DreamFont3D: Personalized Text-to-3D Artistic Font Generation

Xiang Li
Shandong University
Jinan, China
xiangli_@mail.sdu.edu.cn

Lei Meng
Shandong University,
Shandong Research Institute of Industrial Technology
Jinan, China
lmeng@sdu.edu.cn

Lei Wu*
Shandong University
Jinan, China
i_lily@sdu.edu.cn

Manyi Li
Shandong University
Jinan, China
manyili@sdu.edu.cn

Xiangxu Meng
Shandong University
Jinan, China
mxx@sdu.edu.cn

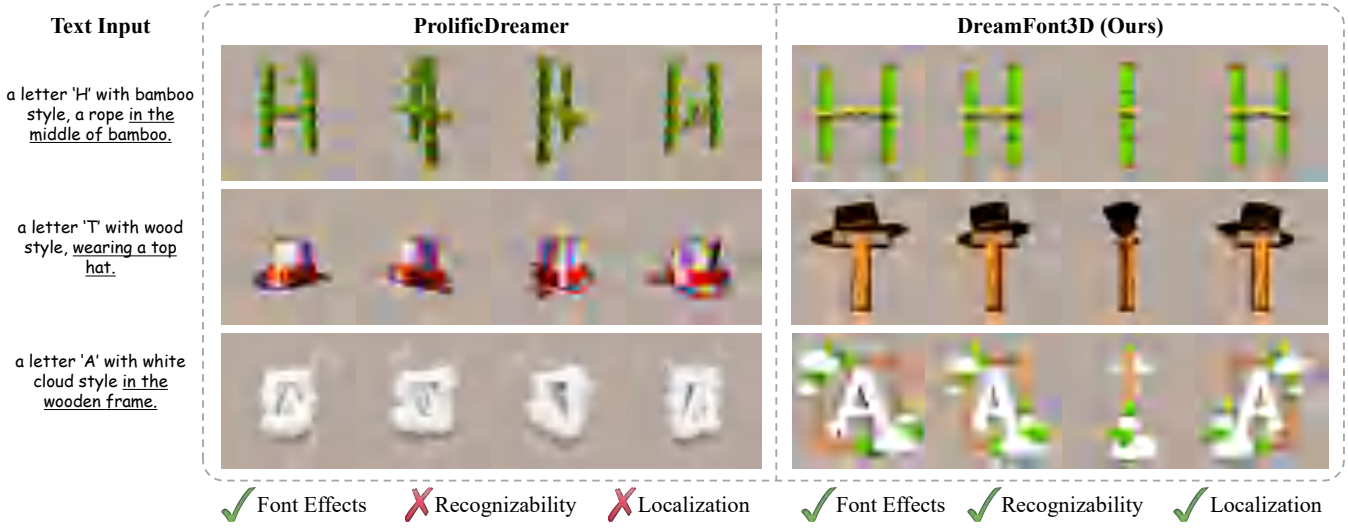


Figure 1: Comparison between our proposed DreamFont3D and ProlificDreamer [Wang et al. 2023b] in the text-to-3D artistic font generation under different text prompts. DreamFont3D can better handle the complex text prompts with multiple objects to generate recognizable fonts, and it performs well when to infer the local position of multiple objects.

ABSTRACT

Text-to-3D artistic font generation aims to assist users for innovative and customized 3D font design by exploring novel concepts and styles. Despite of the advances in the text-to-3D tasks for general objects or scenes, the additional challenge of 3D font generation is to preserve the geometric structures of strokes in an appropriate extent, which determines the generation quality in terms of the recognizability and the local effect control of the 3D fonts. This paper presents a novel approach for text-to-3D artistic font generation, named DreamFont3D, which utilizes multi-view font masks and layout conditions to constrain the 3D font structure and local font effects. Specifically, to enhance the recognizability of 3D fonts, we propose the multi-view mask constraint (MC) to optimize the differentiable 3D representation while preserving the font structure. We also present a progressive mask weighting (MW) module to ensure a trade-off between the text-guided stylization of font effects and the mask-guided preservation of font structure. For precise control over local font effects, we design the multi-view attention modulation (AM) that guides the visual concepts to appear in specific regions according to the provided layout conditions. Compared with existing text-to-3D methods, DreamFont3D shows its own superiority in the consistency between font effects and text prompts, the recognizability, and the localization of font effects. *Code and data at https://moonlight03.github.io/DreamFont3D/.*

*Corresponding author

SIGGRAPH Conference Papers '24, July 27–August 01, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07
https://doi.org/10.1145/3641519.3657476

CCS CONCEPTS

• **Computing methodologies** → **Shape modeling**.

KEYWORDS

text-to-3d generation, font generation, diffusion model, neural radiance fields

ACM Reference Format:

Xiang Li, Lei Meng, Lei Wu, Manyi Li, and Xiangxu Meng. 2024. DreamFont3D: Personalized Text-to-3D Artistic Font Generation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27–August 01, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3641519.3657476

1 INTRODUCTION

Font generation has achieved remarkable success, producing stunning results [Jiang et al. 2017; Wang et al. 2023f, 2020]. However, prior studies are limited to operations within the raster domain and applications in flat design, not extendable to 3D environments such as 3D animations or virtual reality. Recently, text-to-3D technology achieves much attention [Jain et al. 2022; Mohammad Khalid et al. 2022; Wang et al. 2022], which allows the 3D content generation under the guidance of natural language. However, comparing to the generation of 3D objects, the creation of 3D artistic fonts is essentially challenging as it requires not only generating the target font effect described by the text prompt, but also controlling the overall font structure and the position of local font effects. For instance, in the first row of Figure 1, the results need the effects of bamboo and rope, a recognizable letter 'H', as well as a rope in the middle of bamboo. Although existing methods may generate the 3D fonts with reasonable font effects matching with text prompts [Chen et al. 2023a; Lin et al. 2023], their performance is limited in the font recognizability and the localization of font effects. Therefore, the text-to-3D artistic font generation with precise font structure and positional control of font effects remains a challenge.

To achieve the generation of text to arbitrary 3D objects, some researchers propose leveraging 2D priors from pre-trained text-to-image diffusion models [Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022] to guide 3D generation, as demonstrated by a representative work DreamFusion [Poole et al. 2023]. Most existing text-to-3D research follows a unified 2D prior optimization paradigm: they utilize the text-to-image diffusion models as supervision for various views of 3D content and optimize the 3D representation (such as Neural Radiance Fields [Mildenhall et al. 2020]) through score distillation sampling [Poole et al. 2023]. Follow-up works, exemplified by Magic3D [Lin et al. 2023] and ProlificDreamer [Wang et al. 2023b], involve the implementation of multi-stage optimization strategies, optimizing the diffusion prior with the 3D representation simultaneously and the proposal of more effective score distillation algorithms to improve the resolution and details of the generated results. However, since the strength of text-to-image diffusion models is to understand text and image content, they are not adept at accurately converting arbitrary font structure into image space and synthesizing spatially controllable font effects. Therefore, the previous text-to-3D methods [Jain et al. 2022; Wang et al. 2022, 2023a] following the above 2D prior optimization often cause unrecognizable font and unstable local effects.

