Empowering Backbone Models for Visual Text Generation with Input Granularity Control and Glyph-Aware Training

Wenbo Li[1][*], Guohao Li[2], Zhibin Lan[1], Xue Xu[2], Wanru Zhuang[1],
Jiachen Liu[2], Xinyan Xiao[2], Jinsong Su[1,3][†]
[1]School of Informatics, Xiamen University, China,
[2]Baidu Inc., Beijing, China
[3]Shanghai Artificial Intelligence Laboratory

liwenbo@stu.xmu.edu.cn   liguohao@baidu.com   jssu@xmu.edu.cn

Abstract

Diffusion-based text-to-image models have demonstrated impressive achievements in diversity and aesthetics but struggle to generate images with legible visual texts. Existing backbone models have limitations such as misspelling, failing to generate texts, and lack of support for Chinese text, but their development shows promising potential. In this paper, we propose a series of methods, aiming to empower backbone models to generate visual texts in English and Chinese. We first conduct a preliminary study revealing that Byte Pair Encoding (BPE) tokenization and the insufficient learning of cross-attention modules restrict the performance of the backbone models. Based on these observations, we make the following improvements: (1) We design a mixed granularity input strategy to provide more suitable text representations; (2) We propose to augment the conventional training objective with three glyph-aware training losses, which enhance the learning of cross-attention modules and encourage the model to focus on visual texts. Through experiments, we demonstrate that our methods can effectively empower backbone models to generate semantic relevant, aesthetically appealing, and accurate visual text images, while maintaining their fundamental image generation quality.

1 Introduction

Recently, diffusion-based models (Ho et al., 2020; Rombach et al., 2022; Saharia et al., 2022; Balaji et al., 2022; Zhang et al., 2023; Sauer et al., 2023) have revolutionized the field of text-to-image generation, particularly in terms of diversity and aesthetics. Among various text-to-image tasks, visual text generation has attracted much attention due to the growing demand for generating images containing visual texts in the AI art community and

---

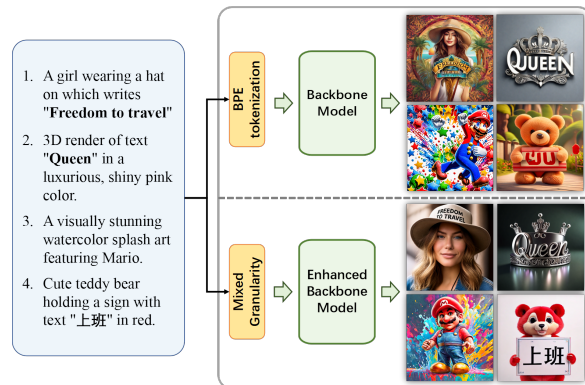* Work done during internship in Baidu.
† Corresponding author.

Figure 1: Comparison between the backbone models (top) and our models (bottom). Our methods can empower the backbone models to generate complex (top left), artistic (top right) visual texts while maintaining fundamental image generation quality (bottom left). Besides, our method can be transferred to Chinese text generation (bottom right).

commercial fields. Despite their attractiveness, this task remains challenging, as most current diffusion models struggle to produce images with precise, readable visual texts. At present, dominant studies on this task can be roughly divided into two categories. Some researchers focus on adding additional conditions to reduce the difficulty of generating images with visual texts, resulting in restricted diversity and visual texts not coherent with the background. (Chen et al., 2023b; Ma et al., 2023; Yang et al., 2023; Tuo et al., 2023; Chen et al., 2023a; Zhao and Lian, 2023). Other researchers directly explore the performance of backbone models on visual text generation, which avoids the limitations of the previous type of methods but suffers from challenges such as misspelling, ignoring, and repeating words. To deal with these issues, early studies (Saharia et al., 2022; Balaji et al., 2022; Liu et al., 2023b) explore various text encoders to address misspelling issues. Recent commercial models such as Dall-E 3 (Betker et al., 2023) and Stable Diffusion 3 (Esser et al., 2024) demonstrate

remarkable performance, further validating the potential of this research direction. However, they lack support for other languages, such as Chinese.

To explore potential avenues for improvements, we first conduct several preliminary experiments and observe that the visual text generation performance of backbone models are mainly constrained for two reasons: First, BPE tokenization requires the model to combine subwords to form complete visual words, increasing the difficulty of generating visual texts. Second, The model is unable to effectively bind visual texts to the corresponding text tokens due to the insufficient learning of the cross-attention modules.

Based on these analyses, we propose a series of methods that significantly improve the visual text generation capability of backbone models, as shown in Figure 1. Specifically, we first introduce a *mixed granularity input* strategy that provides more suitable text representations. Then, we augment the conventional MSE loss with three glyph-aware losses: (1) *attention alignment loss* refines the cross-attention maps, thereby better binding visual texts to their corresponding text tokens; (2) *local MSE loss* highlights the importance of visual text areas; (3) *OCR recognition loss* encourages the model to generate accurate visual texts.

Figure 2 demonstrates that our methods effectively enhance the backbone model's visual text generation ability while maintaining its fundamental capabilities. Particularly, our methods can be transferred to the generation of Chinese texts.

2 Related Work

2.1 Visual Text Generation

Recent studies on visual text generation primarily focus on introducing additional conditions, such as rendered text images, or position coordinates during inference.

Some works concatenate representations of the rendered text image with the latent variable as the model input. For example, TextDiffuser (Chen et al., 2023b) and GlyphDraw (Ma et al., 2023) concatenate the representation of position-aware mask with the latent variable, and utilize pre-trained models to generate positional information. UDiffText (Zhao and Lian, 2023) utilizes an inpainting model that considers concatenation of the position mask, the masked image, and the original image as input. Instead of introducing additional conditions through concatenation, some works also explore

to utilize auxiliary modules. GlyphControl (Yang et al., 2023) use a ControlNet, which receives images with rendered texts as input. Building upon this, AnyText (Tuo et al., 2023) introduces a fusion network that receives position and image masks to support more flexible position control and image editing. Apart from these, several works add special tokens representing additional conditions. For example, TextDiffuser-2 (Chen et al., 2023a) adds additional position tokens into the text encoder to generate text based on the predicted coordinates.

