FontStudio: Shape-Adaptive Diffusion Model for Coherent and Consistent Font Effect Generation

Xinzhi Mu Li Chen Bohan Chen[†] Shuyang Gu
Jianmin Bao Dong Chen Ji Li Yuhui Yuan

Microsoft
{xinzhimu,yuyua}@microsoft.com
https://font-studio.github.io/

Fig. 1: Illustrating the font effect generation results by our FontStudio system. We observe that most concepts are generated in adherence to complex font shapes adaptively. We also notice a coherent 3D structure and depth effect. Refer to the supplementary for a detailed prompt of these generative font effects.

Abstract. Recently, the application of modern diffusion-based text-to-image generation models for creating artistic fonts, traditionally the domain of professional designers, has garnered significant interest. Diverging from the majority of existing studies that concentrate on generating artistic typography, our research aims to tackle a novel and more demanding challenge: the generation of text effects for multilingual fonts. This task essentially requires generating coherent and consistent visual content within the confines of a font-shaped canvas, as opposed to a traditional rectangular canvas. To address this task, we introduce a novel shape-adaptive diffusion model capable of interpreting the given shape and strategically planning pixel distributions within the irregular canvas. To achieve this, we curate a high-quality shape-adaptive image-text dataset and incorporate the segmentation mask as a visual condition to steer the image generation process within the irregular-canvas. This approach enables the traditionally rectangle canvas-based diffusion model

[†] Intern at Microsoft.

to produce the desired concepts in accordance with the provided geometric shapes. Second, to maintain consistency across multiple letters, we also present a training-free, shape-adaptive effect transfer method for transferring textures from a generated reference letter to others. The key insights are building a font effect noise prior and propagating the font effect information in a concatenated latent space. The efficacy of our FontStudio system is confirmed through user preference studies, which show a marked preference (78% win-rates on aesthetics) for our system even when compared to the latest unrivaled commercial product, Adobe Firefly[1].

Keywords: Shape-Adaptive · Diffusion Model · Font Effect

1 Introduction

Recently, models based on diffusion techniques for text-to-image generation have achieved significant success in rendering photorealistic images on standard rectangular canvases [6, 23, 32]. Many follow-up efforts have built many generation-driven exciting applications like subject-driven image generation and spatial conditional image generation. For instance, ControlNet [52] offers a powerful method for integrating spatial conditioning controls, such as edges, depth, segmentation, and more, into pre-trained text-to-image diffusion models, enhancing their versatility and application range.

Despite these advancements, the focus predominantly remains on rectangular canvases, leaving the potential for image generation on non-standard, arbitrarily shaped canvases largely untapped. The task of creative font effect generation essentially requires generating visual contents in non-regular and complex-shaped canvas. It demands not only synthesizing semantic objects or concepts aligning with arbitrary user prompts but also a deep understanding of the geometric shapes of the font canvas. In essence, the visual elements produced must be precisely positioned within the irregular-canvas to ensure visual harmony while also ensuring faithful generation within the specific font canvas following the given text prompt. Our empirical analysis, illustrated in Figure 2, demonstrates the outcomes of directly utilizing conventional diffusion models, including ControlNet, SDXL, and SDXL-Inpainting model, designed for rectangular canvases. From this analysis, it becomes evident that simply adapting models intended for rectangular canvases to generate visual content for the diverse array of font shapes presents a significant and largely uncharted challenge in the field.

To bridge the gap between traditional rectangle-canvas-based diffusion models and the intricate task of comprehending font shapes for font effect generation, we propose an innovative and potent shape-adaptive diffusion model. This model excels in producing high-quality visual content that conforms to any given shape, encompassing multilingual font outlines and even more intricate patterns such as fractal-structured snowflakes. The key idea is to build a high-quality shape-adaptive triplet training data and each instance consists of {irregular-canvas,

[1] `https://firefly.adobe.com/generate/font-styles`

Fig. 2: Comparison with conventional diffusion models designed for rectangular canvas. Most of these methods struggle to generate the appealing visual content within font-shaped canvas. For ControlNet (CN), we find treating the font mask as depth or computing the canny edge map based on font mask suffers various artifacts. Our FontStudio generates much better results in general.

Fig. 3: FontStudio vs. Adobe Firefly. Win-rates accessed by human evaluator preferences in font effect generation.

irregular-image, text prompt} and then train a conditional diffusion model to generate the visual contents within the irregular-canvas. To maintain compatibility with pre-trained diffusion models and ensure efficient training, we choose a rectangular canvas to serve as a placeholder, accommodating both the irregularly shaped canvas and the corresponding irregular image.

The task of generating font effects requires preserving effect consistency across multiple irregular canvases. Merely using the diffusion model in isolation often results in inconsistent outcomes. To address this challenge, we introduce a novel, training-free effect transfer method that combines the effect of a reference letter with the shape mask of a target letter. This method leverages a font effect noise prior to ensure font effect consistency and propagates the reference style and texture from the source to the target image in a concatenated latent space. Our empirical results demonstrate that this approach can effectively serve as a powerful tool for transferring effects or styles.

