Layout-Agnostic Scene Text Image Synthesis with Diffusion Models

Qilong Zhangli[1,2] Jindong Jiang[1] Di Liu[1] Licheng Yu[2] Xiaoliang Dai[2]
Ankit Ramchandani[2] Guan Pang[2] Dimitris N. Metaxas[1] Praveen Krishnan[2]
[1]Rutgers University [2]Meta AI

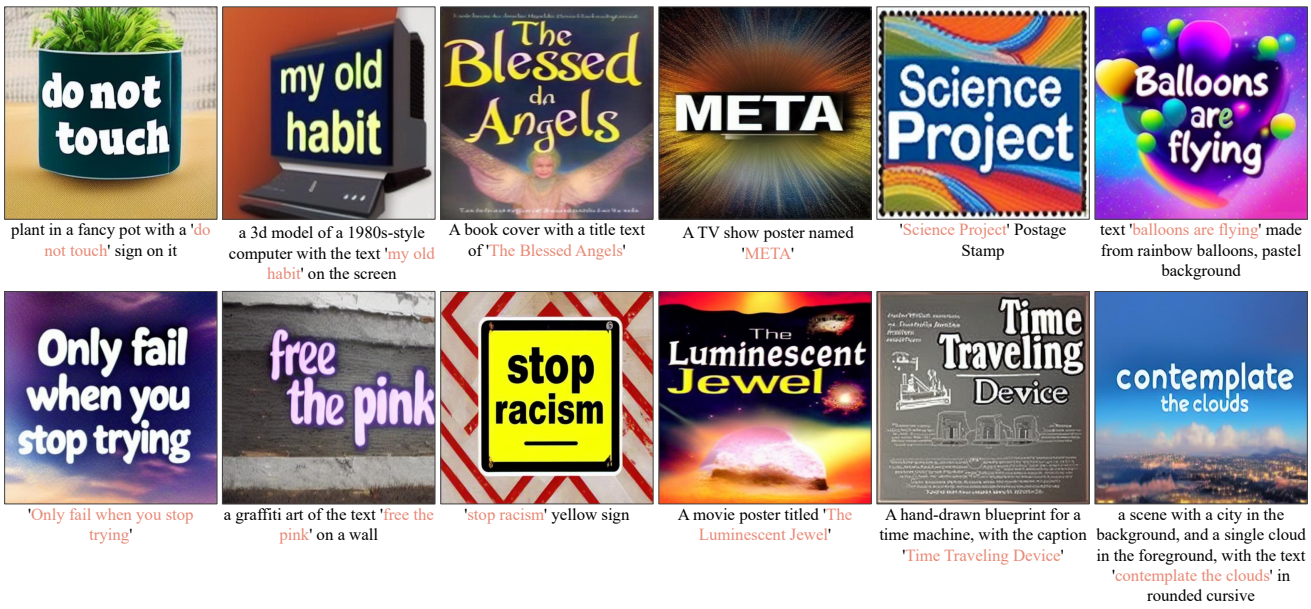

Figure 1. **Scene Text Generation.** Qualitative samples of scene text images generated by our model are presented. These images contain visually appealing texts that are coherent with the background and are created without relying on any spatial information or predefined layouts as input, thereby enhancing the Text-to-Image (T2I) Diffusion Model's capability to generate text.

Abstract

While diffusion models have significantly advanced the quality of image generation, their capability to accurately and coherently render text within these images remains a substantial challenge. Conventional diffusion-based methods for scene text generation are typically limited by their reliance on an intermediate layout output. This dependency often results in a constrained diversity of text styles and fonts, an inherent limitation stemming from the deterministic nature of the layout generation phase. To address these challenges, this paper introduces SceneTextGen, a novel diffusion-based model specifically designed to circumvent the need for a predefined layout stage. By doing so, SceneTextGen facilitates a more natural and varied representation of text. The novelty of SceneTextGen lies in its integration of three key components: a character-level encoder for capturing detailed typographic properties, coupled with a character-level instance segmentation model and a word-level spotting model to address the issues of unwanted text generation and minor character inaccuracies. We validate the performance of our method by demonstrating improved character recognition rates on generated images across different public visual text datasets in comparison to both standard diffusion based methods and text specific methods.

1. Introduction

The text-to-image [50, 57] task has gained popularity with advancements in diffusion models [41, 45, 49], significantly enhancing the quality of image generation. However, seamlessly integrating clear and contextually appropriate text into images is a persistent challenge. Text plays a vital role in many domains, including media content creation and artistic design, yet current diffusion models often struggle to produce text that is lexically correct, in valid font style, and that naturally complements the overall image aesthetics.

(a) Layout-Constrained Models (b) Layout-Agnostic SceneTextGen

Figure 2. Models reliant on predefined text layouts for input exhibit limitations such as constrained font diversity and static text positioning during each inference, leading to a lack of variability in style and arrangement.

Traditional methods [54, 58, 62] of creating scene text images typically formulate this problem as scene text editing which only involves editing or adding text to an existing scene image. These methods have challenges handling complex backgrounds, font styles and lighting variations. While recent models [4, 8, 33, 45, 49] have made strides in addressing these limitations by enhancing text encoding strategies [33] or employing predefined text layouts [8], they still face significant constraints in generating visual text. These constraints become apparent when observing the limited diversity in font styles and the static positioning of text, as shown in Fig. 2. Such rigidity in layout and font selection hampers the capacity of generative models to produce text that is stylistically varied and contextually aligned with the image content.

To address these issues, we propose **SceneTextGen**, a novel framework that capitalizes on the capabilities of latent diffusion models to infuse text into scene images with greater diversity and authenticity. Our approach is specifically engineered to transcend the limitations of predefined layouts, enabling more flexible text placement and an expansive assortment of text styles.