However, the above studies still suffer from the following limitations: (1) The use of these conditions constrains the overall composition of the image, causing issues of restricted diversity and visual texts not coherent with backgrounds; (2) Users are required to provide additional conditions, leading to inconvenience in usage.

2.2 Text-to-Image Backbone Models

Some researchers focus on enhancing the overall capabilities of text-to-image backbone models. Early works in this regard aim at addressing spelling errors by experimenting with various text encoders. For example, Imagen (Saharia et al., 2022) replaces CLIP (Radford et al., 2021) with T5 (Raffel et al., 2019), eDiff-I (Balaji et al., 2022) uses both CLIP and T5.

Additionally, some researchers find that tokenization methods influence the model's ability to generate visual texts. Liu et al. (2023b) believe that the primary reason for spelling errors lies in the lack of character-level glyph information caused by BPE tokenization, and propose to solve this by adopting the character-level text encoder ByT5 (Xue et al., 2021).

Recently, some commercial models, such as Dall-E 3 (Betker et al., 2023) and Stable Diffusion 3 (Esser et al., 2024) show outstanding performance in visual text generation. This demonstrates that with the development of backbone models, the performance of visual text generation is concurrently improving. However, these commercial models only support English, leaving the generation of visual texts in other languages unsolved.

In this work, we propose a series of methods, which empower the backbone models with the ability to generate accurate and aesthetic visual texts in two aspects. First, we propose a mixed granularity input strategy to provide more suitable text representations. Second, we augment the conventional training objective with three glyph-aware losses.

Figure 2: Visual text generation results of our models. Our methods significantly empower the backbone models to generate semantic relevant, visual appealing visual text images generation in English and Chinese.
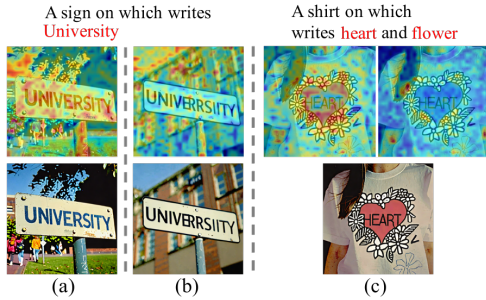

Figure 3: Visualization of the cross-attention maps. (a): "*University*" is correctly spelled, the token has large values on the corresponding areas. (b): "*University*" is not correctly spelled, token "*university</w>*" fails to focus on the corresponding area. (c): The token "*heart</w>*" attends to the corresponding area, thus is correctly generated, while the token "*flower</w>*" highlights irrelevant region and fails to generate the corresponding visual text.

3 Preliminary Study

In this section, we first introduce the basic concepts of the diffusion based text-to-image backbone model, and then conduct experiments to identify potential avenues for improvements.

3.1 Diffusion Based Text-to-Image Backbone Models

Model Architecture. The commonly-used architecture of text-to-image backbone models derives from the latent diffusion model (Rombach et al., 2022), which is composed of three modules:

(1) a VAE (Kingma and Welling, 2014) consists of an encoder to compress images into the latent space, and a decoder to reverse them back; (2) a UNet (Ronneberger et al., 2015) denoiser ϵ_θ performs diffusion denoising process at latent space; (3) a text encoder $\mathcal{T}$ encodes the text prompt into representation c.

Diffusion Process. This process defines a Markov chain of forward diffusion process which continually applies the noise sampled from a Gaussian distribution to the real data $z_0 = \mathcal{E}(x_0)$:

$$q(z_t|z_{t-1}) := \mathcal{N}(z_t; \sqrt{\alpha_t}z_{t-1}, (1 - \alpha_t)I), \quad (1)$$

where α_t is a time-aware schedule. As t increases, z_t asymptotically approaches the noise in a standard Gaussian distribution.

The UNet denoiser ϵ_θ is trained to predict the noise ϵ_t added to the image at timestep t, thereby reversing the Markov chain. A mean squared error (MSE) loss is utilized to supervise the training:

$$\mathcal{L}_{mse} = \mathbb{E}_{z_0,c,\epsilon_t,t} \left[||\epsilon_\theta(z_0, t, c) - \epsilon_t||_2^2\right]. \quad (2)$$

To add conditional guidance, the representation c is fed into each cross-attention block of the UNet model as:

$$Attn(z_t, c) = \text{Softmax}(\frac{Q(z_t) \cdot K(c)^T}{\sqrt{d}})V(c), \quad (3)$$