Last, we established the GenerativeFont benchmark to facilitate a comprehensive evaluation of our methodologies across various dimensions. The results from a user study, depicted in Figure 3, when benchmarked against Adobe Firefly—the leading font effect generation system—reveal a surprising outcome. Our FontStudio system markedly outperforms Adobe Firefly in several key areas. Specifically, thanks to our shape-adaptive generation approach, we observed a remarkable improvement in both shape fidelity and overall aesthetics, with our system achieving win rates of 78% vs. 10% in aesthetics and 66% vs. 6% in shape fidelity. While FontStudio secures these promising achievements, we continue to thoroughly investigate the system's limitations and engage in discussions on emerging challenges that beckon attention from the broader research community.

2 Related Work

Artistic Font Generation. Previous research has explored various facets of font-related tasks, with studies such as [4, 7, 44] concentrating on font creation.

Other methods, including GAN-based approaches [3, 14, 20, 49], stroke-based techniques [5], and statistical approaches [46–48], aim to transfer existing image styles onto font images. Additionally, research on semantic font typography [11, 22, 38, 45, 51, 53] investigates 2D collage generation and reverse challenges, while [19, 39, 40] focus on modifying characters for thematic expression without sacrificing readability. There are also frameworks for glyph design, either leveraging existing assets [50] or large language models [15]. Anything to Glyph [43] parallels our study by suggesting alterations to location features in the diffusion model's denoising phase. However, this method often produces noticeable shadows in both the foreground and background, restricting its utility in further design applications and failing to ensure character consistency throughout generation. Unlike these existing studies, we focus on generating text effects for multilingual fonts, aiming to produce coherent and consistent visual content within the confines of a font-shaped canvas.

Diffusion-based Image Synthesis and Attention. The landscape of text-to-image generation has seen considerable growth in recent times [9, 28, 33], with diffusion-based methodologies [32, 33, 37] at the forefront, pushing the boundaries of image synthesis quality. The scope of investigation has broadened from straightforward text-to-image conversions to encompass a variety of intricate image applications, including conditional image generation [42, 52], image inpainting [2, 36], image-to-image translation [9, 28, 33], image editing [12, 18], and tailored generation [17, 34, 35, 41]. It is noteworthy that these explorations predominantly take place on standard rectangular canvases. The integration of attention mechanisms in diffusion models has spurred a variety of research in areas like image editing [8, 10, 13, 27, 30, 31]. Recent works such as [1, 29] explore attention for style transfer, with StyleAligned [29] closely aligning with our shape-adaptive effect transfer's goals of achieving stylistic consistency through attention-directed generation using reference images. We empirically show that our method performs better in delivering stylistically coherent generated images while preserving diversity.

3 Approach

First, we illustrate the definition and mathematical formulation of the font effect generation task, delve into the primary challenges associated with this task, and outline the foundational insights guiding our methodology. Second, we introduce the key contribution of this work, namely, a shape-adaptive diffusion model, designed to produce visual content on canvases of any shape. Last, we detail the implementation of our shape-adaptive effect transfer method, which utilizes font effect noise prior and font effect propagation to achieve our objectives.

3.1 Preliminary

We use the subscript $\widehat{\ }$ to indicate that the given tensor has a non-rectangular and irregular spatial shape. For example, $\mathbf{X}$ represents a tensor with a rectan-

gular spatial shape like $h \times w$, while $\widehat{\mathbf{X}}$ denotes a tensor of irregular shape with variable dimensions.

Definition of font effect generation. Given a target font effect text prompt $\mathbf{T}$ and a sequence of irregular font shape canvases $\{\widehat{\mathbf{M}}_i | i = 1, ..., n\}$ corresponding to a sequence of letters, the objective is to build a set-to-set mapping function $f(\cdot)$ that can generate a set of coherent and consistent font effect images $\{\widehat{\mathbf{I}}_i | i = 1, ..., n\}$ of the same shape as the given irregular font-shape canvases $\{\widehat{\mathbf{M}}_i | i = 1, ..., n\}$ accordingly. We illustrate the mathematical formulation of font effect generation process as follows:

$$\{\widehat{\mathbf{I}}_i \mid i = 1, ..., n\} = f(\{\widehat{\mathbf{M}}_i \mid i = 1, ..., n\} \mid \mathbf{T}), \tag{1}$$

where we can also use different font effect text for each mask separately. We propose to access the font effect generation quality from the following four critical aspects:

- *Aesthetics*: Each generated image $\widehat{\mathbf{I}}_i$ should be visually attractive.
- *Font Shape Fidelity*: While an exact match isn't necessary, each $\widehat{\mathbf{I}}_i$ should closely resemble its original font shape $\widehat{\mathbf{M}}_i$.
- *Font Style Consistency*: $\widehat{\mathbf{I}}_i$ should exhibit a coherent style for any other image $\widehat{\mathbf{I}}_j$, presenting as a unified design.
- *Prompt Fidelity*: Every $\widehat{\mathbf{I}}_i$ must adhere to the provided target effect prompt.

Primary challenges. The first key challenge in font effect generation is ensuring that the generated visual objects are positioned creatively and coherently on the font-shaped canvas. We have already shown that the results from simply applying diffusion models designed for rectangular canvases are far from satisfactory, as demonstrated in Figure 2. The second challenge involves maintaining font shape fidelity, as the primary goal of generative fonts is to convey messages creatively. Additionally, ensuring consistent font effects across different letters is also a non-trivial and challenging task, considering the canvas shapes vary significantly among different letters.