The two primary contributions of this work are: (i) the integration of character-level encoder to capture the typographic properties of visual text and carefully injecting it into the cross attention layers of the diffusion model, and (ii) introducing a novel word spotting loss using a pre-trained OCR along with a character segmentation loss to make the network more faithful in generating visual text. The character level encoder naturally blends with the existing diffusion model architecture and therefore allows the network to learn the layout of visual text implicitly rather than constraining itself to some pre-defined form. Our comprehensive evaluations confirm that SceneTextGen surpasses contemporary methods, facilitating the generation of images with text that is both aesthetically pleasing and rich in variety.

2. Related Work

2.1. Text 2 Image Generation Models

In the realm of text-to-image generation, diffusion models [11, 16, 27, 37, 41, 44, 49, 63, 64] represent a significant leap forward. Different from GAN-based models [10, 17, 67], diffusion models employ a stochastic process that iteratively adds noise to an image and learn to reverse this process to generate images from textual descriptions. Their capacity to produce high-quality, detailed visual content from text prompts has been well documented. The Latent Diffusion Model (LDM)[42] further enhances this approach by operating in a compressed image latent space, improving both efficiency and image quality. It also facilitates high-quality conditional generation through cross-attention text conditioning, leading to various downstream applications [12, 13, 15, 19, 55]. DALL-E [41], renowned for its novel approach of combining discrete VAEs with transformer language models, has shown remarkable ability in generating diverse and complex images from textual descriptions, showcasing the potential of transformer architectures in creative generative tasks. Deepfloyd [49], utilizing the robust T5 text encoders [40], not only enhances image quality but also facilitates nuanced understanding of complex prompts, thereby allowing for more accurate and context-aware visual representations. ControlNet [63], has demonstrated exceptional performance in conditioned image generation by providing the model with references such as skeleton, canny edge images, and segmentation maps. However, these models frequently encounter difficulties when it comes to the generation of text within images (see Fig. 4), a task requiring the text to be not only visually integrated but also contextually pertinent to the image content. This challenge stems from the complexity of modeling the fine-grained interplay between visual and textual elements, ensuring that the generated text is legible, aesthetically fitting, and semantically in sync with the image. Prior attempts to refine this aspect have led to improvements [8, 49, 63], yet the generation of contextually coherent text in images remains a largely unsolved problem, underscoring the need for more focused research.

2.2. Scene Text Generation

The success of adversarial networks such as GANs [7, 10, 17, 56, 67] for image generation and style transfer [20, 21] gave rise to scene text generation methods which can generate text at the granularity of glyphs [1, 26] or individual words [23, 54, 58]. Many previous works focused on the particular task of scene text editing where the model learns a style from a reference image and renders the target content in that style. Methods such as SRNet [54], SWAPText [58]

decomposes the problem into: (1) learning the foreground text using style and content, (2) background in-painting network to remove existing text, and (3) a blending network to merge foreground and background. These methods often fail in learning the correct style and removing existing text from complex backgrounds. TextStyleBrush [23] proposes a self-supervised approach to disentangle content and style and generate word images in a one-shot manner. All these previous methods are limited to generating individual words, requires reference word style images and does not generalize to generate the entire scene text image.

Spatial fusion GAN (SF-GAN) [62] generates text images by superimposing a foreground content image, transformed to match the style and geometry of a background image. This approach is aimed more at pure text synthesis on image regions without text, whereas in this work we aim to generate both image and text in a manner that reflects its natural appearance in real-world contexts.

With the significant progress of diffusion models in text-to-image generation, recent methods in scene text generation adapt these models towards producing more legible visual text. DIFFSTE [18] enhances scene text editing with a dual encoder design in diffusion models, promoting text legibility and style control. It is adept at mapping text instructions to images, showcasing zero-shot capabilities for rendering text in novel font variations and interpreting informal natural language instructions. This method is however a scene text editing method which generates single keywords on specific regions defined by a mask, whereas we propose a scene text generation network. One of the closest work in this space is TextDiffuser [8] which address the problem of scene text generation by decomposing the problem into two-stages. In the first stage, a transformer model create the layout mask for keywords extracted from text prompts. This is then taken as condition data while formulating the diffusion model to generate scene text images. They also introduce the character segmentation loss which helps in generating legible text. ControlNet [63], though not originally designed for scene text image rendering, has been effectively adapted for this purpose. It utilizes canny edge maps as conditional inputs, sourced from printed text images generated by a layout model, to fine-tune the diffusion model's output. The use of pre-defined layout for visual text generation in [8, 63], seriously limits the diversity of text styles, fonts and even the layouts. We believe this is due to the inherent difficulty in predicting layouts independently without the general image guidance. In our work, we avoid the need of pre-defined layout and make the network implicitly learn layout along with image generation using our novel way of injecting character level features. We also introduce a word spotting loss which augments the character segmentation loss proposed in [8] to generate more legible visual texts.

2.3. Scene Text Recognition

Advances in computer vision have laid a foundation for sophisticated analytical techniques in various domains [2, 9, 14, 30–32, 36, 46, 53, 60, 61, 65]. Especially in scene text image recognition, most existing works split the process into two stages: a text detection [3, 28, 29, 34, 48, 66] module to detect words or characters from complex backgrounds, and a text recognition [2, 24, 47, 52] module which transcribes the text into unicode characters given a cropped word image. More recently, end-2-end methods [6, 25, 38, 43, 51] have become popular due to the benefits of joint training of detector and recognizer to share contextual information.

