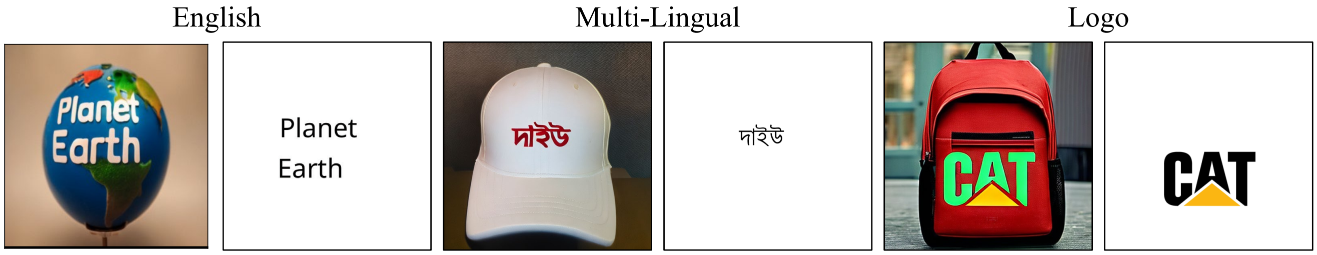
Conditional Text-to-Image Generation with Reference Guidance

Taewook Kim[1], Ze Wang[2], Zhengyuan Yang[3], Jiang Wang[3], Lijuan Wang[3], Zicheng Liu[2], Qiang Qiu[1]

[1]Purdue University [2]AMD [3]Microsoft



A globe with the words 'Planet Earth' written in bold letters

A word printed on a cap

A backpack with a Cat logo

Figure 1. **Samples produced by our method.** The proposed REFDIFFUSER generates accurate English scene-text images, Multi-lingual scene-text images, and Logo images based on reference conditions.

Abstract

Text-to-image diffusion models have demonstrated tremendous success in synthesizing visually stunning images given textual instructions. Despite remarkable progress in creating high-fidelity visuals, text-to-image models can still struggle with precisely rendering subjects, such as text spelling. To address this challenge, this paper explores using additional conditions of an image that provides visual guidance of the particular subjects for diffusion models to generate. In addition, this reference condition empowers the model to be conditioned in ways that the vocabularies of the text tokenizer cannot adequately represent, and further extends the model's generalization to novel capabilities such as generating non-English text spellings. We develop several small-scale expert plugins that efficiently endow a Stable Diffusion model with the capability to take different references. Each plugin is trained with auxiliary networks and loss functions customized for applications such as English scene-text generation, multi-lingual scene-text generation, and logo-image generation. Our expert plugins demonstrate superior results than the existing methods on all tasks, each containing only 28.55M trainable parameters.

1. Introduction

Recent developments in text-to-image models have yielded groundbreaking achievements, showcasing an unprecedented capability in translating natural language descriptions into visually compelling depictions. Despite the remarkable capabilities of generating images given general textual instructions, off-the-shelf text-to-image models usually struggle with precisely generating specific subjects at a high standard. For example, Stable Diffusion (SD) [48] is known for English misspelling.

Several research works have sought to address this issue by substituting conventional CLIP text encoders with a more powerful language model, such as T5 [46]. As demonstrated by DeepFloyd [2], Imagen [52] and eDiff-I [9], large language models enhance both sample fidelity and image-text alignment. While scaling up language models can enhance language comprehension, it comes at a substantial cost of computation and model training. Another line of research strives to resolve this challenge with additional conditions that exhibit spatial correlation with the input. TextDiffuser [12] utilized a segmentation mask predicted by a layout generation module as an additional condition. In a similar vein, GlyphDraw [41] proposed using text renderings of target keywords to condition the model for generating Chinese scene-text images. However, these methods demonstrate limited capability in generating im-

Contact: kim3803@purdue.edu

ages that are beyond the vocabulary of the text encoder they are trained with. For instance, modern text-to-image models universally employ English language models [45] for prompt embedding, hindering their generalization to image generation with text from a different language (Figure 1, middle) or accurate rendition of non-text subjects such as logos (Figure 1, right).

To resolve the aforementioned shortcoming of text-to-image diffusion models, we propose REFDIFFUSER , an approach that uses reference images as an additional source of condition, improving the generation results of particular subjects of interest with the guidance of explicit visual reference of their appearance. Our method is built upon the SD [48], which adopts an UNet for progressive denoising in a learned latent space of images. Instead of training an independent network branch or new layers from scratch to process the new visual condition, we directly encode the reference image into the same latent space using the VAE. More specifically, we augment the first layer of SD UNet to simultaneously accept the noisy image latent and the reference latent as the inputs. We then finetune the UNet to learn the natural blending of the visual references and the textual instruction to generate high-fidelity images following both conditions. Considering the large scales of the state-of-the-art diffusion models, tuning full models for diverse tasks can be prohibitive in terms of both parameter size and memory footprints. We resort to the low-rank adaptation of the SD models and develop a series of small-scale expert plugins, each containing only up to 28.55M parameters. Training objectives and auxiliary networks are customized for various applications, including English scene-text generation, multi-lingual scene-text generation, and logo-image generation. We present comprehensive discussions on the sampling strategies and showcase the distinct impacts of both conditions in different denoising time steps, achieving an English scene-text, multi-lingual scene-text, and logo generation accuracy of 61.73%, 46.88%, 44.07 with high prompt fidelity, outperforming the baselines.

The promising results on diverse applications shed light on a general way of customizing expert text-to-image models for particular subjects of interest.