Formulation of our framework. To address the above challenges, we first reformulate the font effect generation task into the combination of two sub-tasks including *font effect generation for a reference letter* and *font effect transfer from reference letter to each other letter*. The mathematical formulation is summarized as follows:

$$\widehat{\mathbf{I}}_{\text{ref}} = g(\widehat{\mathbf{M}}_{\text{ref}} \mid \mathbf{T}), \tag{2}$$

$$\widehat{\mathbf{I}}_i = h(\widehat{\mathbf{M}}_i \mid \mathbf{T},\ \widehat{\mathbf{M}}_{\text{ref}},\ \widehat{\mathbf{I}}_{\text{ref}}),\ \mathrm{i} \in \{1, \cdots, \mathrm{n}\}, \tag{3}$$

where we use the function $g(\cdot)$ to perform font effect generation based on a single irregular reference canvas, denoted as $\widehat{\mathbf{M}}_{\text{ref}}$. The function $h(\cdot)$ is used to generate consistent font effects, conditioned on the previously generated reference font effect image $\widehat{\mathbf{I}}_{\text{ref}}$, the reference font mask $\widehat{\mathbf{M}}_{\text{ref}}$, and the current font mask $\widehat{\mathbf{M}}_i$. We choose the same reference letter mask for all font effect transfer letters. To implement these two critical functions, we proposed a Shape-Adaptive Diffusion

Model marked as $g(\cdot)$ and a Shape-adaptive Effect Transfer together with Shape-Adaptive Diffusion Model marked as $h(\cdot)$. We will explain the details in the following discussion.

3.2 Shape-Adaptive Diffusion Model

The key challenge in font effect generation arises from the gap between most existing diffusion models, which are trained on rectangular canvases, and the requirement of this task for visual content creation capability on any given irregularly shaped canvas. To close this critical gap, we introduce a shape-adaptive diffusion model that is capable of performing visual content creation on any irregularly shaped canvas as function $g(\cdot)$.

We follow the mathematical formulations outlined in Equation 2 and utilize the transformation function $g(\cdot)$, which is applied to the irregular canvas $\widehat{\mathbf{M}}_i$ conditioned on a given user prompt $\mathbf{T}$, to represent the shape-adaptive diffusion model. The output of the function $g(\cdot)$ is essentially an image $\widehat{\mathbf{I}}_i$ with an irregular shape. Given that directly processing irregular canvases of varying resolutions presents several non-trivial challenges in training standard diffusion models, we propose rasterizing and positioning the irregular canvas mask within a rectangular placeholder, as $\mathbf{M} = \mathsf{Rasterize}(\widehat{\mathbf{M}})$. Essentially, $\mathbf{M}$ is the binary rasterized form of $\widehat{\mathbf{M}}$ where the pixels inside $\widehat{\mathbf{M}}$ are with 1 and the other pixels are with 0. Additionally, we utilize a rectangular image $\mathbf{I}$ to encapsulate the irregular font effect image $\widehat{\mathbf{I}}$ and include an irregular alpha mask layer $\mathbf{M_I}$ to eliminate the regions outside the irregular canvas. Given the irregular shaped canvas mask and image encapsulated within rectangle ones, we reformulate the original Equation 2 as follows:

$$\mathbf{I}, \mathbf{M_I} = \bar{g}(\mathbf{M} \mid \mathbf{T}) \tag{4}$$

where the predicted alpha mask layer $\mathbf{M_I}$ is different from the input conditional font mask $\mathbf{M}$ and it is necessary to ensure coherent and creative effects along the boundary regions. With the alpha mask prediction, we also avoid the necessity to use additional segmentation model to handle the artifacts outside the font-shaped canvas. We elucidate the key that differentiating the refined alpha mask from the conditional canvas mask is achieved through canvas mask augmentation during the training of the subsequent shape-adaptive diffusion model.

Our shape-adaptive diffusion model consists of two sub-models: a shape-adaptive generation model followed by a shape-adaptive refinement model. The shape-adaptive generation model, dubbed SGM, primarily generates content relevant to the prompt within a designated region, utilizing $\mathbf{M}$. Following this, the shape-adaptive refinement model (SRM) takes over, aiming to enhance the initial results by creating an image $\mathbf{I}$ with more defined and natural edges, along with the corresponding mask $\mathbf{M_I}$ for the generated image $\mathbf{I}$. Figure 4 illustrates the entire framework of of our approach.

Shape-adaptive Generation Model. Training a shape-adaptive generation model is non-trivial, and we face two key challenges. The first is the lack of high-