where Q, K and V denote the query, key and value projections, and d denotes the output dimension.
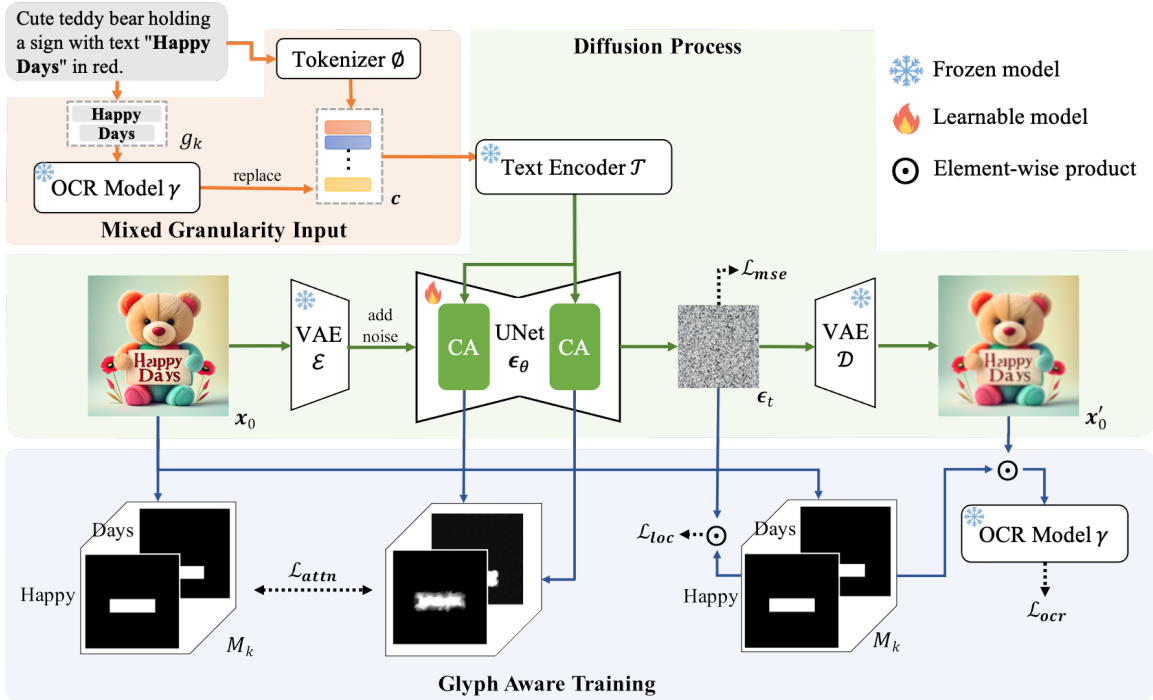
Figure 4: The framework of our methods. The Mixed Granularity Input strategy considers glyph words as whole units to provide more suitable text representations. The Glyph Aware Training includes three losses: (1) the attention alignment loss enhances the learning of cross-attention modules; (2) the local MSE loss highlights the importance of visual text areas; (3) the OCR recognition loss encourages the model to generate accurate visual texts.

3.2 Experimental Analyses

To identify avenues for improvements, we use the commonly-used backbone model–SD-XL (Podell et al., 2023) to conduct two groups of experiments.

In the first group of experiments, we investigate the effect of BPE tokenization on two subsets: (1) S_1, where words are split into subwords by BPE tokenization, and (2) S_2, consisting of words that remains the same after BPE tokenization. To eliminate the impact of word frequency and length, we select 100 words for each subset from 5,000 common words with lengths ranging from 5 to 8 letters[1]. Results show that the model achieves an accuracy of 0.3 in S_1, compared to 0.46 in S_2, indicating that **BPE tokenization increases the difficulty for the model in generating visual texts**, as it splits a word into subwords and requires the model to combine them into a complete visual word.

As stated in previous works (Hertz et al., 2023; Chefer et al., 2023), the cross-attention maps of the UNet can reflect the relevance between generated objects and corresponding text tokens. Similarly, visual texts can also be treated as objects, and texts to be generated, which we refer to as *glyph texts*,

should therefore have a robust relationship with the corresponding visual texts in the image. In the second group of experiments, we extract and visualize the cross-attention maps for glyph tokens at the last timestep, as depicted in Figure 3. We can clearly observe that cross-attention maps with corresponding visual texts generated are correctly localized, while the maps that do not have corresponding visual texts generated highlights irrelevant regions. Thus, we conclude that **glyph tokens indeed have a robust relationship with visual text areas through cross-attention mechanism**, which the model fails to effectively capture.

In summary, based on our experimental analyses, we believe that BPE tokenization and the insufficient learning of cross-attention modules constrain the model's ability to correctly generate visual texts.

4 Methods

Based on the observations from our preliminary study, we propose a series of methods to improve the visual text generation capability of backbone models. As shown in Figure 4, our improvements mainly involve two aspects: (1) we introduce a mixed granularity input strategy to replace the BPE
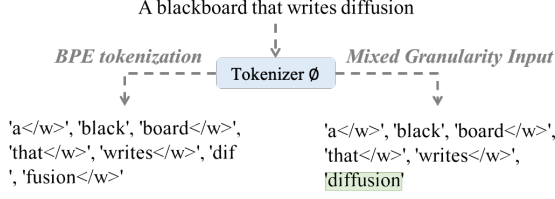
---

[1] https://github.com/first20hours/google-10000-english

8004

A blackboard that writes diffusion

BPE tokenization   Tokenizer ∅   *Mixed Granularity Input*

'a</w>', 'black', 'board</w>', 'that</w>', 'writes</w>', 'dif ', 'fusion</w>'

'a</w>', 'black', 'board</w>', 'that</w>', 'writes</w>', 'diffusion'

Figure 5: Mixed granularity input. The word "*diffusion*" is considered as a whole instead of being tokenized.

subword input; (2) we augment the conventional training objective with three glyph-aware training losses, which regulates the cross-attention maps and encourage the model to focus on visual texts.

4.1 Mixed Granularity Input

Our preliminary study reveals that BPE tokenization constrains the performance of the model, highlighting the necessity to represent glyph texts in a more suitable granularity. In this regard, previous studies (Liu et al., 2023b; Chen et al., 2023a; Zhao and Lian, 2023) commonly utilize character-level tokenization, which splits words into characters. However, as stated in our preliminary study, this split challenges the model to combine characters into a complete visual word. To deal with this issue, we consider each glyph word as a whole within the model, as shown in Figure 5. Given the impracticality of including every word in the vocabulary, a method is needed to get the embedding for every word. Therefore, we extract intermediate features from the OCR model as new text embeddings following Tuo et al. (2023), which inherently possess sufficient glyph information. Specifically, for a user prompt y containing N glyph words $g_1, g_2, \ldots, g_N$, we render each glyph word into an image without providing positional information, resulting in an image sequence I_g. Then, we feed them into the OCR model γ, where the text embedding c is refined as follows:

$$c = \mathcal{T}(\phi(y), \xi(\gamma(I_g))), \quad (4)$$

where $\mathcal{T}$ is the CLIP text encoder, ϕ is the BPE tokenizer, and ξ is a linear module.