2. Related Works

Diffusion models. Diffusion models Surpassing the prior family of generative models such as GAN [7, 18, 27, 44] VAE [33, 47, 57, 62], diffusion models [22, 56, 59] have demonstrated remarkable capabilities in generating images with both high quality and diversity, either in pixel space [22, 28, 58] or a learned latent space [10, 48, 64]. With the help of advanced pretrained language models [45, 46] and sampling techniques [21, 58], text-to-image diffusion models [9, 11, 42, 43, 48, 52, 66] show unprecedented results on image generation following textual instruction. They generate high-resolution images by either operating in latent space [19, 48] or using cascaded models [9, 52] to progressively scale up resolution. Motivated by the known issue of misspelling for almost all the public text-to-image diffusion models, several works have been proposed to improve text drawing by additional conditions such as masks [12] and glyph [41]. However, the model design customized for a particular language [12] prevents them from generalizing to general visual references. Recent research has been striving to improve the generation quality of particular subjects by model tuning. Specifically, model customization [8, 13, 31, 35, 51, 55, 65] focuses on a transfer learning approach that tunes model parameters to fix new concepts given examples. Textual inversion [17] learns a new concept by learning a new word vector. These methods fit well to particular subjects but usually fail to generalize to other similar ones. A line of research works [16, 36, 37, 68, 71] such as ControlNets [71] have been proposed to enhance the controllability of the generation using a new input condition.

Visual text generation. Despite the progress in generating images in high-quality images, the existing generative model has been noted to generate visually inaccurate images (i.e., misspellings). Several research studies have been proposed to mitigate this issue [2, 12, 39, 41, 52]. One line of research has shown that the visual accuracy of the text renderings can be improved by deploying a language model of larger capacity [2, 41, 52, 67]. Imagen [52] has demonstrated that encoding text prompts with the generic text encoder T5 [46] pretrained on text-only corpora can improve both the fidelity and image-text alignment [52]. To evaluate the accuracy of the generated visual text images, several papers have constructed their own benchmark sets [12, 39, 41] using off-the-shelf OCR models [24, 25, 29, 38, 70].

While some existing works rely on text renderings to improve English and multi-lingual generation [12, 60, 67], or focus on personalized object generation [35, 51], our work is distinguished by introducing a single, unified framework that bridges these capabilities. By leveraging lightweight, task-specific expert plugins, our method achieves high fidelity in producing accurate multi-lingual text as well as complex brand logos, thereby validating its broad applicability.

3. Method

The proposed REFDIFFUSER generates images conditioned on both a text prompt and a reference image. The image condition gives the model an uncurated visual reference for the generation targets, such as character shapes in scene-text image generation. We train the model to blend the reference concept naturally into the generated images without violating the text conditions. This reference condition empowers the model to produce contents that the original diffusion model fails to generate precisely, and can even help extend the generation to concepts not included in the language model vocabulary. Our proposed model is constructed on the foundation of the pre-trained Stable Diffusion model

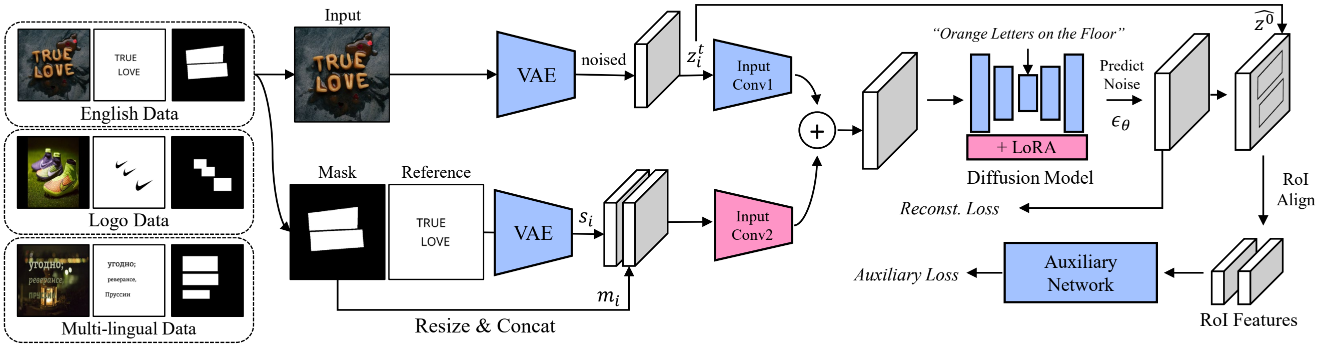


Figure 2. **Overview of the model training.** Expert plugin modules are trained on each dataset. The input and reference conditions are encoded using a VAE, and the mask is resized and concatenated with the reference. An additional convolutional layer (Input Conv2) processes these concatenated features, and its output is combined with the existing input convolution layers. The regional features are obtained from denoised latents, and an auxiliary network receives them. The auxiliary network is trained online with each expert plugin to encourage accurate drawing of the target subjects. The Input Conv2 and the LoRA parameters are updated during training.

[48]. We employ the low-rank adaptation technique [23] to fine-tune the model for specific tasks, such as scene-text image generation, and develop a series of small-scale plug-ins, each seamlessly converting a Stable Diffusion model into an expert model for different applications with references.

3.1. Overview

The proposed REFDIFFUSER builds upon the foundation of the pre-trained SD models [48] with a pre-trained CLIP [45] text encoder to encode text prompts, and is trained to incorporate additional image conditions. Training additional network branches to process the reference images introduces considerable cost [71]. Exploiting the fact that the references reside in the same image modality, we directly encode the reference images into the same latent space using the VAE [30] encoder of SD, and augment the input layer of the diffusion UNet [50], expanding the inputs with the additional reference latent and a spatial mask indicating the location of the targets of interest in the reference image.