In the nexus of scene text generation and recognition, leveraging pre-trained scene text recognition or word spotting models as guidance during diffusion-based text-to-image synthesis has surfaced as an innovative strategy. For simplicity, in this paper we refer to a scene text recognition module as OCR (optical character recognition). The integration of OCR-derived losses enables the refinement of generative models to produce text that is not just visually coherent but also contextually accurate. This confluence of generative modeling prowess with OCR accuracy paves the way for novel research avenues to generate images with text that is both authentic to read and visually integrated.

3. Methodology

3.1. Motivation

Our objective is to facilitate layout-free text image generation with diverse layouts and styles. A straightforward approach would be to directly fine-tune an existing latent diffusion model with text images. However, our early investigation suggests that this strategy did not yield significant improvements compared to the original LDM. We hypothesize that this limitation is due to two primary factors. First, the language encoder of the LDM, primarily designed for semantic interpretation, fails to capture character-specific information adequately for text rendering. This encoder tends to provide more semantic than structural information about the text, necessitating a dedicated network for encoding the text and guiding the model on its visual representation. Second, the conventional denoising loss used in diffusion models seems insufficient for accurately rendering text in images, often leading to text regions resembling textual patterns without the distinct features of text strokes. To address these challenges, we propose integrating a character-level encoder and a hierarchical cross-attention mechanism to learn character-level context information. Additionally, we introduce two auxiliary losses at the word and character levels to emphasize the text presentation. The subsequent sections will detail each of these components and their contribution to enhancing our model's performance.

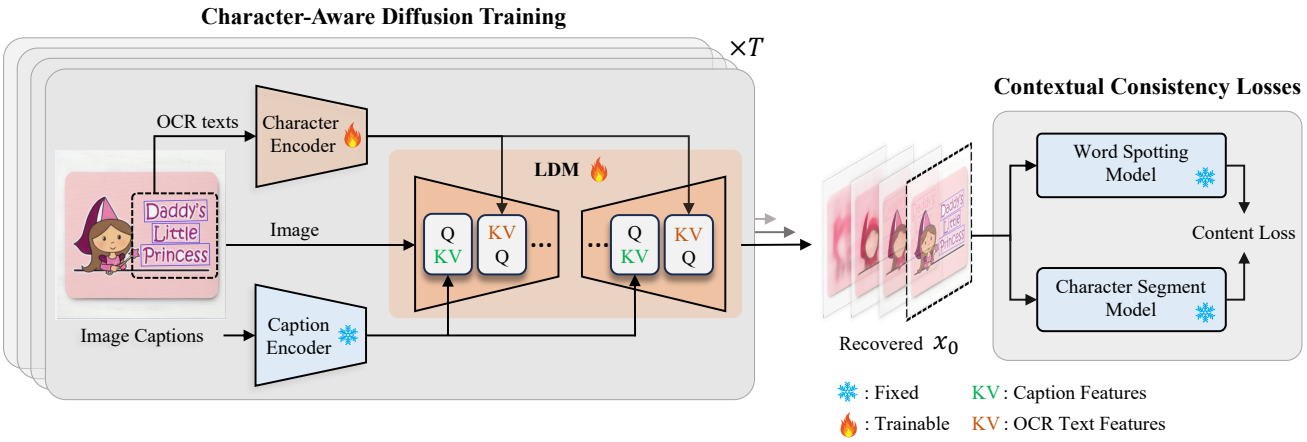

Figure 3. Model Framework: SceneTextGen employs a character-level encoder to extract detailed character-specific features. During loss computation, the model leverages both word-level and character-level supervisions to guide the recovery of the image, in addition to the standard denoising loss. This dual-level supervision enhances the model's ability to accurately generate and refine text within scenes.

3.2. Preliminaries

The proposed method is trained on a corpus of visual text images. Given an image $\mathcal{I}$ with textual caption c, we denote w_i, where $i \in [0, N]$, as the visual text (words) present in this image. We also assume that for each word, we know its location, given in terms of the bounding box. For all practical purposes, we assume this can be pre-computed using a pre-trained OCR detection [28] and recognition network [5]. Please note, the OCR bounding box and transcriptions are typically noisy. Fig. 3 presents an overview of our proposed method. A latent diffusion model has three key components: a CLIP encoder for semantic embedding of text and images, an autoencoder for dimensionality reduction and feature extraction, and a UNet-based structure for effective image-text synthesis and manipulation. The core architecture of our method follows the latent diffusion model, which is adapted to incorporate the proposed character-level features and content-based loss functions.

3.3. Character-Level Encoding

Our approach begins by extracting and ordering the visual text words from the image based on its locations as given by the bounding box. We sort the word boxes following the conventional reading pattern, i.e., from left to right and top to bottom. This provides us the naive ordering of visual text which is then tokenized at the character level to capture its typographic information. Then, a character encoder is used to encode these tokenized characters into a high-dimensional feature space, allowing an accurate transcription of the text's spelling and appearance. The derived text features encapsulate the typographical details and are poised for the subsequent integration with the image's latent features, which have already been contextually primed by the initial cross-attention with the encoded caption features obtained from the CLIP [39] text encoder.

3.4. Hierarchical Text Integration Process

We have employed a sequential cross-attention mechanism to integrate character information into the U-Net architecture effectively. This method is inspired by the observation that the cross-attention of captioning features is pivotal in shaping the overall spatial structure of an image, a notion supported by prior studies [15]. Leveraging this concept, our model initially constructs the general spatial layout of the image content, guided by the caption, using the first cross-attention layer. It then focuses on the precise rendering of characters in a subsequent cross-attention phase. This approach establishes a hierarchical text integration process, facilitating the development of a preliminary visual scaffold that is both thematically and contextually coherent. It ensures that the textual elements are accurately positioned and seamlessly integrated into the overall image structure.