4.2 Glyph-Aware Training

Formally, the overall training objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_{mse} + \lambda_1 \cdot \mathcal{L}_{attn} + \lambda_2 \cdot \mathcal{L}_{loc} + (1 - \lambda_1 - \lambda_2) \cdot \mathcal{L}_{ocr}, \quad (5)$$

where $\mathcal{L}_{attn}$, $\mathcal{L}_{loc}$ and $\mathcal{L}_{ocr}$ denote the attention alignment loss, the local MSE loss, and the OCR recognition loss, respectively.

4.2.1 Attention Alignment Loss $\mathcal{L}_{attn}$

To enhance the learning of cross-attention modules, we introduce an attention alignment loss, which encourages the model to ensure that each visual text mainly attends to the corresponding glyph token. Specifically, the cross-attention map between the intermediate feature of the noisy latent variable z_t and the refined representation c_g of glyph tokens can be calculated as:

$$CA(z_t, c_g) = \text{Softmax}(\frac{Q(z_t) \cdot K(c_g)^T}{\sqrt{d}}). \quad (6)$$

To encourage that each visual text has large values in the corresponding area, we minimize the distance between the cross-attention maps and the corresponding segmentation masks of visual texts, which is defined as follows:

$$\mathcal{L}_{attn} = \frac{1}{N} \sum_{k=1}^{N} \left\| CA(z_t, c_g^k) - M_k \right\|_2^2, \quad (7)$$

where M denotes the segmentation mask of the k-th visual text corresponding to its glyph token.

Through this training process, the model can effectively capture a more robust understanding of the relationships between the visual texts and glyph tokens, thus faithfully generating the desired visual texts.

4.2.2 Local MSE Loss $\mathcal{L}_{loc}$

Since the MSE loss only measures pixel-wise distance and lacks additional focus on visual text areas, we apply a weighting strategy to the MSE loss following Ma et al. (2023), which we refer to as the *local MSE loss*. To mitigate the impact of visual text area size, we add a weighting term w which is the ratio of the image area to the visual text area. Formally, the local MSE loss can be formulated as:

$$\mathcal{L}_{loc} = \frac{1}{N} \sum_{k=1}^{N} w^k \cdot \mathcal{L}_{loc}^k,$$

$$\mathcal{L}_{loc}^k = \mathbb{E}_{z_0, \epsilon_t, t} \left[M_k \odot \left\| \epsilon_\theta(z_0, t, c) - \epsilon_t \right\|_2^2 \right]. \quad (8)$$

4.3 OCR Recognition Loss $\mathcal{L}_{ocr}$

To further encourage the model to generate accurate visual texts, we introduce an OCR recognition task. At each training step, we can estimate the fully denoised image latent variable z_0', as implemented in DDPM (Ho et al., 2020). We then input this latent variable into the VAE decoder to obtain an approximate image x_0', which is subsequently

Metrics	DeepFloyd	SD-XL	SDXL-Turbo	LCM-LoRA	SD Cascade	Ours(XL)	Ours(Turbo)
Quantitative Results							
CLIP score	28.256	28.056	30.001	26.631	**30.102**	29.168	29.394
OCR (Precision)	0.153	0.116	0.315	0.032	0.298	**0.348**	**0.360**
OCR (Recall)	0.210	0.138	**0.337**	0.038	0.308	0.283	0.307
OCR (F1 score)	0.177	0.125	0.326	0.035	0.303	0.312	**0.331**
User Study Results							
Text Aesthetics	0.100	0.050	0.100	0.016	0.200	**0.259**	**0.275**
Text Accuracy	0.131	0.041	0.134	0.006	0.188	**0.234**	**0.266**
Semantic Relevance	0.066	0.022	0.059	0.009	0.178	**0.303**	**0.363**
Image Aesthetics	0.038	0.034	0.053	0.031	0.178	**0.287**	**0.378**

Table 1: Quantitative results of English text generation compared with other backbone models. 'XL', 'Turbo' denotes SD-XL, SDXL-Turbo. Our models achieve the best results in terms of most metrics.

fed into the OCR model for recognition. As implemented in the training of the OCR model, we use the CTC loss (Graves et al., 2006) to refine the predicted results. Since this estimation introduces more distortion as t increases, we add a weighting term related to t, which is set as $\bar{\alpha}_t$ following Tuo et al. (2023). The OCR recognition loss can be formulated as:

$$\mathcal{L}_{ocr} = \frac{1}{N} \sum_{k=1}^{N} \bar{\alpha}_t \cdot CTC(\boldsymbol{x}_0' \odot M_k, g_k), \quad (9)$$

where $CTC(\cdot)$ denotes the CTC loss function.

5 Experiments

5.1 Dataset

To better unleash the potential of the model, we require a large-scale, high-quality dataset that satisfies the following criteria:

- The dataset should contain images with clear and recognizable visual texts.
- The visual texts in the images should occupy a prominent area and be coherent with the background.
- The captions should include detailed descriptions of the visual texts.
- The aesthetic quality of the images should be comparable to those used for pre-training.

Following the aforementioned criteria, we construct an English dataset consisting of 240K samples by filtering internal datasets, and a Chinese dataset containing 50K synthetic samples using

image-to-image models and rendering tools[2]. More details are introduced in Appendix A.

5.2 Experimental Setting

Evaluation Metrics. We quantify the visual text generation quality from two aspects: (1) **CLIP score** (Hessel et al., 2021) measures the semantic relevance between the generated image and the input prompt by calculating the cosine similarity of their representations from CLIP image and text model (Radford et al., 2021). (2) **OCR Accuracy** detects the texts in the generated images utilizing OCR tools. We calculate the precision, recall and F1 score between the detected texts and the ground truths. Furthermore, we evaluate the model 's fundamental capability through **FID** (Heusel et al., 2017) score, which compares the distribution of generated images with that of real images. Note that we are unable to calculate the FID score in our main experiments, due to the lack of source images.