Specifically, the input to the diffusion UNet is changed from the original $\mathbf{z}_i^t \in \mathbb{R}^{c \times h \times w}$, which is the VAE encoded representation of the input image i with noise injected at time step t, to $\mathbf{x}_i^t$:

$$\mathbf{x}_i^t = \text{concat}(\mathbf{z}_i^t, \mathbf{s}_i, \mathbf{m}_i), \tag{1}$$

where $\mathbf{s} \in \mathbb{R}^{c \times h \times w}$ is a VAE encoded representation of the spatial reference image, and $t \sim [1, T]$ is the step within the T total time steps. The location mask $\mathbf{m} \in \mathbb{R}^{h \times w}$ in (Equation (1)) is a binary array indicating the location and desired shape of the reference objects. The binary location mask $\mathbf{m}$ is downsampled to the same spatial size as the latent. The mask $\mathbf{m}$ and reference latent $\mathbf{s}$ remain constant across all time steps and are not injected with any noise in contrast to $\mathbf{z}$. With this input composition, we now have $\mathbf{x}^t \in \mathbb{R}^{(2c+1) \times h \times w}$.

To handle the additional $c + 1$ channels, we introduce an extra input convolutional layer (*Input Conv2*, Figure 2). Its output is added feature-wise to that of the original input convolution, and the combined feature map is then passed to the subsequent layers of the diffusion model for reconstruction. The reconstruction loss of diffusion training is defined as:

$$\mathcal{L}_{\text{diff}} = \sum_{i=1}^{N} ||\epsilon_i - \epsilon_\theta(\mathbf{x}_i^t, \mathbf{c}_i, t)||^2, \tag{2}$$

where ϵ_θ represents the diffusion model parametrized by θ, N is the batch size, ϵ denotes the noise prediction target, and $\mathbf{c}_i$ is the CLIP embedding of the text prompt.

Directly tuning θ introduces a high cost in terms of both parameter sizes and memory footprints. Therefore, we resort to the low-rank adaptation method [23] and develop expert plug-ins for different applications, each containing only a small number of parameters.

3.2. Example Plugin 1: Text Image Generation

In scene-text image generation, our goal is to produce high-quality images that not only align with the text prompt describing a scene but also accurately generate the target text. To achieve precise text generation, we provide the model with an additional reference image in which the desired characters are pre-rendered at the corresponding location in the scene, as illustrated in Figure 2. We use OCR detection results [12] to locate texts within the images, which can easily be obtained with the standard OCR detectors [24]. Specifically, given the OCR labels, we generate the reference image by rendering the text to the corresponding region on a blank canvas. We also generate the binary location mask, where positive values indicate the precise size and shape of the desired text to draw in the image. With the reference image encoded using the SD VAE for $\mathbf{s}$ and binary mask $\mathbf{m}$ resized to the same spatial size as the latent, we construct the model input as in (Equation (1)).

To ensure the spelling accuracy of the character sequence within the mask region, we facilitate the training with an online-trained lightweight text recognition network. We detail the network architecture of the recognition network in Appendix A.1. Given the noise estimation predicted by the diffusion network, we reconstruct the denoised latent at timestep 0, denoted as $\hat{\mathbf{z}}^0$, from the corrupted latent $\mathbf{z}^t$ [22]. We employ RoIAlign [20] to extract text regions into a fixed size, and these pooled regions are fed into an online-trained text recognition network $\psi_{\text{recog}}(\cdot)$ to compute the text recognition loss. The recognition loss is computed as,

$$\mathbf{r}_k = \text{RoIAlign}(\hat{\mathbf{z}}^0, \mathbf{B}_k) \tag{3}$$

$$\mathbf{o}_k = \psi_{\text{recog}}(\mathbf{r}_k) \tag{4}$$

$$\mathcal{L}_{\text{recog}} = \frac{-1}{L}\sum_{j=1}^{L} y_j \log \mathbf{o}_{k,j}, \tag{5}$$

where $\mathbf{B}_k$ is the bounding box label for k-th region in the image, $\mathbf{o}_k$ is the network output. L denotes the length of the k-th word and j is an enumerator for each character in the word. Here, $\mathbf{o}_{k,j} \in \mathbb{R}^{|C|}$ is the predicted probability over the character set at each position j, and $y_j \in \mathbb{R}^{|C|}$ is the corresponding one-hot label. We provide more details of the loss in Appendix A.1. Note that, unlike TextDiffuser, our method only requires word-level information, instead of character-level segmentation supervision [12]. This network can be discarded after training, and the final model consists of only the original SD UNet and a lightweight expert plug-in. The final learning objective of the diffusion network, $\mathcal{L}_\epsilon$, combines the reconstruction loss and the recognition loss through a weighted sum as:

$$\mathcal{L}_\epsilon = \lambda_{\text{recog}}\mathcal{L}_{\text{recog}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}}. \tag{6}$$

Multi-Lingual Text Image Generation. The flexibility of the conditions allows us to extend the generation to concepts beyond the vocabulary of the language model SD trained with. To show this, we develop an expert plug-in for multi-lingual text (MLT) image generation that covers multiple languages. Due to the difficulties in collecting large-scale datasets for MLT generation, we first pretrain the plugin module and recognition network on the large-scale English OCR dataset [12], then proceed with the fine-tuning. The overall learning process is mostly the same as the English text image generation training, except that the size of the alphabet is expanded to cover all the target languages simultaneously. Due to insufficient images in the existing MLT OCR datasets, we augment training data with synthesized images by manually collecting text-free background images [5], fonts in different scripts, and text corpora of each language, then rendering text words in random regions inside the background images (e.g., Figure 3, Greek, Russian, Thai.). We then use a merged dataset with both real images from MLT OCR datasets [14] and the synthesized image. Directly training the model with the mixed dataset compromises the quality of the generated images, as they can exhibit noticeable artifacts inherited from the synthetic images. To remedy this, we introduce an additional scaler α to scale the diffusion reconstruction loss:

$$\begin{aligned}\widetilde{\mathcal{L}}_{\text{diff}} = &\sum_{i\in\text{synth}}^{N} \alpha||\epsilon_i - \epsilon_\theta(\mathbf{x}_i^t, \mathbf{c}_i, t)||^2 \\ &+ \sum_{i\notin\text{synth}}^{N} ||\epsilon_i - \epsilon_\theta(\mathbf{x}_i^t, \mathbf{c}_i, t)||^2.\end{aligned} \tag{7}$$