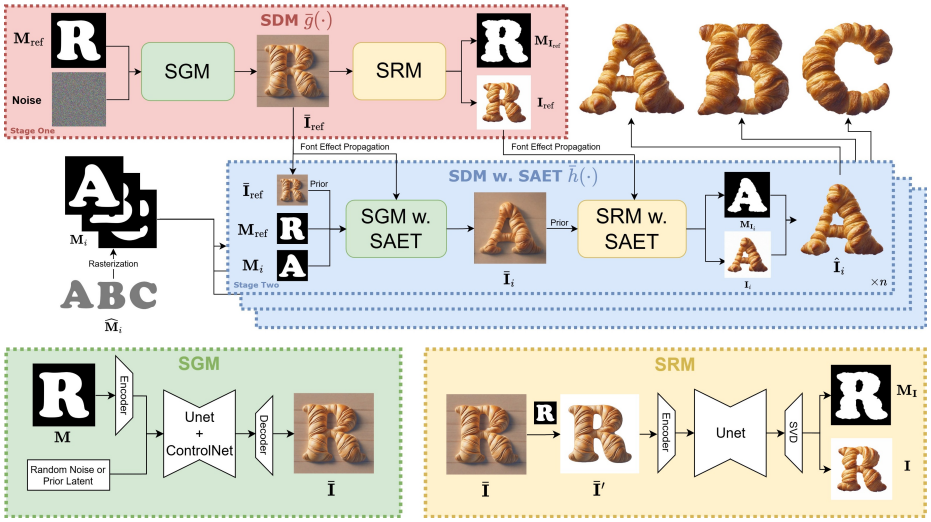

Fig. 4: Overall framework of our approach. The shape-adaptive diffusion model (SDM) consists of two components: the shape-adaptive generation model (SGM) and the shape-adaptive refinement model (SRM). The SGM generates content within a rasterized shape, whereas the SRM refines content edges and produces a refined shape alpha mask using our shape-adaptive VAE decoder (SVD). In stage one, we use SDM to generate reference images and in stage two, by employing shape-adaptive effect transfer (SAET), we transfer the style of reference images to target images to ensure style consistency between $\widehat{\mathbf{I}}_i$. Prior indicates font effect noise prior used in SAET.

quality training data that aligns text with images encapsulated within an irregularly shaped canvas. The second challenge arises from the default self-attention and cross-attention schemes, which directly map text information across the entire rectangular canvas. This approach inadvertently allows for visual content generation in regions outside the irregularly shaped canvas, which is also rasterized into a rectangle, thereby diminishing the effectiveness of targeted content generation within the desired canvas region. To overcome these challenges, we propose two key contributions: constructing high-quality shape-adaptive image-text data and implementing a shape-adaptive attention scheme. We elaborate more details on these two techniques in the following discussion.

Shape-adaptive Image-Text Data Generation. To construct high-quality, shape-adaptive image-mask-text triplets for training our shape-adaptive generation models, we have chosen BLIP [24] to generate a text prompt set, DALL·E3 as the engine for generating our training images according to prompts and SAM [21] to generate foreground masks. For details on data generation, please see the supplementary. This process has resulted in approximately $80,000$ prompts, with each prompt yielding three unique images and corresponding masks, culminating in a total of $240,000$ high-quality training instances. Examples are shown in Figure 5.

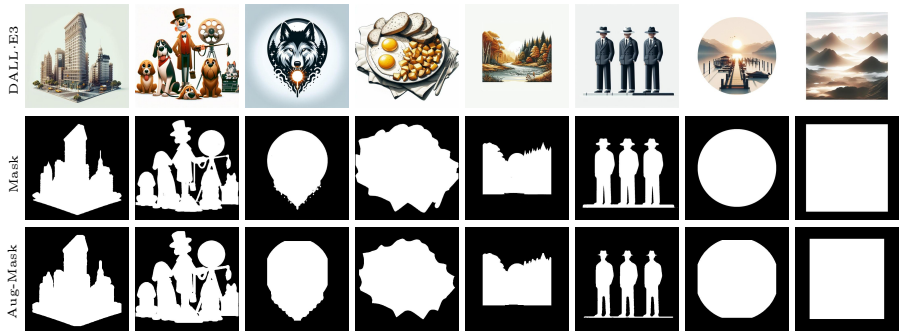

Fig. 5: Illustrating examples of our shape-adaptive images generated with DALL·E3 (first row) for training the shape-adaptive generation model(SGM) and shape-adaptive VAE decoder(SVD). We show the SAM-based segmentation masks (left six columns) and the human-designed canvas masks (right two columns) for training SGM in the second row. The last row displays the augmented masks used as input conditions during SVD training, ensuring that the model learns to refine the augmented masks into the segmentation masks.

Shape-adaptive Attention. We use $\mathbf{\Phi} \in \mathbb{R}^{c_{\text{in}} \times h \times w}$ to represent the image latent features extracted by a VAE encoder before they are sent into the UNet of the diffusion model. We use $\mathbf{\Phi}' \in \mathbb{R}^{n \times c}$ to represent the reshaped and transformed latent features that are sent into the multi-head cross-attention mechanism. By applying different linear projections, we transform $\mathbf{\Phi}'$ into the query embedding space $\mathbf{Q}$, and the text prompt embedding (or pixel embedding) into the key embedding space $\mathbf{K}$ and value embedding space $\mathbf{V}$ for cross-attention (or self-attention). To accommodate our irregularly shaped canvas, we introduce a specialized variant: shape-adaptive attention scheme.

The key insight involves partitioning the entire image's feature maps into two groups: the foreground and the background. We use $\mathbf{M}_A$ to denote the foreground pixels, the subscript fg to label the key and value embeddings associated with the regions inside the irregular canvas, and the subscript bg to label the key and value embeddings associated with the regions outside the irregular canvas. The mathematical formulation is shown as follows:

$$\begin{aligned} &\mathsf{ShapeAdaptive\text{-}MultiHeadAttention}(\mathbf{Q}, \mathbf{K}_{fg}, \mathbf{K}_{bg}, \mathbf{V}_{fg}, \mathbf{V}_{bg}) = \\ &\qquad \mathbf{M}_A \cdot \mathsf{MultiHeadAttention}(\mathbf{Q}, \mathbf{K}_{fg}, \mathbf{V}_{fg}) \\ &\qquad +(1 - \mathbf{M}_A) \cdot \mathsf{MultiHeadAttention}(\mathbf{Q}, \mathbf{K}_{bg}, \mathbf{K}_{bg}), \end{aligned} \tag{5}$$

where we empirically discover that our shape-adaptive attention scheme can effectively minimize content creation outside the irregular canvas.