Implementation Details. We train our models based on SDXL-base-1.0 and SDXL-Turbo. We utilize the PaddleOCR v4 model[3] to extract intermediate features, perform the OCR recognition task, and conduct evaluation. We set λ_1 and λ_2, to 0.4, 0.2, respectively, determined by a grid search on the validation set, which are varied from 0.1 to 1.0 with an interval of 0.1. We set the learning rate to 2e-5 and conduct a total of 10K steps of training. The overall training process takes 7 hours and 50 minutes on 8 A800 GPUs.
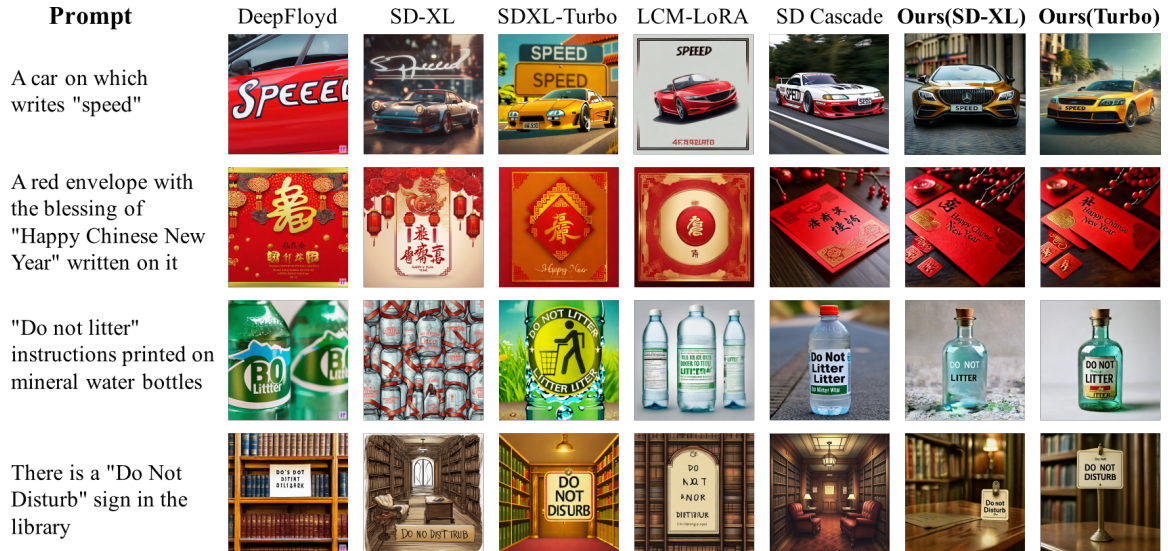
---

[2]https://pypi.org/project/pillow/
[3]https://github.com/PaddlePaddle/PaddleOCR

Figure 6: Visualization of generating English texts compared with other backbone models.

5.3 Quantitative Results

As shown in Table 1, we conduct quantitative comparison with existing backbone models on the ChineseDrawText benchmark (Ma et al., 2023). We compare our models with DeepFloyd (DeepFloyd-Lab, 2023), SD-XL, SDXL-Turbo (Sauer et al., 2023), LCM-LoRA (Luo et al., 2023), and SD-Cascade (Pernias et al., 2023). Results show that our models outperform other baseline models under the majority of metrics.

5.4 User Study

To further validate the effectiveness of our proposed methods, we conduct a human evaluation comparing our English models with other baseline models on ChineseDrawText benchmark. Three raters are asked to compare these images from four dimensions including text aesthetics, text accuracy, semantic relevance, and image aesthetics, and then select the images they prefer. Throughout the process, all raters are unaware of which model the image is generated from. The results in Table 1 show that human raters greatly prefer our models on all aspects, which further validates the effectiveness of our approaches in generating high-quality and visual text images. The detailed participant instruction are listed in Appendix B.

5.5 Qualitative Results

To provide more straightforward comparison, we provide some visualization samples from the test set in Figure 6. We can clearly observe that our models better capture the semantic relevance, thus

Figure 7: Comparison between Ours(Turbo) at the top and SDXL-Turbo at the bottom.

generate visual texts at the reasonable place. For example, our models generate visual text "*Speed*" at the front of the car (line 1), while some baselines (SD-XL, SDXL-Turbo, LCM-LoRA) fail to capture the guidance "*write on the car*". Besides, as shown in line 3, our models effectively avoid issues like misspelling (Deepfloyd, SDXL-Turbo, LCM-LoRA), repeating (SD Cascade, SDXL-Turbo) and ignoring words (Deepfloyd), or failing to understand the instruction (SD-XL). We provide visualization comparison and more showcases in Appendix C.

Regarding the failure cases, we analyze why the recall of our model is lower than that of the SDXL-Turbo model. As shown in Figure 7, the SDXL-Turbo model tends to generate repeated words. For example, in the first column, the SDXL-Turbo model generates "*do*" and "*not*" twice and spells them correctly once, leading to a higher recall. However, this repetition fails to align with human reference, resulting in a lower precision score.

Method	FID
Ours(SD-XL)	**53.4**
Ours(SDXL-Turbo)	**52.9**
SD-XL (Podell et al., 2023)	73.6
SDXL-Turbo (Sauer et al., 2023)	61.9

Table 2: COCO zero-shot FID_{5k}(FID) comparison.

Figure 8: Visualization of generating images without visual texts.

5.6 Comparison of Generating Images without Visual Texts

In order to evaluate the fundamental image generation performance of our models, we use FID to quantify the image quality without visual texts on 5K samples from COCO2017 (Lin et al., 2014), as shown in Table 2. Furthermore, we visualize some image generation examples in Figure 8. The quantitative and qualitative results indicate that our models maintain the fundamental capability to generate visual appealing and semantic relevant images.