We scale down α for the synthetic images to prevent the diffusion model from overfitting to the synthetic artifacts when learning to compose the scene and exploit the synthetic data mainly for improving the spelling correctness with the help of the recognition network. We observe that this simple loss scaling effectively improves the generated image quality. Loss terms related to the recognition network are identical to the ones used in the English scene-text generation.

3.3. Example Plugin 2: Logo Image Generation

We further show that the proposed framework can be generalized to non-text objects and use logo image generation as an example to show the versatility of the proposed conditional image generation with reference guidance.

For the reference images of logo image generation, we paste the standard logo we collected from the Internet onto a blank canvas and create the reference images. We visualize one example reference image in Figure 2. Similar to the MLT case, we synthesize logo images to augment the training set (Figure 4), and we deploy an auxiliary network to classify logos. To compute the loss, we predict the denoised latent features at timestep 0, $\hat{\mathbf{z}}^0$, and apply the RoIAlign [20] to extract the region $\mathbf{r}_k = \text{RoIAlign}(\hat{\mathbf{z}}^0, \mathbf{B}_k)$. The RoI features are then provided to the auxiliary network $\psi_{\text{logo}}(\cdot)$ to do the classification:

$$\mathcal{L}_{\text{logo}} = \frac{-1}{K}\sum_{k=1}^{K} y_k \log \psi_{\text{logo}}(\mathbf{r}_k), \tag{8}$$

where K denotes the total number of RoI instances in the batch enumerated by k, $y_k \in \mathbb{R}^{|M|}$ is the one-hot labels for $|M|$ logos we use for training. The overall learning objective can be formulated as follows:

$$\mathcal{L} = \lambda_{\text{logo}}\mathcal{L}_{\text{logo}} + \lambda_{\text{diff}}\widetilde{\mathcal{L}}_{\text{diff}}. \tag{9}$$

Although we train the model with a closed set of $|M|$ logos, the model can generalize to the ones that are unseen during training. We show visual examples to validate this in Figure 7, and Figure I in the Appendix.

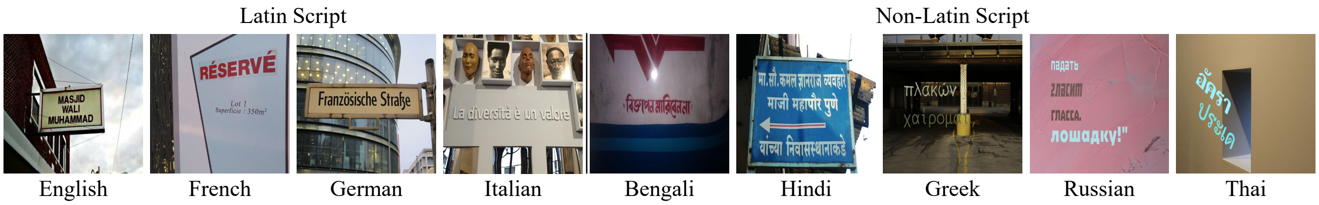


Figure 3. **Example images from the MLT training dataset. Left:** Languages in Latin scripts. **Right:** Languages in Non-Latin script. We use synthetic images for Greek, Russian, and Thai languages.

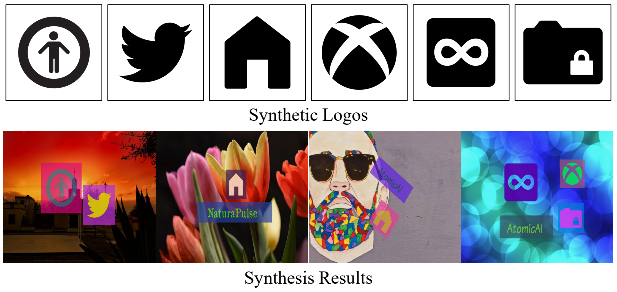


Figure 4. **Example synthesis result of the log images. Top:** Example logo images that are embedded into background images. **Bottom:** The resulting synthesized images using the logos.

3.4. Scheduled Classifier-Free Guidance

Classifier-free guidance [21] has become the standard technique for sampling with conditional diffusion models. It substantially improves the generated image quality and alignment with the condition prompt by extrapolating the predicted ϵ toward the direction of the condition:

$$\hat{\epsilon_\theta}^t := \epsilon_\theta(\mathbf{z}^t, \mathbf{c}, t) + \omega(\epsilon_\theta(\mathbf{z}^t, \mathbf{c}, t) - \epsilon_\theta(\mathbf{z}^t, \varnothing, t)), \qquad (10)$$

where $\epsilon_\theta(\mathbf{z}^t, \varnothing, t)$ denotes unconditional ϵ prediction with an empty text prompt.