3.5. OCR-Guided Diffusion for Text Accuracy

To ensure the textual accuracy of generated images, our model incorporates an OCR loss, utilizing the predictions from the end to end pre-trained GLASS model [43]. Upon each iteration of the diffusion process, the UNet predicts a denoising step, from which we derive an estimation of clean image x_0. This estimated x_0 is then decoded through a Variational Autoencoder (VAE) [22] to reconstruct an image.

The GLASS OCR model performs inference on this reconstructed image to produce word-level recognition results, which are represented as a tensor P with shape $[N, L, K]$, where N is the number of words detected, L is the maximum length of any word, and K is the size of the

character set. The ground truth for these detections is represented as a tensor G with shape $[N, L]$, where each entry is the character index for the corresponding position, and non-character positions are marked with zero.

The OCR loss ($\mathcal{L}_{\text{OCR}}$) is computed using a masked cross-entropy function, which is formulated as follows:

$$\mathcal{L}_{\text{OCR}} = -\frac{1}{\sum M} \sum_{i=1}^{N} \sum_{j=1}^{L} M_{ij} \cdot \log \left(\frac{\exp(P_{ij}[G_{ij}])}{\sum_{k=1}^{K} \exp(P_{ij}[k])} \right) \tag{1}$$

Here, M is a binary mask tensor that has the same shape as G and indicates the valid character positions (i.e., where $G_{ij} \neq 0$). M_{ij} represents the binary value of the mask at the i-th word and j-th character position, P_{ij} is the predicted probability distribution over the character set, and G_{ij} is the ground truth character index at that position.

By integrating this OCR loss into the training regime, we guide the diffusion model to produce text that is not only visually coherent but also textually accurate, as recognized by the GLASS [43] OCR model, thereby enhancing the overall fidelity of the generated images.

3.6. Refinement of Text Generation with Character-Level Constraints

During the iterative refinement of our diffusion model, we observed an unintended consequence of the OCR loss; the model tended to generate images with repetitive words. This issue is potentially attributed to the blurry nature of images at higher noise levels during training, which could render the OCR loss counterproductive. The word-level OCR loss, while ensuring textual accuracy, imposes no explicit constraint on the quantity of text within the image, inadvertently encouraging the model to generate excessive text.

To address this, we augmented our loss function with a character-level segmentation loss, which acts directly on the latent space rather than the recovered image. After obtaining the predicted latent features of a image x_0 from the UNet, we proceed in two directions: we decode x_0 using the VAE to compute the word-level OCR loss on the recovered image (as explained earlier), and we also apply x_0 to a pre-trained character-level segmentation model based on U-Net adapted from [8]. This model outputs a 96-dimensional feature map (corresponding to the length of the alphabet plus one for non-character pixels) with a spatial resolution of 64×64. The character-aware loss is then computed via cross-entropy between this feature map and a resized character-level segmentation mask C.

Thus, the total loss function is a composite of the denoising loss, the word-level recognition loss, and the character-level segmentation loss, expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{denoising}} + \lambda_{\text{word}} \mathcal{L}_{\text{OCR-word}} + \lambda_{\text{char}} \mathcal{L}_{\text{OCR-char}} \tag{2}$$

where $\mathcal{L}_{\text{denoising}}$ is the denoising loss, $\mathcal{L}_{\text{OCR-word}}$ is the word-level OCR loss, $\mathcal{L}_{\text{OCR-char}}$ is the character-level segmentation loss, and λ_{word} and λ_{char} are weighting coefficients balancing the contribution of each term.

The character-level segmentation loss is formulated as:

$$\mathcal{L}_{\text{OCR-char}} = -\frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{96} y_{hwc} \log \left(\hat{y}_{hwc} \right) \tag{3}$$

where H and W are the height and width of the feature map, y is the ground truth character segmentation mask, and $\hat{y}$ is the predicted character probability map from the segmentation model.

By integrating character-level information directly in the latent space, we impose a structured constraint on text generation, promoting both the accuracy and the appropriate quantity of text in the generated images.

4. Experiments

4.1. Implementation Details

Datasets For the training of our model, we utilized the publicly available MARIO dataset from [8], excluding MARIO-TMDB, and MARIO-OpenLibrary subsets as they are not publicly accessible. Upon removing any corrupted images, our final dataset comprised 7,249,449 image-caption pairs. In addition, to bolster our model's capability in generating a broader range of concepts (beyond text-centric images), we integrated 2,110,745 non-text images. These additional images, accompanied by text pairs, were selected based on a minimum predicted aesthetics score of 6.25, allowing for joint training to enhance overall performance.

Baselines We conducted quantitative comparisons of our SceneTextGen method against several leading approaches, including LDM[42], ControlNet[63], TextDiffuser[8], GlyphControl[59], and DeepFloyd[49], utilizing the publicly available code and pre-trained models for fairness. Notably, DeepFloyd is distinguished by its dual super-resolution modules, enabling it to produce high-resolution images at 1024x1024 pixels, in contrast to the 512x512 pixel images generated by the other models. For ControlNet comparisons, we employed Canny edge maps of printed text images created by the initial model stage of TextDiffuser as conditioning inputs. However, due to the unavailability of APIs, open-source code or checkpoints, we could not extend our comparative analysis to include Imagen[45], eDiff-i[4], or GlyphDraw[35].