In this paper, we propose DreamFont3D, a novel text-to-3D artistic font approach designed to improve font recognizability and localization of font effects. The essence of our model is the employment of font masks and layout as input control conditions, which guide the 3D shape and local font effects from multiple perspectives. Specifically, DreamFont3D leverages a pre-trained diffusion model as a 2D prior to refine the parameters of a NeRF-based 3D volume, facilitating the generation of text-specified font effects. To preserve the target font structure, we propose multi-view mask constraint to prevent over-deformation of the 3D fonts. This is facilitated by the utilization of multi-view font masks that are readily accessible and offer structure information into the 3D fonts in different viewpoints. We additionally present a progressive mask weighting module to achieve a trade-off between text-guided stylization of font effects and mask-guided preservation of font structure. To achieve localization of font effects, we introduce layout conditions, e.g., freehand font image with several colors, and propose multi-view attention modulation. Since the layout of the generated image is related to the attention maps [Hertz et al. 2023; Kim et al. 2023], we dynamically modulate the multi-view attention maps in the pre-trained diffusion model according to the layout conditions to guide objects to appear in specific regions. Due to the text-to-3D artistic font generation is a new task, we collect a new evaluation dataset to validate our method. Extensive experiments have proven the superiority of DreamFont3D in text-to-3D artistic font generation task.

Our contributions are as follows:

- To the best of our knowledge, this paper presents a novel approach designed for the text-to-3D artistic font generation. This approach not only ensures text-consistent font effects and higher font recognizability but also provides positional control over the font effects.
- We propose the multi-view mask constraint to improve the recognizability of generated 3D artistic font, and a progressive mask weighting module to achieve a trade-off between text-guided stylization and font structure.
- We propose the multi-view attention modulation, which combine attention modulation with text-to-3D techniques to achieve the localization of font effects. It is the first attempt to control spatial position when generating creative 3D content.

2 RELATED WORK

2.1 Artistic Font Image Generation

Artistic font image generation [Gao et al. 2019; Li et al. 2022b, 2023; Wang et al. 2023d] can be defined as transferring the style of reference images to another artistic font image or a binary font source image [Yang et al. 2019a]. The overall style of the generated result, e.g., texture, conceptual style, etc., remains consistent with the reference images, while the font structure remains consistent with the original image. The previous works have adopted methods based on patch-based texture synthesis [Yang et al. 2022, 2019b] or Generative Adversarial Networks [Li et al. 2020; Yang et al. 2019a] to ensure that the generated style aligns with the correct style distribution. Recently, Anything2Glyph [Wang et al. 2023e] leverages the advantage of Stable Diffusion in generating images of any object, and uses masks to constrain the approximate skeleton-level positioning of objects within images. However, these methods have been confined to the 2D space, making their results inapplicable to 3D scenes. In contrast, our proposed model is capable of generating

realistic 3D artistic fonts. It allows for the creation of artistic fonts with any font effect through text prompts, opening new frontiers in 3D artistic font design.

2.2 Diffusion Methods with Spatial Control

Due to the advantages of diffusion models [Ma et al. 2024; Song et al. 2020] over GANs [Dong et al. 2022; Karras et al. 2021] in terms of generating higher image quality and better training stability, text-to-image diffusion models [Chefer et al. 2023; Gal et al. 2023; Tewel et al. 2023] have achieved tremendous success in the synthesis of diverse, photorealistic images. However, since text-to-image models are typically trained on datasets with short text captions, they often struggle to capture all the details in dense text prompts composed of several phrases, especially when there are complex relative positional relationships [Kim et al. 2023]. A series of recent works [Epstein et al. 2023; He et al. 2023b; Rombach et al. 2022; Voynov et al. 2023] have addressed this issue by introducing spatial layout conditions and attention mechanism [Guo et al. 2020; Wang et al. 2023c]. For example, Composable Diffusion [Liu et al. 2022] and MultiDiffusion [Bar-Tal et al. 2023] execute a separate denoising process for each phrase at every timestep. DenseDiffusion [Kim et al. 2023] introduces the attention modulation and semantic segmentation map as layout conditions to guide objects to appear in specific areas without the need for additional training. Despite significant advancements in spatial position control within text-to-image generation, the spatial position control research in text-to-3D generation remains notably underexplored.

2.3 Text-to-3D Generation

Due to the early 3D generation works [Chang et al. 2015; Schwarz et al. 2020; Sun et al. 2022] rely on large-scale 3D data, researchers have begun to explore the potential of leveraging powerful text-to-image diffusion models for 3D content creation. A pioneering work, DreamFusion [Poole et al. 2023], utilize the text-to-image diffusion model acts as a critic, ensuring that the differentiable 3D representation match the distribution of real images when rendered from any viewpoints. However, DreamFusion often produces unrealistic and rough results. Following this, a series of works [Metzer et al. 2023; Raj et al. 2023; Xu et al. 2023b] have significantly advanced in producing more photorealistic and text-consistency 3D content. Recently, Control3D [Chen et al. 2023b] incorporates ControlNet [Zhang et al. 2023] and leverages its sketch-to-image capabilities to promote geometric controllability. MVDream [Shi et al. 2023] utilizes a multi-view diffusion model as 2D prior to achieve more consistent 3D generation across multi-views. However, these methods exhibit deficiencies in maintaining the structural integrity of 3D font structures and in exerting precise control over local effects. This results in inaccuracies in font structures and instability in the positioning of font effects.

3 METHOD

3.1 Overview

As shown in Figure 2, given a freehand font image and a descriptive text prompt as the conditions, our goal is to automatically produce the 3D font satisfying both the target font structure and the specified font effect by utilizing the prior knowledge of the pre-trained text-to-image diffusion models. Specifically, to obtain the 3D font effect consistent with the text prompt, we employ NeRF (Neural Radiance Fields [Mildenhall et al. 2020]) as the 3D representation, and adopt the score distillation sampling technique to optimize the parameters of 3D representation based on the frozen text-to-image model. Then, to preserve the font structure during the optimization, we propose the multi-view mask constraint to prevent the over-deformation of the 3D shape. At the same time, we utilize a progressive mask weighting module to achieve a trade-off between the text-guided stylization of the font effect and the mask-guided preservation of the font structure. Finally, to enhance the control ability for creative design, we propose the multi-view attention modulation which allows to specify different local font effects for different regions of the font by manipulating the intermediate attention maps.

3.2 Score Distillation Sampling

To generate font effects consistent with text prompts, we introduce the score distillation sampling (SDS [Poole et al. 2023]) technique. It iteratively optimize the parameters θ of a differentiable 3D representation utilizing a text-to-image diffusion model ϕ as the 2D prior. At each iteration, it selects a random camera pose and uses the volumetric renderer g to generate the image $x = g(\theta)$. x is injected with random Gaussian noise ϵ and then fed into the text-to-image model, which is used to predict the noise to generate the image corresponding to the text prompt. Subtracting the injected noise ϵ from the predicted noise $\hat{\epsilon}_\phi(z_t \mid y, t)$ produces the gradient that is used to update the parameters θ of NeRF. Therefore, the gradient of the SDS loss with respect to 3D representation θ is:

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, x) = \mathbb{E}_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(z_t \mid y, t) - \epsilon)\frac{\partial x}{\partial \theta}], \tag{1}$$

z_t refers to the input of the diffusion model. $w(t)$ is a weighting function depending on the timestep t. By doing so, the SDS loss enables the rendered images to align with the text prompts, thereby ensuring that the font effects can match the textual descriptions.

3.3 Multi-View Mask Constraint

Due to the diversity of font structures, existing text-to-image diffusion models struggle to predict the structure of arbitrary characters based on textual descriptions [Wang et al. 2023e]. Even with additional text prompts to describe perspective information, as DreamFusion [Poole et al. 2023] did, generating recognizable 3D artistic fonts still remains a challenge.