5.7 Ablation Study

To investigate the effectiveness of each design, we further compare our SD-XL model with the following variants in Table 3:

(1) ⇒*Char+BPE tokenization, and*⇒*BPE tokenization*. In the first variant, we replace our mixed granularity input with the mixture of character-level and BPE tokenization. In the second variant, we only utilize BPE tokenization. As shown in line 2, our mixed granularity input strategy outperforms the mixture of character-level and BPE tokenization. We hypothesize that this is because the model struggles to combine the glyphs of characters to form a complete visual word. The result in line 3 shows that the mixture of character-level and BPE tokenization achieves better results comparing to BPE tokenization, which demonstrate the effectiveness of providing character-level glyph information.

Method	Pre.	Rec.	F1.
Ours	**0.348**	**0.283**	**0.312**
⇒ Char+BPE tokenization	0.231	0.193	0.210
⇒ BPE tokenization	0.227	0.182	0.202
w/o $\mathcal{L}_{attn}$	0.278	0.236	0.250
w/o $\mathcal{L}_{loc}$	0.254	0.236	0.244
w/o $\mathcal{L}_{ocr}$	0.270	0.218	0.241

Table 3: Ablation on the generation of English texts. ⇒* means replacing the input granularity with *. $\mathcal{L}_{loc}$, $\mathcal{L}_{attn}$ and $\mathcal{L}_{ocr}$ denote the local MSE loss, the attention alignment loss, and the OCR recognition loss, respectively. 'Pre.' and 'Rec.' denote Precision and Recall respectively.

Figure 9: Visualization of generating Chinese texts compared with other backbone models. * denote the baseline model that is trained on our Chinese dataset for 10K steps.

(2) *w/o* $\mathcal{L}_{attn}$. In this variant, the attention alignment loss $\mathcal{L}_{ocr}$ is removed. The result in line 4 shows a significant performance drop, which confirms our previous assumption that the insufficient learning of cross-attention modules constrains the visual text generation capability of backbone models.

(3) *w/o* $\mathcal{L}_{loc}$. We remove the local MSE loss $\mathcal{L}_{loc}$ from this variant. The result in line 5 demonstrates a significantly decline in OCR accuracy, indicating that focusing on visual text areas is indeed helpful for generating correctly spelled visual texts.

(4) *w/o* $\mathcal{L}_{ocr}$. We remove the OCR recognition loss $\mathcal{L}_{ocr}$, and observe from line 6 that the performance suffers from a great decline, demonstrating the effectiveness of OCR recognition loss.

5.8 Chinese Text Generation

We further explore the effectiveness of our methods to generate Chinese visual texts. Instead of considering words as whole units, we use the mixture of

Method	CLIP score	Pre.	Rec.	F1
SD-XL*	29.569	0.299	0.319	0.308
SDXL-Turbo*	29.827	0.275	0.295	0.284
SD Cascade*	29.734	0.146	0.154	0.150
Ours(SD-XL)	29.258	0.582	0.608	0.595
Ours(Turbo)	**29.919**	**0.598**	**0.613**	**0.606**

Table 4: Quantitative results of Chinese text generation. * denotes that these baselines are trained on our Chinese dataset for 10K steps before comparison.

character-level and BPE tokenization for Chinese texts due to two reasons. First, Chinese glyphs are excessively complex, resulting in the intermediate features of each character being too similar in distribution to be effectively distinguished. Second, fewer characters are included in each Chinese word, thus is easier to be combined into a complete visual word.

Note that due to the lack of open source Chinese backbone text-to-image models for comparison, we train both our models and the baseline models on our Chinese dataset for 10K steps. We choose SD-XL, SDXL-Turbo and SD Cascade, which achieve relatively better performance in English, as baseline models, and use the prompt templates from the ChineseDrawText benchmark with texts included in our Chinese dataset as test set. Quantitative results in Table 4 show that our models greatly outperform other baseline models. As for qualitative comparison, we visualize some samples from our test set, as shown in Figure 9. Our model generates accurate visual texts, while other baseline models fails to correctly generate Chinese texts, indicating that our methods enhance the learning of Chinese texts. We provide the ablation study for Chinese text generation in Appendix D.

6 Conclusion

In this paper, we conduct a preliminary study and find that BPE tokenization, as well as the model's insufficient learning of cross-attention modules, constrains the visual text generation performance of diffusion-based backbone models. Based on these insights, we propose a series of methods, aiming to empower the backbone model with the ability to generate accurate and aesthetically appealing visual text images, while maintaining fundamental image generation quality. Specifically, we introduce a mixed granularity input strategy to provide

more suitable text representations. Besides, we augment the conventional training objective with three glyph-aware training losses, which enhance the learning of the cross-attention modules and encourage the model to focus on visual texts. Experiments demonstrate the effectiveness of our methods. Typically, our methods can be transferred to Chinese text generation.

In the future, we intend to explore visual text generation for more languages, and generate texts in different styles (Liu et al., 2023a). Besides, we also plan to explore utilizing glyph enhanced diffusion models for image-to-image translation (Lan et al., 2024).

Limitations

While our methods enhance the visual text generation capability of the backbone models, several limitations still remain. First, our methods require to train the diffusion backbone model, which may be time consuming and expensive. Besides, our methods are unable to completely solve the issue of misspelling, ignoring and repeating words.

Ethics Statement

This research paper rigorously addresses the ethical considerations associated with text-to-image models, ensuring that all methods used in this study are conducted responsibly and ethically. Our models are trained using open-source backbone models. To address concerns related to training data, we implement a strict filtering process to exclude inappropriate content, such as NSFW images and offensive visual text. The evaluation experiments are conducted using widely recognized public benchmarks, and participants involved in the user studies are systematically trained.

Acknowledgments

References

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *CoRR*, abs/2211.01324.