Evaluation Criteria We assess text rendering quality using the MARIO-7M-Eval, MARIO-TMDB-Eval, and

MARIO-OpenLibrary-Eval datasets. Our evaluation is twofold: firstly, through CLIPScore, which measures the cosine similarity between the image and text representations from CLIP; and secondly, via OCR Evaluation, which leverages existing OCR tools to detect and recognize text regions within the generated images. Metrics such as Average Precision, Average Recall, F1 Score, and Accuracy are employed to determine the presence of keywords in the generated images. During training, the input of the character encoder is the ground truth OCR texts. During the inference process, the character encoder receives as its input the text from captions provided by the user. These captions are enclosed in quotation marks as specified in the user's prompts, and the input of the CLIP text encoder is the full caption. For each generated and ground truth image pair, we utilize the easy-ocr library for OCR detection and recognition, followed by Hungarian Matching between the sets of texts, applying the Levenshtein distance (or edit distance) on the matched text pairs for the OCR evaluation.

Algorithm 1 T2I Visual Text Performance Evaluation

function METRICS(M)
  $P \leftarrow$ PRECISION(M)
  $R \leftarrow$ RECALL(M)
  $F \leftarrow$ F1(P, R)
  $A \leftarrow$ ACCURACY(M)
  return $\{P, R, F, A\}$
end function

$GT \leftarrow$ OCR(Ground Truth Images)
$Gen \leftarrow$ OCR(Generated Images)
$Scores \leftarrow$ an empty list
for each $pair$ in ZIP(GT, Gen) **do**
  $C \leftarrow$ COSTMAT($pair$)
  $M \leftarrow$ HUNGARIAN(C)
  $Scores \leftarrow Scores \cup \{$METRICS($M$)$\}$
end for
return AVERAGE($Scores$)

Pseudo Code for OCR Performance Evaluation The methodology described in Algorithm 1 demonstrates the steps taken to assess the performance of the text-to-image conversion. In the initial step, the algorithm applies OCR to both the ground truth and the generated images. This process results in two sets of text outputs, which are then paired for comparison. The Hungarian algorithm is employed here to find the optimal matching between elements (words) of these two sets, minimizing the overall difference between the matched pairs. This is crucial for an objective and accurate comparison. For each matched pair, we calculate a cost matrix, which serves as the input for the Hungarian algorithm. The output of this step is a matching matrix M which represents the best possible alignment between the text elements in the ground truth and the generated images. Subsequently, the algorithm computes key performance metrics for each pair: Precision, Recall, F1 Score, and Accuracy. Precision focuses on the accuracy of the replicated text, recall measures the completeness, F1 score provides a balance between precision and recall, and accuracy gives an overall effectiveness of the text replication. Finally, the algorithm averages these scores across all image pairs to provide an overall performance evaluation of the text-to-image conversion process.

4.2. Quantitative Results

Our experimental analysis in Tab. 1 provides a direct comparison of the OCR based recognition scores among various models as measured in terms of Precision, Recall, F1 scores, and Accuracy. Our results indicate that SceneTextGen consistently outperforms competing models in most metrics. Latent Diffusion Model, lacking a sophisticated mechanism for text comprehension, typically under-performs, leading to lower OCR scores. In contrast, DeepFloyd[49] incorporates a T5 encoder which aids in textual understanding, thereby enhancing the quality of the generated text. However, its performance is still limited due to an insufficient character-level understanding.

ControlNet[63], TextDiffuser[8], and GlyphControl[59], which utilize predefined text layouts or spatial information, show mixed results. While the explicit introduction of text information allows ControlNet to achieve high OCR scores, the resulting text often appears artificial and lacks seamless integration within the images (see Fig. 4).

Cross-dataset Generalization Ability As demonstrated in Tab. 1, SceneTextGen-7M, despite being trained solely on the MARIO-7M dataset, exhibits strong generalization capabilities. It maintains robust OCR scores across evaluation sets from both the TMDB and OpenLibrary datasets, underscoring its adaptability and the efficacy of its training methodology.

4.3. Measuring Font Style Diversity

To quantitatively assess the diversity of font styles generated by SceneTextGen, we utilized a pretrained VGG-based font recognition model trained on synthesized text images. This approach involved first extracting text image patches using a pretrained Optical Character Recognition (OCR) model. These patches were then processed through the font recognition model to retrieve features from the penultimate layer. By applying t-SNE for dimensionality reduction, we visualized the feature space to examine the proximity and diversity of text generated by different methods.

As illustrated in Fig. 5, the rendered text images — which serve as a baseline — were created by printing text onto a white canvas using the layout generator from [8] with

	Pre-defined Text Layout	MARIO-7M				TMDB				OpenLibrary			
		AP(↑)	AR(↑)	F1(↑)	AC(↑)	AP(↑)	AR(↑)	F1(↑)	AC(↑)	AP(↑)	AR(↑)	F1(↑)	AC(↑)
TextDiffuser-10M [8]	✓	0.6135	0.4289	0.4683	0.3425	0.4401	0.4243	0.3994	0.2889	0.5617	0.3816	0.4242	0.2916
GlyphControl-10M* [59]	✓	0.5075	0.4118	0.4140	0.2883	0.3635	0.4082	0.3457	0.2362	0.4734	0.3762	0.3836	0.2534
Latent Diffusion Model		0.1482	0.1690	0.1296	0.0753	0.1717	0.2579	0.1703	0.0974	0.1911	0.2873	0.1923	0.1106
DeepFloyd [49]		0.2467	0.2788	0.2206	0.1366	0.2284	0.3738	0.2449	0.1500	0.2555	0.3910	0.2635	0.1610
ControlNet [63]	✓	0.5102	**0.4444**	0.4238	0.2981	0.3075	0.4667	0.3284	0.2194	0.4050	0.4273	0.3690	0.2415
TextDiffuser-7M [8]	✓	0.4778	0.3447	0.3682	0.2512	0.3198	0.3362	0.2961	0.1930	**0.4257**	0.3146	0.3318	0.2108
SceneTextGen-7M(Ours)		**0.5274**	0.4420	**0.4424**	**0.3088**	**0.3813**	**0.4716**	**0.3790**	**0.2602**	0.4136	**0.4519**	**0.3945**	**0.2571**