Our key idea is to compute multi-view binary masks as constraints to preserve the font structure. First, compared to the detailed text prompts or multi-view RGB images, the binary font masks are easily accessible and can describe the font structure precisely. Second, it is easy to apply perspective transformation on the mask image to obtain multi-view constraints for 3D generation, without the need to formulate the constraints in 3D format.

To implement the multi-view mask constraint, we first uniformly sample a series of azimuth P, and perform perspective transformation on the font mask to obtain multi-view font masks M_f. Then we add two rectangular-shaped masks for the cases azimuth 90° and 270°, where the length is determined by the projection of the input font mask in the vertical direction, and width customized.

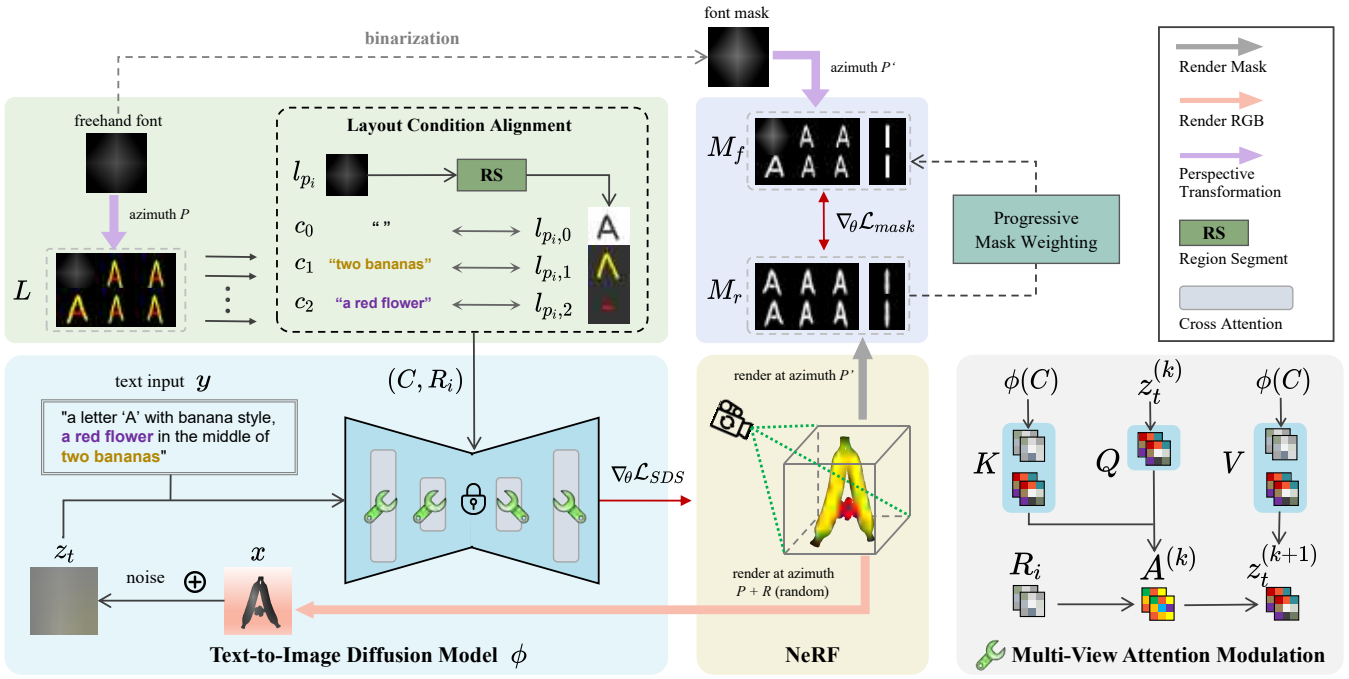


Figure 2: Overview of the DreamFont3D. We use a frozen text-to-image diffusion model to optimize font effects of 3D font with $\mathcal{L}_{SDS}$. The input freehand font is used to create multi-view layouts $L = \{l_{p_i}\}_{i=1}^{X}$ by perspective transformation, and each l_{p_i} is segmented into multi-view layout regions $R_i = \{l_{p_i,j}\}_{j=0}^{Y-1}$. The transformation rules follow a set of uniformly sampled azimuth $P = \{p_i\}_{i=1}^{X}$, $p_i \neq 90° \wedge p_i \neq 270°$. The text phrases $C = \{c_j\}_{j=0}^{Y-1}$ and R_i align according to the annotations, and are packaged as (C, R_i) input into the cross attention layer for multi-view attention modulation to control the local font effects (k is the index of layer). In another branch, the input freehand font is binarized into a font mask. Then, we obtain multi-view font mask $M_f = \{m_{f_i}\}_{i=1}^{X+2}$ through perspective transformation, $P' = P \cup \{90°, 270°\}$. The M_f is used together with rendered mask M_r to calculate $\mathcal{L}_{mask}$ to constraint the font structure.

The purpose of the above operation is to better approximate the thickness of a z-dimensional 3D font.

Next, we need to obtain the mask rendered by NeRF to calculate the mask loss and constrain the geometric structure of the 3D font. Specifically, we first define a set of rays for rendering the mask, and the directions of these rays are calculated based on the azimuth P'. The volume density accumulation along a ray r for mask rendering can be formulated as follows:

$$m_r = 1 - \exp\left(-\int_{t_n}^{t_f} \sigma(r(t))\, dt\right), \tag{2}$$

where m_r represents the mask value along the ray r. $\sigma(r(t))$ is the volume density at position $r(t)$ on the ray path. t_n and t_f are the starting and ending range of the integration respectively, corresponding to the positions where the light enters and leaves the scene. The process of rendering a mask can be explained as calculating the cumulative volume density along the path of the ray and converting it into a mask value that reflects the probability of encountering an opaque object along that path. As a result, we obtain the rendered masks M_r, and the loss function that uses multi-view font masks to constrain the structure of 3D font is defined as:

$$\mathcal{L}_{mask} = \|M_f - M_r\|_2^2. \tag{3}$$

3.4 Trade-off of Font Structure and Stylization

The appearance of artistic fonts is determined by a combination of the font structure and font effects, e.g., texture, conceptual style, etc. During our optimization of the NeRF representation, the multi-view mask constraint encourages the 3D font to strictly align with the font structure of masks, while eliminating the necessary deformation to obtain the reasonable font effect described by the text prompt. On the other hand, the SDS loss focuses on the font effect, but is not aware of the font structure and often cause the loss of the font structure. Additionally, the persistent imposition of mask constraints can cause the loss of detailed font effects and the emergence of the foggy phenomenon.

We propose the progressive mask weighting to achieve a flexible trade-off between the text-guided stylization of font effect and the mask-guided preservation of the font structure. Specifically, we select the NeRF's checkpoint at h-th iteration, and render the masks of 3D representation. The rendered masks $M_{r,h}$ is slightly deformed from the original multi-view font masks, but remains the key features of the font structure. We perform sharpening (Laplacian filter) and anti-aliasing (Gaussian blur) on $M_{r,h}$ to improve clarity and

edge smoothness. Then, the $M_{r,h}$ and M_f are used together to further optimize the font structure:

$$\mathcal{L}_{mask} = \alpha \left\| M_{r,h} - M_r \right\|_2^2 + (1-\alpha) \left\| M_f - M_r \right\|_{2.}^2 \tag{4}$$

where α is a linearly increasing weight parameter as the iteration progresses ($0 < \alpha <= 1$). In summary, the progressive mask weighting is a technique that enhances model's compatibility between text-guided stylization of font effect and mask-guided preservation of font structure by weighting the multi-view font masks and the rendered masks.