James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † Wesam-Manassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. 2023. Improving image generation with better captions.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2023a. Textdiffuser-2: Unleashing the power of language models for text rendering. *CoRR*, abs/2311.16465.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2023b. Textdiffuser: Diffusion models as text painters. In *NeurIPS*.

DeepFloyd-Lab. 2023. Deepfloyd if.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. 2020. PP-OCR: A practical ultra lightweight OCR system. *CoRR*, abs/2009.09941.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *CoRR*, abs/2403.03206.

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. 2022. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and A foundation framework. *CoRR*, abs/2202.06767.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-prompt image editing with cross-attention control. In *ICLR*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *ICLR*.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, Min Zhang, and Jinsong Su. 2024. Translatotron-v(ison): An end-to-end model for in-image machine translation. In *ACL*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *ECCV*.

Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. 2023a. On the cultural gap in text-to-image generation. *CoRR*.

Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. 2023b. Character-aware models improve visual text rendering. In *ACL*.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. 2023. Lcm-lora: A universal stable-diffusion acceleration module. *CoRR*, abs/2311.05556.

Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. 2023. Glyphdraw: Learning to draw chinese characters in image synthesis models coherently. *CoRR*, abs/2303.17870.

Pablo Pernias, Dominic Rampas, Mats L. Richter, Christopher J. Pal, and Marc Aubreville. 2023. Wuerstchen: An efficient architecture for large-scale text-to-image diffusion models. *CoRR*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. 2023. Adversarial diffusion distillation. *CoRR*, abs/2311.17042.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. In *NeurIPS*.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. 2023. Anytext: Multilingual visual text generation and editing. *CoRR*, abs/2311.03054.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *CoRR*, abs/2105.13626.

Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. 2023. Glyphcontrol: Glyph conditional control for visual text generation. In *NeurIPS*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Yiming Zhao and Zhouhui Lian. 2023. Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. *CoRR*, abs/2312.04884.

A Dataset

A.1 Data Collection

To obtain suitable English training data, we initially survey academic datasets such as LAION-5B (Schuhmann et al., 2022) and WuKong (Gu et al., 2022), and find that these academic datasets have issues with low image resolution, which significantly degrades the overall quality of the images when used for training. Therefore, we utilize our internal dataset and employ the high-precision OCR model PaddleOCR (Du et al., 2020) to filter data with texts. To get high quality captions including visual text information, we utilize a Multimodal Large Language Model (MLLM) and include the OCR results in the prompt to improve the accuracy of the generated captions. Following the above steps, we construct a English dataset of 240,000 high-aesthetic image-caption pairs.

Regarding Chinese data, we identify several issues with both web-crawled and academic datasets upon sampling: an excessive amount of text, overly complex glyphs, and text not occupying a prominent area in the images. These issues increase the difficulty of generating Chinese visual texts. Consequently, we explore constructing synthetic data using rendering and image-to-image models. We first select Chinese phrases that consists of two characters with no more than 10 strokes for each character [4], and then apply manual deduplication to prevent overfitting to some characters, resulting in 255 phrases. We then design 10 templates to stipulate positional information and render the characters onto background images according to the templates. Finally, we apply image-to-image models to generate the backgrounds and conduct post-filtering with the OCR model to ensure that the aforementioned issues are avoided. However, due to the use of predefined rules, there are significant limitations in the diversity and overall aesthetic quality of the data, which impact the quality of the images generated by the model.
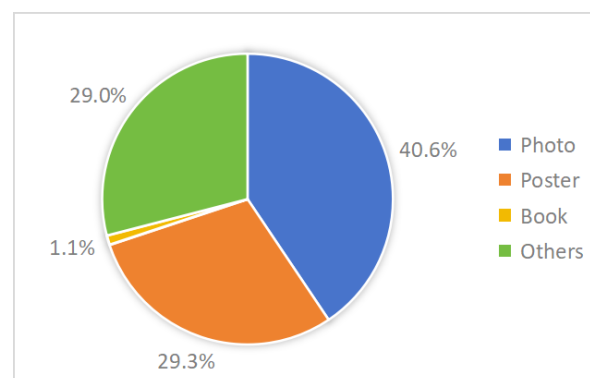
---

[4] https://github.com/thunlp/THUOCL/tree/master

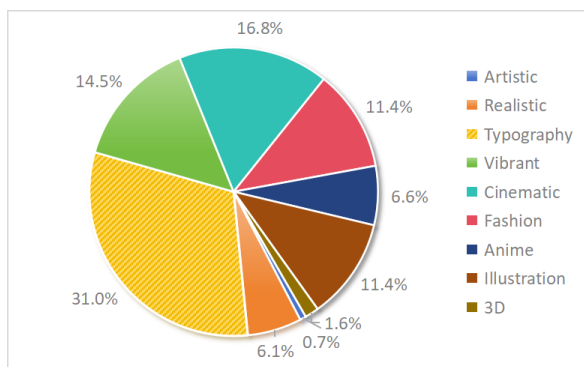Figure 10: Statistics of data sources of our dataset.

Figure 11: Statistics of data style of our dataset.

A.2 OCR Filtering Rules

We filter our collected data with the following criteria:

- **Height and width larger than 1024 for English dataset, and 512 for Chinese dataset.** We find that low-resolution sample has a negative impact for the training.

- **Area size for each visual text are more than 10% of the whole image.** Visual texts being too small will increase the error rate of OCR recognition during training, thus introduce noise into data, and images with small texts often contain watermarks.

- **At least one detected text appears in the caption.** MLLMs would reject to describe when there is no visual text included in the image, we mark these images as invalid.

- **Text areas are at least 10% away from border.** Texts too close to image boarder are more likely to be pruned when regulating images within a batch during training.

- **Number of texts should be no more than 5.** Samples that include too much texts typically have small areas for each text.

A.3 Data Statistics

We further provide the statistics of our data source and style, as listed in Figure 10 and 11.