Augmented Classifier-Free Guidance. With the additional reference condition included, we opt to reformulate the CFG as below to enable more detailed guidance of each condition,

$$\begin{aligned}\hat{\epsilon}^t &= \epsilon_\theta(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \varnothing, t)\\ &+ \omega_{\text{prompt}}\Big[\epsilon_\theta(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \mathbf{c}, t) - \epsilon_\theta(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \varnothing, t)\Big]\\ &+ \omega_{\text{ref}}\Big[\epsilon_\theta(\text{concat}(\mathbf{z}^t, \mathbf{s}, \mathbf{m}), \varnothing, t) - \epsilon_\theta(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \varnothing, t)\Big]\\ &+ \omega_{\text{all}}\Big[\epsilon(\text{concat}(\mathbf{z}^t, \mathbf{s}, \mathbf{m}), \mathbf{c}, t) - \epsilon(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \varnothing, t)\Big],\end{aligned} \qquad (11)$$

where the reference condition is set to be full zero in $\text{concat}(\mathbf{z}^t, \varnothing, \varnothing)$ to omit the reference condition. To enable the updated classifier-free guidance, we randomly drop the text condition $\mathbf{c}$ and the reference condition $\{\mathbf{s}, \mathbf{m}\}$ at a 10% chance during training. We provide additional details of the augmented CFG in Appendix A.2.

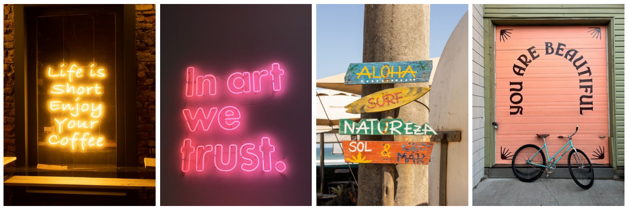


Figure 5. **Example images in the AText dataset we collect.**

CFG-scale scheduling. Standard classifier-free guidance adopts consistent scale ω across all sampling steps. We empirically observe that when an additional condition is introduced, dynamically adjusting the scale for each condition across sampling steps can further improve generated results. Specifically, we observe that setting a higher guidance scale ω_{ref} for reference conditions at the earlier sampling steps benefits the accurate visual text rendition, as it helps to establish the overall layout of the text elements. Hence, we dynamically adjust the guidance scales as,

$$\begin{aligned}\omega^t_{\text{ref}} &= \gamma \frac{t}{T}^{\rho_{\text{speed}}},\\ \omega^t_{\text{prompt}} &= \gamma(1 - \frac{t}{T}^{\rho_{\text{speed}}}).\end{aligned} \qquad (12)$$

We set γ and ω_{all} as constant values, and ρ_{speed} is also a constant that controls the speed of increase or decrease in the guidance scale.

4. Experiments

4.1. Datasets

We explain the dataset that we use for training. Additional details are available in Appendices A.3 and A.4.

English. For training English plug-in models, we first train the model with MARIO-10M [12] and further fine-tune the model by including samples from LAION-Aesthetic [54] and additional aesthetic text images (AText) that we collect (Figure 5). We use MARIO-Eval [12] benchmark for evaluation.

Multi-Lingual Text. For the MLT plug-in, we finetune the English plug-in model trained with MARIO-10M using the merged set of real images from ICDAR2019 [14], and syn-

Method	Acc↑	CLIP↑
SD	0.00	0.2902
TextDiffuser	0.00	0.3045
ControlNet	15.00	**0.3054**
Ours$^-$	**22.50**	0.3037

Table 1. **Evaluation results of zero-shot Russian generation**. Ours$^-$ denotes the model trained *without* seeing any Russian images during training. The accuracy is denoted in [%].

Method	Acc↑	F-1↑
w/o Reference	2.00	27.93
w/ Reference	**46.13**	**74.58**

Table 2. **Ablation studies on the impact of reference image.** Results are denoted in [%].

Method	# Params (M)↓
SD	859.52
TextDiffuser	876.86
ControlNet	1220.80
Ours	**28.55**

Table 3. **Comparison results of the number of trainable parameters.**

thetic images that we generated. Due to the lack of pre-existing benchmarks for MLT image generation, we developed a new benchmark for model evaluation on nine different languages, having 2,500 prompts in total. Details of the benchmark are available in Appendix A.4.

Logos. Similar to the MLT plug-in, we also finetune the English plug-in model trained with MARIO-10M using the merged set of real images from the merged set of FlickrLogos-32 [26] and Logos in the wild (LITW) [61], and the synthetic images that we generated. For evaluation

4.2. Evaluation Metrics

We use **Fréchet inception distance (FID)** to assess the image quality, and we evaluate the semantic alignment by computing the **CLIP Score** between the text prompts and the generated images and measuring the aesthetic scores of the generated images obtained from a pretrained model [54].

Text images. To measure the accuracy of the generated text, we use **Accuracy** and **F1-score**. We apply different OCR Engines for English generation tasks and MLT image generation tasks. We use the Microsoft Read API for English, and the Google Cloud Vision API for MLT image generation tasks. We consider a detected text correct if it is exactly matched with a keyword, and the F-1 score is defined as the harmonic mean of precision and recall.

Logo images. To measure the accuracy of the logo image generation, we train a Mask-RCNN [20] on the merged set of FlickrLogos-32 [49] and LITW dataset [61]. We report the Accuracy and F1-score of the logo detection model applied to an image set generated by models.

4.3. Implementation Details.

We first fine-tune the pre-trained Stable Diffusion (SD) V2.1 [48] on the MARIO-10M dataset [12]. For English image generation, we further train the model using the AText dataset we collect. For MLT and Logo image generation, we train the model with merged sets of synthetic images and real images. We set batch size=10 for each of the 8 GPUs. We use AdamW [40] with a learning rate of 1e-4 to tune the parameter. For MLT image generation, we use a character set containing 847 characters from the nine languages. We use a rank of 32 for all LoRA. We use 100 steps of DDIM [58] sampling. For all the experiments, we set the recognition loss weight λ as 0.025.