3.5 Localization of Font Effects

The localization of font effects refers to specifying unique font effects for different local spatial regions. However, it is difficult for the text-to-image models to infer the layouts and generate the correct composition of the multiple objects, not to mention the multi-view consistency problem in the text-to-3D setting. As a consequence, the existing text-to-3D methods often predict inconsistent positional relationships or redundant font effects.

DenseDiffusion [Kim et al. 2023] achieves layout control based on attention modulation in text-to-image generation, and states a significant correlation between the generated image layout and attention maps. The attention modulation can dynamically adjust the attention maps using layout and text phrases as input, ultimately influencing the predicted noise. In this paper, we propose the multi-view attention modulation to guide the prediction of noise and control the layout of 3D font. Unlike [Kim et al. 2023], we use the diffusion model to separately predict layout-related noise at multiple views; SDS iteratively updates the 3D representation using the gradient information computed from the noise.

Layout condition alignment. First, we use different regions of the freehand font to represent different font effects. Then, we perform a perspective transformation on the freehand font to obtain the multi-view font layout L, where each element l_{p_i} needs to undergo region segmentation to obtain layout regions $R_i = \{l_{p_{i,j}}\}_{j=0}^{Y-1}$, i and j are used to specify views and regions, respectively. We align each layout region $l_{p_{i,j}}$ with the text phrase c_j according to annotations and pack them as (C, R_i). Then, we input (C, R_i) into the cross attention layer. Note that we leave c_0 as an empty string, corresponding to the white region of $l_{p_i,0}$ to represent the background region.

Multi-view attention modulation. The paired (C, R_i) are the input conditions of attention modulation by providing spatial layout of font effects. Specifically, we dynamically adjust the attention maps based on each paired (C, R_i) to obtain a higher value, so that the object described by c_j can be generated in the corresponding region $l_{p_{i,j}}$. The key is to calculate the value of A, which is used to indicate the degree of modulation, and its calculation takes into account the range of the original value and the area size of the region:

$$A = R_i \odot A_{max} \odot S - (1 - R_i) \odot A_{min} \odot S, \tag{5}$$

where R_i indicates whether to increase or decrease the attention score for a particular region. The matrix S represents the segment area size, which is used to automatically adjust the modulation degree based on the area size of each region. In order to maintain the original generative ability of the pre-trained model, we also introduce the matrix $A_{max}, A_{min} \in \mathbb{R}^{|queries| \times |keys|}$, which identify each query's maximum and minimum values, ensuring the modulated values stay close to the original range:

$$A_{max} = \max(QK^\top) - QK^\top, \tag{6}$$

$$A_{min} = QK^\top - \min(QK^\top). \tag{7}$$

Ultimately, we modulate the attention maps as below,

$$z_t^{(k+1)} = A^{(k)} V = softmax\left(\frac{QK^\top + A}{\sqrt{d}}\right) V, \tag{8}$$

where $Q = \{z_t^{(k)}\}$, $K = V = \{\phi(C), z_t^{(k)}\}$, k represents the k-th attention layer and $\phi(C)$ represents the encoded word embeddings. The text phrase C and text prompt y use the same text encoder.

3.6 Overall Optimization

Our method requires the alternating execution of three optimization strategies: the SDS at random azimuth R to optimize font effects, the multi-view mask constraint at azimuth P' to optimize font structure, and the layout-guided SDS at azimuth P to control local text effects. In summary, the overall objective function can be formulated as:

$$\theta^* = arg\min_{\theta} \mathcal{L}(\phi, x = g(\theta), L, M_f), \tag{9}$$

where $\mathcal{L}$ represents the two loss functions of the optimization process. Since the optimization process is performed alternately, $\mathcal{L}_{SDS}$ and $\mathcal{L}_{mask}$ are used alternately to update NeRF parameters during training.

4 EXPERIMENTS

Figure 3 shows the 3D artistic fonts generated by our method both with and without the use of layout conditions. Next, we will introduce the implementation details and experimental results.

4.1 Implementation Details

Data collection. We collect a new evaluation dataset comprising freehand font images, font masks, and text prompts. As shown in Figure 5, we craft 50 freehand font images, featuring English, Chinese, numerals, and special symbols, each has 3-4 distinct color regions and text phrase annotations. GPT-4 [OpenAI 2023] provides support for writing text prompts. Font masks can be obtained from the grayscale of freehand fonts and batch generation. We obtain 260 font masks through batch generation using font library files. Font library files can be downloaded from Foundertype website.

Parameter setting. Our method employs 64×64 rgb views as supervision conditions for the 2D prior, 64×64 font mask and freehand font image. The azimuth $P' = \{(i-1) \times 45°\}_{i=1}^{8}$. We render the result views at 800×800 resolution for display. For score distillation sampling, we leverage random sampling of the camera radius and field-of-view angles, aligning with stable-dreamfusion [Tang 2022]. The implementation of our baseline also use stable-dreamfusion [Tang 2022]. The weight coefficient of the $\mathcal{L}_{SDS}$ and $\mathcal{L}_{mask}$, λ_{sds}=1 and λ_{mask}=500. All experiments are conducted on a single NVIDIA 3090 GPU. We train our model for 10,000 iterations and the whole optimization process takes up about 30 minutes, 18G GPU memory for each 3D artistic font. In order to preserve the details of text-guided font effects, SDS is applied throughout the entire optimization process. Multi-view mask constraint is executed from

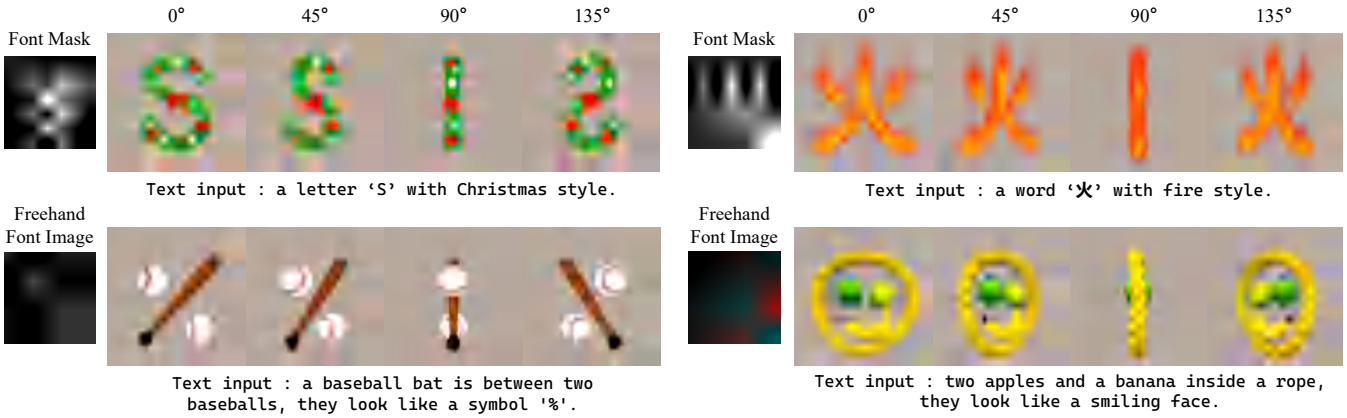


Figure 3: The Results of the Text-to-3D Artistic Font Generation with and without Layout Conditions. The DreamFont3D can generate 3D forms of English, Chinese, special symbols and emoticons. Our results faithfully adhere to both textual descriptions and required font structure.

