $\xi_{style} + D_{line} + D_{word}$	If you can dream it, you can achieve it, start believing.

Figure 19. More visual ablation results of the style module ξ_{style} and content discriminators (i.e., D_{line} and D_{word}). ϵ_{single} denotes a single CNN-Transformer style encoder. The red boxes highlight failures of structure preservation.

Style module in DiffBrush	So t his is an invitatury from my good friends of the TUC that
Single style encoder	small steps can lead to big changes, every step counts.
Ver. enhancing head only	<u>small steps can lead to big changes, every step counts.</u>
Hor. enhancing head only	small s teps can lead to big changes, every step counts.
Our style module	small steps can lead to big changes, every step counts.
Style module in DiffBrush	Federal G overnment that the financial burden
Single style encoder	Every accomplishment starts with the decision to try.
Ver. enhancing head only	<u>Every accomplishment starts with the decision to try.</u>
Hor. enhancing head only	Every a ccomplishment starts with the decision to try.
Our style module	Every accomplishment starts with the decision to try.
Style module in DiffBrush	for h im, although some think his position
Single style encoder	Champions keep playing until they get it right, practice counts.
Ver. enhancing head only	<u>Champions keep playing until they get it right, practice counts.</u>
Hor. enhancing head only	Champions k ee playing until they get it right, practice counts.
Our style module	Champions keep playing until they get it right, practice counts.

Figure 20. More visual ablation results of vertical enhancing and horizontal enhancing heads. The red lines indicate alignment of words along the vertical axis, while blue boxes indicate word spacing.

Text content	Challenges make life interesting,
Style sample	
VATr	
One-DM	
Ours	
Text content	overcoming them makes it meaningful.
Style sample	
VATr	
One-DM	
Ours	
Text content	and this matched reasonably well with
Style sample	
VATr	
One-DM	
Ours	
Text content	who seems to have had charge of the Mule when
Style sample	
VATr	
One-DM	
Ours	

Figure 21. Generated handwritten text-line images conditioned on style samples with more complex and realistic backgrounds.

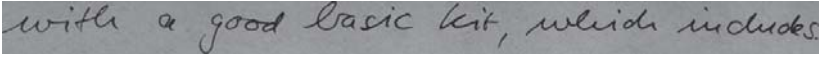
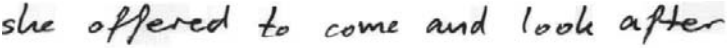


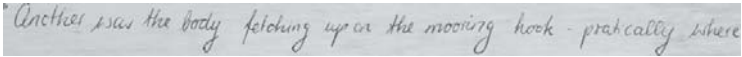
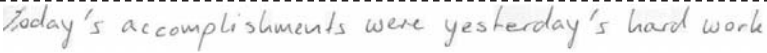
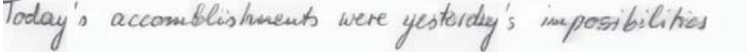
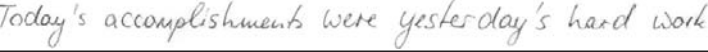
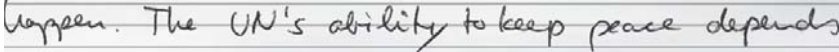



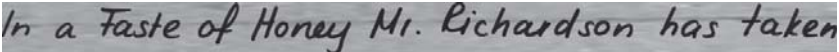
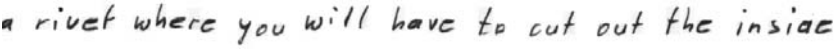
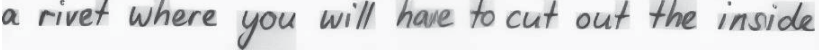

Text content	she offered to come and look after
Style sample	
VATr	
One-DM	
Ours	
Text content	Today's accomplishments were yesterday's hard work
Style sample	
VATr	
One-DM	
Ours	
Text content	married, and so repeated the enquiry,
Style sample	
VATr	
One-DM	
Ours	
Text content	a rivet where you will have to cut out the inside
Style sample	
VATr	
One-DM	
Ours	

Figure 22. Generated handwritten text-line images conditioned on style samples with more complex and realistic backgrounds.

Text Content	A year from now, you may wish you had started today.
Style sample	<i>for negotiations in a fortnight's time, these Commonwealth</i>
TS-GAN	<i>A year from now, you may wish you had started today.</i>
CSA-GAN	<i>A year from now, you may wish you had started today.</i>
VATr	<i>A year from now, you may wish you had started today.</i>
DiffusionPen	<i>A year <u>to firm</u> now, started now, you wish you <u>stoday</u>.</i>
One-DM	<i>A <u>year</u> from now you may wish you had <u>started</u> today.</i>
Ours	<i>A year from now, you may wish you had started today.</i>
Text Content	Success is not final, failure is not fatal, it's the courage.
Style sample	<i>Mr. Thorneycroft's main purpose will be to</i>
TS-GAN	<i>Success is not final, failure is not fatal, it's the courage.</i>
CSA-GAN	<i>Success is not final, failure is not fatal, it's the courage.</i>
VATr	<i>Success is not final, failure is not fatal, it's the courage.</i>
DiffusionPen	<i><u>Success</u> is <u>not</u> is not final finally, it's the <u>courage</u>.</i>
One-DM	<i>Success is not final, <u>failure</u>, is not fatal, it's <u>the courage</u>.</i>
Ours	<i>Success is not final, failure is not fatal, it's the courage.</i>
Text Content	You have the power to create the life you want, starting now.
Style sample	<i>down. a few minutes later, Mr. Fell got up</i>
TS-GAN	<i>You have the power to create the life you want, starting now.</i>
CSA-GAN	<i>You have the power to create the life you want, starting now.</i>
VATr	<i>You have the power to create the life you want, starting now.</i>
DiffusionPen	<i>You <u>have</u> you want the power to create the <u>have the courage</u> now.</i>
One-DM	<i>You have the power to <u>create the</u> life you want, starting now.</i>
Ours	<i>You have the power to create the life you want, starting now.</i>

Figure 24. Comparisons with state-of-the-art methods for English handwritten text-line generation. The red circles highlight incorrect content structure.