Table 1. Comparative analysis of OCR based text recognition scores across different models. Note, 7M denotes models that were trained on the MARIO-7M dataset (7 million images with texts). In contrast, the TextDiffuser-10M category includes models trained on an expanded dataset collection that encompasses MARIO-7M, MARIO-TMDB, and MARIO-OpenLibrary. * denotes LAION-Glyph dataset.

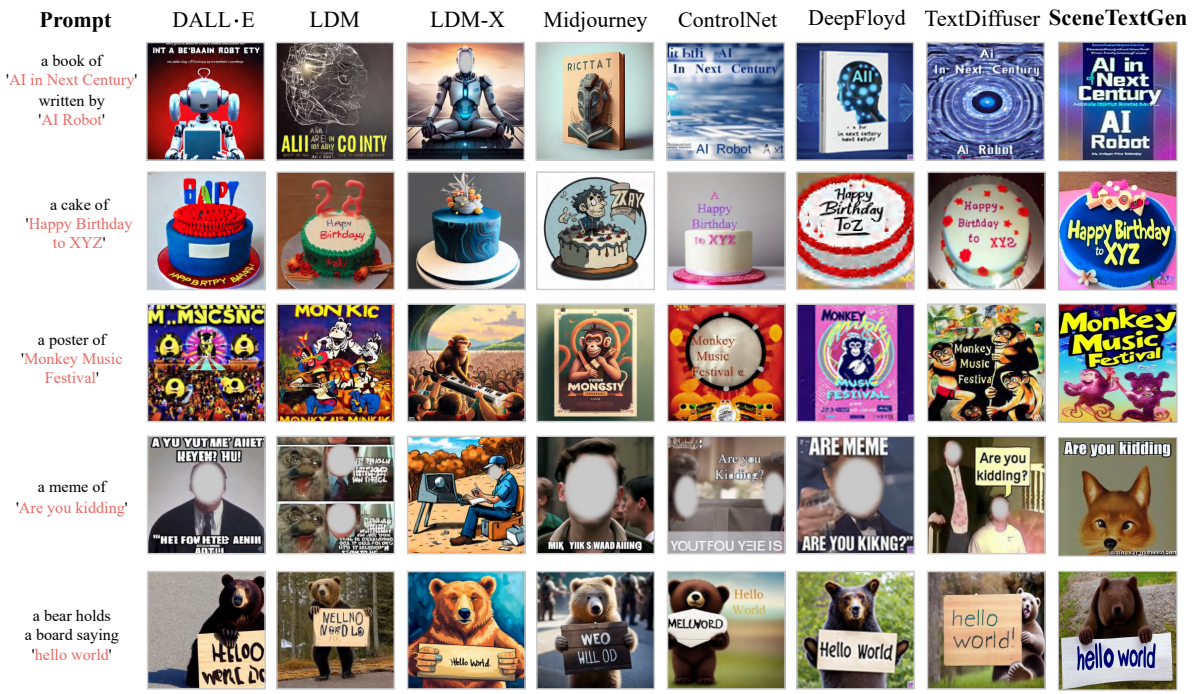

Figure 4. Comparative visualization of generated images. We present a side-by-side comparison of images generated from the same text prompt across different existing methods (with results generated by [8] as a proxy. Human faces are blurred for ethical considerations.). Each row corresponds to a unique prompt, showcasing the visual quality, text clarity, and contextual coherence achieved by each method.

the default 'Arial' font. This process establishes a reference for the feature distribution of 'Artificial Texts.' ControlNet [63], constrained by predefined text canny edges, exhibits a feature distribution closely aligned with the rendered text images. This proximity suggests its limitations in integrating text seamlessly within images, a finding corroborated by the qualitative results shown in Fig. 4. TextDiffuser [8], although performing well in most instances, still shares some feature space with ControlNet, indicating occasional production of artificial or unnatural texts. In contrast, SceneTextGen — operating independently of any predefined canny edges or layouts — demonstrates a distinct distribution with minimal overlap with the 'rendered text images,' signifying its robustness in generating naturalistic text within the context of scene images. Example visualizations are in Fig. 6. In addition, to better understand the text layout given by each model, we also show in Fig. 7 the overall distribution of visual text in the generated images and ground truth images.

4.4. Weight of Each Loss

In this section, we present two ablation studies to evaluate the impact of different loss function configurations on OCR performance. Table 3 explores the effects of using combinations of word-level and character-level losses. Table 4 examines the impact of varying the weights of these loss func-

	Pre-defined Text Layout	CLIP Score(↑)
TextDiffuser-10M [8]	✓	0.3436
GlyphControl-10M* [59]	✓	0.3450
Latent Diffusion Model		0.3015
DeepFloyd [49]		0.3267
ControlNet [63]	✓	0.3424
TextDiffuser-7M [8]	✓	0.3385
SceneTextGen-7M(Ours)		**0.3455**

Table 2. Comparison of CLIP scores reflecting the overall image quality generated by various models. SceneTextGen-7M achieves the highest CLIP score, indicating superior text-image alignment without relying on text layouts. * denotes LAION-Glyph dataset.