B Participant Instruction

Objective: Evaluate the images based on the following four criteria. Select one or multiple images that you prefer; you can select all or none of them. Each sample will be evaluated by three raters.

- **Text Accuracy**:

  Metric: The visual texts should be correctly spelled and easily recognizable. The following errors should be considered in descending order of importance: neglecting or repeating words, misspelling words, generating words that are not requested.

  Further Consideration: A repeated word may not be exactly the same as others. For example, if the prompt asks to generate the text 'apple', texts like 'aple', 'apple', 'appple' should all be considered as repeated words.

- **Text Aesthetics**:

  Metric: The color of the visual texts should be coherent with the background. The font style of the visual texts should match the current scenario. For example, in paintings, the texts should be artistic; in posters, the texts should be eye-catching and occupy a prominent area. The positions of the visual texts should be reasonable.

  Further Consideration: If the image does not contain any recognizable visual texts, it should be less preferred. The accuracy of the visual texts is not considered in this metric.

- **Semantic Relevance**:

  Metric: The image should depict a scenario that matches the requirements of the user prompt.

  Further Consideration: Note that the relevance between the visual texts and the user prompt is considered in text accuracy; the overall semantic relevance is more important. If the image contains noise from visual texts, it should be less preferred. For example, if the user prompt asks the image to contain the word 'apple' without needing to draw a real apple, images containing a real apple should be considered as bad cases.

- **Image Aesthetics**:

  Metric: The image should be visually appealing.

  Further Consideration: The aesthetics of the visual texts are less important than the background.

A photo of a mug of coffee with the word "Peace" on the cup.

A vintage hot rod, cruising along a sunny street line with palm trees. The text "Summer Fun" is displayed in a retro, distressed style.

An imaginative city skyline with blue text "NEWYORK" in bold letters.

A cute chubby cat sitting with a dialogue bubble, which says " Summer Fun" in a handwritten-style font.

A green door with word "DO NOT DISTURB" written on it.

An antique glass kettle filled with boiling water, on which writes text "Good Morning" in a mystical style.

Figure 12: More visualizations of visual text generation results.

Variant	CLIP score	Pre.	Rec.	F1
240K / 10K	29.168	0.348	0.283	0.312
120K / 5K	28.584	0.206	0.203	0.205
48K / 2K	28.075	0.127	0.173	0.146

Table 5: Quantitative results of scalability of our methods, A / B denotes the model is trained on A samples for B steps.

B.1 Scalability

To investigate the scalability of our methods, we train three variants based on SD-XL, as detailed in Table 5. The results demonstrate that as data scale and training steps increase, the model's performance consistently improves.

C More Visualization Results

As depicted in Figure 12 and Figure 13, we showcase more visualizations results of our models on visual text generation task. Our model can generate visual appealing, style diverse, and legible visual text images, while maintaining basic capability to generate images without visual texts.

D Ablation Study for Chinese Text Generation

We further conduct ablation study for Chinese texts and compare the results with English texts.

| Method	Pre.	Rec.	F1
Ours	**0.598**	**0.613**	**0.606**
$\Rightarrow$ BPE tokenization	0.0	0.0	0.0
w/o $\mathcal{L}_{attn}$	0.578	0.604	0.591
w/o $\mathcal{L}_{loc}$	0.116	0.120	0.118
w/o $\mathcal{L}_{ocr}$	0.411	0.420	0.416

Table 6: Ablation on the generation of Chinese texts. $\Rightarrow$ BPE tokenization means using only BPE tokenization. $\mathcal{L}_{loc}$, $\mathcal{L}_{attn}$ and $\mathcal{L}_{ocr}$ denotes the local MSE loss , the attention alignment loss , and the OCR recognition loss, respectively

(1) *Char+BPE tokenization⇒BPE tokenization.* We replace the mixture of character-level and BPE tokenization with BPE tokenization. While considering words as whole units achieves the best results for English texts, we find that utilizing character-level tokenization for Chinese texts yields the best performance. We indicate that this is because the glyph information for each Chinese character is more complex, and fewer characters are included in each phrase.

(2) *w/o $\mathcal{L}_{attn}$.* We remove the attention alignment loss $\mathcal{L}_{attn}$ in this variant. From line 3, we observe a greater performance decline of OCR accuracy for English texts than Chinese texts without the attention alignment loss, indicating that English texts are more susceptible to cross-attention scores.

Watercolor painting colorful seascape. At the center, there's a boat with a bright red sail, floating on calm waters.
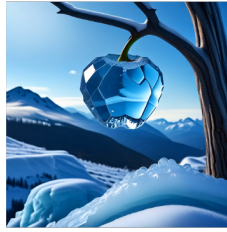
a crystal-clear ice apple on a tree branch, with a snow-capped mountain in the distance.

A captivating sumi-e inspired masterpiece illustration featuring a close-up of a menacing wolf.

A delightful black and white coloring page featuring a cute, cartoon-style giraffe.

A unique cubism-inspired portrait of a cat.

A skull with glowing blood-red eyes. The top half of the skull's forehead seamlessly transitions into a whippy ice cream.

Figure 13: More visualizations of image generation without visual texts.

We assume this is because more English words are included in each input image during training, making it harder to bind each visual text to its text token.

(3) *w/o* $\mathcal{L}_{loc}$. The local MSE loss $\mathcal{L}_{loc}$ is not included in this variant. As shown in line 4, we observe a greater performance decline for Chinese texts when local MSE loss is not incorporated. This indicates that Chinese glyphs are harder to learn and should receive more attention during training.

(4) *w/o* $\mathcal{L}_{ocr}$ We remove the OCR recognition loss $\mathcal{L}_{ocr}$. As shown in line 5, similar to results for English texts, there is a significant performance decline when OCR recognition loss is not included, the which indicates that the OCR recognition loss does have a positive effect for both English and Chinese texts.