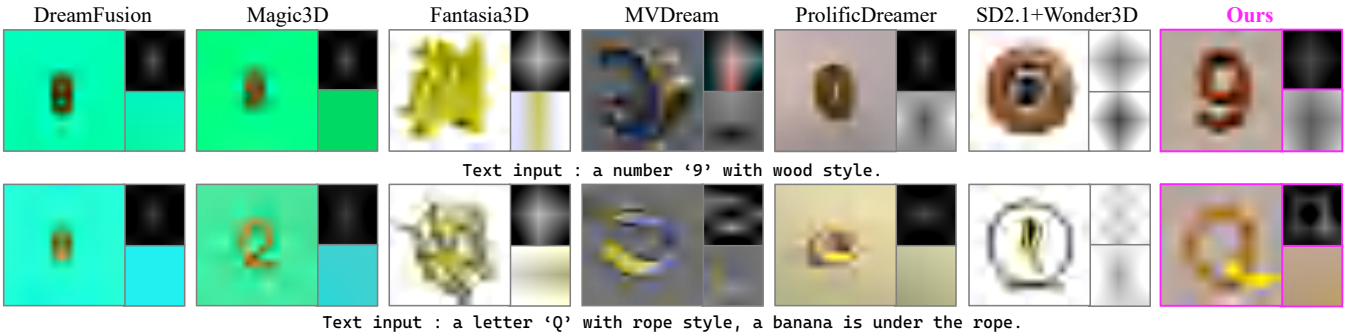


Figure 4: Qualitative Performance Comparison. We use the same text prompts to compare existing text-to-3D models and a text-to-image and image-to-3D baseline, and show the results in front view, side view (azimuth = -45°) and normal. These results indicate that DreamFont3D has achieved better performance in terms of the consistency between font effects and text prompts, font recognizability, and localization of font effects.

(a) Freehand font images (b) Batch-generated font masks

Figure 5: Examples of the Freehand Font Images and Font Masks.

0 to 9,000 iterations, and performing progressive mask weighting at h=4,000. Multi-view attention modulation is executed from 0 to 7,000 iterations.

4.2 Evaluation Metrics

We use CLIP [Radford et al. 2021], quality assessment and alignment assessment for quantitatively comparison, where the last two metrics come from T^3 Bench [He et al. 2023a]. CLIP loss quantifies the correlation between a single view of the generated 3D artistic font and the input text, leveraging the CLIP encoder. The quality assessment combines multi-view text-image scores (ImageReward [Xu et al. 2023a]) and regional convolution to detect consistency of text prompts and multi-views. The alignment assessment uses multi-view captioning (BLIP [Li et al. 2022a]) and Large Language Model (GPT4 [OpenAI 2023]) to measure whether the text in the captioning is consistent with the text input (from 1-5 score).

Table 1: Quantitative Comparisons. We use CLIP, quality assessment, and alignment assessment to evaluate quantitative performance.

Methods	CLIP ↑	Quality ↑	Alignment↑
DreamFusion	0.3903	18.25	2.45
Magic3D	0.4319	25.69	2.60
Fantasia3D	0.4321	19.55	1.70
MVDream	0.4071	28.81	2.95
ProlificDreamer	0.4324	31.04	2.80
SD2.1+Wonder3D	0.5794	40.05	2.70
Ours	**0.7714**	**47.71**	**4.10**

Table 2: User Study. A total of 30 volunteers evaluated the 10 groups of results of 7 methods across 3 metrics. The best results are shown in bold.

Method	Font Effect ↑	Recognizability↑	Localization↑
DreamFusion	3.45	2.38	2.52
Magic3D	3.72	2.57	2.66
Fantasia3D	2.91	1.76	1.89
MVDream	3.37	2.32	2.28
ProlificDreamer	3.65	2.19	2.62
SD2.1+Wonder3D	4.07	2.64	2.92
Ours	**4.56**	**4.57**	**4.60**

4.3 Performance Comparison

We compare five text-to-3D methods, using the threestudio library [Yuanchen et al. 2023] for code implementation. DreamFusion [Poole et al. 2023], Magic3D [Lin et al. 2023], and our model use DeepFloyd [DeepFloyd 2022] as diffusion model. However, the first two are inconsistent with the original papers as they use private diffusion models. ProlificDreamer[Wang et al. 2023b], MVDream [Shi et al. 2023], Fantasia3D [Chen et al. 2023a] use SD2.1 [Rombach et al. 2022] as the diffusion model, which is consistent with the original papers. All text-to-3D methods, except Fantasia3D (uses DMTET [Shen et al. 2021]), adopt Instant-NGP [Müller et al. 2022] as NeRF backbone. In addition, we construct a two-stage baseline by combining the SD2.1 and an image-to-3D model (Wonder3D [Long et al. 2024]) for comparison.

Quantitative comparison. Table 1 reports the average values of twenty sets of generated results. Since existing methods fail to generate views of precise font structure and infer the suitable layout based on the textual description, the CLIP, quality assessment scores, and alignment assessment scores for these methods are relatively low. In contrast, DreamFont3D focuses on font structure optimization and the control of local font effects, thus outperforming existing methods in the three mentioned metrics.

Qualitative comparison. Figure 4 illustrates that the current text-to-3D model is capable of producing some font effects that align with the text prompts, including wood textures and yellow cylinders resembling bananas. However, they generate unrecognizable 3D font and cannot satisfy the positional relationships specified in text prompts. In contrast to existing methods, DreamFont3D generates results that not only exhibit accurate font effects and better recognizability, but also successfully achieve the localization of font effects, as exemplified by phrases like 'a banana is under the rope'.

User study. Although the mentioned quantitative metrics can evaluate the consistency between generated views of 3D representation and text prompts, the text-to-3D artistic font generation is an open-ended task. To more comprehensively evaluate the font effect, recognizability, and localization of font effects, we conducted a user study. A total of 30 volunteers participated in this evaluation, which involve 15 graduate students in related fields and 15 undergraduate students without related research foundation. Specifically, we anonymously show the 10 groups of text prompts and four views of generated results from DreamFont3D and other methods. Users are asked to rate the displayed views. The scoring scale will be based on confidence (1-5 scores) for the following three questions. (1) Font effect: How confident are you that the results have the font effects described in the text prompts? (2) Recognizability: How confident are you that the results is a recognizable character? (3) Localization: How confident are you that the positional relationship of the font effects is consistent with the text prompts?

The user feedbacks are summarized in Table 2, where a significant gap exists between ours and the others. Although the other text-to-3D models can generate corresponding font effects based on text input, however, they obtain lower scores in terms of font recognizability and localization of font effects. In contrast, the generated results of our method are appreciated in terms of all three metrics.

4.4 Ablation Study

In this subsection, we verified the effectiveness of the proposed modules, respectively.

Multi-view mask constraint (MC). As shown in Figure 6, the baseline's generated results exhibit a distinct cluster of bananas. This is because the optimization objective of the original SDS was to make the 3D representation appear as the letter 'A' from any angle, which led to the generation of unrecognizable 3D fonts. After applying MC, the recognizability of 3D fonts has been improved.