Shape-adaptive Generation Model Training. Based on the above prepared $240,000$ shape-adaptive image-text pairs generated by DALL·E3 and the proposed shape-adaptive attention scheme, we conduct the training of the shape-adaptive generation model following the controlnet-depth-sdxl-1.0 [25] by replacing the original depth map condition with the generated or hand-crafted canvas

masks. During training, we fix the UNet part of the model and only fine-tune the ControlNet components. We conduct the training on a cluster with 16× A100 GPUs, set the batch size as 256, and maintain a constant learning rate of 1e-6 throughout the training process, which spanned $60,000$ steps.

Shape-adaptive Refinement Model. Shape-adaptive generation model can generate user specified content within a designated area. However, there are a few drawbacks. First, there may still be solid color backgrounds and object shadows that interfere with the generation of the alpha mask (See $\bar{\mathbf{I}}$ in Figure 4). Second, the generated font effects are usually hard-edged which may not be preferred by the user. To further improve the visual appealing of the content within the irregular canvas, suppress the undesired artifacts outside the canvas and offer a flexible control between readability and text-effect strength, we propose to apply an additional shape-adaptive refinement model to refine the object's edges for a more natural appearance and generating a precise post-refinement alpha mask.

Regeneration Strategy of Shape-adaptive Refinement Model. We first crop the output $\bar{\mathbf{I}}$ predicted by the shape-adaptive generation model following the font-shaped canvas mask $\mathbf{M}$, and then paste the segmented font-shaped canvas region onto a rectangle white-board, resulting in $\bar{\mathbf{I}}'$. Next, we extract its latent representation $\mathbf{z}_0{}'$ and add noise to get $\mathbf{z}_t$ for $t < T$. We implement a regeneration strategy to start with $\bar{\mathbf{I}}'$ as the starting image and introducing a small amount of noise, disrupt the high-frequency signals while preserving the low-frequency components. We find the diffusion model struggles to alter low-frequency signals during the denoising process, but concentrates on refining high-frequency elements to smooth out the object's edges.

Our shape-adaptive refinement model supports flexible control of readability and text-effect strength via noise strength. By using a larger noise strength value, the model tends to generate results with stronger text effect and vice versa. In our default setting, we set noise strength of shape-adaptive refinement model to 0.8. This value provides the model with enough flexibility to modify the character boundaries while ensuring the characters remain readable.

Shape-adaptive VAE Decoder (SVD). By applying a decoder to the denoised estimation $\mathbf{z}_0$, we can generate a font effect image with refined edges, which may not confront to the given font-shaped canvas. This necessitates refining the alpha mask to enhance visual quality. To this end, we propose fine-tuning a shape-adaptive VAE decoder capable of predicting an additional refined alpha mask associated with the decoded font effect image. We simply augment the original VAE decoder with an additional input and output channel to facilitate mask conditioning and prediction. During fine-tuning, we apply alpha mask augmentation to the original segmentation masks predicted with SAM [21], as shown in the third row of Figure 5. In summary, the fine-tuned VAE decoder is capable of predicting a refined alpha mask in addition to decoding the image.

3.3 Shape-adaptive Effect Transfer

As we have ensured the creation of high-quality visual content on any given irregular font-shaped canvas, the next critical challenge is ensuring a consistent

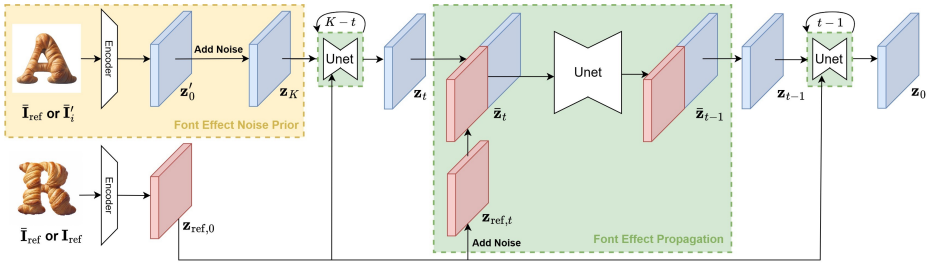

Fig. 6: Illustrating font effect noise prior and font effect propagation within shape-adaptive effect transfer (SAET) scheme. SAET can be applied on both shape-adaptive generation model (SGM) and shape-adaptive refinement model (SRM). When SAET is applied to SGM, we use $\bar{\mathbf{I}}_{\text{ref}}$ for both font effect noise prior and font effect propagation. When SAET is applied to SRM (shown in figure), we use $\bar{\mathbf{I}}'_i$ for font effect noise prior and $\mathbf{I}_{\text{ref}}$ for font effect propagation.

font effect across multiple characters. We propose a shape-adaptive effect transfer (SAET) scheme to transfer the reference font effect from one image to all target letter font images. SAET can be applied to any diffusion-like models. The key idea involves modulating the inputs and outputs of the diffusion model as well as influencing the latent feature of the denoising process, denoted as $\mathbf{z}_t$. In our case, we applied SAET to shape-adaptive diffusion model including both SGM and SRM. Therefore, we can reformulate the original Equation 3 as follows:

$$\mathbf{I}_i, \mathbf{M}_{\mathbf{I}_i} = \bar{h}(\mathbf{M}_i \mid \mathbf{T},\ \mathbf{M}_{\text{ref}},\ \mathbf{I}_{\text{ref}},\ \mathbf{M}_{\mathbf{I}_{\text{ref}}}),\ \mathrm{i} \in \{1, \cdots, \mathrm{n}\}, \tag{6}$$

In the following, we differentiate the style source (reference image) from the style recipient (target image) for clarity.