Method	FID↓	CLIP↑	Acc↑	F-1↑.
SD	51.29	0.3015	0.03	2.14
ControlNet	51.48	0.3424	23.90	58.65
DeepFloyd	**34.90**	0.3267	2.62	17.62
TextDiffuser	38.75	0.3436	56.09	78.24
SD3	37.21	0.3424	9.62	49.55
SDXL	58.54	0.3242	2.16	20.33
Flux	41.25	0.3198	54.98	52.35
GlyphControl	61.10	0.3411	27.04	64.00
AnyText	52.39	0.3426	18.03	60.74
Ours	38.59	**0.3454**	58.26	79.15
Ours$^+$	42.19	0.3434	**61.73**	**80.08**

Table 4. **Comparison results of English generation on MARIO-Eval benchmark**. Acc and F-1 denote OCR Accuracy and F1-score, respectively. + denotes the results with scheduled classifier guidance. Accuracy and F-1 are denoted in [%].

4.4. Ablation Study

In this section, we present ablation studies to gain deeper insights into the proposed method. We provide additional ablation studies in Appendix B.

Impact of Reference Image. We examine the influence of reference image guidance. Notably, we observe a huge gap in the performance between the model trained to be conditioned with additional reference guidance and the model without it (Table 2). This observation indicates the significant role that visual references play in both proper conditioning and providing valuable information about the targets.

Zero-shot Text Image Generation. As reference guidance enables our model to handle conditions beyond the vocabularies encountered during training, we explore its zero-shot capability. We evaluate the zero-shot MLT results for Russians by excluding all Russian text images during training Table 1. Notably, our zero-shot model outperforms the baselines in terms of OCR accuracy, demonstrating its effectiveness.

Method	Latin				Non-latin			
	FID↓	**CLIP↑**	**Acc↑**	**F-1↑**	**FID↓**	**CLIP↑**	**Acc↑**	**F-1↑**
SD	114.34	0.3032	0.38	1.87	113.07	0.3067	0.00	0.00
GlyphControl	115.51	0.2922	58.25	68.73	130.44	0.2879	0.00	0.00
AnyText	115.51	0.3100	10.38	48.87	130.44	0.3155	4.50	17.42
SD3	134.31	0.2983	25.06	46.13	133.24	0.3090	0.11	0.11
SDXL	126.49	0.3041	16.18	13.90	132.13	0.3101	0.00	0.00
ControlNet	140.03	0.3010	10.63	43.03	136.54	0.3038	12.00	40.09
Ours	117.20	0.2952	56.38	76.68	116.79	0.3059	23.80	38.96
Ours$^+$	117.96	0.2803	64.56	79.09	117.09	0.2940	29.20	41.78

Table 5. **Evaluation results of the MLT image generation.** **Latin languages**: English, German, French, and Italian. **Non-Latin languages**: Bengali, Hindi, Greek, Russian, and Thai. Result denoted by $+$ applies CFG scheduling with $\rho_{\text{speed}} = 0.2$, $\gamma = 3.5$ and $\omega_{\text{all}} = 4$. Accuracy and F-1 are denoted in [%].

Efficiency Analysis. We report the parameter count with comparison with baseline approaches in Table 3. Notably, our method requires significantly fewer parameters, demonstrating that it is both effective and resource-efficient, enabling efficient integration into existing pipelines.

4.5. Quantitative Results

We provide comparison results with SD [48], SD3 [15], SDXL [43], Flux [32], GlyphControl [67], AnyText [60], ControlNet [71], DeepFloyd [2], and TextDiffuser [12]. For ControlNet, we use the model trained to be conditioned on the Canny Edge map, and we provide Canny Edge maps of rendered text images or logo images as input conditions during inference. For log generation, we also compare with BLIP-Diffusion [34], IP-Adapter [69], and MS-Diffusion [63]. We present results for our method both with and without the scheduled CFG guidance described in Section 3.4.

English Images. For the English image generation task, our method achieves the best performance on all the OCR-related metrics, including OCR Accuracy and F-1 Score (Table 4). Additionally, our method also achieves the best CLIP score, demonstrating that our model is not only capable of generating accurate images but also best aligns with the text prompts. Besides, our method achieves comparable results of FID with DeepFloyd, which achieves lower accuracy than ours by a large margin. When the proposed CFG scheduling scheme is applied, we obtain further improvement in accuracy, with a decrease in FID performance. We provide additional analysis on CFG scheduling at Appendix B.

Multi-Lingual Images. For multi-lingual text (MLT) image generation, we report results in the Latin language and the non-Latin language sets (Table 5). There are five different languages in *Latin languages*: English, German, French, and Italian. There are four different languages in *Non-Latin languages*: Bengali, Hindi, Greek, and Russian. For all language sets, we achieve the best accuracy result. Similar to English results, we obtain improvements in accuracy when CFG scheduling is applied. We report the detailed per-language results in Appendix B.1 due to the space limit.

Method	FID↓	CLIP↑	Acc↑	F-1↑
SD	**74.40**	0.3469	10.33	13.71
ControlNet	110.77	0.3553	34.20	39.36
TextDiffuser	88.97	0.3183	9.13	10.06
BLIP-Diffusion	104.18	0.3418	12.20	18.27
IP-Adapter	117.72	0.3509	22.47	25.44
MS-Diffusion	83.27	0.3769	34.73	41.23
Ours	88.49	0.3759	42.87	48.86
Ours$^+$	89.49	**0.3770**	**44.07**	**48.91**

Table 6. **Evaluation results for logo image generation.** Result denoted by $+$ applies CFG scheduling with $\rho_{\text{speed}} = 0.2$, $\gamma = 3.5$ and $\omega_{\text{all}} = 4$ for CFG scheduling. Accuracy and F-1 are denoted in [%].