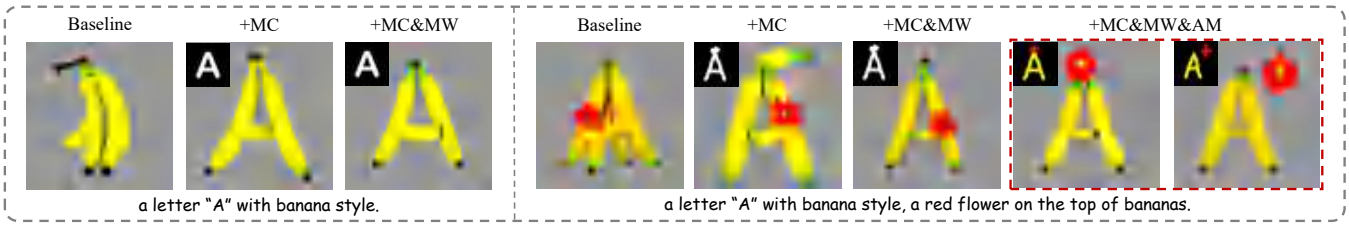
We also verified the impact of the number of masks. In Figure 7, while the views become denser, the stylization effect decreases, e.g., some bamboo leaves and branches disappear, and the font structure becomes neater and more regular. These results indicate that the multi-view mask constraint play a key role in the optimization of font structure, and more masks increase the strength of constraint on font structure. To balance stylization, font structure, and optimization efficiency, +8×MC is an empirical choice.

Progressive mask weighting (MW). The text-guided stylization of the font effect inevitably leads to over-deformation of fonts due to the natural geometric properties of certain objects, such as the curvature of bananas as seen +MC in Figure 6. Although the MC enhances the font recognizability, its similarity to the provided font mask decreases. To preserve the font structure within the mask, we apply MW to achieve a trade-off between the text-guided stylization of the font effect and the mask-guided preservation of the font structure, see +MC&MW in Figure 6.

Multi-view attention modulation (AM). As shown in Figure 6(b), controlling the local effects of 3D artistic fonts solely based on positional descriptions in the text prompts is challenging and often leads to incorrect spatial relationships and redundant attributes. We apply two different freehand font as layout conditions and AM to achieve the precise localization of font effects. These results prove that AM can effectively control the local font effects according to layout conditions.

5 CONCLUSIONS AND LIMITATIONS

We introduce a novel approach to constrain the geometry and control the local effects of 3D content using multi-view masks and layouts. This technique is employed in the creation of 3D artistic



(a) Unified font effect generation (b) Locally controlled font effects generation

Figure 6: Ablation Study. (a) Unified font effect generation. We validate the effectiveness of multi-view mask constraints (MC) and progressive mask weighting (MW) without using layout conditions. (b) Locally controlled font effects generation. We also validate the effectiveness of the above two modules and multi-view attention modulation (AM) when using layout conditions.

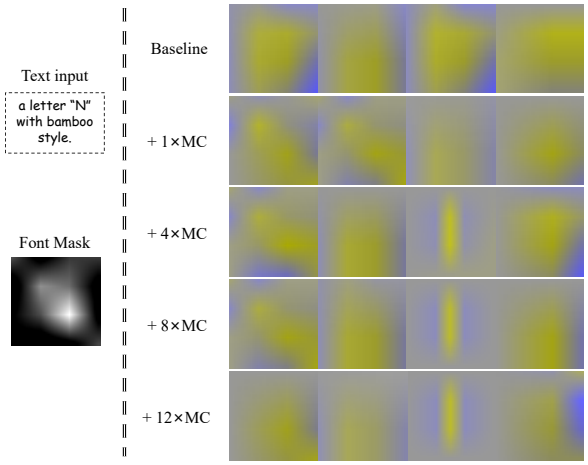


Figure 7: The Impact of the Number of Masks on Font Structure. +1×MC means that only the front view mask is used. +4×MC, +8×MC, and +12×MC respectively perform mask constraints every 90°, 45°, and 30°.

fonts, providing a convenient tool for both novice and professional designers to craft personalized 3D artistic fonts. The effectiveness and superiority of our proposed method are demonstrated through comparative experiments and a user study.

Limitations and future work. Our method also has certain limitations, see Figure 12. Due to the stereotypical representations of object geometry by pre-trained diffusion models, e.g., bamboo being straight and bananas being curved, preserving the structural integrity of complex font structure is a challenge. This can be alleviated by choosing suitable font effects such as 'rope.' Secondly, our approach primarily emphasizes the font structure and localization of font effects in 3D artistic fonts, rather than achieving higher resolution and photorealistic details. Future enhancements could incorporate multi-stage optimization and high-resolution latent diffusion models to enhance both the quality and level of detail.

ACKNOWLEDGMENTS

This work is supported in part by the National Key R&D Program of China (Grant no. 2021YFC3300203), the TaiShan Scholars Program (Grant no. tsqn202211289), the Shandong Province Excellent Young Scientists Fund Program (Overseas) (Grant no. 2022HWYQ-048), and the Oversea Innovation Team Project of the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073).

REFERENCES

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. Article 74, 16 pages.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

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 Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023a. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yang Chen, Yingwei Pan, Yehao Li, Ting Yao, and Tao Mei. 2023b. Control3d: Towards controllable text-to-3d generation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1148–1156.

DeepFloyd. 2022. DeepFloyd IF. https://github.com/deep-floyd/IF.

Pei Dong, Lei Wu, Lei Meng, and Xiangxu Meng. 2022. Hr-prgan: High-resolution story visualization with progressive generative adversarial networks. *Information Sciences* 614 (2022), 548–562.

Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. 2023. Diffusion Self-Guidance for Controllable Image Generation. (2023).

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–13.

Yue Gao, Yuan Guo, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. 2019. Artistic glyph image synthesis via one-stage few-shot learning. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–12.

Wenya Guo, Ying Zhang, Xiangrui Cai, Lei Meng, Jufeng Yang, and Xiaojie Yuan. 2020. LD-MAN: Layout-driven multimodal attention network for online news sentiment recognition. *IEEE Transactions on Multimedia* 23 (2020), 1785–1798.

Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. 2023a. T^3Bench: Benchmarking Current Progress in Text-to-3D Generation. arXiv:2310.02977 [cs.CV]

Yutong He, Ruslan Salakhutdinov, and J. Zico Kolter. 2023b. Localized Text-to-Image Generation for Free via Cross Attention Control. arXiv:2306.14636 [cs.CV]

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 *The Eleventh International Conference on Learning Representations*.

Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 867–876.

Yue Jiang, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. 2017. DCFont: an end-to-end deep chinese font generation system. *SIGGRAPH Asia 2017 Technical Briefs* (2017).

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in neural information processing systems* 34 (2021), 852–863.

Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. 2023. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7701–7711.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Wei Li, Yongxing He, Yanwei Qi, Zejian Li, and Yongchuan Tang. 2020. FET-GAN: Font and effect transfer via k-shot adaptive instance normalization. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 1717–1724.

Xiang Li, Lei Wu, Xu Chen, Lei Meng, and Xiangxu Meng. 2022b. Dse-net: Artistic font image synthesis via disentangled style encoding. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

Xiang Li, Lei Wu, Changshuo Wang, Lei Meng, and Xiangxu Meng. 2023. Compositional zero-shot artistic font synthesis. In *Proceedings of IJCAI*.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 300–309.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*. Springer, 423–439.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2024. Wonder3D: Single Image to 3D using Cross-Domain Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Zhanhui Kang. 2024. Plug-in Diffusion Model for Sequential Recommendation. *arXiv preprint arXiv:2401.02913* (2024).