Text Content	as love awakens our souls to new beginnings.
Style sample	is to be made at a meeting at Labour
TS-GAN	as love awakens our souls to new beginnings.
CSA-GAN	as love awakens our souls to new beginnings.
VATr	as love awakens our souls to new beginnings.
DiffusionPen	as love awakens our souls to new beginnings.
One-DM	as love awakens our souls to new beginnings.
Ours	as love awakens our souls to new beginnings.
Text Content	Do not wait for leaders, do it alone, person to person.
Style sample	Delegates form Mr. Kenneth Kaunda's United National Independence
TS-GAN	Do not wait for leaders, do it alone, person to person.
CSA-GAN	Do not wait for leaders, do it alone, person to person.
VATr	Do not wait for leaders, do it alone, person to person.
DiffusionPen	Do not wait for leaders, do it alone, person to person.
One-DM	Do not wait for leaders, do it alone, person to person.
Ours	Do not wait for leaders, do it alone, person to person.
Text Content	The only way to do great work is to love what you do.
Style sample	Mr. Brown, passionate and warm-hearted, led
TS-GAN	The only way to do great work is to love what you do.
CSA-GAN	The only way to do great work is to love what you do.
VATr	The only way to do great work is to love what you do.
DiffusionPen	The only way to do great work is to love what you do.
One-DM	The only way to do great work is to love what you do.
Ours	The only way to do great work is to love what you do.

Figure 25. Comparisons with the state-of-the-art methods for English handwritten text-line generation. The red circles highlight incorrect content structure.

Text content	西藏自治区主席向巴平措今天在国新办发布会上表示
Style sample	在国务院新闻办今日举行的新闻发布会上,万钢表示
One-DM	西藏自治区主 ^席 向巴平措今天在 ^国 新办发布会上表示
Ours	西藏自治区主席向巴平措今天在国新办发布会上表示
Text content	武和平19日在公安部召开的新闻发布会上透露
Style sample	实行居住地户口登记制度的改革目标。这一
One-DM	武和平19日在公安部召开的新 ^闻 发布会 ^{上透} 露
Ours	武和平19日在公安部召开的新闻发布会上透露
Text content	NASA当天在华盛顿国家航空航天博物馆门前展示
Style sample	姆斯韦布"太空望远镜的同尺寸模型,这一大型望远镜
One-DM	NASA当天在 ^华 盛顿国家航空航天 ^博 物馆门前展示
Ours	NASA当天在华盛顿国家航空航天博物馆门前展示
Text content	国家旅游局质量规范与管理司司长满宏卫介绍,
Style sample	着排洪道涌入黄河,从小西湖大桥附近开始,黄河
One-DM	国家旅游局质量规范与管理司司长满宏卫介绍,
Ours	国家旅游局质量规范与管理司司长满宏卫介绍,
Text content	负责人指出,教育部要求各地要切实做好网上录取各项
Style sample	有关部门和学校要严格执行招生计划和录取标准,严禁
One-DM	负责人指出,教 ^育 部要求各地要切实做 ^好 网上录取各项
Ours	负责人指出,教育部要求各地要切实做好网上录取各项

Figure 26. Comparisons with One-DM [7] on Chinese handwritten text-line generation. The blue boxes highlight the character spacing, while the red circles emphasize the incorrect character structures.

Text content	周小川出席会议时表示，中国愿在亚洲经济合作中发挥更加积极
Style sample	与非洲开发银行等区域开发银行之间的合作，积极参与非洲开发基金等金融
One-DM	周小川出席会议时表示，中国愿在亚洲经济合作中发挥更加积极
Ours	周小川出席会议时表示，中国愿在亚洲经济合作中发挥更加积极
Text content	可亲的散文使我觉得他是个脾气最好的人；然而专杀微弱的人类为
Style sample	可惜的是，我们不能目睹黄龙全貌。倘若乘一架低空穿行的飞
One-DM	可亲的散文使我觉得他是个脾气最好的人，然而专以杀微弱的人类为
Ours	可亲的散文使我觉得他是个脾气最好的人；然而专杀微弱的人类为
Text content	一支所用的类铲，民兵卢学桂、李金亭用来砍死敌人的大刀，
Style sample	当我夸他们那块“海绵”地的时候，他“叭”地磕
One-DM	一支所用的类铲，民兵卢学桂、李金亭用来砍死敌人的大刀，
Ours	一支所用的类铲，民兵卢学桂、李金亭用来砍死敌人的大刀，
Text content	正常消化机能有一定作用。而且，其中所含纤维素有一
Style sample	的玛曲湿地干涸面积已经达到10.2万平方公里。
One-DM	正常消化机能有一定作用。而且，其中所含纤维素有一
Ours	正常消化机能有一定作用。而且，其中所含纤维素有一
Text content	增加投资，尤其是中、小民营企业应在非洲寻找更多的投资合
Style sample	量子基金创始人吉姆·罗杰斯在《热门商品投资》一书用
One-DM	增加投资，尤其是中、小民营企业应在非洲寻找更多的投资合
Ours	增加投资，尤其是中、小民营企业应在非洲寻找更多的投资合

Figure 27. Comparisons with One-DM [7] on Chinese handwritten text-line generation. The blue boxes highlight the character spacing, while the red circles emphasize the incorrect character structures.