Logo Images. Our approach also attains notable results in producing logo images. Our model ranks second in the FID measurement, while achieving the best results on CLIP score, Accuracy, and F-1 score (Table 6). Similar to English and MLT generation, the model achieves improvement in accuracy when CFG scheduling is applied. To further demonstrate the generalization of the trained plug-in for logo image generation, we fed the network with novel logos and icons that are not included in the dataset. Apart from offering natural and faithful blending of the logos to the described scenes, we further demonstrate in Figure 7 that our method generalizes to novel logos and icons unseen during model training.

4.6. Qualitative Results

We present side-by-side results in Figure 6 for direct comparisons of the image quality of different methods. Our method achieves both faithful blending of target subjects to the scene and strong correspondence to the text prompt in all experiments. We also provide logo image generation results when unseen logo images are given (Figure 7), which further validates the effectiveness of the proposed method and the model's ability to generalize across novel instances. Moreover, our method is highly flexible and can be easily extended to text editing tasks (Figure 8). Following [12], we augment the channel of the input latent by concatenating

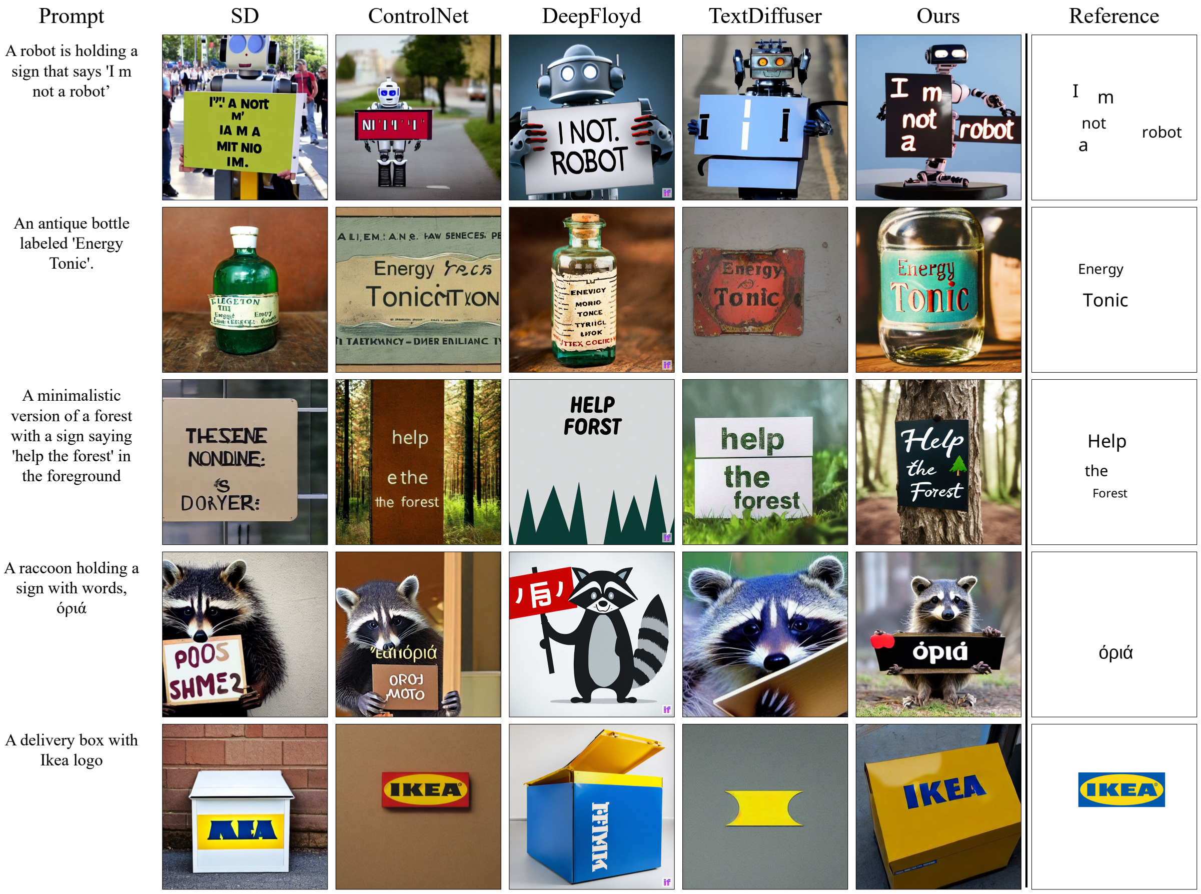

Prompt
SD
ControlNet
DeepFloyd
TextDiffuser
Ours
Reference
A robot is holding a sign that says 'I m not a robot'
I m not a robot
An antique bottle labeled 'Energy Tonic'.
Energy Tonic
A minimalistic version of a forest with a sign saying 'help the forest' in the foreground
Help the Forest
A raccoon holding a sign with words, όpιά
όpιά
A delivery box with Ikea logo
IKEA

Figure 6. **Side-by-side qualitative comparison with baselines.**

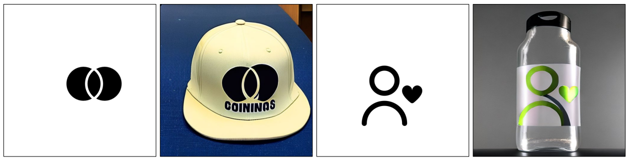
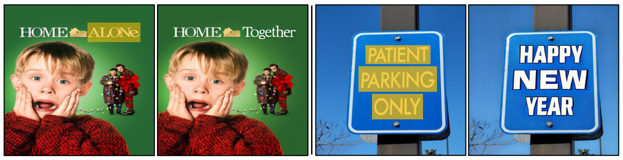

"A cap with a logo."
"a water bottle with an icon."