Framework Overview. The efficacy of shape-adaptive effect transfer scheme is attributed to two pivotal factors: first, it provides the target image with an effect prior based on the reference image; second, it iteratively integrates effect information from the reference image throughout the denoising process, resulting in a target image with a consistent font effect. Figure 4 also illustrates the overall framework of our shape-adaptive effect transfer approach.

Font Effect Noise Prior. Drawing inspiration from SDEdit [26], we devise a font effect noise prior scheme by initializing target font images with partially noised latents derived from the original reference font effect image. This approach enhances the model's ability to generate styles consistently. The overall implementation is depicted in Figure 6.

Font Effect Propagation. We further propose to propagate the font effect information from the reference font image to the target font image like following: at any denoising stage t within a UNet, given the target image's latent $\mathbf{z}_t$ and the reference image's latent $\mathbf{z}_{\text{ref},0}$, we escalate $\mathbf{z}_{\text{ref},0}$ to the same noise level as $\mathbf{z}_t$, yielding $\mathbf{z}_{\text{ref},t}$. We then concatenate $\mathbf{z}_t$ with $\mathbf{z}_{\text{ref},t}$ to obtain $\bar{\mathbf{z}}_t = \text{Concat}(\mathbf{z}_{\text{ref},t}, \mathbf{z}_t)$, which is then processed through UNet for denoising. After deducing the noise component, we selectively utilize the noise pertaining to $\mathbf{z}_t$

Table 1: Ablation results of SGM

Model	M-CLIP-Int ↑	M-CLIP-Ext ↓
SDXL-ControlNet-Canny	26.03	21.52
SDXL-ControlNet-Depth	24.11	23.24
SGM trained w. Est-depth	24.51	18.28
SGM trained w. Cropped Est-depth	24.11	18.22
SGM	**27.26**	**18.11**

Table 2: Shape-Adaptive Effect Transfer vs. StyleAligned

Model	CLIP-I↑	DINO↑
FontStudio w.o. SAET	81.02	54.27
FontStudio w. StyleAligned	82.77	60.79
FontStudio	**84.63**	**67.07**

Table 3: Comparison with Adobe Firefly

Model	CLIP↑	CLIP-I↑
Firefly	28.48	81.74
FontStudio	**29.44**	**84.63**

for denoising $\mathbf{z}_t$ to achieve $\mathbf{z}_{t-1}$, iterating this step until reaching $\mathbf{z}_0$. The effect propagation between the source latents and the target latents mainly happen within the self-attention modules. Figure 6 illustrates the detailed process.

We modify both the shape-adaptive generation model and shape-adaptive refinement model to support processing the concatenated latent representations of a source font effect image and a target font image with font effect prior. We empirically find setting the noise strengths with different values within SGM and SRM achieves the best results. Refer to the supplementary for more details.

Discussion. Our empirical findings suggest that our method is resilient to variations in the reference font shape, yielding consistent results across a wide range of reference font shapes. We have observed that choosing a reference character with a larger foreground area is beneficial. This is interpreted as the larger foreground providing more informative units for the self-attention mechanism, thereby enhancing the generation of new characters. In practice, we often use the letter 'R' from the specified font as our reference for generation due to its typically large foreground area. Refer to supplementary material for more details. Additionally, our approach demonstrates flexibility across different language scripts, having been successfully applied to fonts in Chinese, Japanese, and Korean in our extended experiments.

4 Experiments

4.1 GenerativeFont benchmark

We introduce the GenerativeFont benchmark, which comprises 145 test cases, to enable comprehensive comparisons. These prompts vary in length and are categorized into five themes: Nature, Material, Food, Animal, and Landscape. The character sets extend beyond English, incorporating Chinese, Japanese, and Korean characters, offering a diverse linguistic and cultural representation. This benchmark serves as the foundation for all data analyses and comparative studies conducted in this work. For detailed information on its construction, please refer to the supplementary material.

4.2 Ablation Study on Shape-Adaptive Diffusion Model

To assess the ability of models to accurately generate content within the font canvas area in accordance with provided prompts, we introduced the M-CLIP-Int and M-CLIP-Ext metrics. These metrics make use of an additional mask to direct the evaluation towards the intended areas, both inside and outside the canvas. In the calculations for M-CLIP-Int and M-CLIP-Ext, we mask areas outside the canvas in white and subsequently average these altered CLIP similarity scores across the benchmark.