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2023. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12663–12673.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. 2022. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*. 1–8.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4 (2022), 102:1–102:15.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2209.14988* (2023).

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Ben Mildenhall, Nataniel Ruiz, Shiran Zada, Kfir Aberman, Michael Rubenstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. 2023. DreamBooth3D: Subject-Driven Text-to-3D Generation. *ICCV* (2023).

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems* 33 (2020), 20154–20166.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. 2023. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512* (2023).

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Weilin Sun, Xiangxian Li, Manyi Li, Yuqing Wang, Yuze Zheng, Xiangxu Meng, and Lei Meng. 2022. Sequential fusion of multi-view video frames for 3D scene generation. In *CAAI International Conference on Artificial Intelligence*. Springer, 597–608.

Jiaxiang Tang. 2022. Stable-dreamfusion: Text-to-3D with Stable-diffusion. https://github.com/ashawkey/stable-dreamfusion.

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. 2023. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. 2023. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2022. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3835–3844.

Changshuo Wang, Lei Wu, Xu Chen, Xiang Li, Lei Meng, and Xiangxu Meng. 2023d. Letter Embedding Guidance Diffusion Model for Scene Text Editing. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 588–593.

Changshuo Wang, Lei Wu, Xiaole Liu, Xiang Li, Lei Meng, and Xiangxu Meng. 2023e. Anything to Glyph: Artistic Font Synthesis via Text-to-Image Diffusion Model. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.

Chi Wang, Min Zhou, Tiezheng Ge, Yuning Jiang, Hujun Bao, and Weiwei Xu. 2023f. CF-Font: Content Fusion for Few-Shot Font Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1858–1867.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. 2023a. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12619–12629.

Yizhi Wang, Yue Gao, and Zhouhui Lian. 2020. Attribute2font: Creating fonts you want from attributes. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 69–1.

Yuqing Wang, Zhuang Qi, Xiangxian Li, Jinxing Liu, Xiangxu Meng, and Lei Meng. 2023c. Multi-channel attentive weighting of visual frames for multimodal video classification. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2023b. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023a. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. arXiv:2304.05977 [cs.CV]

Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. 2023b. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20908–20918.

Shuai Yang, Jiaying Liu, Wenjing Wang, and Zongming Guo. 2019a. Tet-gan: Text effects transfer via stylization and destylization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1238–1245.

Shuai Yang, Zhangyang Wang, and Jiaying Liu. 2022. Shape-Matching GAN++: Scale Controllable Dynamic Artistic Text Style Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 7 (2022), 3807–3820. https://doi.org/10.1109/TPAMI.2021.3055211

Shuai Yang, Zhangyang Wang, Zhaowen Wang, Ning Xu, Jiaying Liu, and Zongming Guo. 2019b. Controllable artistic text style transfer via shape-matching gan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4442–4451.

Guo Yuanchen, Liu Yingtian, and Shao et al. Ruizhi. 2023. Threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

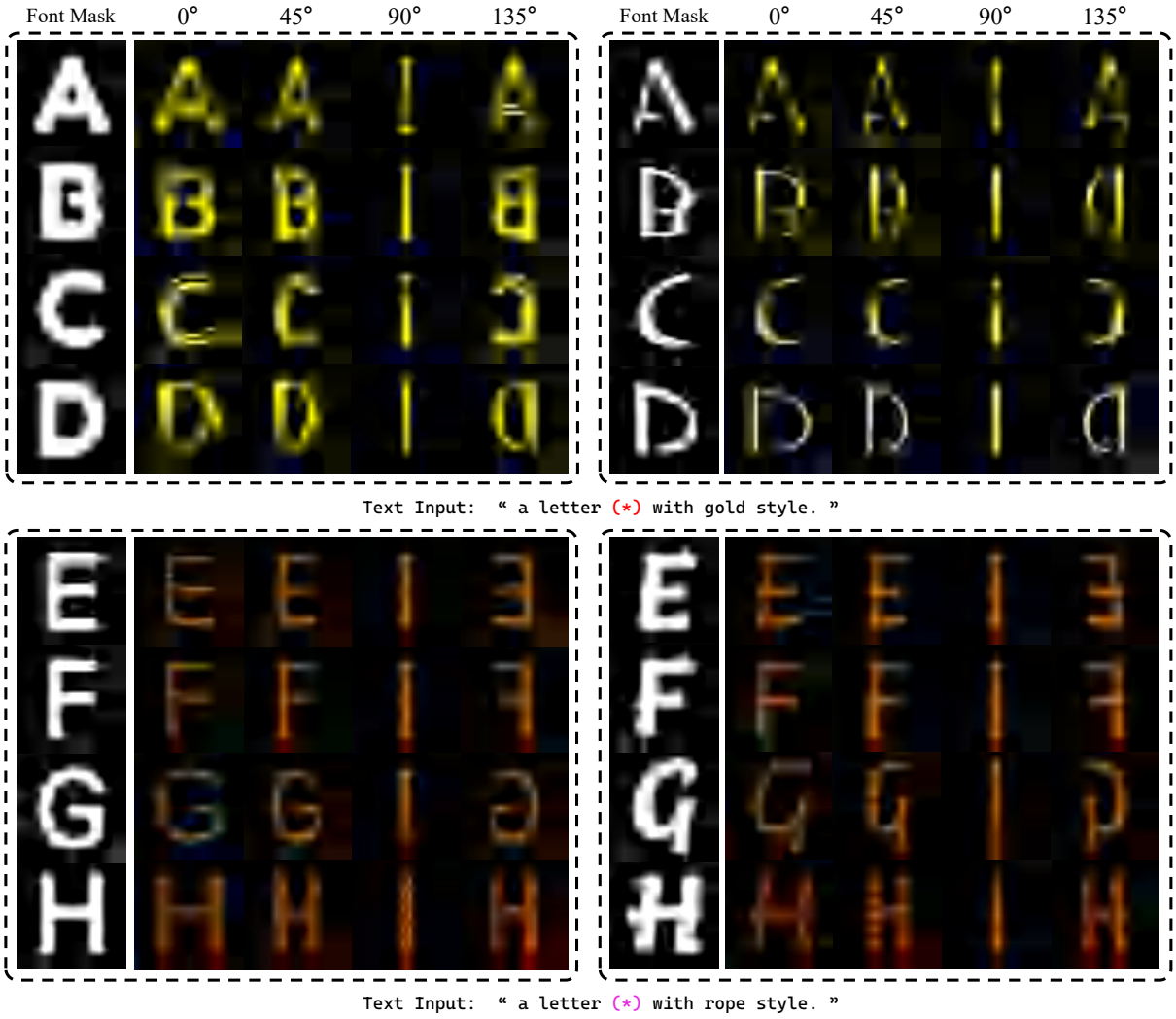


Figure 8: Results with Different Font Structure. We show the generated results with different font structure and letters. These examples we present use the same text prompt paradigm, only the font masks are different. '(*)' indicates a specified letter.