Figure 7. **Unseen logo image generation results.** The logo images in the reference are not included in the training logo set and hence, are unseen during the model training. More visualization results of generated unseen logo images are provided in Appendix C.



Masked Input
Output
Masked Input
Output

Figure 8. **Text editing results.** Regions denoted in yellow are masked and edited.

the encoded latent of the masked image. We provide more details and additional qualitative results in Appendix C.

5. Conclusion

In this paper, we introduced REFDIFFUSER, a text-to-image diffusion model based on a visual reference guide. We expanded a pretrained SD model to accept an additional reference image as input, which provides the model with visual guidance to the appearance of the generation target, allowing for the precise generation of concepts even beyond the text-encoder vocabulary. With lightweight expert plugins efficiently tuned by applying low-rank adaptation, and training methods adapted for each task, we demonstrated expert plugins for applications including English scene-text generation, multi-lingual scene-text generation, and logo-image generation. Experimental results validated the superiority of the proposed method, as our model achieves superior results on every task. This research shed light on a general framework for providing additional visual references to text-to-image models for precise generation.

References

[1] Project gutenberg's adventures of sherlock holmes. https://www.gutenberg.org/files/48320/48320-h/48320-h.htm. 2

[2] Deepfloyd. https://www.gutenberg.org/files/48320/48320-h/48320-h.htm. 1, 2, 7

[3] Foundation icon fonts 2. https://zurb.com/playground/foundation-icons. 2

[4] LAION-ASTHETIC. https://laion.ai/blog/laion-aesthetics/. 2

[5] Pexels. https://www.pexels.com/. 4, 2

[6] Tmdb movie metadata. https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata. 2

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. 2017. 2

[8] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. 2023. 2

[9] 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. 1, 2

[10] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *ECCV*, 2022. 2

[11] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[12] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. 2023. 1, 2, 3, 4, 5, 6, 7

[13] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. 2024. 2

[14] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. ICDAR2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *International Conference on Document Analysis and Recognition*, 2019. 4, 5, 2, 6

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. 7

[16] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*, 2022. 2

[17] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. 2014. 2

[19] 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 *CVPR*, 2022. 2

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 4, 6

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 5, 1

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020. 2, 4

[23] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 3, 4, 5

[24] Mingxin Huang, Yuliang Liu, Zhenghao Peng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Yuan, Kai Ding, and Lianwen Jin. Swintextspotter: Scene text spotting via better synergy between text detection and text recognition. In *CVPR*, 2022. 2, 3

[25] Mingxin Huang, Jiaxin Zhang, Dezhi Peng, Hao Lu, Can Huang, Yuliang Liu, Xiang Bai, and Lianwen Jin. Estextspotter: Towards better scene text spotting with explicit synergy in transformer. In *CVPR*, 2023. 2

[26] Yannis Kalantidis, Lluis Garcia Pueyo, Michele Trevisiol, Roelof van Zwol, and Yannis Avrithis. Scalable triangulation-based logo recognition. In *Proceedings of ACM International Conference on Multimedia Retrieval*, 2011. 6, 2

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[28] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. 2022. 2

[29] Taeho Kil, Seonghyeon Kim, Sukmin Seo, Yoonsik Kim, and Daehee Kim. Towards unified scene text spotting based on sequence generation. In *CVPR*, 2023. 2

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[31] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2

[32] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2023. 7

[33] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. 2016. 2

[34] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *NeurIPS*, 2023. 7

[35] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[36] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet ++:

Improving conditional controls with efficient consistency feedback. In *ECCV*, 2024. 2

[37] 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 *CVPR*, 2023. 2

[38] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *ECCV*, 2020. 2

[39] 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. In *arXiv preprint arXiv:2212.10562*, 2022. 2

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[41] 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. 1, 2

[42] Alexander Quinn 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. 2022. 2

[43] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 7

[44] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arxiv 2015. *ICLR*, 2016. 2

[45] 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 *ICLR*, 2021. 2, 3

[46] 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. 21(1):5485–5551, 2020. 1, 2

[47] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. 2015. 2

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 6, 7

[49] Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof Van Zwol. Scalable logo recognition in real-world images. In *Proceedings of ACM International Conference on Multimedia Retrieval*, 2011. 6, 2

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015. 3

[51] 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 *CVPR*, 2023. 2

[52] 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. 2022. 1, 2

[53] C Schuhmann, R Vencu, R Beaumont, R Kaczmarczyk, C Mullis, A Katta, T Coombes, J Jitsev, and A LAION Komatsuzaki. 400m: Open dataset of clip-filtered 400 million image-text pairs. arxiv 2021. *arXiv preprint arXiv:2111.02114*. 2

[54] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. 2022. 5, 6, 2

[55] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 2

[56] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. 2015. 2

[57] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. 2015. 2

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021. 2, 6

[59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020. 2

[60] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *ICLR*, 2023. 2, 7

[61] Andras Tüzkö, Christian Herrmann, Daniel Manger, and Jürgen Beyerer. Open Set Logo Detection and Retrieval. In *Proceedings of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications: VISAPP*, 2018. 6, 2

[62] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. 30, 2017. 2

[63] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *ICLR*, 2025. 7

[64] Ze Wang, Jiang Wang, Zicheng Liu, and Qiang Qiu. Binary latent diffusion. In *CVPR*, 2023. 2

[65] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, 2023. 2

[66] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. 2024. 2

[67] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. 2024. 2, 7

[68] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael

Zeng, et al. Reco: Region-controlled text-to-image generation. In *CVPR*, 2023. 2

[69] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 7

[70] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Tongliang Liu, Bo Du, and Dacheng Tao. Deepsolo: Let transformer decoder with explicit points solo for text spotting. In *CVPR*, 2023. 2

[71] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 7