Comparison between Shape-adaptive Generation Model and Conventional Rectangle-canvas based Diffusion Models. Figure 2 showcases the qualitative results from conventional diffusion models trained for rectangle canvas. SDXL faces challenges in performing the font effect generation task due to missing shape-specific guidance. Conversely, SDXL-Inpaint, while not tailored to fill the entire area with designated content, often produces barely recognizable shapes. Both SDXL-ControlNet-Canny and SDXL-ControlNet-Depth are capable of processing masked inputs; however, their training primarily focuses on matching prompts with the entire rectangle image canvas, inadvertently causing prompt content to appear outside the intended shape area. This misalignment adversely affects their M-CLIP-Ext scores, as detailed in Table 1. Additionally, the lack of targeted control guidance within the shape leads to diminished M-CLIP-Int scores for these models. We also note that it is impractical to apply SDXL-ControlNet-Segmentation to our task. The reason is that ControlNet requires a precomputed segmentation map of finite number of classes, and it is hard to estimate a reasonable segmentation map that fits the irregular font shape while following the complex semantics of user prompts.

Training Objective. We fine-tuned two depth models using our image dataset: the first model employed estimated depth maps, while the second utilized depth maps that were cropped according to the shape mask. Table 1 shows that both models underperformed in all metrics, underscoring the difficulty depth models face in generating content within specified areas, despite being trained with our data. However, our training approach significantly increases the models' flexibility, a key factor in the superior performance of our model.

4.3 Ablation Study on Shape-adaptive Effect Transfer

In this section, we employ the CLIP-I score and DINO score to assess the visual font effect similarity across the generated characters as in [34].

Comparison with Baseline and StyleAligned [16]. Our baseline for comparison involves the shape-adaptive diffusion model without SAET, where each character is generated independently using uniform seed. We also substitute StyleAligned for our SAET to evaluate its performance. The outcomes, illustrated in Table 2, reveal that models utilizing SAET significantly outperform those that do not in terms of both CLIP-I and DINO score. Figure 7 highlights that, despite a fixed generation seed, maintaining style consistency across different shapes proves challenging for models without SAET.

Fig. 7: Qualitative comparison results: FontStudio vs. StyleAligned.

Fig. 8: Qualitative comparison results: FontStudio vs. Adobe Firefly Text Effect.

Fig. 9: Qualitative font-effect results generated with our FontStudio.

4.4 Comparison with State-of-the-Art

Comparison with Adobe Firefly. Figure 8 shows outputs from both frameworks. Firefly's outputs feature high contrast and a consistent style but often include mismatched patterns, reducing character clarity and aesthetic value. In contrast, FontStudio presents outputs with cohesive colors, diverse styles, and clear linework, enhancing letter integration. For shape fidelity, both frameworks maintain character legibility, though Firefly's can appear fragmented with missing strokes due to its aesthetic issues. Stylistically, both are largely consistent, though Firefly occasionally shows minor discrepancies. Regarding prompt fidelity, both generally follow prompt instructions, but Firefly struggles more with style-related prompts. Table 3 delineates the quantitative comparison on GenerativeFont benchmark between our results and those by Firefly, focusing on the CLIP Score and CLIP-I Score, reflective of Prompt Fidelity and Style Consistency, respectively. Our analysis underscores our methodology's superior performance over Firefly across these metrics. Moreover, we provides more visualization results in Figure 9.

User Study and GPT-4V evaluation. We engaged 25 evaluators, including 10 professionals, to assess the benchmark results, and similar assessments were conducted for GPT-4V. Participants rated the outcomes using four metrics to determine which were superior. The findings, displayed in Figure 3, confirm the superiority of our FontStudio over Adobe Firefly in every category. We also have similar results for GPT-4V with a 65% win rate in aesthetics, 76% in shape fidelity and 74% in style consistency. Refer to the supplementary for more details.

5 Conclusion

We introduced FontStudio, an innovative system crafted for generating coherent and consistent visual content specifically designed for font shapes. The system consists of two principal components: a shape-adaptive diffusion model that tackles the challenge of creating content on irregular canvases, and a shape-adaptive effect transfer scheme ensuring uniformity across characters. Furthermore, we present the GenerativeFont benchmark, a tool developed for the quantitative evaluation of our method's efficacy. Our empirical studies demonstrate that FontStudio adeptly responds to user prompts, creating high-quality and aesthetically pleasing font effects. Notably, it surpasses both previous studies and the commercial solution Adobe Firefly in all metrics assessed.

References

1. Alaluf, Y., Garibi, D., Patashnik, O., Averbuch-Elor, H., Cohen-Or, D.: Cross-image attention for zero-shot appearance transfer. arXiv preprint arXiv:2311.03335 (2023)
2. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (TOG) **42**(4), 1–11 (2023)