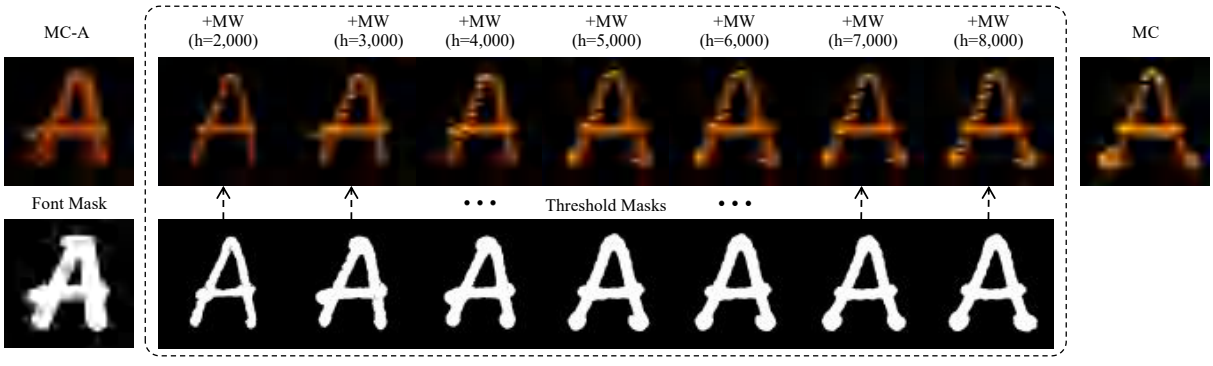


Figure 9: Influence of Progressive Mask Weighting (MW) on the Results. MC: Stopping multi-view mask constraint midway, the implementation details of MC are mentioned in Section 4.1. MC-A: Always using multi-view mask constraint without MW. We also show the generated results using different threshold masks from 2,000 to 8,000 iterations. We found that when h=4,000 iterations, the font structure and font effect details are closest to a balanced state.

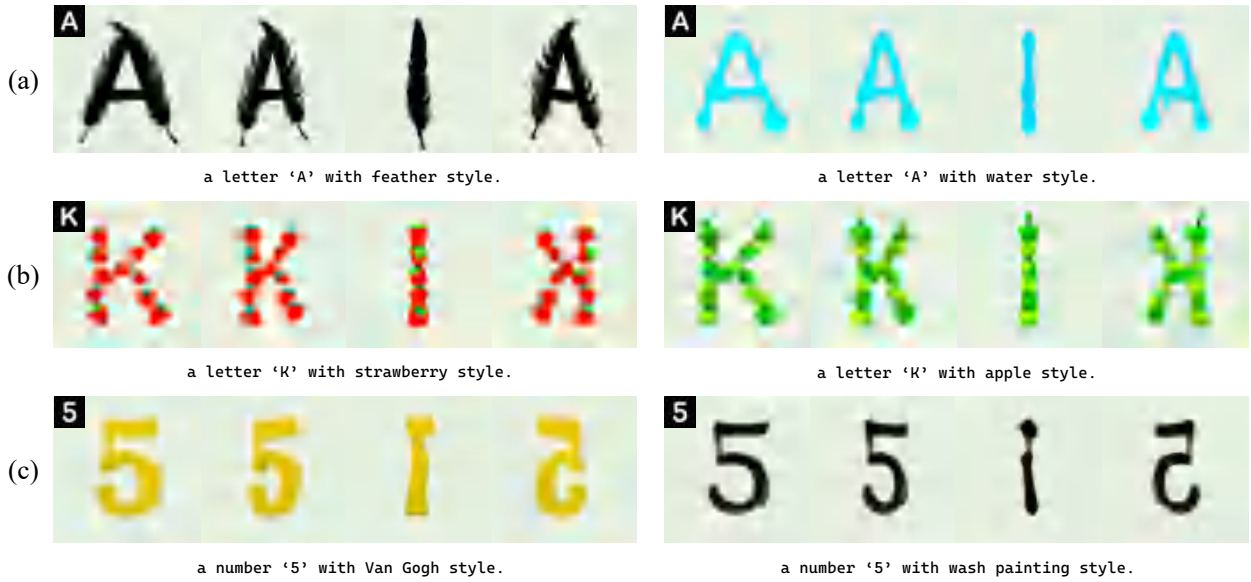


Figure 10: Text-to-3D Artistic Font Generation with Different Font Effect Style. We present three lines of generated results, each with the same character but different font effects. (a) Composed font effects. (b) Arranged font effects. (c) Abstract font effects.

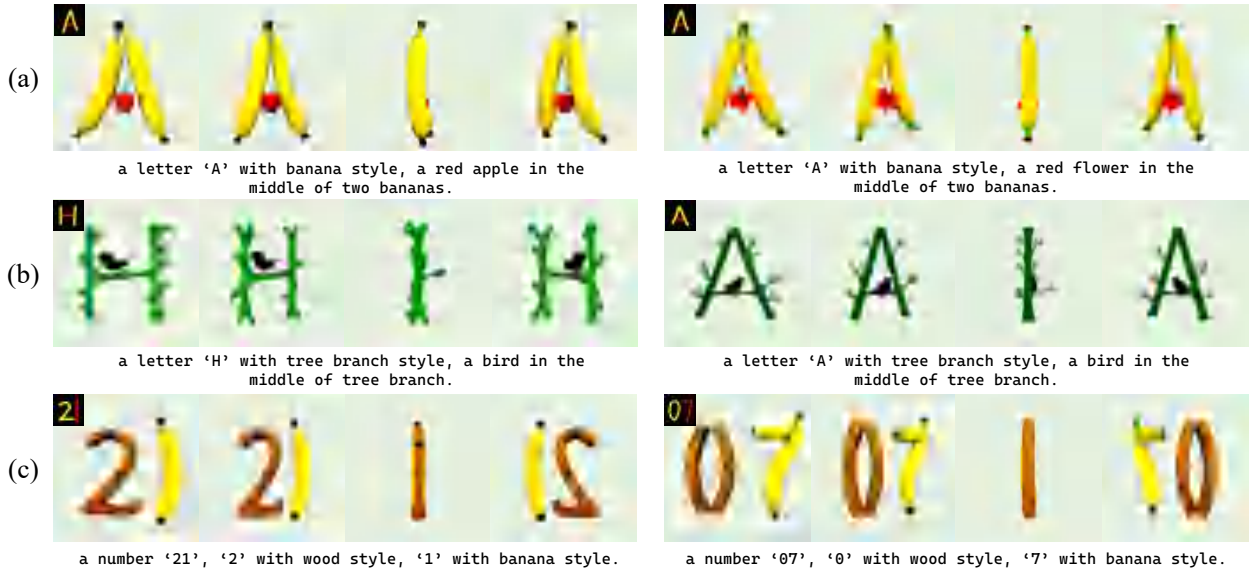


Figure 11: Text-to-3D Artistic Font Generation with Localization of Font Effects. (a) Localized font effects alteration through text prompts. (b) Localized font effects control for different characters. (c) Character-specific font effects control.

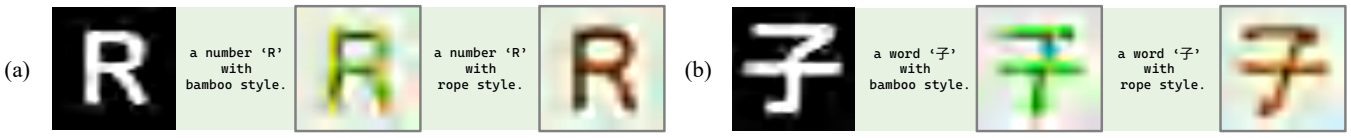


Figure 12: Limitations of font structure generation. (a) The phenomenon of bamboo's inherent inflexibility. (b) The phenomenon of lost strokes in Chinese font. Using materials such as bamboo to achieve significant deformations that approximate the font structure presents a challenge to our model. This issue can be mitigated by selecting suitable materials, such as rope.