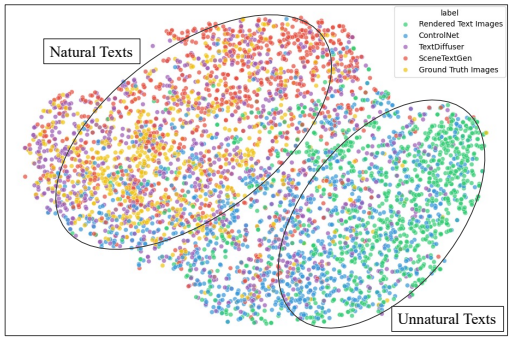

Figure 5. t-SNE representation of text region embeddings derived from the penultimate layer features of a font recognition model.

Caption: A sign that says 'Hello World'

Figure 6. Diversity of font styles. SceneTextGen is able to generate visually appealing and diverse font styles and layouts for text, without any style or layout prompts.

tions. In both studies, we assess the Average Precision (AP) and Accuracy (AC) scores to understand how these configurations influence the model's ability to accurately recognize and generate text in images.

4.5. Limitations

While SceneTextGen demonstrates superior performance in scene text image generation, challenges remain. Limited scene complexity: Models struggle to accurately generate complex elements in conjunction with text, such as a car with a 'green' sticker specifically in the back window.(Fig. 8, Case 1). Long text handling: Text accuracy and coherence decrease with increasing length (Fig. 8, Case 2). These areas require further development.

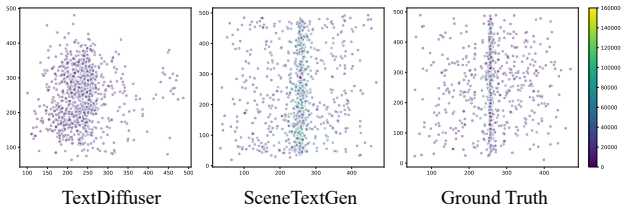

Figure 7. Distribution of visual text coordinates in the generated image w.r.t. the area of each box. As expected from real-life observations, the visual text in the ground truth images tends to be located at the center of each image.

L_{word}	L_{char}	AP	AC
✓		0.5082	**0.3109**
	✓	0.5083	0.3035
✓	✓	**0.5274**	0.3088

Table 3. Effects of loss function combinations on the MARIO-7M-Eval dataset

W_{lam}	C_{lam}	AP	AC
1	2	0.4107	0.2782
0.1	0.2	0.4587	0.2895
0.01	0.02	**0.5274**	**0.3088**

Table 4. Impact of varying loss weights on the MARIO-7M-eval dataset

Figure 8. Failure cases in complex scene interpretation and processing of lengthy prompts for text.

5. Conclusion

SceneTextGen, incorporating a character-level encoder and hierarchical text integration, offers advancements in scene text image generation. Despite improved text rendering, limitations arise in generating complex visuals and handling lengthy text. These challenges highlight the ongoing difficulty in reconciling textual accuracy with broader image synthesis. This work furthers our understanding of text-image generation, paving the way for future exploration.

Acknowledgements This research project has been partially funded by research grants to Dimitris N. Metaxas through NSF: 2310966, 2235405, 2212301, 2003874, and FA9550-23-1-0417.

References

[1] Samaneh Azadi, Matthew Fisher, Vladimir G Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. Multi-content gan for few-shot font style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7564–7573, 2018. 2

[2] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4715–4723, 2019. 3

[3] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9365–9374, 2019. 3

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 5

[5] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 4

[6] Michal Bušta, Yash Patel, and Jiri Matas. E2e-mlt-an unconstrained end-to-end method for multi-language scene text. In *Computer Vision–ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14*, pages 127–143. Springer, 2019. 3

[7] Qi Chang, Zhennan Yan, Mu Zhou, Di Liu, Khalid Sawalha, Meng Ye, Qilong Zhangli, Mikael Kanski, Subhi Al'Aref, Leon Axel, et al. Deeprecon: Joint 2d cardiac segmentation and 3d volume reconstruction via a structure-specific generative method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 567–577. Springer, 2022. 2

[8] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *arXiv preprint arXiv:2305.10855*, 2023. 2, 3, 5, 6, 7, 8

[9] Yunhe Gao, Mu Zhou, Di Liu, Zhennan Yan, Shaoting Zhang, and Dimitris N Metaxas. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. *arXiv preprint arXiv:2203.00131*, 2022. 3

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[11] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[12] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023. 2

[13] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, Di Liu, Qilong Zhangli, Jindong Jiang, Zhaoyang Xia, Akash Srivastava, and Dimitris Metaxas. Proxedit: Improving tuning-free real image editing with proximal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4291–4301, 2024. 2

[14] Xiaoxiao He, Chaowei Tan, Bo Liu, Liping Si, Weiwu Yao, Liang Zhao, Di Liu, Qilong Zhangli, Qi Chang, Kang Li, et al. Dealing with heterogeneous 3d mr knee images: A federated few-shot learning method with dual knowledge distillation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023. 3

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 4

[16] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 2

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[18] Jiabao Ji, Guanhua Zhang, Zhaowen Wang, Bairu Hou, Zhifei Zhang, Brian Price, and Shiyu Chang. Improving diffusion models for scene text editing with dual encoders. *arXiv preprint arXiv:2304.05568*, 2023. 3

[19] Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. In *Advances in Neural Information Processing Systems*, pages 8563–8601, 2023. 2

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[23] Praveen Krishnan, Rama Kovvuri, Guan Pang, Boris Vassilev, and Tal Hassner. Textstylebrush: Transfer of text aesthetics from a single example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 3