3. Azadi, S., Fisher, M., Kim, V.G., Wang, Z., Shechtman, E., Darrell, T.: Multi-content gan for few-shot font style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7564–7573 (2018)
4. Balashova, E., Bermano, A.H., Kim, V.G., DiVerdi, S., Hertzmann, A., Funkhouser, T.: Learning a stroke-based representation for fonts. In: Computer Graphics Forum. vol. 38, pp. 429–442. Wiley Online Library (2019)
5. Berio, D., Leymarie, F.F., Asente, P., Echevarria, J.: Strokestyles: Stroke-based segmentation and stylization of fonts. ACM Transactions on Graphics (TOG) **41**(3), 1–21 (2022)
6. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., Ramesh, A.: Improving image generation with better captions (2023), `https://cdn.openai.com/papers/dall-e-3.pdf`
7. Campbell, N.D., Kautz, J.: Learning a manifold of fonts. ACM Transactions on Graphics (ToG) **33**(4), 1–11 (2014)
8. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023)
9. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
10. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)
11. Chen, M., Xu, F., Lu, L.: Manufacturable pattern collage along a boundary. Computational Visual Media **5**, 293–302 (2019)
12. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
13. Epstein, D., Jabri, A., Poole, B., Efros, A., Holynski, A.: Diffusion self-guidance for controllable image generation. Advances in Neural Information Processing Systems **36** (2024)
14. Gao, Y., Guo, Y., Lian, Z., Tang, Y., Xiao, J.: Artistic glyph image synthesis via one-stage few-shot learning. ACM Transactions on Graphics (TOG) **38**(6), 1–12 (2019)
15. He, J.Y., Cheng, Z.Q., Li, C., Sun, J., Xiang, W., Lin, X., Kang, X., Jin, Z., Hu, Y., Luo, B., et al.: Wordart designer: User-driven artistic typography synthesis using large language models. arXiv preprint arXiv:2310.18332 (2023)
16. Hertz, A., Voynov, A., Fruchter, S., Cohen-Or, D.: Style aligned image generation via shared attention. arXiv preprint arXiv:2312.02133 (2023)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
18. Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An edit friendly ddpm noise space: Inversion and manipulations. arXiv preprint arXiv:2304.06140 (2023)
19. Iluz, S., Vinker, Y., Hertz, A., Berio, D., Cohen-Or, D., Shamir, A.: Word-as-image for semantic typography. arXiv preprint arXiv:2303.01818 (2023)
20. Jiang, Y., Lian, Z., Tang, Y., Xiao, J.: Scfont: Structure-guided chinese font generation via deep stacked networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 4015–4022 (2019)

21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
22. Kwan, K.C., Sinn, L.T., Han, C., Wong, T.T., Fu, C.W.: Pyramid of arclength descriptor for generating collage of shapes. ACM Trans. Graph. **35**(6), 229–1 (2016)
23. Lab, D.: Deepfloyd if. `https://github.com/deep-floyd/IF` (2023)
24. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
25. library, D.: Sdxl-controlnet: Depth (2023), `https://huggingface.co/diffusers/controlnet-depth-sdxl-1.0`
26. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
27. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
28. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
29. Park, D.H., Luo, G., Toste, C., Azadi, S., Liu, X., Karalashvili, M., Rohrbach, A., Darrell, T.: Shape-guided diffusion with inside-outside attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4198–4207 (2024)
30. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
31. Patashnik, O., Garibi, D., Azuri, I., Averbuch-Elor, H., Cohen-Or, D.: Localizing object-level shape variations with text-to-image diffusion models supplementary materials
32. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
34. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
35. Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. arXiv preprint arXiv:2307.06949 (2023)
36. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
37. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487 **4**

38. Saputra, R.A., Kaplan, C.S., Asente, P.: Improved deformation-driven element packing with repulsionpak. IEEE transactions on visualization and computer graphics **27**(4), 2396–2408 (2019)
39. Tanveer, M., Wang, Y., Mahdavi-Amiri, A., Zhang, H.: Ds-fusion: Artistic typography via discriminated and stylized diffusion. arXiv preprint arXiv:2303.09604 (2023)
40. Tendulkar, P., Krishna, K., Selvaraju, R.R., Parikh, D.: Trick or treat: Thematic reinforcement for artistic typography. arXiv preprint arXiv:1903.07820 (2019)
41. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
42. Voynov, A., Aberman, K., Cohen-Or, D.: Sketch-guided text-to-image diffusion models. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
43. Wang, C., Wu, L., Liu, X., Li, X., Meng, L., Meng, X.: Anything to glyph: Artistic font synthesis via text-to-image diffusion model. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023)
44. Wang, Y., Lian, Z.: Deepvecfont: Synthesizing high-quality vector fonts via dual-modality learning. ACM Transactions on Graphics (TOG) **40**(6), 1–15 (2021)
45. Xu, J., Kaplan, C.S.: Calligraphic packing. In: Proceedings of graphics interface 2007. pp. 43–50 (2007)
46. Yang, S., Liu, J., Lian, Z., Guo, Z.: Awesome typography: Statistics-based text effects transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7464–7473 (2017)
47. Yang, S., Liu, J., Yang, W., Guo, Z.: Context-aware text-based binary image stylization and synthesis. IEEE Transactions on Image Processing **28**(2), 952–964 (2018)
48. Yang, S., Liu, J., Yang, W., Guo, Z.: Context-aware unsupervised text stylization. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 1688–1696 (2018)
49. Yang, S., Wang, Z., Wang, Z., Xu, N., Liu, J., Guo, Z.: Controllable artistic text style transfer via shape-matching gan. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
50. Zhang, J., Wang, Y., Xiao, W., Luo, Z.: Synthesizing ornamental typefaces. In: Computer Graphics Forum. vol. 36, pp. 64–75. Wiley Online Library (2017)
51. Zhang, J., Yang, Z., Jin, L., Lu, Z., Yu, J.: Creating word paintings jointly considering semantics, attention, and aesthetics. ACM Transactions on Applied Perceptions (TAP) **19**(3), 1–21 (2022)
52. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
53. Zou, C., Cao, J., Ranaweera, W., Alhashim, I., Tan, P., Sheffer, A., Zhang, H.: Legible compact calligrams. ACM Transactions on Graphics (TOG) **35**(4), 1–12 (2016)