[24] Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2231–2239, 2016. 3

[25] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13094–13102, 2023. 3

[26] Wei Li, Yongxing He, Yanwei Qi, Zejian Li, and Yongchuan Tang. Fet-gan: Font and effect transfer via k-shot adaptive instance normalization. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1717–1724, 2020. 2

[27] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[28] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11474–11481, 2020. 3, 4

[29] Ron Litman, Oron Anschel, Shahar Tsiper, Roee Litman, Shai Mazor, and R Manmatha. Scatter: selective context attentional scene text recognizer. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11962–11972, 2020. 3

[30] Di Liu, Yunhe Gao, Qilong Zhangli, Ligong Han, Xiaoxiao He, Zhaoyang Xia, Song Wen, Qi Chang, Zhennan Yan, Mu Zhou, et al. Transfusion: multi-view divergent fusion for medical image segmentation with transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 485–495. Springer, 2022. 3

[31] Di Liu, Xiang Yu, Meng Ye, Qilong Zhangli, Zhuowei Li, Zhixing Zhang, and Dimitris N Metaxas. Deformer: Integrating transformers with deformable models for 3d shape abstraction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14236–14246, 2023.

[32] Di Liu, Anastasis Stathopoulos, Qilong Zhangli, Yunhe Gao, and Dimitris Metaxas. Lepard: Learning explicit part discovery for 3d articulated shape reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[33] Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. Character-aware models improve visual text rendering. *arXiv preprint arXiv:2212.10562*, 2022. 2

[34] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE transactions on multimedia*, 20(11):3111–3122, 2018. 3

[35] Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. Glyphdraw: Learning to draw chinese characters in image synthesis models coherently. *arXiv preprint arXiv:2303.17870*, 2023. 5

[36] Carlos Martín-Isla, Víctor M Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J Fulton, Tewodros Weldebirhan Arega, Kumaradevan Punithakumar, et al. Deep learning segmentation of the right ventricle in cardiac mri: The m&ms challenge. *IEEE Journal of Biomedical and Health Informatics*, 2023. 3

[37] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[38] Siyang Qin, Alessandro Bissacco, Michalis Raptis, Yasuhisa Fujii, and Ying Xiao. Towards unconstrained end-to-end text spotting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4704–4714, 2019. 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2

[41] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1, 2

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 5

[43] Roi Ronen, Shahar Tsiper, Oron Anschel, Inbal Lavi, Amir Markovitz, and R Manmatha. Glass: Global to local attention for scene-text spotting. In *European Conference on Computer Vision*, pages 249–266. Springer, 2022. 3, 4, 5

[44] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023. 2

[45] 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. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2, 5

[46] Lohrasb Ross Sayadi, Usama S Hamdan, Qilong Zhangli, and Raj M Vyas. Harnessing the power of artificial intelligence to teach cleft lip surgery. *Plastic and Reconstructive Surgery–Global Open*, 10(7):e4451, 2022. 3

[47] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4168–4176, 2016. 3

[48] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048, 2018. 3

[49] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd-if. https://github.com/deep-floyd/IF, 2023. GitHub Repository. 1, 2, 5, 6, 7, 8

[50] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiaoyuan Jing, Fei Wu, and Bingkun Bao. Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020. 1

[51] Peng Wang, Hui Li, and Chunhua Shen. Towards end-to-end text spotting in natural scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7266–7281, 2021. 3

[52] Peng Wang, Cheng Da, and Cong Yao. Multi-granularity prediction for scene text recognition. In *European Conference on Computer Vision*, pages 339–355. Springer, 2022. 3

[53] Song Wen, Hao Wang, Di Liu, Qilong Zhangli, and Dimitris Metaxas. Second-order graph odes for multi-agent trajectory forecasting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5101–5110, 2024. 3

[54] Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Editing text in the wild. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1500–1508, 2019. 2

[55] Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. Slotdiffusion: Object-centric generative modeling with diffusion models. *Advances in Neural Information Processing Systems*, 36:50932–50958, 2023. 2

[56] Zhaoyang Xia, Yuxiao Chen, Qilong Zhangli, Matt Huenerfauth, Carol Neidle, and Dimitris Metaxas. Sign language video anonymization. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources, Marseille, France, 25 June 2022*, 2022. 2

[57] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 1

[58] Qiangpeng Yang, Jun Huang, and Wei Lin. Swaptext: Image based texts transfer in scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14700–14709, 2020. 2

[59] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36, 2024. 5, 6, 7, 8

[60] Jiachi Ye, Haoyan Kang, Hao Wang, Chen Shen, Belal Jahannia, Elham Heidari, Navid Asadizanjani, Mohammad-Ali Miri, Volker J Sorger, and Hamed Dalir. Demultiplexing oam beams via fourier optical convolutional neural network. In *Laser Beam Shaping XXIII*, pages 16–33. SPIE, 2023. 3

[61] Jiachi Ye, Maria Solyanik, Zibo Hu, Hamed Dalir, Behrouz Movahhed Nouri, and Volker J Sorger. Free-space optical multiplexed orbital angular momentum beam identification system using fourier optical convolutional layer based on 4f system. In *Complex Light and Optical Forces XVII*, pages 70–80. SPIE, 2023. 3

[62] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3653–3662, 2019. 2, 3

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 5, 6, 7, 8

[64] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037, 2023. 2

[65] Qilong Zhangli, Jingru Yi, Di Liu, Xiaoxiao He, Zhaoyang Xia, Qi Chang, Ligong Han, Yunhe Gao, Song Wen, Haiming Tang, et al. Region proposal rectification towards robust instance segmentation of biological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–139. Springer, 2022. 3

[66] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017. 3

[67] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2